PROCEEDINGS OF PAPERS

Zbornik radova

(Ic)ETRAN 2019

6th International Conference on Electrical, Electronic and Computing Engineering

in conjunction with

ETRAN

63rd National Conference on Electrical, Electronic and Computing Engineering

Proceedings of Papers – 6th International Conference on Electrical, Electronic and Computing Engineering, IcETRAN 2019, Silver Lake, Serbia, June 03 – 06, 2019

Zbornik radova - 63. Konferencija za elektroniku, telekomunikacije, računarstvo, automatiku i nuklearnu tehniku, Srebrno jezero, 03 – 06. juna, 2019. godine

Main Editor / Glavni urednik
Dejan Popović

Editors / Urednici
Slobodan Vukosavić, Boris Lončar

Published by / **ETRAN Society, Belgrade, Academic Mind, Belgrade**
Izdavači / **Društvo za ETRAN, Beograd i Akademska misao, Beograd**

Production / Izrada
Academic Mind, Belgrade / Akademska misao, Beograd

Place and year of publication / Mesto i godina izdanja
Belgrade, 2019. / Beograd, 2019.

Circulation / Tiraž
200 primeraka / 200 copies

ISBN 978-86-7466-785-9

www.etran.rs

ORGANIZERS

ETRAN Society, Belgrade,
Faculty of Technology and Metallurgy, University of Belgrade, Belgrade, Serbia
School of Electrical Engineering, University of Belgrade.

Conferences were organized under the auspices of Ministry of Education, Science and Technological Development of the Republic of Serbia with the support of IEEE – Institute of Electrical and Electronics Engineers, USA and technical support of Academic MInd, Serbia.

ORGANIZING COMMITTEE ICETRAN 2019

Chair
Prof. Boris Loncar, Faculty of Technology and Metallurgy, Belgrade (Serbia)
Vice-chair
Prof. Slobodan Vukosavić, corresponding member SASA, Faculty of Electrical Engineering, Belgrade (Serbia)
Members
Prof. Boban Veselić, Faculty of Electronic Engineering, Niš (Serbia)
Prof. Milica Janković, Faculty of Electrical Engineering, Belgrade (Serbia)
Prof. Miroslav Lutovac, The Academy of Enginnering Sciences of Serbia (AESS), Belgrade (Serbia)
Prof. Dragan Mančić, Faculty of Electronic Engineering, Niš (Serbia)
Prof. Vera Marković, Faculty of Electronic Engineering, Niš (Serbia)
Prof. Ivan Milentijević, Faculty of Electronic Engineering, Niš (Serbia)
Prof. Milan Milosavljević, Singidunum University, Belgrade (Serbia)
PhD Dragana Nikolić, Institute Vinča, Belgrade (Serbia)
Prof. Zorica Nikolić, Faculty of Electronic Engineering, Niš (Serbia)
Prof. Vesna Paunović, Faculty of Electronic Engineering, Niš (Serbia)
Prof. Tatjana Pešić-Brđanin, University of Banja Luka (Bosnia and Hertzegovina)
Prof. Miroslav Popović, Faculty of Technical Science, Novi Sad (Serbia)
Prof. Aneta Prijić, Faculty of Electronic Engineering, Niš (Serbia)
Prof. Milan Rapaić, vice-chair, Faculty of Technical Science, Novi Sad (Serbia)
PhD Aleksandar Rodić, Full Research Professor, Institut Mihajlo Pupin, Belgrade (Serbia)
Prof. Marko Rosić, Faculty of Technical Sciences, Čačak (Serbia)
Prof. Lazar Saranovac, Faculty of Electrical Engineering, Belgrade (Serbia)
Prof. Aleksandra Smiljanić, Faculty of Electrical Engineering, Belgrade (Serbia)
Prof. Platon Sovilj, Faculty of Technical Sciences, Novi Sad (Serbia)
Prof. Petar Spalević, Faculty of technical Sciences, Kosovska Mitrovica (Serbia)
Assist. prof. Kovilja Stanković, Faculty of Electrical Engineering, Belgrade (Serbia)

Assist. prof. Miodrag Tasić, Faculty of Electrical Engineering, Belgrade (Serbia)
Prof. Dejan Ćirić, Faculty of Electronic Engineering, Niš (Serbia)

Support to promotion of the awarded papers

Prof. Milić Đekić
Faculty of Technical Sciences, Čačak (Serbia)

Local Organizing Committee of IcETRAN 2018

Antonijević Dunja, University of Belgrade, Serbia
Dimović Slavko, University of Belgrade, Serbia
Marković Maja, University of Belgrade, Serbia
Miljojčić Tatjana, University of Belgrade, Serbia
Pešić Ivan, University of Belgrade, Serbia
Radisavljević Anđela, University of Belgrade, Serbia
Rubinjoni Luka, University of Belgrade, Serbia
Ugrinović Vukašin, University of Belgrade, Serbia

Program and Technical support

Mirjana Jovanić, ETRAN
Zlatko Jarnević, ETRAN
Marko Vujadinović, Academic Mind, Belgrade
Aleksandar Rašković, Academic mind, Belgrade
Boban Milijić, Academic mind, Belgrade

This volume contains the papers presented at , 6th International Conference on Electrical, Electronic and Computing , (Ic)ETRAN 2019 in conjunction of ETRAN 63rd National Conference on Electrical, Electronic and Computing Engineering held on June 3-6, 2019 in Veliko Gradište.

There were 247 submissions. Each submission was reviewed by at least 1, and on the average 2, reviewers. The committee decided to accept papers as follows. Invited papers are included and presented as papers in Sessions.

During the Conferences were held three Special Sessions:

- **Special Session dedicated to academician Rajko Tomović**

Moderators:
Academician SASA Dejan Popović, Professor Emeritus Srđan Stanković and Prof. Srbijanka Turajlić

- **Special Session: New Materials**

Moderators:
Corresponding member SASA Velimir Radmilović: **Decoupling of Electrical and Thermal Properties in Nanostructured Materials;**

Prof. Vladimir Srdić: **Ultrafast Spin Dynamics in Multiferroic Oxides;**

Prof. Petar Uskoković: **Synthesis and Supercapacitive performances of Electrospun Carbon Nanofibers Decorated with Spinel $Co_{1.5}Mn_{1.5}O_4$ Nanocrystals**

- **Special Session: Centers of Excellence**

Moderators:
Corresponding Member SASA Branislav Jelenković, **Center of Excellence: Photonics Center at the Institute of Physics**

Prof. Đorđe Janaćković, **Centre of Research Excellence: Nanotechnology and Functional Materials Centre**

Dr Branko Matović, **Center of Excellence: Multifunctional Materials Centre at the "Vinča" Institute for Nuclear Sciences**

August 27, 2019

Belgrade

Boris Loncar

TABLE OF CONTENTS / SADRŽAJ

Acoustics / Akustika

Antennas and propagation / Antene i prostiranje

Control Systems / Automatika

Biomedical Engineering / Biomedicinska tehnika

Power Engineering / Elektroenergetika

Electric circuits and systems, and signal processing /
Električna kola, električni sistemi i obrada signala

Electronics / Elektronika

Metrology / Metrologija

Microwave technique, technologies and systems / Mikrotalasna tehnika, tehnologije i sistemi

New Materials in Electrical and Electronic Engineering / Novi Materijali

Computing and information engineering / Računarska tehnika i informatika

Telecommunications/ Telekomunikacije

Artificial Intelligence / Veštačka Inteligencija

SOUND EVENT DETECTION FROM PARTIALLY ANNOTATED DATA: TRENDS AND CHALLENGES

Romain Serizel, Nicolas Turpault

Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France

ABSTRACT

This paper proposes an overview of the latest advances and challenges in sound event detection and classification with systems trained on partially annotated data. The paper focuses on the scientific aspects highlighted by the task 4 of DCASE 2018 challenge: large-scale weakly labeled semi-supervised sound event detection in domestic environments. Given a small training set composed of weakly labeled audio clips (without timestamps) and a larger training set composed of unlabeled audio clips, the target of the task is to provide not only the event class but also the event time boundaries given that multiple events can be present in an audio clip. This paper proposes a detailed analysis of the impact of the time segmentation, the event classification and the methods used to exploit unlabeled data on the final performance of sound event detection systems.

Index Terms— Sound event detection, Weakly labeled data, Semi-supervised learning, Audio segmentation, DCASE 2018

1. INTRODUCTION

We are constantly surrounded by sounds and we rely heavily on these sounds to obtain important information about what is happening around us. Ambient sound analysis aims at automatically extracting information from these sounds. It encompasses disciplines such as sound scene classification (in which context does this happen?) or sound event detection and classification (SED) (what happens during this recording?) [1]. This area of research has been attracting a continuously growing attention during the past years as it can have a great impact in many applications including smart cities, autonomous cars or ambient assisted living.

DCASE 2018 task 4 (large-scale weakly labeled semi-supervised sound event detection in domestic environments) focused on SED with time boundaries in domestic applications [2]. The systems submitted had to detect when an sound event occurred in an audio clip and what was the class of the event (as opposed to audio tagging where only the presence of a sound event is important regardless of when it happened). We proposed to investigate the scenario where a large scale corpus is available but only a small amount of the data is labeled. Task 4 corpus was derived from the Audioset corpus [3] targeting classes of sound events related to domestic applications. The labels are provided at clip level (an event is present or not within a sound clip) but without the time boundaries (weak labels, that can also be referred to as tags) in order to decrease the annotation time. These constraints indeed correspond to constraints faced in many real applications where the budget allocated to annotating is limited.

In order to fully exploit this dataset, the submitted systems had to tackle two different problems. The first problem is related to the exploitation of the unlabeled part of the dataset either in unsupervised approaches [4, 5] or together with the labeled subset in semi-supervised approaches [6, 7, 8]. The second problem was related to the detection of the time boundaries and how to train a system that can detect these boundaries from weakly labeled data [9, 10]. The evaluation metric chosen was selected because it was penalizing these boundary estimation errors heavily. The goal was to encourage participants to focus on the time localization aspect.

Through a detailed overview of the systems submitted to DCASE 2018 task 4 we propose an overview of some recent advances in SED with partially annotated data[1]. We will first briefly describe task 4 and the related audio corpus in Section 2. Systems performance over all classes will be presented and analyzed in Section 3. We will present a class-wise analyze in Section 4 and discuss the impact of the metric chosen in Section 5. Section 6 will draw the conclusions of the paper and present some perspectives for SED.

This work was made with the support of the French National Research Agency, in the framework of the project LEAUDS Learning to understand audio scenes (ANR-18-CE23-0020) and the French region Grand-Est.Experiments presented in this paper were carried out using the Grid5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000)

[1]Additional result plots and analysis can be be found at `https://turpaultn.github.io/dcase2018-results/`

2. DCASE 2018 TASK 4

2.1. Audio dataset

The task focuses on a subset of Audioset that focuses on 10 classes of sound events [2]. Audioset consists in 10-second audio clips extracted from youtube videos[3]. The development set provided for task 4 is split into a training set and a test set.

2.1.1. Training set

In order to reflect what could possibly happen in a real-world scenario, we provide three different splits of training data in task 4 training set: a labeled training set, an unlabeled in domain training set and an unlabeled out of domain training set (clips that do not contain any of the target classes):

Labeled training set: contains 1,578 audio clips (2,244 class occurrences) for which weak labels provided in Audioset have been verified and corrected by human annotators. One-third of the audio clips in this set contain at least two different classes of sound events.

Unlabeled in domain training set: contains 14,412 audio clips. The audio clips are selected such that the distribution per class of sound event (based on Audioset labels) is close to the distribution in the labeled set.

Unlabeled out of domain training set: is composed of 39,999 audio clips extracted from classes of sound events that are not considered in the task (according to unverified Audioset labels).

2.2. Test set

The test set is designed such that the distribution in term of clips per class of sound event is similar to that of the weakly labeled training set. The test set contains 288 audio clips (906 events). The test set is annotated with strong labels, with time boundaries (obtained by human annotators).

2.3. Evaluation set

The evaluation set contains 880 audio clips (3,187 events). The process to select the audio clips was similar to the process applied to select audio clips in the training set and the test set, in order to obtain a set with comparable classes distribution (See also Table 1). Labels with time boundaries are obtained by human annotators.

The duration distribution for each sound event class is presented on Figure 1. One of the focus of this task is the development of approaches that can provide fine time-level segmentation while learning on weakly labeled data. The observation of the event duration distribution confirms that in order to perform well it is essential to design approaches that are efficient at detecting both short events and events that have a longer duration.

Class	Test	Eval
Alarm/bell/ringing	112	306
Blender	40	56
Cat	97	243
Dishes	122	370
Dog	127	450
Electric shaver/toothbrush	28	37
Frying	24	67
Running water	76	154
Speech	261	1401
Vacuum cleaner	36	56
Total	906	3187

Table 1: Number of sound events per class in the test set and the evaluation set.

2.4. Task description

The task consists of detecting sound events within web videos using weakly labeled training data. The detection within a 10-seconds clip should be performed with start and end timestamps.

2.4.1. Task evaluation

Submissions were evaluated with event-based measures for which the system output is compared to the reference labels event by event [11] (see also Figure 2). The correspondence between sound event boundaries are estimated with a 200 ms tolerance collar on onsets and a tolerance collar on offsets that is the maximum of 200 ms and 20 % of the duration of the sound event.

- True positives are the occurrences when a sound event present in the system output corresponds to a sound event in the reference annotations.
- False positives are obtained when a sound event is present in the system output but not in the reference annotations (or not within the tolerance collars on the onset or the offset).
- False negatives are obtained when a sound event is present in the reference annotations but not in the system output (or not within the tolerance collars).

Submissions were ranked according to the event-based F1-score. The F1-score was first computed class-wise over the whole evaluation set:

$$F1_c = \frac{2TP_c}{2TP_c + FP_c + FN_c}, \quad (1)$$

where TP_c, FP_c and FN_c are the number of true positives, false positives and false negative for sound event class c over the whole evaluation set, respectively.

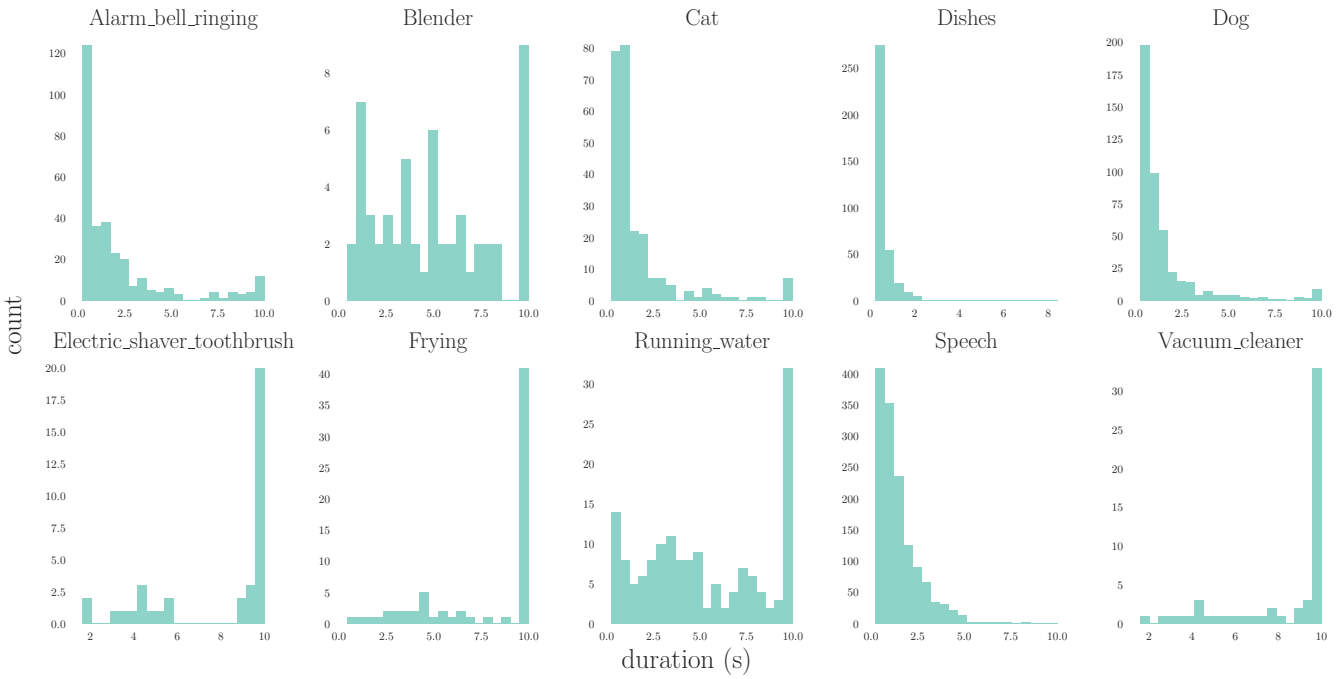


Fig. 1: Duration distribution by class of sound events on the evaluation set.

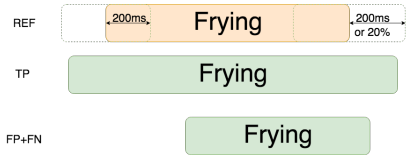


Fig. 2: Event-based F1-score.

The final score is the F1-score average over sound event classes regardless of the number of sound events per class (macro-average):

$$F1_{\text{macro}} = \frac{\sum_{c \in \mathcal{C}} F1_c}{n_{\mathcal{C}}}, \tag{2}$$

where $\mathcal{C}$ is the sound event classes ensemble and $n_{\mathcal{C}}$ the number of sound event classes.

3. ANALYSIS OF THE PERFORMANCE OVER ALL SOUND EVENT CLASSES

In this section we present and analyze submissions performance regardless of the sound event classes.

3.1. Task submissions and results overview

DCASE 2018 task 4 gathered 50 submissions from 16 different research teams involving 57 researchers overall. The official team ranking and some characteristics of the submitted systems are presented in Table 2. The best two submissions quite clearly stand out from other submissions. They also go beyond the rather standard approaches based convolutional neural networks (CNN) or stacked CNN and recurrent neural networks (RNN) also denoted as CRNN. The best system, submitted by JiaKai (**jiakai_psh**) [12], relies on a mean-teacher model that exploits unlabeled data to regularize the classifier learned on the weakly labeled data [28]. The system submitted by Liu et al. (**liu_ustc**) [13] that ranked second relies on an energy based sound event detection as a pre-processing to a capsule network [29]. The output of the network is then post processed to ensure that silence between events and events themselves are longer than a minimum duration.

Other notable submissions include the system from Kothinti et al. (**kothinti_jhu**) [15] that relies on a sound event detection based on restricted Boltzmann machines (RBM) as a pre-processing. This solution performs well at detecting onsets but not so much for offset detection (see also Section 4.1). Dinkel et al. proposed a system (**dinkel_sjtu**) that uses Gaussian mixture models (GMM) and hidden Markov models (HMM) to perform sound event alignment [25]. Gaussian filtering is then used as post-processing. Pellegrini et al. proposed a system (**pellegrini_irit**) that relies on multiple instance learning (MIL) to exploit weakly labeled data [23]. Both these systems perform pretty decently on segmentation (see also Section 3.2) but they suffer from pretty poor sound event classification performance (see also Figure 8).

Rank	System	Features	Classifier	Parameters	F1 (%)
1	**jiakai_psh [12]**	log-mel energies	CRNN	1M	**32.4**
2	**liu_ustc [13]**	log-mel energies	CRNN, Capsule-RNN	4M	**29.9**
3	**kong_surrey [14]**	log-mel energies	VGGish 8 layer CNN	4M	**24.0**
4	**kothinti_jhu [15]**	log-mel energies, auditory spectrogram	CRNN, RBM, cRBM, PCA	1M	**22.4**
5	**harb_tug [16]**	log-mel energies	CRNN, VAT	497k	**21.6**
6	**koutini_jku [17]**	log-mel energies	CRNN	126k	**21.5**
7	**guo_thu [18]**	log-mel energies	multi-scale CRNN	970k	**21.3**
8	**hou_bupt [19]**	log-mel energies & MFCC	CRNN	1M	**21.1**
9	**lim_etri [20]**	log-mel energies	CRNN	239k	**20.4**
10	**avdeeva_itmo [21]**	log-mel energies & MFCC	CRNN, CNN	200k	**20.1**
11	**wangjun_bupt [22]**	log-mel energies	RNN	1M	**17.9**
12	**pellegrini_irit [23]**	log-mel energies	CNN, CRNN with MIL	200k	**16.6**
13	**moon_yonsei [24]**	Raw wavforms	RseNet, SENet	10M	**15.9**
14	**dinkel_sjtu [25]**	log-mel energies & MFCC	CRNN, HMM-GMM	126k	**13.4**
15	**wang_nudt [26]**	log-mel energies & Δ features	CRNN	24M	**12.6**
	baseline [2]	log-mel energies	CRNN	126k	**10.8**
16	**raj_iit [27]**	CQT	CRNN	215k	**9.4**

Table 2: Team ranking and submitted systems characteristics.

3.2. Segmentation

In this section, we focus on the segmentation performance. That is, the ability of the submitted systems to localize sound events in time without having to predict the class. Figures 3, 4 and 5 present the event-based F1-score computed without taking the sound event class labels into account and for a tolerance collar of 200 ms, 1 s and 5 s, respectively. The fact that there is only little performance difference between the sound event detection performance (Table 2) and the segmentation performance tends to indicates that segmentation is possibly the main limiting factor in overall performance. This is actually confirmed by the rather high tagging performance of most systems presented on Figure 8.

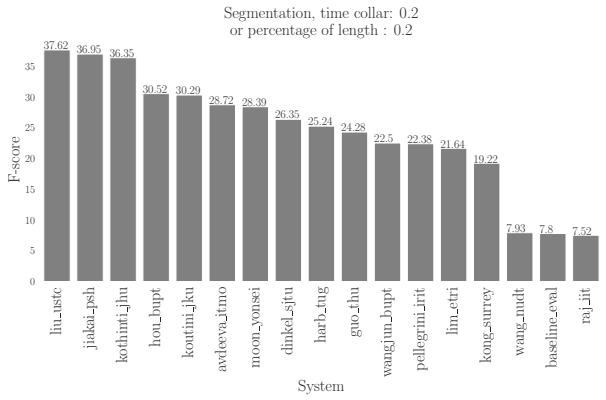


Fig. 3: Segmentation performance (tolerance collar on onsets is 200 ms and tolerance collar on offsets is the maximum of 200 ms and 20 % of the event length).

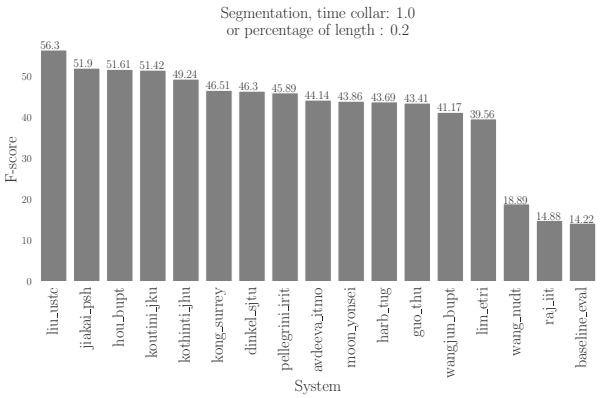


Fig. 4: Segmentation performance (tolerance collar on onsets is 1 s and tolerance collar on offsets is the maximum of 1 s and 20 % of the event length).

Currently, most of the systems are able to detect if an event occurred within a rather crude time area (see also Figure 5 but are not able to properly segment the audio clips in terms of sound events (see also Figure 3). The systems that performed best in terms of segmentation are the systems that actually implemented some sort of segmentation among which **liu_ustc** [13] and **kothinti_jhu** [15]. The winning system is ranked second in term of segmentation and owe its first overall rank to a much better classification than competing systems (see also Figure 8).

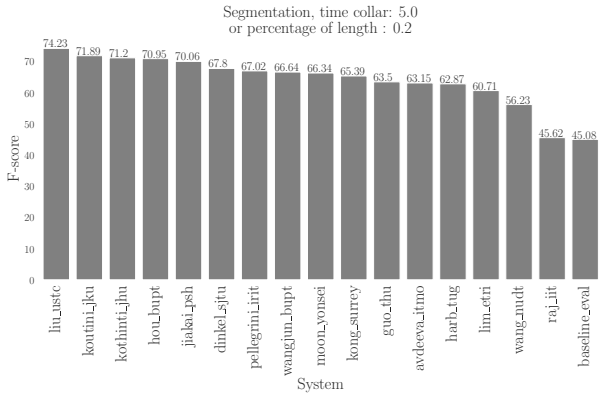


Fig. 5: Segmentation performance (tolerance collar on onsets is 5 s and tolerance collar on offsets is the maximum of 5 s and 20 % of the event length).

3.3. Use of unlabeled data

One of the challenges proposed by DCASE 2018 task 4 was to exploit a large amount of unlabeled data. In the section we analyze the approaches proposed by participants. Most of the systems submitted used a pseudo-labeling approach where a first system trained on the labeled data is used to obtain labels for the unlabeled set (**liu_ustc**) [13], **hou_bupt** [19]). Variations on this included setting a confidence threshold to decide to keep the label or not (**koutini_jku** [17], **wang_nudt** [26], **pellegrini_irit** [23], **harb_tug** [16], **moon_yonsei** [24]) and gradually introducing new audio clips with these pseudo labels (**wangjun_bupt** [22]).

The winning system (**jiakai_psh** [12]) used the unlabeled data within a mean-teacher scheme [28]. It is composed of two models: a student model and a mean-teacher model whose weights are the exponential average of the student's weights. On labeled data, the student model weights are updated to optimize a classification cost on the sound event classes. Additionally, consistency costs are computed to compare the output of the student model and the mean-teacher model on both the labeled and the unlabeled data. Kothinti et al. (**kothinti_jhu** [15]) proposed to use both the weakly labeled and unlabeled in-domain data to train several RBM that are used to detect sound event boundaries.

3.4. Complexity

The complexity of the submitted systems (in terms of number of parameters) is presented in Table 2. The only system that used raw waveforms as input (**moon_yonsei** [24]) is among the most complex systems yet it is not even among the top 10 systems. This tends to indicate that the dataset proposed for task 4 is too small to train SED systems using raw waveforms that are usually known to require a lot of training data. The most complex system (**wang_nudt** [26]) is about 200 times more complex than the baseline in particular because it combines several complex models. However it performs only slightly better than the baseline. The winning system (**jiakai_psh** [12]) is about 10 times more complex than the baseline and the best performing system that has a number of parameters similar to that of the baseline (**koutini_jku** [17]) improves the baseline F1-score performance by more than 10 % absolute.

3.5. Duration of events

It has been shown above that the systems performance largely depends on the systems ability to properly segment the audio clips in terms of sound events. Figure 1 presents the duration distribution for each class of sound events on the evaluation set. From this distribution we can separate the sound events into two categories of events: short sound events ('Alarm/bell/ringing', 'Cat', 'Dishes', 'Dog' and 'Speech") and long sound events ('Blender', 'Electric shaver/toothbrush', 'Frying', 'Running water' and 'Vacuum cleaner').

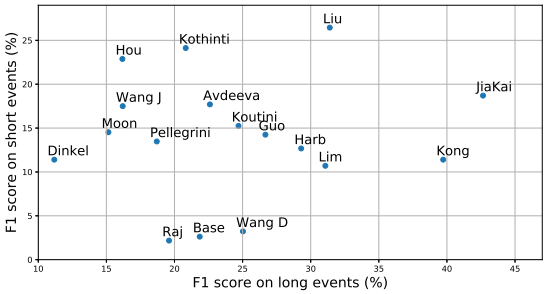


Fig. 6: Systems performance on short sound events depending on their performance on long sound events.

System	Short	Long	All	Rank
liu_ustc [13]	**26.4**	31.4	29.9	2
kothinti_jhu [15]	**24.1**	20.8	22.4	4
hou_bupt [19]	**22.9**	16.2	21.1	8
jiakai_psh [12]	**18.7**	42.6	32.4	1
avdeeva_itmo [21]	**17.7**	22.6	20.1	10
baseline [2]	**2.6**	21.8	10.8	16

Table 3: Top 5 systems on short events ('Alarm/bell/ringing', 'Cat', 'Dishes', 'Dog' and 'Speech").

Figure 6 presents the performance of the submitted systems on short sound events depending on their performance on long sound events. No system is clearly outperforming the others on both short and long sound events. This is confirmed when looking at the top performing systems on short sound events (Table 3) and on long sound events (Table 4). These rankings tend to show that the approaches proposed

System	Long	Short	All	Rank
jiakai_psh [12]	**42.6**	18.7	32.4	1
kong_surrey [14]	**39.7**	11.4	24	3
liu_ustc [13]	**31.4**	26.4	29.9	2
lim_etri [20]	**31.1**	10.7	20.4	9
harb_tug [16]	**29.3**	12.7	21.6	5
baseline [2]	21.8	**2.6**	10.8	16

Table 4: Top 5 systems on long events ('Blender', 'Electric shaver/toothbrush', 'Frying', 'Running water' and 'Vacuum cleaner').

were either tailored to perform well on short sound events (top systems are also the systems that performed best in terms of segmentation, see also Figure 3) or on long sound events (top systems are also among the best systems in terms of tagging, see also Figure 8). However, in order to perform well on the SED task systems had to perform reasonably well on both short and long sound events. This is the case for the top two systems (**jiakai_psh** [12] and **liu_ustc**) [13]) that are in the top five both short and long sound events.

4. ANALYSIS OF THE CLASS-WISE PERFORMANCE

It have been shown above that systems performance can vary to a great extent depending on the sound events duration that is tightly related to the sound event class itself. Therefore, in this section we focus on the performance of the submitted systems depending on the sound event classes. Table 5 presents the class-wise event-based F1-score for the 10 best performing submitted systems. The best system (**jiakai_psh** [12]) outperforms other systems on five sound event classes upon ten (mainly long sound events). However, it performs rather poorly on some of the remaining sound event classes (mainly short sound events). On the other hand, the second best system (**liu_ustc** [13]) outperforms other systems on a single sound event class ('Dog') but is generally not too far from the best performance on several other sound event class. This explains why it can still compare with the winning system in terms of overall performance.

In general 'Speech' and 'Alarm bell ringing' seem to be the easiest sound event classes to detect and classify. This could be explained by the fact that sound events from these classes are not too short (with a median duration of 1.17 s and 0.57 s, respectively), occurs many times in the training set (in 550 clips and 205 clips, respectively) and generally have rather clear onsets and offsets (see also Section 4.1). There is a clear separation between 'Cat', 'Dishes' and 'Dog' and other sound event classes. The former seems more difficult to detect and classify than the latter. This can be due to the fact that sound events in these classes are short and present a large acoustic variability. Interestingly, the submitted systems that perform best on these sound event classes are not necessarily among the top three systems. For example **hou_bupt** [19] obtains the best performance on 'Dishes' and clearly outperforms other submissions with 23.5 % F1-score. However, it ranked eighth overall (but was among the top five systems on short sound events, see also Table 3). The best system on 'Cat' (by a rather large margin) with 25.3 % F1-score is **pellegrini_irit** [23] that relies on MIL and that is not even in the top 10 in terms of overall performance.

4.1. Performance on onset and offset detection

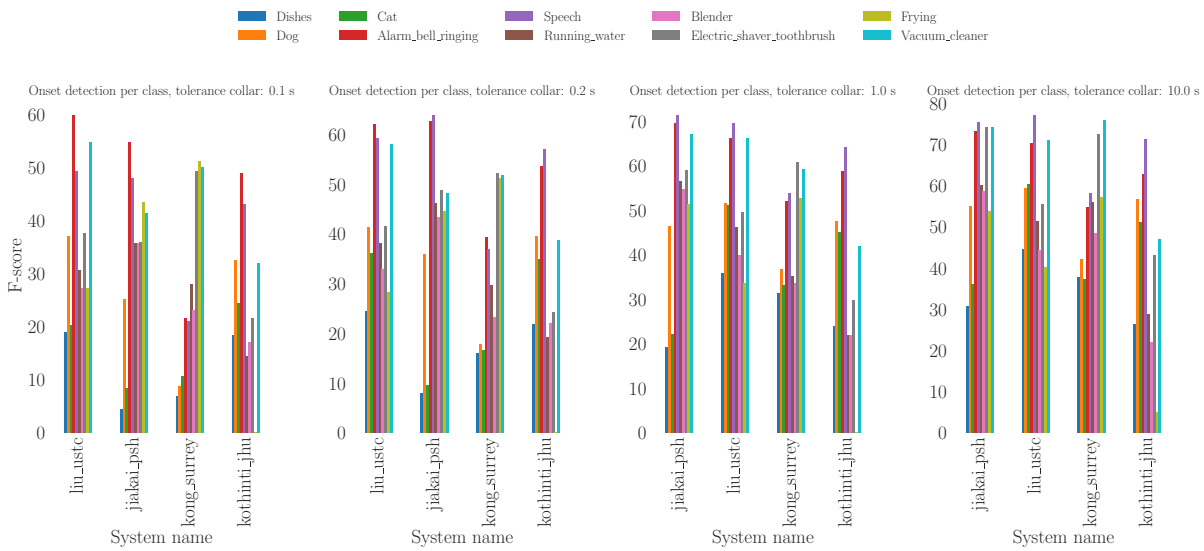
For some sound event classes that slowly decay the time location of offsets can be difficult to locate (and the concept of offset itself can even become ambiguous in reverberant scenarios). Therefore, we now focus on the detection of onsets and offsets separately. In the plots presented in this section (see also Figure 7), sound events are classified from the shortest (on the left) to the longest (on the right) according to their median duration. Additionally, for the sake of clarity, only the systems among the top four in overall performance are presented here. Systems are presented in decaying overall onset or offset detection performance (the best system is on the left side).

4.1.1. Onset

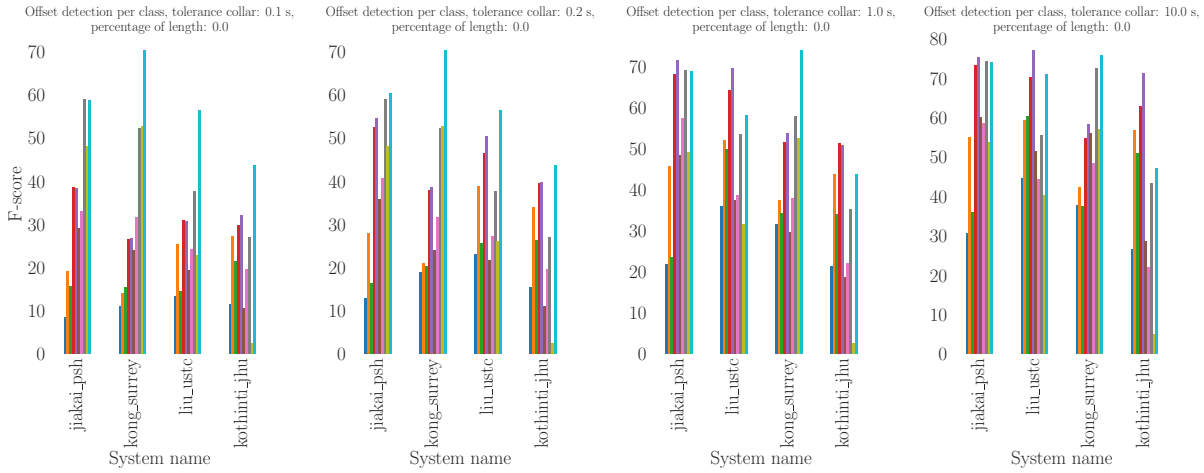
Figure 7a presents F1-score for onset detection for varying tolerance collars (in seconds). Performance generally increases when the tolerance collar is increased. For small tolerance collars, **liu_ustc** [13] performs best which confirms previous analysis about the relatively good segmentation of their system. When the tolerance collar is larger than 0.5 s **jiakai_psh** [12] outperforms other system which also confirm that the proposed segmentation is a bit too coarse.

The remaining errors for a 10 sec tolerance collar indicate that the systems were not able to predict how many onsets for the specific sound event class occurred within the audio clip. In most cases this could also corresponds to the case where the sound event was not detected at all (see also Figure 7b).

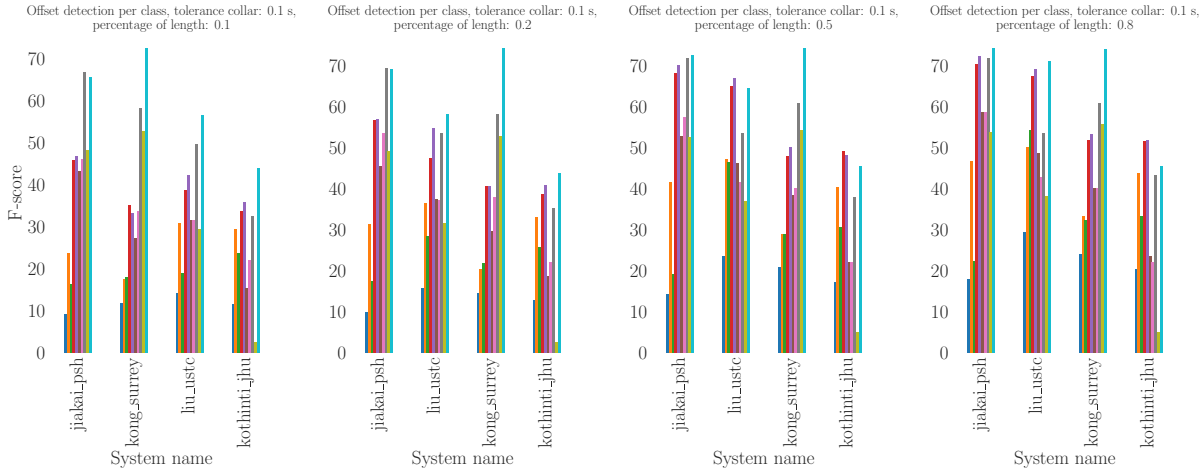
When looking at particular sound event classes, in general systems exhibit good onset detection performance for 'Speech' and 'Alarm bell ringing'. As mentioned above, this can be due to the fact that these sound events occur frequently in the training set but it can also be related to the fact that the sound events from these classes indeed have rather clear onsets that appear to be easier to detect. On the other hand, sound event classes as 'Cat' and 'Dishes' seem to be difficult to detect. For the former it is probably due to the fact that the onsets are not always clear as for the latter it is most generally related to sound events that are simply missed by the systems because they are too short. For the remaining sound event classes, the performance varies a lot from one system to another and seems to be affected by the segmentation strategy implemented.



(a) F1-score for onset detection with absolute tolerance collars.



(b) F1-score for offset detection with absolute tolerance collars.



(c) F1-score for offset detection with tolerance collars relative to event duration.

Fig. 7: Event-based F1-score for onset and offset detection with varying tolerance.

System	Alarm	Blender	Cat	Dishes	Dog	Electric	Frying	Water	Speech	Vacuum
jiakai_psh [12]	**49.9**	**38.2**	3.6	3.2	18.1	**48.7**	35.4	**31.2**	**46.8**	48.3
liu_ustc [13]	46.0	27.1	20.3	13.0	**26.5**	37.6	10.9	23.9	43.1	50.0
kong_surrey [14]	24.5	18.9	7.8	7.7	5.6	46.4	**43.6**	15.2	19.9	50.0
kothinti_jhu [15]	36.7	22.0	**20.5**	12.8	**26.5**	24.3	0.0	9.6	34.3	37.0
harb_tug [16]	15.4	30.0	8.1	17.5	9.7	21.0	34.7	17.3	31.1	31.5
koutini_jku [17]	30.0	16.4	13.1	9.5	8.4	23.5	18.1	12.6	42.9	40.8
guo_thu [18]	35.3	31.8	7.8	4.0	9.9	17.4	32.7	18.3	31.0	24.8
hou_bupt [19]	41.4	16.4	6.4	**23.5**	20.2	9.8	6.2	14.0	40.6	32.3
lim_etri [20]	11.6	21.6	7.9	5.9	17.4	27.8	14.9	15.5	21.0	**60.0**
avdeeva_itmo [21]	33.3	15.2	14.9	6.3	16.3	15.8	24.6	13.3	27.2	34.8
baseline [2]	4.8	12.7	2.9	0.4	2.4	20.0	24.5	10.1	0.1	30.2

Table 5: Class-wise event-based F1-score for the top 10 submitted systems.

4.1.2. Offset

Figure 7b presents F1-score for offset detection for varying tolerance collars (in seconds). When comparing with Figure 7a it appears that offsets are indeed more difficult to detect. The high F1-score for some sound event classes such as ('Electric shaver toothbrush', 'Frying' or 'Vacuum cleaner') is mainly due to the fact that many of the sound events in these classes do not have an offset within the audio clips and therefore the offset to be detected is simply the final boundary of the audio clip.

It is generally admitted that penalizing offset detection based on an absolute time tolerance collar is not a reasonable choice specially for long sound events. In particular because this type of tolerance collar might be affecting long sound events (with longer decay) much more than short (possibly percussive) sound events. Therefore, the metric retained for DCASE 2018 task 4 include both an absolute time tolerance collar and a tolerance collar that was computed as a percentage of the sound event duration (the maximum of these two values was retained). With this approach, the absolute time tolerance collar usually applies to short sound events while the tolerance collar relative to event length applies to longer sound events.

Figure 7c presents F1-score for offset detection for varying tolerance collars (in percent of the sound event duration). Note that the absolute time tolerance collar is kept to 0.1 s here in order to avoid unreasonably small tolerance collars for short sound events. As expected, this kind of tolerance collar has less effect than absolute time tolerance collar on offset detection of short sound events such as 'Cat', 'Dishes' or 'Dog' but can affect greatly the offset detection performance on long sound events such as 'Running water' or 'Blender'.

Quite surprisingly, **jiakai_psh** [12] outperforms the other submitted systems (even those which had demonstrated a better segmentation performance until now) including with low tolerance collars. When looking at particular sound event classes, in general the submitted systems exhibit good offset detection performance for 'Speech' and 'Alarm bell ringing' even if in this case offsets are usually not as well defined as onsets were.

5. IMPACT OF THE METRIC

For DCASE 2018, the F1-score was computed in an event-based fashion in order to put on strong focus on the sound event segmentation. Class-wise performance was averaged in order to discard the effects of the sound event classes imbalance (2). In this section, we study the impact of these choices on the performance evaluation of the submitted systems.

5.1. F1-score computation relatively to events or segments

As opposed to event-based metrics, segmented-based metrics are computed by comparing the system outputs and the reference on short segments. The sound event classes are then considered to be active or not on the full segment. The final metric is computed on all the segments [11]. This approach reports if a system is able to detect if a sound event class is active with a specific time resolution (the segment length) and can prove more robust than event-based metrics to phenomena such as short pauses between consecutive sound events. Figure 8 presents a comparison between the event-based F1-scores (on the left) and the segment-based F1-scores (on the right) for varying tolerance collars and time resolutions, respectively.

As expected, segmented-based metrics are more permissive to errors in the detection of the sound event boundaries. Indeed the reported segment-based F1-scores (from 40 % to 70 % depending on the time resolution) are much higher than their event-based counterpart (from 5 % to 60 % depending on the tolerance collar). Additionally, the segment-based F1-score seems to be favoring systems that are good at tagging while event-based F1-score favors systems that have good segmentation performance. This is particularly clear for sys-

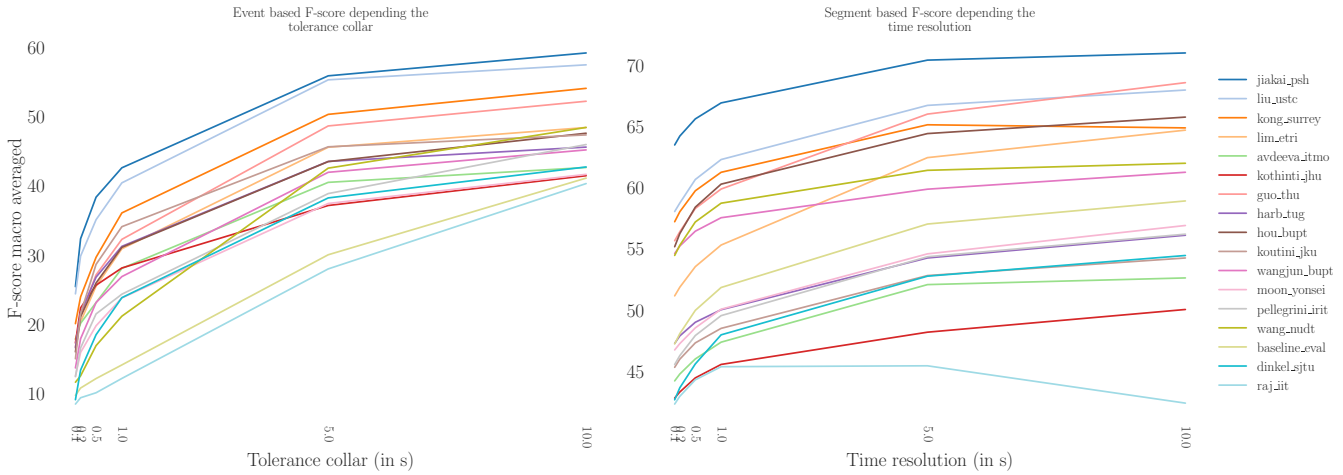


Fig. 8: Comparison between event-based and segmented-based F1-scores depending on the tolerance collar and time resolution, respectively.

tems like **hou_bupt** [19], **guo_thu** [18] and the task baseline [2] which perform much better in terms of segment-based F1-score and for **kothinti_jhu** [15] that performs much better in terms of event-based F1-score.

When the time resolution for the segment-based F1-scores is 10 s the reported performance is actually that of a tagging task. The tagging ranking is then rather different than the general ranking (see also Table 2) and the ranking for segmentation (see also Figure 3). This emphasizes once again that none of the submitted systems is actually outperforming others in both segmentation and tagging but that in order to perform well on the task, systems had to perform at least decently on both. This is the case for **jiakai_psh** [12] and **liu_ustc** [13] that clearly stand out in the final ranking.

As the choice of the metric is tightly related to the targeted application, some approaches can be better suited when you need to know exactly when a sound event from a specific class did occur (in which case you might select a system that performs well in terms event-based F1-score) some other approaches can be suited to monitor the activity within a time period (approximately when was each sound event class active, depending on the time resolution, in which case we might select systems that perform well in terms segment-based F1-score)

5.2. Micro average

While macro-averaging (used in task 4) computes the final F1-score as the average across sound event classes (regardless of the number of events for each class), micro-averaging computes the final F1-score as the average of each single decision. It therefore gives more importance to sound event classes that occur more frequently (see also Table 1 for the distribution). For example, 'Speech' events will account for almost half of the performance when using micro-averaged F1-score.

Figure 9 presents event-based F1-score depending on the averaging method. We can observe a clear performance improvement between macro-averaged and micro-average F1-score for the systems that performed well the most frequent sound event classes ('Alarm bell ringing', 'Dishes', 'Dog' or 'Speech') such as **lim_etri** [20]. One the other hand the systems that were able to perform well on less frequent sound event classes ('Electric shaver/toothbrush', 'Frying'...) but not on frequent sound event classes can see their performance decreased between macro-averaged and micro-averaged F1-score as this is the case for **kong_surrey** [14]. The top two systems (**jiakai_psh** [12] and **liu_ustc**) [13]) were performing reasonably well on the most frequent sound event classes and therefore still outperform other systems in terms of micro-averaged F1-score.

Once again, the choice of the metric is related to the targeted application. If you want to detect mainly the sound event classes that occur the most frequently and that missing rare sound event classes is not really a problem then you should select approaches that perform well in terms of micro-averaged F1-score. On the contrary if detecting rare sound event classes is important then approaches that perform well in terms of macro-averaged F1-score seem better suited.

6. CONCLUSION

In this paper we proposed an overview of some of the latest advances and challenges in sound event detection with systems trained on partially annotated data through the analysis of the results of DCASE 2018 challenge task 4. The paper focused on the scientific aspects highlighted by the task: exploiting both unlabeled and weakly labeled data to train a system that provides not only the event class but also the event time boundaries. It has been shown that both the segmen-

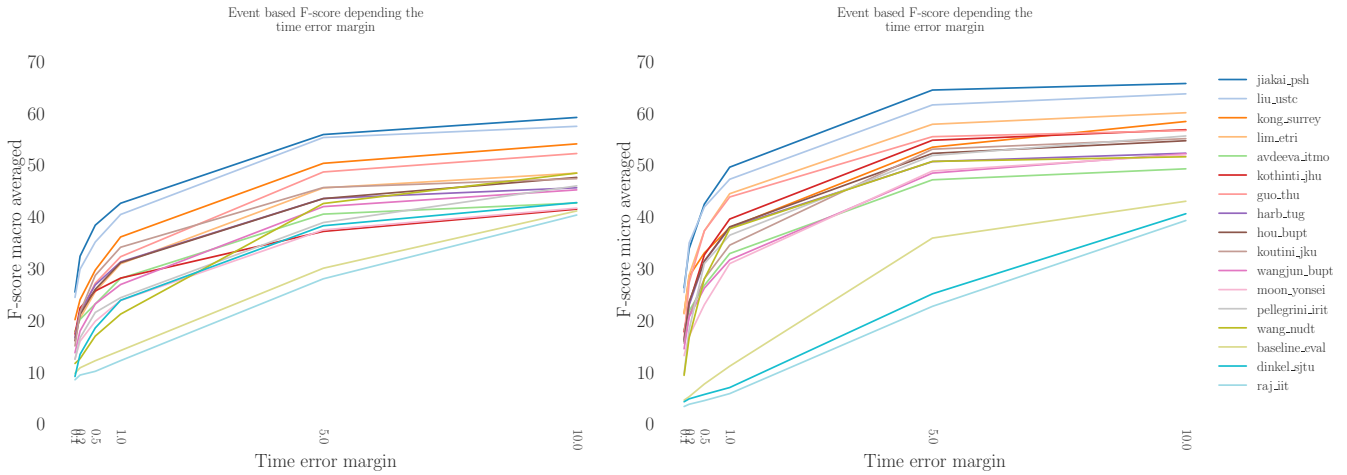


Fig. 9: Event-based F1-score depending on the class averaging method.

tation and the classification ability play an important role in the final performance. However whereas the tagging performance (related to the classification ability) is generally rather good for many systems, only few systems did implement an explicit segmentation strategy. This aspect actually remains quite challenging as training a system to detect sound events and predict their time localization from weakly labeled data is far from trivial. Therefore, one question for future works is to investigate if strongly labeled data that is generated synthetically can help solving this issue. This is one of the challenges investigated in the task 4 of DCASE 2019 challenge.

7. ACKNOWLEDGMENT

The authors would like to thank the other organizers of DCASE 2018 task 4 (Hamid Eghbal-zadeh from Johannes Kepler University – Austria and Ankit Parag Shah from Carnegie Mellon University –United States) as well as all participants to the task. They also would like to thank the DCASE 2018 challenge organization team (Toni Heittola, Annamaria Mesaros and Tuomas Virtanen from Tampere University of Technology – Finland) for they support while organizing the task.

8. REFERENCES

[1] Tuomas Virtanen, Mark D Plumbley, and Dan Ellis, *Computational analysis of sound scenes and events*, Springer, 2018.

[2] Romain Serizel, Nicolas Turpault, Hamid Eghbal-Zadeh, and Ankit Parag Shah, "Large-scale weakly labeled semi-supervised sound event detection in domestic environments," in *Proc. DCASE2018*, November 2018, pp. 19–23.

[3] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. ICASSP*, 2017.

[4] J. Salamon and J. P. Bello, "Unsupervised feature learning for urban sound classification," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[5] Aren Jansen, Manoj Plakal, Ratheet Pandya, Dan Ellis, Shawn Hershey, Jiayang Liu, Channing Moore, and Rif A. Saurous, "Unsupervised learning of semantic audio representations," in *Proc. ICASSP*, 2018.

[6] Z. Zhang and B. Schuller, "Semi-supervised learning helps in sound event classification," in *Proc. ICASSP*, 2012, pp. 333–336.

[7] Tatsuya Komatsu, Takahiro Toizumi, Reishi Kondo, and Yuzo Senda, "Acoustic event detection method using semi-supervised non-negative matrix factorization with a mixture of local dictionaries," in *Proc. DCASE)*, 2016, pp. 45–49.

[8] B. Elizalde, A. Shah, S. Dalmia, M. H. Lee, R. Badlani, A. Kumar, B. Raj, and I. Lane, "An approach for self-training audio event detectors using web data," in *Proc. EUSIPCO*, 2017, pp. 1863–1867.

[9] A. Kumar and B. Raj, "Audio event detection using weakly labeled data," *CoRR*, vol. abs/1605.02401, 2016.

[10] A. Kumar and B. Raj, "Audio event and scene recognition: A unified approach using strongly and weakly labeled data," in *Proc. IJCNN*. IEEE, 2017, pp. 3475–3482.

[11] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, pp. 162, May 2016.

[12] Lu JiaKai, "Mean teacher convolution system for dcase 2018 task 4," Tech. Rep., DCASE2018 Challenge, September 2018.

[13] Yaming Liu Liu, Jie Yan, Yan Song, and Jun Du, "Ustc-nelslip system for dcase 2018 challenge task 4," Tech. Rep., DCASE2018 Challenge, September 2018.

[14] Qiuqiang Kong, Iqbal Turab, Xu Yong, Wenwu Wang, and Mark D. Plumbley, "DCASE 2018 challenge baseline with convolutional neural networks," Tech. Rep., DCASE2018 Challenge, September 2018.

[15] Sandeep Kothinti, Keisuke Imoto, Debmalya Chakrabarty, Sell Gregory, Shinji Watanabe, and Mounya Elhilali, "Joint acoustic and class inference for weakly supervised sound event detection," Tech. Rep., DCASE2018 Challenge, September 2018.

[16] Robert Harb and Franz Pernkopf, "Sound event detection using weakly labeled semi-supervised data with gcrnns, vat and self-adaptive label refinement," Tech. Rep., DCASE2018 Challenge, September 2018.

[17] Khaled Koutini, Hamid Eghbal-zadeh, and Gerhard Widmer, "Iterative knowledge distillation in r-cnns for weakly-labeled semi-supervised sound event detection," Tech. Rep., DCASE2018 Challenge, September 2018.

[18] Yingmei Guo, Mingxing Xu, Jianming Wu, Yanan Wang, and Keiichiro Hoashi, "Multi-scale convolutional recurrent neural network with ensemble method for weakly labeled sound event detection," Tech. Rep., DCASE2018 Challenge, September 2018.

[19] Yuanbo Hou and Shengchen Li, "Semi-supervised sound event detection with convolutional recurrent neural network using weakly labelled data," Tech. Rep., DCASE2018 Challenge, September 2018.

[20] Wootaek Lim, Sangwon Suh, and Youngho Jeong, "Weakly labeled semi-supervised sound event detection using crnn with inception module," Tech. Rep., DCASE2018 Challenge, September 2018.

[21] Anastasia Avdeeva and Iurii Agafonov, "Sound event detection using weakly labeled dataset with convolutional recurrent neural network," Tech. Rep., DCASE2018 Challenge, September 2018.

[22] Wang Jun and Li Shengchen, "Self-attention mechanism based system for dcase2018 challenge task1 and task4," Tech. Rep., DCASE2018 Challenge, September 2018.

[23] leo Cances, Thomas Pellegrini, and Patrice Guyot, "Sound event detection from weak annotations: Weighted gru versus multi-instance learning," Tech. Rep., DCASE2018 Challenge, September 2018.

[24] Moon Hyeongi, Byun Joon, Kim Bum-Jun, Jeon Shin-hyuk, Jeong Youngho, Park Young-cheol, and Park Sung-wook, "End-to-end crnn architectures for weakly supervised sound event detection," Tech. Rep., DCASE2018 Challenge, September 2018.

[25] Heinrich Dinkel, Yanmin Qiand, and Kai Yu, "A hybrid asr model approach on weakly labeled scene classification," Tech. Rep., DCASE2018 Challenge, September 2018.

[26] Dezhi Wang, Kele Xu, Boqing Zhu, Lilun Zhang, Yuxing Peng, and Huaimin Wang, "A crnn-based system with mixup technique for large-scale weakly labeled sound event detection," Tech. Rep., DCASE2018 Challenge, September 2018.

[27] Rojin Raj, Shefali Waldekar, and Goutam Saha, "Large-scale weakly labelled semi-supervised cqt based sound event detection in domestic environments," Tech. Rep., DCASE2018 Challenge, September 2018.

[28] Antti Tarvainen and Harri Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. NIPS*, 2017, pp. 1195–1204.

[29] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton, "Dynamic routing between capsules," in *Proc. NIPS*, 2017, pp. 3856–3866.

Visualization and optimization of features in classification of motor sound

Ana Đorđević, Dejan Ćirić and Marko Ličanin

Abstract— Sound waves conduct a lot of important information about sound sources and environment, from individual physical events to sound scenes as a whole. Hitherto, in sound analysis and processing, two or three dimensional plots have been sufficient to correctly represent needed information, e.g. frequency response plot or spectrogram plot. With emerging machine and deep learning technologies, choosing the right technique for visualization of multidimensional matrices is of great significance for abstracting out the right information, understanding and interpreting the results clearly and easily. In this paper, various options for representation of acoustic features are analyzed in order to improve machine learning model for classification. This analysis is based on audio signals (sounds) of DC motors. Various types of visualizations have been used in order to optimize feature vector for classification.

Index Terms—acoustic features, features visualization and optimization, machine learning, DC motor sounds.

I. Introduction

A sound event is a segment of audio that a listener can consistently label and distinguish in an acoustic environment [1]. Analysis of a large variety of sounds requires calculation of a larger number of parameters from sound signals, and usage of automatic methods like machine learning [2] to differentiate between various types of audio signals. The applications of such automatic sound event detection are various. Any embedded system with capability to record sound signals can become more aware of sound in its environment [3]. In industrial and environmental surveillance systems, smart homes and cities, devices can automatically detect and/or classify sound events of interest.

Automatic annotation of sound can significantly improve for example, quality control and predictive maintenance in various industries. Here, break down or faulty product (part) can be detected in an automated way. By analyzing a large variety of sounds of a particular industrial product in good and bad conditions, a model for supervised learning can be made. Based on this model, bad or faulty products (parts) can be detected and eliminated early in production.

Ana Đorđević is with the Faculty of Electronic Engineering, University of Niš, Aleksandra Medvedeva 14, 18000 Niš, Serbia (e-mail: ana.djordjevic@elfak.rs); Ana Đorđević is also with the innSonoFaculty of Electronic Engineering, University of Niš, Aleksandra Medvedeva 14, 18000 Niš, Serbia (e-mail: ana.djordjevic@elfak.rs).

Dejan Ćirić is with the Faculty of Electronic Engineering, University of Niš, Aleksandra Medvedeva 14, 18000 Niš, Serbia (e-mail: dejan.ciric@elfak.ni.ac.rs).

Marko Ličanin is with the Faculty of Occupational Safety, University of Niš, Čarnojevićeva 14, 18000 Niš, Serbia (e-mail: marko.licanin@znrfak.ni.ac.rs).

There are several challenges in computational sound analysis needed for the mentioned purpose. First, acoustic characteristics of even a single class of sounds can be highly diverse. For example, in the case of “bed DC motor” (faulty motor), the acoustics can vary enormously depending on the type of error in the motor. Second, in industrial environments, there are almost always multiple sources producing sound (noise) simultaneously. In such environments, many different types of sounds and unwanted noise can be present. Acoustic characteristics of some of them may be very close to the target sounds. Generally speaking, the captured audio is a superposition of contributions of all the sources present, which again distorts the signal. Third, an audio signal captured by a microphone is affected by impulse response between the source and microphone, which may alter the signal sufficiently to prevent matching of models developed to recognize the sound.

To overcome the present situation, spatial attention must be paid to choose the right sound signal features. [1] In this way, the aim is to generate a reliable and large enough data base that should be used for building the correct model.

Fig.1 presents the block diagram of a typical audio classification system based on supervised machine learning. The input of this system is an audio signal, either recorded in real-time or loaded from an audio recording file from storage. The audio processing block consists of different processing stages such as filtering, noise reduction or similar.

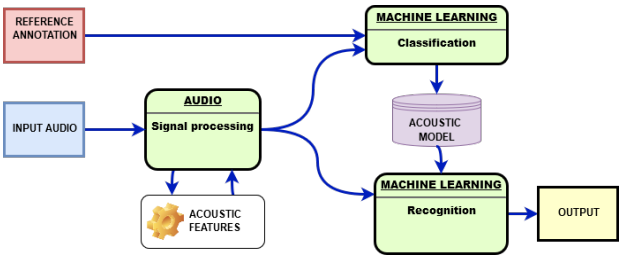


Fig. 1. Block diagram of a typical audio classification system.

The purpose of the feature extraction block is to obtain information sufficient for classifying the target sounds. This also includes making the subsequent modeling stage more computationally efficient and also easier to be realized with limited amount of development material. In other words, by feature extraction, the signal is transformed into a representation (feature matrix) that maximizes the sound recognition performance of the analysis system [4]. The acoustic features provide a numerical representation of the audio content relevant for machine learning. They characterize the signal containing information about its

physical properties, i.e., signal energy, distribution in frequency, and change over time. Acoustic features can be categorized in five main groups: temporal features, spectral shape features, cepstral features, perceptually motivated features and spectrogram image based features [1].

II. Acoustic Features

Temporal features are computed directly from the temporal waveform. Mostly used temporal features are temporal envelope, zero crossing rate, temporal waveform moments, and autocorrelation coefficients. Human perception of sound widely relies on its frequency content. Consequently, it is a natural choice to derive acoustic features from frequency representation of the signal. Among others, spectral features include energy, entropy of energy, spectral envelope, central frequency and spectral irregularity features. In audio analytics, temporal and spectral features are rarely used separately. Those are the simplest features extracted to model specific aspects of the signal, usually considered and evaluated together as one set of features sometimes referred as low-level features.

On the other hand, the cepstral features are powerful enough to be used on their own as input for classification or machine learning problem. Mel frequency cepstral coefficients (MFCC) are the most common cepstral coefficients [5]. They are obtained as the inverse discrete cosine transform of the log energy in mel frequency bands. Besides MFFC, linear prediction cepstral coefficients (LPCC) based on LPC coefficients, the gammatone feature cepstral coefficients (GFCC) or constant-Q cepstral coefficients (CQCC) are also very popular.

Good representatives of perceptually motivated features are loudness as the subjective impression of the intensity of a sound, then sharpness and perceptual spread as a measure of the timbral width of a given sound.

Acoustic features can also be extracted from the time-frequency representation of a sound signal. Spectrogram image-based features are inspired by computer vision to characterize the shape, texture and evolution of the time-frequency content in a sound scene. Such features have proven to be competitive with more traditional above mentioned acoustic features on classification tasks [6].

The obtained acoustic features together with labels of the training examples are used to learn models for the sound classes of interest. The labels contain information about the presence of target sound classes in the training data. They are used as a reference information to automatically learn a difference between acoustic features of different classes. At the final step, the learned acoustic model is made.

Analysis of various acoustic features, feature selection and process of optimization of the acoustic model for classification of DC motors has been presented in this paper. DC motors are divided (labeled) in two classes - non-faulty motors and faulty motors, depending on their working conditions. Labeling is based on subjective opinion of expert listeners working for the DC motor manufacturer. Data base is made of sound samples of a number of DC motors. Recording has been made in the semi anechoic chamber, where only the floor is reflective. Processes of collecting audio data is aspired to obtain sufficient amount of representative examples of both sound classes necessary for as accurate as possible classification. After recording and signal processing, feature extraction and visualization has been performed. All of the extracted features has been analyzed from various aspects in order to optimize the final feature vector, which will be used for classification. Furthermore, results from this analysis has been used to determine cause of error in classification.

III. Data Processing and Feature Extraction

A. Data acquisition and pre-processing of audio signals

In the process of designing and building machine learning system, data acquisition plays an important role since the performance of the final application highly depends on the data used to develop it. Audio signals that make the data base for this research are recorded in the semi anechoic chamber. The measurement microphone and tested DC motor samples are held on exactly the same positions throughout the measurement sessions. The microphone faces the center of tested DC motor with the distance of 42 cm. To cover intra-class variability, nearly 200 samples of both non-faulty and faulty DC motors are recorded.

After recording, pre-processing is applied to the audio signals as a step before the acoustic feature extraction. The main aim of this stage is to reduce the effects of disturbances and potentially enhance certain characteristics of the audio signal in order to maximize performance of learning algorithm. When DC motors are tested in industrial environment, low frequency noise can significantly affect the recordings. To address this issue, some noise suppression techniques can be used to reduce interference of ambient noise [1]. In this study, only high-pass filtering of the signal with cutoff frequency of 300 Hz is applied.

B. Acoustic feature extraction

Feature extraction can also be considered as a procedure of data rate reduction. Features extracted must contain relevant information with respect to the desired properties of the analyzed sound. Analysis algorithms should be based on a relatively small number of features. An audio signal is voluminous, and as such, it is hard to be processed directly in any analysis task. Therefore, it is needed to transform the initial data representation to a more suitable one. Properties of original signal can be typified by acoustic features which will result in the reduction of data volume. Good knowledge of the application domain is necessary in order to choose the best features.

In Fig. 2, the block diagram of acoustic feature extraction is shown. For that purpose, *short term* processing technique is applied. Thus, an audio signal is divided into overlapping short-term windows (frames) and the analysis is carried out on a frame basis. Length of frame used here is 50 ms, which can be often met in audio analytic applications. Naturally audio signals are non-stationary, their properties vary over time. But in a short time frame, sound from the DC motor can be considered to be stationary. This type of processing generates a sequence, one feature vector per audio signal. The dimensionality of feature vector depends on nature of the adopted features. It is not uncommon to use one-dimensional

features, like the energy of a signal. However, in most sophisticated audio analysis applications several features are extracted and combined to form feature vectors of increased dimensionality. The extracted sequence(s) of feature vectors can then be used for subsequent processing/analysis of the audio data.

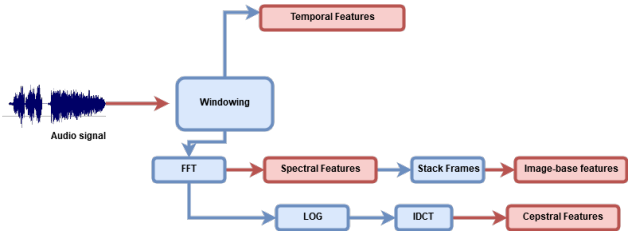


Fig. 2. Block diagram of extraction of a DC motor acoustic features.

For this research, a large set of 212 acoustic features is generated from each short term frame. In the first step, matrix of dimensions 46×212, where 46 represents the number of short term overlapping frames, and 212 is the number of features. This situation is not preferable, since the number of features is rather large. This can lead to a learning model that tends to be over–fitted. As a result, their performance degenerates.

IV. Feature Vector Visualization and Optimization

As it is previously mentioned, DC motor sound signal can be considered to be stationary. This can be observed in Fig.3, where energy in time as an audio feature is presented. The value of this feature fluctuates to a certain extent over the time (over the frames). These fluctuations are larger for some sound samples, while they are smaller for some others.

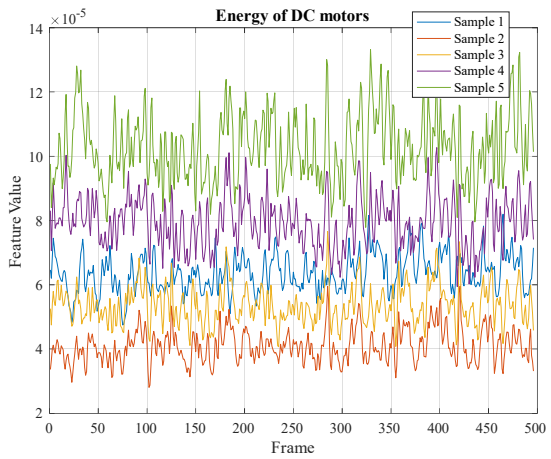


Fig. 3. Energy of five different samples of DC motors over the time (frames).

Stationary behavior of a signal is used as the first step to reduce dimensionality of the feature vector. For each feature, its mean value is calculated from all short term frames. By this procedure, each signal in the data base is represented by a single value for each feature. Also, values in the feature vector are normalized in order to overcome the differences in ranges of values of the features of different nature.

The total number of 212 extracted features is rather large and computationally demanding. This is why the next logical step is to optimize the feature vector by reducing the number of features needed for correct classification. In order to optimize the data set and to reduce complexity of feature matrix, few different visualization and feature selection techniques are used.

First, a 1D representation of normalized feature vector is made (the 1D is related to mean feature value). This representation for three different samples of non-faulty and faulty motors is given in Fig 4. The difference between classes is observable from the 20th to 40th feature, and from 140th to 180th feature. The rest of the features does not show significant variation between classes, which means that they do not contribute to a significant extent to the classification process. A very basic optimization of feature vector can be done by using this kind of plot for data analysis. A drawback of this kind of plot is that the only the small number of audio signals (samples) can be observed at once.

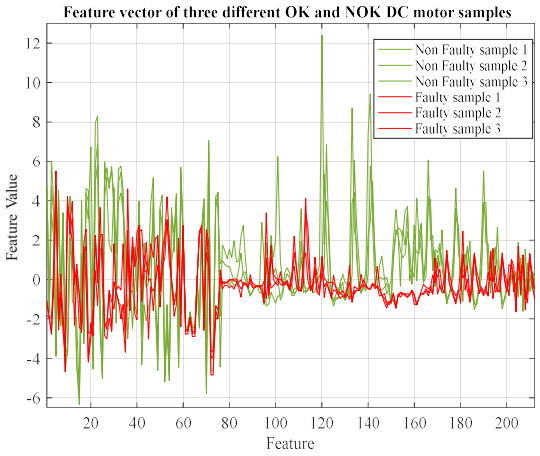


Fig. 4 1D representation of feature vector of three different samples of non-faulty and faulty DC motors.

Another option for visualization is to divide features into groups of two. The next step is to make graph where values of feature 1 (F1) are given in the *x*-axis, while values of feature 2 (F2) are given in the *y*-axis. In this way, a 2D graph of features is generated, that is, a field of features in two-feature domain. Representative examples of two such graphs are presented in Fig. 5. On the first pair of features represented on Fig 5 subplot a) the distinction between Non Faulty and Faulty motors is greater than in the second pair. A similar conclusion can be made for lot of other combinations. Even though this process has been slow, unnecessary features can be detected and removed from further classification process.

In the same way as previously done, the features can be divided into groups of three features. Then, 3D graphical representations of these groups of features are generated. Here, the values of feature 1 (F1) are shown in *x*-axis, the values of feature 2 (F2) are shown in *y*-axis, and the values of feature 3 (F3) are shown in *z*-axis. The effects of features are more prominent in spaces of three different features than in

1D or 2D spaces.

Besides, increase of graph dimensionality provides some new insights in the analysis of differences between classes. This kind of graph is presented in the figures below.

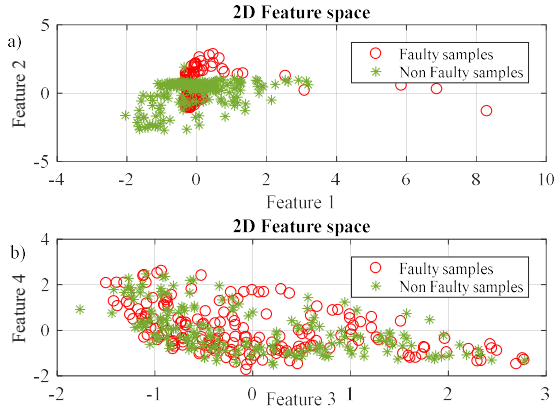


Fig. 5 2D representation of features in two-feature domain a) 2D space made of Feature 1 and 2 as representation of good distinction between classes b) 2D space made of Feature 1 and 2 as representation of bed distinction between classes

In Fig. 6, the values of the 22nd, 23rd and 24th features differ significantly between the classes, so it is easy to make distinction between them. These three features carry important information beneficial to the classifier.

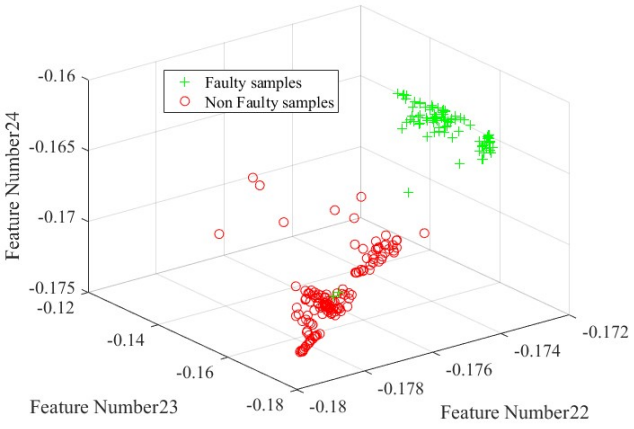


Fig. 6 3D visualization of class difference between non-faulty (OK) and faulty (NOK) DC motors using features 22, 23 and 24.

Opposite to the situation presented in Fig 6, the features given in Fig 7 do not show significant difference between the classes, so they do not contribute to classification process. This kind of visualization has proved to be a useful tool in optimization process where features relevant for classification are selected. Unfortunately, this visualization is slow for analysis, since there are a number of combinations of features to go through.

Another approach to cope with the large dimensionality of the feature space is to use transformation techniques such as Principle Component Analysis (PCA). This is low-dimensional linear approximation of the original data, which can be viewed as a projection of data on the new coordinate axes. In this procedure, the variances on the axes are maximized. In Fig.8, the results of the PCA for non-faulty and faulty motors are shown.

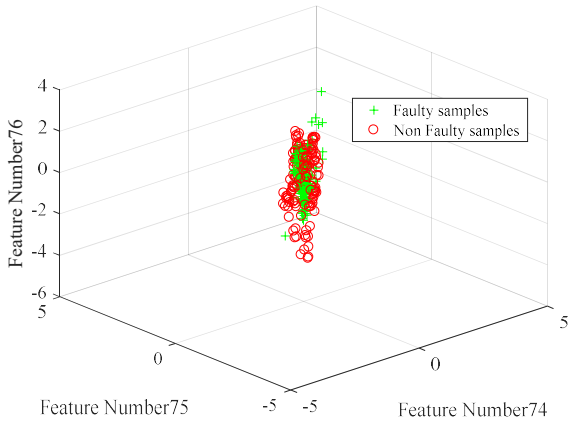


Fig. 7 3D visualization of features using features 74, 75 and 76, where the class difference between non-faulty (OK) and faulty (NOK) DC motors is hardly visible.

As it can be observed, there are some differences between the classes but not that prominent to improve the classification process. The PCA is shown to be good for visualizing the classes. However, in this particular case, especially for optimization of used features, it does not seem to be promising - better results are obtained by simple visualization of samples in three features space. By employing the PCA, there is no guarantee that unwanted or unneeded features will be eliminated, since the features may exhibit high variance. The transformed features could not be easily interpreted, which is one of major drawbacks in a process of understanding the quantities that best describe the classes of DC motors.

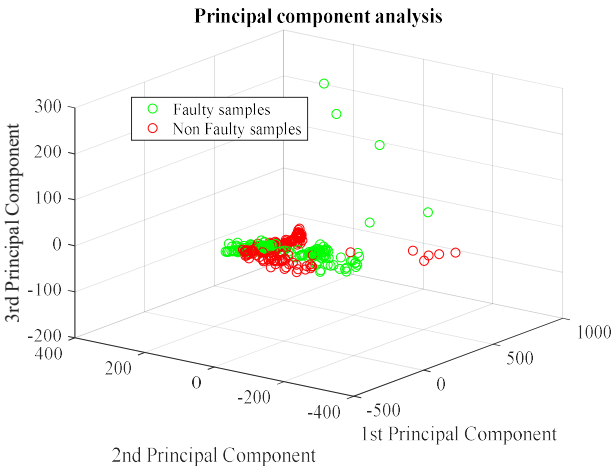


Fig. 8 Principle component analysis (PCA) of feature matrix of non-faulty (OK) DC motors and faulty (NOK) DC motors.

An alternative to feature transformation technique is feature selection approach for feature vector optimization. This technique is based on search through the subsets of extracted features with a goal to find the best ones based on evaluation function among the competing candidate subsets. For this purpose, several different algorithms for selection are tested. [7

] Based on rank in the selection, the extracted features are graded, only the k best ones are used in the machine learning process.

One of the tested methods for feature selection is Infinite Latent Feature Selection, ILFS, the algorithm that allows the investigation of importance of a feature when injected into an arbitrary set of cues. The results from this algorithm are tested in the liner Support Vector Machine (SVM) for classification of DC motors. When only two best ranked features are employed, accuracy of the used algorithm is 95.97%. This is plotted in 2D space of those two features and shown in Fig. 9.

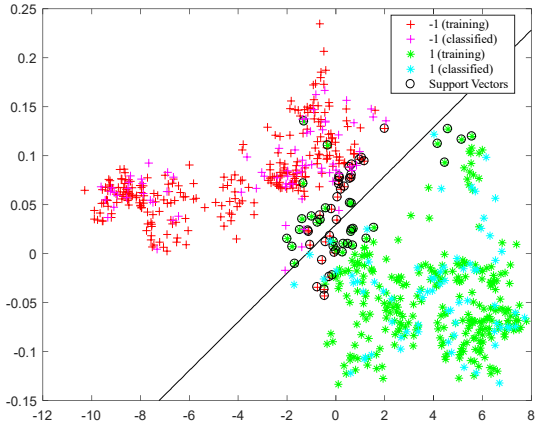


Fig. 9 Visualization of results of SVM trained with 2 best features based on infinite latent feature selection algorithm.

The Laplacian sore algorithm for feature selection is also studied. In this algorithm, for each feature, its Laplacian score is computed to reflect its locality preserving power. The Laplacian score is based on the observation that two data points are probably related to the same topic if they are close to each other.

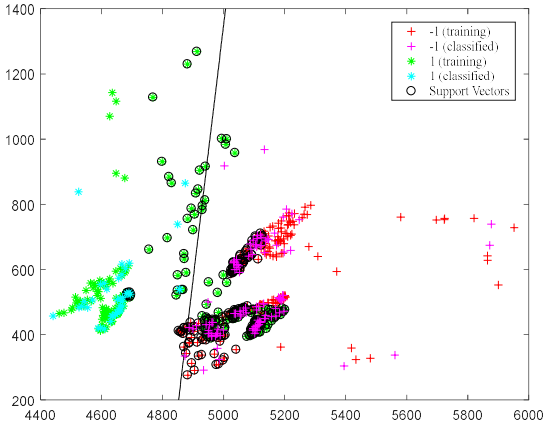


Fig. 10 Visualization of Support Vector Machine trained with 2 best features based on Laplacian score

The machine trained with two best ranked features from this algorithm is less accurate than the previous mentioned one. Here, the accuracy is only 75.84 %, and visualization of the results are shown in the Figure 10. More accurate machine can be obtained by using more than two features in the learning process. In the case of features ranked with the Laplacian score, 15 best features are needed to obtain the same accuracy as in the case of infinitive latent feature selection.

Some other feature selection algorithms are studied, too, and the results are similar to the previously mentioned two. Decision which features will be used for the final classification algorithm is based on the presented visualization and optimization techniques. In the end, in the tested example of DC motors, dimension of the feature vector is reduced from 212 to 35 features. Thus, the computational demand for the machine learning process is significantly reduced and optimized.

V. Conclusion

A large number of potentially useful features can be considered in the design of sound classification systems. Though in some cases it is practicable to use all those features for the classification, it may not be optimal to do so. Beside good audio data base preparation, feature visualization and selection are two very important steps in designing an optimal machine learning algorithm for classification. Visualization and optimization techniques presented in this paper can significantly improve the efficiency of the classification problem. Also, they play a decisive role in the process of designing and building the machine learning algorithm for automatic classification.

Acknowledgment

This research is supported by the Ministry of Education, Science and Technological Development of Serbia through the project no.36026.

References

[1] S. Chu, S. Narayanan, and C. J. Kuo, "Environmental sound recognition with time frequency audio features," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 17, no. 6, 2009, p. 1142

[2] T.Virtanen, M. D. Plumbley, D.Ellis, "Computational Analysis of Sound Scenes and Events", *1st* ed. New York City, USA, Springer International Publishing, 2018.

[3] C.M Bishop, "Pattern Recognition and Machine Learning ", *1st* ed. New York City, USA, Springer International Publishing, 2006

[4] S. Chu, S. Narayanan, C. C. J. Kuo, and M. J. Mataric, "Where am I? Scene recognition for mobile robots using audio features," in IEEE Int. Conf. Multimedia and Expo (ICME), 2006, p. 885.

[5] Peltonen, V., Tuomi, J., Klapuri, A., Huopaniemi, J., Sorsa, T.: Computational auditory scene recognition. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing, vol. 2, p. 1941 (2002)

[6] Kobayashi, T., Ye, J.: Acoustic feature extraction by statistics based local binary pattern for environmental sound classification. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3052–3056. IEEE, New York (2014)

[7] M. Dash and H. Liu. Feature selection for classification. Intelligent data analysis, 1(1-4):131–156, 1997.

Analysis of DC Motor Sounds Using Wavelet-Based Features

Đorđe Damnjanović, Dejan Ćirić and Zoran Perić

***Abstract*—Widespread usage of wavelets in signal processing nowadays confirms effectiveness of this method for different applications such as de-noising of audio signals. The wavelet method can be considered to be either the Fourier transform replacement or its complement and it is able to overcome some important disadvantages of Fourier transform. The wavelet transform typically provides the detail and approximation coefficients as results. They can be used for further processing, as it is the case in de-noising, but these coefficients can be used for some other applications, too. This paper presents the application of wavelet decomposition into detail and approximation coefficients as features of audio signals for their classification. Sounds of direct current (DC) motors with and without fault in both directions of rotation are recorded and used as test audio signals. The effects of wavelet parameters such as wavelet function and decomposition level on the features distinction is investigated. Time and frequency analysis is also done for the tested DC motors.**

***Index Terms*—Wavelets; Detail coefficients; Approximation coefficients; Audio features; DC motors.**

I. Introduction

Wavelets were primarily introduced to solve problems of the Fourier transform [1]. The first official term wavelet dated back to 1909 when Alfred Haar mentioned wavelet in his research [2]. Looking from the historical point of view, the roots of wavelet transform appeared in the Fourier's papers [2]. During the decades, group of scientists developed different types of wavelet families and explained wavelet as a small wave [2,3]. More correct definition states that wavelet is a waveform of effectively limited duration that has an average value of zero [2,3]. The most of the families are named by these scientists: Gabor, Morlet, Daubechies, Haar, etc. Majority of newer wavelet families are upgraded and improved version of the old ones.

Nowadays, the wavelets are widespread. They have found their place in many applications including signal and image de-noising. When special acoustical signals are in focus (for example room impulse responses), wavelets can be applied for their processing (de-noising), too [4]. In this way, it is possible to significantly increase the dynamic range of room impulse responses [4]. Besides image and audio signal de-noising, wavelets can also be used for de-noising of other signals including biomedical signals (electromyogram, electrocardiograph and electroencephalogram) [5] and for remote sensing of very low frequency signals [6]. Another wavelet usage, which is quite new in research, is estimation of truncation time of an RIR [7].

Recently, an increased interest of both academia and industry is shown for audio signal classification and auditory scene recognition. This kind of audio signal processing plays an important role in various areas: biomedical engineering, mechanical engineering, telecommunications, acoustics, etc. In that regards, different features of audio signals are extracted and used for classification and recognition purpose [8,9]. Some of those features are based on wavelets [10,11].

One real problem that occurs in industry is how to assess the quality of produced motors or how to detect a faulty motor and recognize the failure type. A number of different approaches are already presented [12]. Unfortunately, it seems that there is no established unique - "the best" approach. Instead, since in most of situations there are some specific characteristics of the classification or recognition problem, a logical solution would be to apply a customized set of features and appropriate classification algorithm.

One of results of symbiosis of knowledge from signal processing, audio and acoustics is usage of wavelet transform as a tool for extrication of audio features. More specifically, audio features can be generated by signal decomposition into detail and approximation coefficients by applying wavelets [8]. This possibility is explored and analyzed here. The examples of audio signals of recorded DC motors both faulty and non-faulty are used as test samples. Since there are several parameters of wavelets, these parameters are changed and their effects on set of audio features consisting of detail and approximation coefficients are investigated. The focus is on possibility to differentiate faulty from non-faulty motors using sound that they generate. The processing is done in Matlab software package. The results for a few representative examples are presented here, although the research included more than 40 DC motors.

II. Related Work

There are different types of fault detection in motors that include vibration monitoring, motor current signature analysis (MCSA), electromagnetic field monitoring, chemical analysis, temperature measurement, infrared measurement, acoustic

Đorđe Damnjanović is with the Faculty of Technical Sciences Čačak, University of Kragujevac, Svetog Save St. 65, 32000 Čačak, Serbia (e-mail: djordje.damnjanovic@ftn.kg.ac.rs).

Dejan Ćirić is with the Faculty of Electronic Engineering, University of Niš, Aleksandra Medvedeva 14, 18000 Niš, Serbia (e-mail: dejan.ciric@elfak.ni.ac.rs).

Zoran Perić is with the Faculty of Electronic Engineering, University of Niš, Aleksandra Medvedeva 14, 18000 Niš, Serbia (e-mail: zoran.peric@elfak.ni.ac.rs).

noise analysis, and partial discharge measurement [13]. Current signature, vibration signature and sound signature analysis are the most common in use because of certain advantages in relation to these methods [14]. MCSA is a rather popular and reliable technique providing good results, but in some cases it is not sensitive enough, it has low-frequency resolution and installation can be more complicated [14]. Vibration analysis requires appropriate sensors (accelerometers), which can be an additional expense, and sometimes it is not easy to correctly place the sensor in a right place [13,14]. Sound analysis is contactless, low-cost and easy installation approach, where problem typically comes from disturbances of industrial environment (e.g., noise). MCSA is often used in combination with sound signature analysis [14,15].

Rotor faults, bearing faults and unbalanced faults in wings are the most common mechanical faults that can be detected and analyzed with sound signature analysis [13]. There are different methods for extracting relevant sound based features for classification purpose. The authors of Ref. 9 proposed extraction of the following audio features: the average zero-crossing rate, the fundamental frequency and spectral peaks. There are some recent studies where wavelet based features are applied for motor faults detection and motor classification [10,11,15,16]. An example is Ref. 16 where experimental results prove that wavelets can be used for this purpose as simple, easy and fast method.

Decomposition of audio signals with wavelet transform into approximation and detail coefficients represents one of possibilities for feature extraction. Typically, pre-processing stage precedes the wavelet transform. In this stage, audio signals are segmented in smaller frames, although in literature it can be found that wavelet transform is implemented either on longer frames or whole signals [10,15,16]. Since audio signals usually contain noise, de-noising in pre-processing can often be beneficial De-noising can be done by applying wavelets, Notch filtering, moving average filtering, etc [10,15].

For fault analysis in different types of motors, the most common wavelets are Haar and Daubechies, although Coiflets, Symlet and Mayer wavelets are also proposed for that purpose [10,15]. Selection of wavelet type can be very tricky, especially in some cases. For example, certain wavelet families, like Daubechies, have a large number of different types (there are up to 45 Daubechies wavelets when Matlab is used). In Ref. 11, a complete research about the choice of right wavelet in classification of percussive sounds is done. Fig. 1 presents the classification success rate depending on the chosen wavelet from this reference.

Regarding detection of faulty motors, besides selection of appropriate audio features, it is also important to choose right classifier. Similar to the situation with features, different classifiers have been proposed and used in studies [13-15].

For majority of authors, level of decomposition in wavelet transform is one of the most important parameters whose effects need to be investigated. Some authors use decomposition up to the level 8 or 9 [10-12], while some other authors can provide suitable results with much smaller level, for example 3, 4 or 5 [13,17].

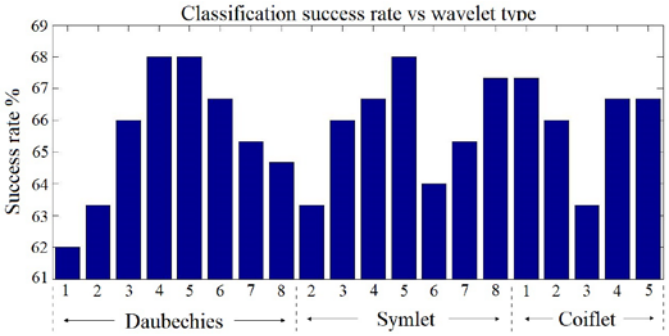


Fig. 1. Classification success vs. wavelet type from Ref. 11.

III. METHODS OF ANALYSIS

More than 40 different DC motors are used in the analysis. All motors are divided in two groups: motors with certain faults (faulty motors) and motors with good characteristics (non-faulty motors). Sound of each motor was recorded in semi-anechoic conditions (only the floor was reflective). The motors were placed on a test bench provided by the motor manufacturer. This test bench is able to drive the motors with adequate force and to apply an adequate load simulating real conditions. The motors were driven in two directions of rotation, where operation in each direction lasted approximately 8 s. The measuring microphone was placed about 40 cm from the tested motor.

In order not to use the transition regions in the beginning and at the end of the recorded signals, the medium parts of duration of 5 s of signals are extracted and further processed.

Wavelet-based feature extraction starts with decomposition of audio signal. The decomposition is usually done using discrete wavelet transform (DWT) rather than continuous wavelet transform (CWT) because of its easier implementation in multilevel signal decomposition [2]. Signal is decomposed into approximation and detail coefficients (high and low frequency components) at each level. This is why the DWT is equivalent to low and high pass filtering [3]. The described procedure is presented in Fig. 2 (down to level 3 of decomposition), where LP is a low-pass filter, HP is a high-pass filter, A stands for approximation coefficients, D stands for detailed coefficients, and 2↓ is down-sampling.

Different wavelets are applied on recorded signals with the main purpose to provide an adequate set of wavelet-based features being able to make difference between faulty and non-faulty motors. Generally speaking, the most common wavelet used in audio signal processing is Haar wavelet. This wavelet is proved to be the most stable in signal de-noising, although Daubechies wavelets provide somewhat better results than Haar in this particular application [4]. Other wavelets families used here are: Coiflets, Symlet, biorthogonal, reverse biorthogonal and Mayer.

The level of decomposition is an important part of analysis too, especially because many authors use different values for the level. Here, the levels of decomposition from 2 to 8 are

used.

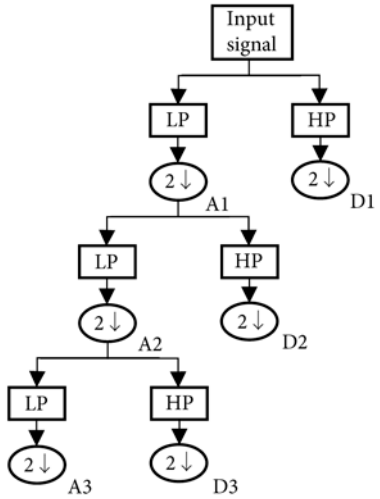


Fig. 2. Block diagram of wavelet transform decomposition into detail and approximation coefficients.

The described processing is done in Matlab software package using the corresponding toolkits. The application is fully automated. The selected wavelet and its level of decomposition is applied on the defined signal for decomposition process by the command *wavedec*. Approximation and detail coefficients are obtained with two commands - *appcoef* and *detcoef*, respectively according to the level of decomposition.

Regarding the analysis, the recorded signals are first observed in time and frequency domain separately, but also their spectrograms are investigated. Then, the wavelet-based features obtained by different combination of wavelet parameters are analyzed in detail. Special attention is paid to differences between faulty and non-faulty motors present in the time and frequency domains (if any), but also in the domain of wavelet-based features.

IV. Results

In this section, the illustrative results for some characteristic samples of tested motors are presented. Thus, Fig. 3 presents the recorded sounds of both faulty and non-faulty DC motors in both directions of rotation. As can be seen from these figures, there are certain fluctuations of amplitude (levels) in time. In some cases (signals), the amplitude fluctuations in time are more prominent. The fluctuations can have shape similar to low frequency pattern or some sudden onset of high amplitude can appear in certain time moments. Regarding faulty motors detection, considering only the time domain presentation, it is rather difficult to make a distinction between faulty and non-faulty motors.

Spectra of signals shown in Fig. 3 are presented in Fig. 4. There are some prominent low frequency components (below 100 Hz), that are mainly consequence of ambient noise. In rest of the frequency range, some peaks and dips appear. Distinction between directions of rotation can be made in an easier way in the spectral domain instead in the time domain. However, even in the spectral domain, it is not an easy task to differentiate the faulty motors from the non-faulty motors.

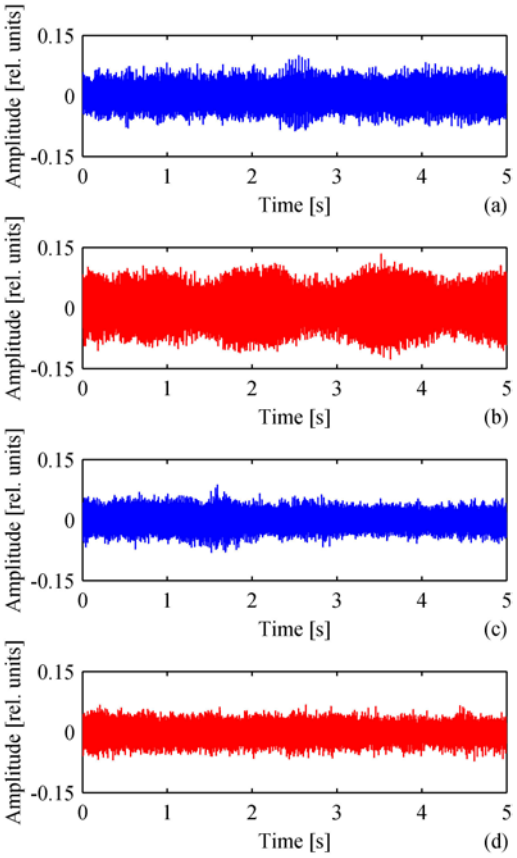


Fig. 3. Audio signals (sounds) of DC motors in time domain: a) non-faulty motor (direction of rotation 1) b) faulty motor (direction 1) c) non-faulty motor (direction 2) d) faulty motor (direction 2).

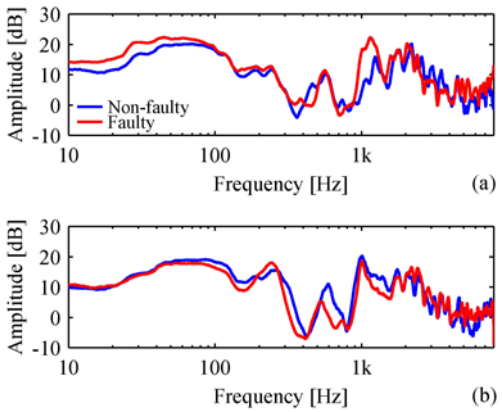


Fig. 4. Spectra of sounds of faulty and non-faulty DC motors: a) non-faulty motor (blue) and faulty motor (red), direction of rotation 1; b) non-faulty (blue) and faulty motor (red), direction of rotation 2.

Spectrograms of motor sounds in direction 1, whose time domain amplitudes are given in Fig. 3(a) are shown in Fig. 5. The peaks of spectra are visible as prominent horizontal lines in the spectrograms. These lines (spectral components) can be considered to be sound harmonics. In the shown case, higher sound levels are present for non-faulty motor, Fig. 6(a). These higher levels for non-faulty motor represent the main difference in the spectrograms for non-faulty and faulty

motors. Unfortunately, this situation is not the general one. Amplitude values (overall levels) changes from motor to motor independently whether the motor is faulty or non-faulty. The amplitude value could not reflect faults in motor.

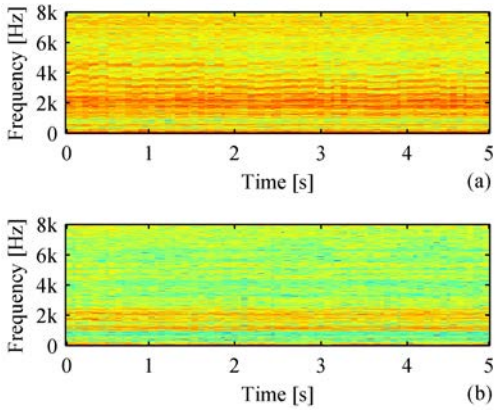


Fig. 5. Spectra of sounds of motors: a) non-faulty b) faulty in direction 1.

The presented characteristics of audio signals in time domain, spectral domain as well as in combined time-frequency domain show that it is rather difficult to make difference between faulty and non-faulty motors. In that regards, it sounds very logical to try to find an alternative approach able to make a clearer distinction between these motors.

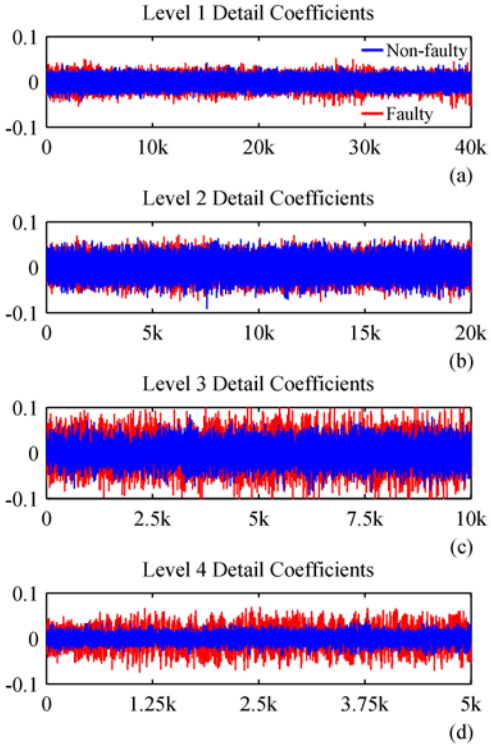


Fig. 6. Detail wavelet coefficients after applying Daubechies 2 wavelet for faulty (red) and non-faulty motors (blue) in direction 1 of rotation: a) level 1, b) level 2, c) level 3 and d) level 4.

The next step in finding adequate set of audio features is the wavelet transform, where Daubechies 2 wavelet (db2 in Matlab) is applied with different decomposition levels on the whole signal. The detail coefficients up to the level 4 for direction of rotation 1 are presented in Fig. 6. In this case, the detail coefficients for faulty and non-faulty motors for level of decomposition of 1, 2 and 3 seem to be rather similar to each other. This is not the case for the level of decomposition of 4 where the values of detail coefficients for the faulty motor are greater than those for the non-faulty motor.

For the levels of decomposition between 5 and 8, the values of detail coefficients for faulty and non-faulty motors are rather similar, see Fig. 7. Looking at both Figs. 6 and 7, it can be seen that the length of detail coefficient array becomes shorter with increase of the level of decomposition.The detail coefficients for the same motors (one faulty and one non-faulty), but for opposite direction of rotation (direction 2) are presented in Figs. 8 and 9. Comparing the results for faulty and non-faulty motors here, the situation is a bit different. The patterns of detail coefficients for these motors are rather similar except for the level 1 and partly for the level 7.

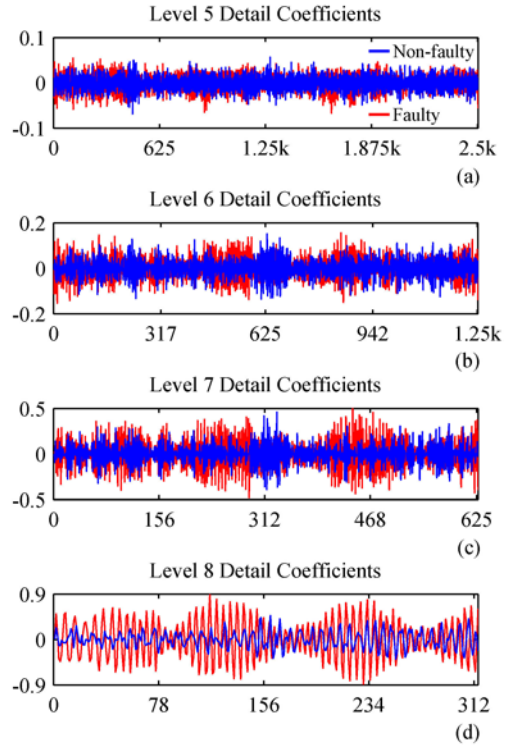


Fig. 7. Detail wavelet coefficients after applying Daubechies 2 wavelet for faulty (red) and non-faulty motors (blue) in direction 1 of rotation: a) level 5, b) level 6, c) level 7 and d) level 8.

Different situation appears when the approximation coefficients are observed. Wherever the approximation coefficients are considered in research, these coefficients from the highest level of decomposition are used. This is one of main differences between analysis of detail and approximation coefficients.

For the tested faulty and non-faulty motors from previous figures, the approximation coefficients for the level of decomposition 8 are presented in Fig. 10. Generally speaking, in some cases, the approximation coefficients can provide difference between faulty and non-faulty motors, as is shown in Fig. 10(a), but there are also examples where these coefficients are similar for faulty and non-faulty motors, as it

is the case given in Fig. 10(b).

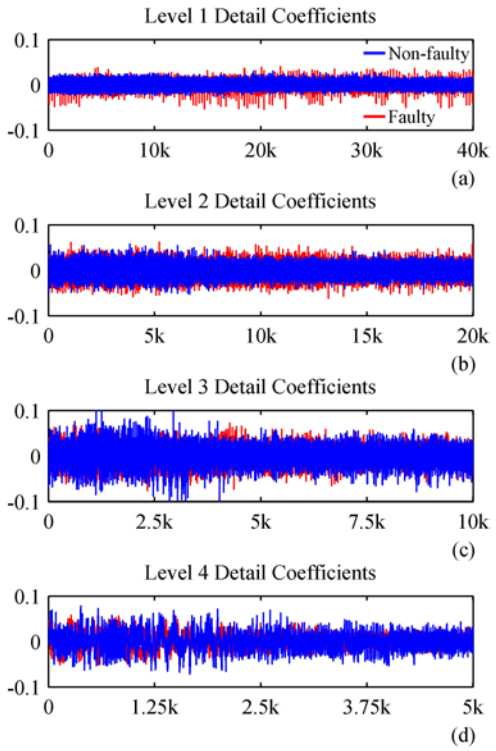


Fig. 8. Detail wavelet coefficients after applying Daubechies 2 wavelet for faulty (red) and non-faulty motors (blue) in direction 2 of rotation: a) level 1, b) level 2, c) level 3 and d) level 4.

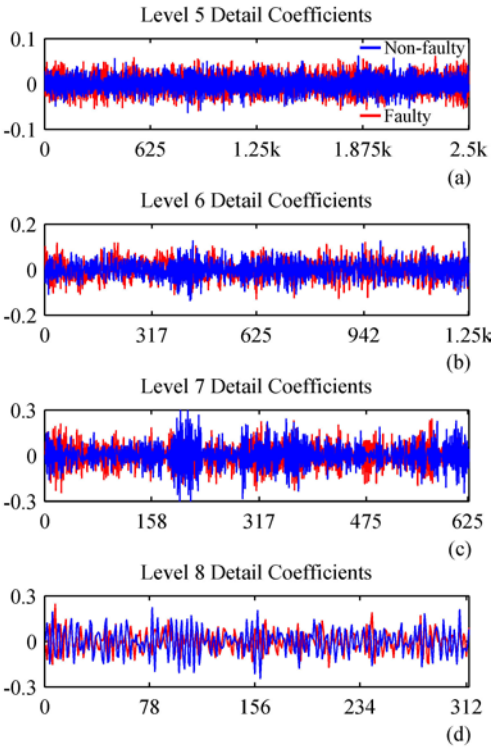


Fig. 9. Detail wavelet coefficients after applying Daubechies 2 wavelet for faulty (red) and non-faulty motors (blue) in direction 2 of rotation: a) level 5, b) level 6, c) level 7 and d) level 8

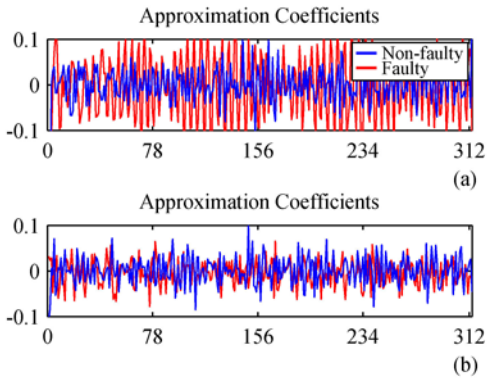


Fig. 10. Approximation wavelet coefficients after applying Daubechies 2 wavelet for faulty (red) and non-faulty motors (blue) a) in direction 1 of rotation, and b) in direction 2 of rotation.

When the wavelet function is changed from Daubechies 2 to other functions (Haar, Symlet, Coiflet, biorthogonal, reverse biorthogonal, Mayer), the patterns of detail and approximation coefficients are also changed to a certain extent. Figure 11 presents 4 cases of 4 different wavelet functions (Haar, Simlet 8, Coiflet 5 and Meyer) applied to the faulty and non-faulty DC motor in direction 1 of rotation. Detail coefficients of the level 4 are presented. Situation is rather similar to the case of Daubechies 2 wavelet. Mayer wavelet leads to slightly poorer result.

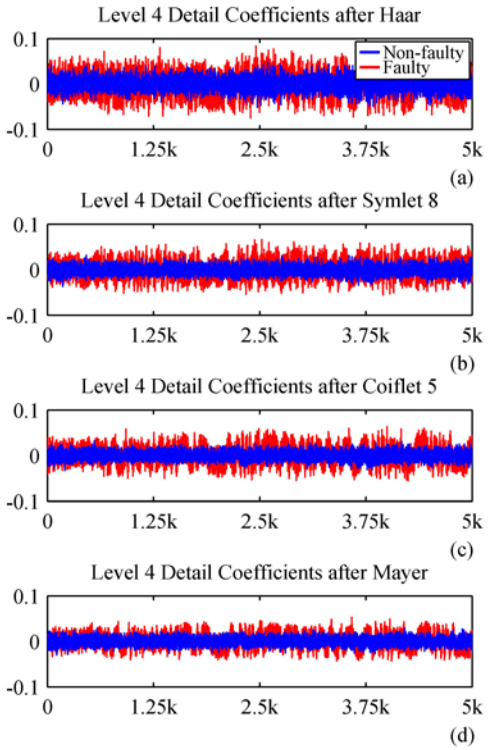


Fig. 11. Detail coefficients of the level 4 for faulty (red) and non-faulty motors (blue) in direction 1 of rotation after applying wavelets: a) Haar b) Symlet 8 c) Coiflet 5 and d) Meyer.

When the effects of decomposition level are observed, it can be concluded that the patterns for detail and approximation coefficients are changed by changing the level of decomposition. In some cases, the best results (the best distinction between faulty and non-faulty motors) are obtained

applying the level 4 (see Fig. 6), but in some other cases (other motors), some other levels would be preferable. Fig. 12 presents cases where differences between faulty and non-faulty motors are larger for some other levels.

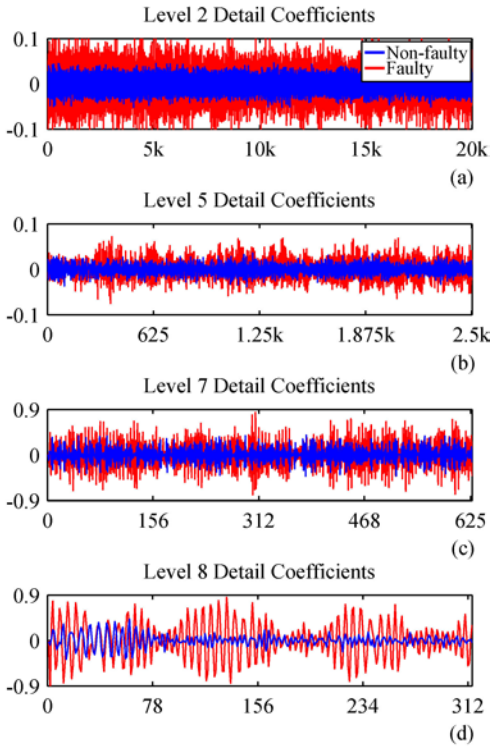


Fig. 12. Detail wavelet coefficients after applying Daubechies 2 wavelet for faulty (red) and non-faulty motors (blue): a) level 2, b) level 5, c) level 7 and d) level 8, where the results for different motors are given in figures from (a) to (d)

V. Conclusion

Depending on the motor fault, sound generated by the motor can be significantly changed from the sound generated by a non-faulty motor. However, in many cases, especially where the fault is considered to be a minor one, the sounds of faulty and non-faulty motors are perceptually rather similar to each other. For naive listeners, it is not easy to make a distinction between these two types of motors.

Wavelet technique is one of options to extract usable features when sound signature analysis is applied for motor estimation. Decomposition of an audio signal into detail and approximation coefficients is the main task in this type of wavelet analysis. With the right choice of wavelet function and level of decomposition, differences in wavelet coefficients for faulty and non-faulty motors can become prominent.

Regarding the wavelet parameters, different wavelets provide rather similar results, although there are cases that are slightly different. One of the main observations of this research is that there is no unique decomposition level leading to the largest differences between faulty and non-faulty motors. If the most preferable level has to be chosen, than this level is either 4 or 8. In the continuation of the research, other options will be explored, such as first to segment the tested audio signal in frames, and then to apply wavelet decomposition.

Acknowledgment

This research is supported by the Ministry of Science and Technological Development of Serbia through the projects No. 44009 and III-47003.

References

[1] M. Sifuzzaman, M.R. Islam, M.Z. Ali, “Application of wavelet transform and its advantages compared to Fourier transform”, *Journal of Physical Sciences*, vol. 13, pp. 121-134, 2009.

[2] R.J.E. Merry, “Wavelet Theory and Applications: a literature study”, Technische Universiteit Eindhoven, Eindhoven, June 7, 2005.

[3] B. Ergen, “Signal and image denoising using wavelet transform”, Chapter 21 in: Advances in wavelet theory and their applications in engineering. Physics and Technology, In-Tech (2012), pp. 495–515.

[4] Đ. M. Damnjanović, D. G. Ćirić, B. B. Predić, “De-Noising of a Room Impulse Response by Applying Wavelets, Acta Acustica United With Acustica”, *Journal of the European Acoustics Association (EAA) - International Journal on Acoustics*, Vol. 104, No. 3, pp. 452 – 463, May/June 2018.

[5] G. Kaushik, H.P. Sinha, L. Dewan, “Biomedical signals analysis by dwt signal denoising with neural networks”, *Journal of Theoretical and Applied Information Technology*, Vol. 62, No.1, pp. 184 – 198, 10th April 2014.

[6] E. Güzel, M. Canyılmaz, M. Türk, “Application of wavelet based denoising techniques to remote sensing very low frequency signals”, *Radio Science*, Vol. 46, Issue 2, pp. 1 – 9, April 2011.

[7] Damnjanović, D. Ćirić, “Usage of Wavelet De-noising for Estimation of Room Impulse Response Truncation Time”, Proceedings of 5th International Conference on Electrical, Electronic and Computing Engineering “IcETRAN 2018”, pp. 565-570, Palić, Serbia, June 11 – 14, 2018.

[8] G. Tzanetakis, G. Essl, P. Cook, “Audio Analysis using the Discrete Wavelet Transform”, Proceedings of the WSES International Conference Acoustics and Music: Theory and Applications (AMTA 2001), pp. 318-323, Skiathos, Greece, January 2001.

[9] T. Zhang, C.-C. Jay Kuo, “Content-Based Audio Classification and Retrieval for Audiovisual Data Parsing”, Chapter 3 in: Audio Feature Analysis, Springer, Boston, MA, 2001

[10] A. Glowacz, “Diagnostics of direct current machine based on analysis of acoustic signals with the use of symlet wavelet transform and modified classifier based on words”, *Eksploatacja i Niezawodnosc – Maintenance and Reliability*, Vol. 16, No. 4, pp. 554 – 558, 2014.

[11] M. Daniels, “Classification of Percussive Sounds Using Wavelet-Based Features”, Ph.D. dissertation, CCRMA, Stanford University, 2010.

[12] C. da Costa, M. Kashiwagi, M. H. Mathias, “Rotor failure detection of induction motors by wavelet transform and Fourier transform in non-stationary condition”, *Case Studies in Mechanical Systems and Signal Processing*, Vol. 1, pp. 15 – 26, Elsevier, 2015.

[13] P. Sharma, N. Saraswat, “Diagnosis of Motor Faults Using Sound Signature Analysis”, *International journal of innovative research in electrical, electronics, instrumentation and control engineering*, Vol. 3, Issue 5, pp. 80 – 83, May 2015.

[14] P. A. Delgado-Arredondo, D. Morinigo-Sotelo, R. A. Osornio-Rios, J. G. Avina-Cervantes, H. Rostro-Gonzalez, R. de J. Romero-Troncoso, “Methodology for fault detection in induction motors via sound and vibration signals”, *Mechanical Systems and Signal Processing*, Vol. 83, pp. 568–589, Elsevier, January 2017.

[15] A. Glowacz, “DC Motor Fault Analysis with the Use of Acoustic Signals, Coiflet Wavelet Transform, and K-Nearest Neighbor Classifier”, *Archives of Acoustics*, Vol. 40, No. 3, pp. 321–327, 2015.

[16] R. S. S. Kumari, D. Sugumar, “Wavelet Based Feature Vector Formation for Audio Signal Classification”, ICACC 2007 International Conference, Madurai, India, pp. 752-755, 9-10 Feb, 2007.

[17] P. de Chazal, B. G. Celler and R. B. Reilly, “Using Wavelet Coefficients for the Classification of the Electrocardiogram”, Proceedings of the 22nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 64 – 67, Chicago, IL, USA, 23-28 July 2000.

The Numerical Study of Atmospheric Attenuation of Outdoor Sound Propagation

Milan Mišković, Miomir Mijić and Miljko Erić

***Abstract* — It is well known that atmospheric attenuation of outdoor sound propagation dominantly depends on meteorological conditions - ambient atmospheric parameters such as temperature, pressure, humidity, etc. The atmospheric attenuation is a function of frequency and propagation distance. In engineering practice as well as in theoretical and applied researches, numerical modeling of atmospheric attenuation of outdoor sound propagation is needed. So, such numerical study is a subject of this paper. Some results of numerical modeling of atmospheric attenuation are presented which quantitatively illustrate the influence of different meteorological parameters on atmospheric attenuation. In order to share their effort with readers, the authors attached MATLAB code for numerical modeling of atmospheric attenuation of sound during propagation outdoors in appendix.**

***Index Terms* — Outdoor sound propagation, attenuation of sound, atmospheric sound attenuation, geometrical divergence attenuation.**

I. Introduction

Motivation of authors for numerical study of atmospheric attenuation is twofold:

- Although qualitative influence of meteorological factors on atmospheric attenuation in outdoor sound propagation is theoretically known and standardized [1], in engineering practice, such as planning of measurement campaign of acoustical events or deployment elements of distributed system for acoustical monitoring, a quick tool for numerical modeling and prediction of acoustical attenuation is needed.
- A key research focus of authors is to find the answer to the question if it is possible to extract signal features needed for automatic classification/identification of acoustic sources (signals), especially sources of impulse acoustic signals which are independent on distance or if such signal features need to be dependent on distance. For such study numerical modeling of atmospheric attenuation is needed.

Different phenomena occur in outdoor sound propagation. In the past, it was a subject of many published papers [1-9]. Phenomena such as attenuation, refraction, diffraction, reflection, etc. are the consequence of the air properties of physical sound propagation media and obstacles that are found on the propagation path. Sound energy is dissipated in air by two major mechanisms. The first is viscous loss due to friction between air molecules which results in heat generation "classical absorption". The second is relaxation process – sound energy is momentarily absorbed in the air molecules and causes the molecules to vibrate and rotate. These molecules can then re-radiate sound at a later instant, like small echo chambers, which can partially interfere with the incoming sound [1].

Outdoor sound propagation is affected by three basic factors: meteorological conditions, ground characteristics and the presence of obstacles [2,3,4].

Meteorological conditions in which sound propagation occurs, such as temperature changes, atmospheric turbulence, precipitation, air humidity, lead to an increase or decrease in sound attenuation during propagation. Changing meteorological conditions can easily cause fluctuations in sound levels by 10-20 dB over time periods of minutes. The longer is the transmission path, the larger are the fluctuations in levels [6]. With sound attenuation, gradients of wind or temperature refract waves either upwards or downwards. Upwards, if sound propagation takes places along upwind or in temperature lapse. Downwards, if sound propagation takes places along downwind or in temperature inversion.

Propagation reality is more complicate then geometrical spreading above flat ground. Some grounds are acoustically hard like concrete, and others soft as snow. The surface over which sound propagates can seldom be considered perfectly rigid or totally reflective with possible exceptions of open water, ice, or concrete. Typical soil surfaces with or without vegetation tend to absorb energy from incident acoustic waves. Accurate prediction of ground effects requires knowledge of the absorptive and reflective properties of the surface respectively the acoustic impedance [7].

Starting from theoretical foundation of atmospheric attenuation given in standard and published papers, a numerical study of outdoor sound attenuation is done in this paper. Some results of numerical modeling are presented which quantitatively illustrate influence of different meteorological parameters on atmospheric attenuation. The authors attached MATLAB code for numerical modeling of atmospheric attenuation in appendix. The authors believe that this code could be useful in readers engineering and research activities.

II. Modeling of Attenuation of Outdoor Sound Propagation

The total attenuation A [dB] of outdoor sound

Milan Mišković – Military Technical Institute, 1 Ratka Resanovića, 11030 Belgrade, Serbia and School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail: milan.miskovic@mod.gov.rs and miskomsm@gmail.com).

Miomir Mijić and Miljko Erić – School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail: emijic@etf.rs, miljko.eric@etf.rs).

propagation, without wind, can be defined by the equation [8]:

$$A = A_{div} + A_{atm} + A_{gr} + A_{bar} + A_{misc} \quad (1)$$

where:

A_{div} – denotes the attenuation due to geometrical divergence,

A_{atm} – denotes the attenuation due to atmospheric absorption,

A_{gr} – denotes the attenuation due to the ground effect,

A_{bar} – denotes the attenuation due to a barrier,

A_{misc} – denotes the attenuation due to miscellaneous other effects.

If sound propagates on flat terrain without obstacles on the propagation path (barriers, trees, settlements, industrial zones, etc.) and the influence of the soil is neglected, the equation (1) can be written in the following form:

$$A = A_{div} + A_{atm} \quad (2)$$

The geometrical divergence accounts for spherical spreading in the free field from a point sound source, making the attenuation A_{div} [dB], equal to:

$$A_{div} = 20\log_{10}(d/d_0) + 11 \quad (3)$$

where:

d [m] – denotes the distance from the source to receiver,

d_0 = 1 m – denotes the reference distance.

The attenuation due to atmospheric absorption A_{atm} [dB] at distance d, is given by equation:

$$A_{atm} = \alpha d \quad (4)$$

where α [dB/m] is the atmospheric attenuation coefficient for each octave band at the mid band frequency. The atmospheric attenuation coefficient may be defined [1] with equations (5), (6), (7) and (8):

$$\alpha = 8.686 f^2 \left(\left[1.84\times10^{-11} (p_a/p_r)^{-1} (T/T_r)^{\frac{1}{2}} \right] + x \right) \quad (5)$$

$$x = (T/T_r)^{-\frac{5}{2}} (y + z) \quad (6)$$

$$y = 0.01275 e^{-\frac{2239.1}{T}} \left[f_{ro} / (f_{rO}^2 + f^2) \right] \quad (7)$$

$$z = 0.1068 e^{-\frac{3352}{T}} \left[f_{rN} / (f_{rN}^2 + f^2) \right] \quad (8)$$

where:

f [Hz] – denotes the frequency of the sound,

p_a [kPa] – denotes the ambient atmospheric pressure,

T [K] – denotes the ambient atmospheric temperature,

f_{rO} [Hz] – denotes the relaxation frequency of oxygen,

f_{rN} [Hz] – denotes the relaxation frequency of nitrogen,

p_r = *101.325* kPa – denotes the reference ambient atmospheric pressure,

T_r = 293.15 K – denotes the reference air temperature.

The relaxation frequency of oxygen and nitrogen are defined with equations (9) and (10), respectively.

$$f_{rO} = \frac{p_a}{p_r} \left\{ 24 + \left[\frac{(4.04\cdot10^4 h)(0.02 + h)}{0.391 + h} \right] \right\} \quad (9)$$

$$f_{rN} = \frac{p_a}{p_r} \left(\frac{T}{T_r} \right)^{-\frac{1}{2}} \left(9 + 280 h e^{-4.17\times\left[\left(\frac{T}{T_r} \right)^{-\frac{1}{3}} - 1 \right]} \right) \quad (10)$$

In equations (9) and (10), h [%] is the molar concentration of water vapor. The molar concentration of water vapor is defined with equations (11), (12) and (13):

$$h = h_r (p_{sat}/p_a) \quad (11)$$

$$p_{sat} = p_r 10^C \quad (12)$$

$$C = -6.8346 (T_{01}/T)^{1.261} + 4.6151 \quad (13)$$

where:

h_r [%] – denotes relative humidity,

p_{sat} – denotes saturation vapor pressure,

T_{01} = 273.16 K – denotes triple point isotherm temperature.

The sound attenuation due to geometric divergence is a consequence of the nature of the sound propagation i.e. the spherical spreading of the sound wave from the point source. When it is necessary to do the exact calculation of the absolute sound level at the distance d from the source, the attenuation A_{div} has to be taken into account.

The attenuation due to atmospheric absorption is the consequence of environment in which the sound propagates. It is linearly dependent on the distance from the sound source and the atmospheric attenuation coefficient. The atmospheric attenuation coefficient depends on: the frequency of the sound, the ambient atmospheric temperature, the relative humidity and the ambient atmospheric pressure.

III. RESULTS OF NUMERICAL MODELING

The mathematical model was used to show the dependence of the attenuation coefficient on the meteorological parameters and the frequency of sound. Numerical results, obtained using the presented mathematical model, confirm the well-known theoretical results on sound attenuation. Additionally, the significance of the simulation results is to show the effect of the change of individual meteorological parameters on the attenuation coefficient.

The results presented in Fig.1. show how the attenuation coefficient depend on relative humidity and frequency of the sound. Ambient pressure and reference ambient atmospheric are equal. The ambient temperature is constant and it is 20 °C.

As it can be seen in Fig. 1, the value of attenuation coefficient increases with the increase of frequency.

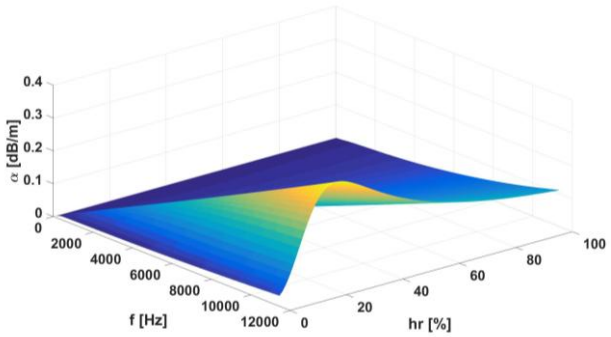


Fig. 1. The atmospheric attenuation coefficient depending of frequency, relative humidity and ambient temperature of 20 °C. The ambient pressure and the reference ambient atmospheric pressure are equal.

The results presented in Fig. 2. show the attenuation coefficient for different ambient temperatures. The ambient pressure and the reference ambient pressure are equal all the time.

It can be seen in Fig. 2. that there is the increase in the attenuation coefficient as the ambient temperature rise. In the numerical modeling, four values of ambient temperature were taken. The highest value of the attenuation coefficient of sound is at the highest frequency and the lowest values of relative humidity.

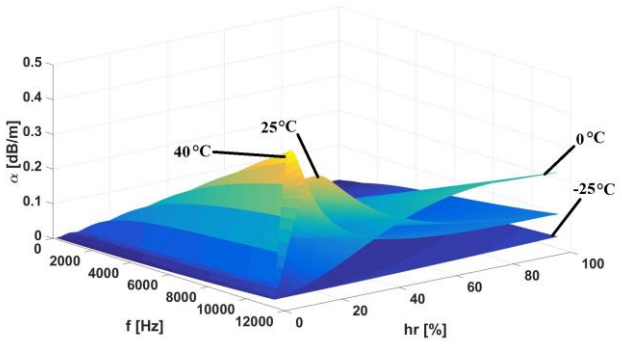


Fig. 2. The atmospheric attenuation coefficient for different ambient temperatures. The ambient pressure and reference ambient pressure are equal.

It cannot be clearly seen in Fig. 2. whether the attenuation coefficient changes with frequency and relative humidity at ambient temperature of -25 °C. In order to eliminate the confusion, the Fig. 3. shows the dependence of the attenuation coefficient on frequency and relative humidity at ambient temperature of -25 °C.

As it can be seen in Fig. 3. the level of the attenuation coefficient is about 10 times lower than the level of the attenuation coefficient when the ambient temperature is +40 °C. Additionally, the attenuation coefficient is the highest at the highest frequencies and relative humidity values. So, the attenuation coefficient also changes at low temperatures but that is much less pronounced than at higher temperatures.

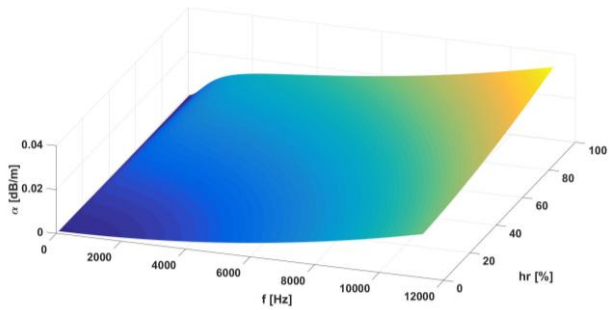


Fig. 3. The atmospheric attenuation coefficient of sound for ambient temperature of –25 °C. The ambient pressure and the reference ambient pressure are equal

The results presented in Fig. 4. show that the attenuation coefficient when ambient pressure is 913.25 mbar and 1213.25 mbar. The ambient temperature of 25 °C remain unchanged.

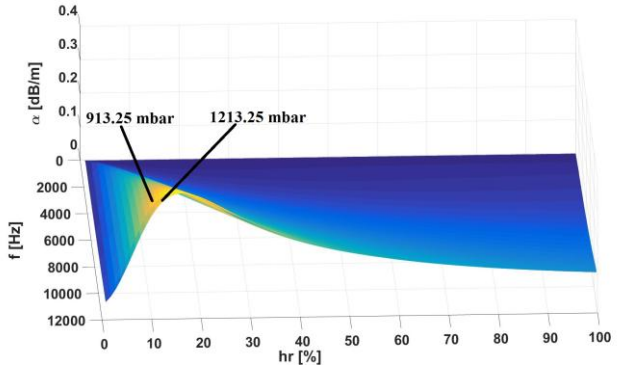


Fig. 4. The atmospheric attenuation coefficient for different ambient pressure. The ambient temperature is constant 25 °C.

The change in the attenuation coefficient is low when ambient pressure varies. Compared to the changes that occur when ambient temperature and relative humidity vary, this change in attenuation coefficient can be neglected. In order to achieve a greater change in the attenuation coefficient, the ambient pressure has to drastically deviate from the reference value of ambient pressure of 1013.25 mbar.

The results presented in Fig. 5. show the attenuation coefficient for the ambient pressure values of 413.25 mbar, 1013.25 mbar and 2413.25 mbar at constant ambient temperature of 25 °C. In the Fig. 5. it can be seen that a drastic deviation of ambient pressure from the reference value is required in order to achieve a greater change in the attenuation coefficient.

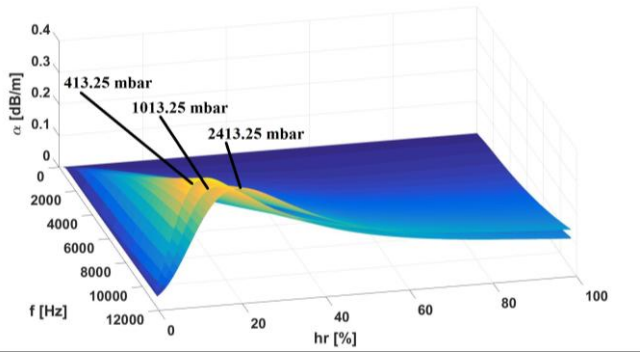


Fig. 5. The atmospheric attenuation coefficient for different ambient pressure of 413.25 mbar, 1013.25 mbar and 2413.25 mbar. The ambient temperature is constant 25 °C.

IV. Conclusion

The numerical study of atmospheric attenuation of sound during propagation outdoors is done in this paper. The described mathematical model indicates the clear dependence of sound attenuation on the meteorological conditions, metrological parameters (ambient atmospheric temperature, ambient atmospheric pressure, relative humidity) and the sound frequency. Several conclusions can be made from the simulation results which fully match the well-known theory.

During the outdoor sound propagation not all the spectral components are equally weakened in the sound wave. It can be concluded from the simulation results that the attenuation coefficient increases with the increase in frequency, i.e. the sound attenuation is highest at the highest frequencies. In this way, the higher spectral components are lost from the sound wave with the increase in distance from the source, and the air acts as a low-pass filter.

The increase in temperature increases the attenuation coefficient of sound in the air. If the ambient atmospheric temperature is -20 ° C and then increases on +40 ° C, the attenuation coefficient rises approximately 10 times.

In normal ambient conditions, the effect of ambient pressure fluctuations on the attenuation coefficient can be ignored in comparison with the effect of fluctuations in temperature and relative humidity.

In order to achieve a significant change in attenuation coefficient due to pressure change a major disturbance is required in the propagation channel or in the air, which in normal conditions cannot occur.

In the case that it is not necessary to carry out calculation of the absolute sound level at a specific point in space, the attenuation of the sound due to geometrical divergence can be neglected.

Appendix

```
function [Aatm,f]=atmospheric_attenuation_IcETRAN2019
(t1,hr,Pam,d,model,fig)
%% Atmospheric attenuation; Milan Miskovic, SRB, 2018.
%% Literature:
% 2.ANSI S1.26-1995: Method for Calculation of the
Absorption of Sound by the Atmosphere
% + Aatm - is the attenuation due to *Atmospheric
absorption*  on propagation path
%% function call
% 1. [Aatm,f] = atmospheric_attenuation_IcETRAN2019
(20,50,1013.25,200,2,1);
% [Aatm,f] = atmospheric_attenuation_IcETRAN2019
(t1,hr,Pam,d,model,fig);
% 2. press run
%% Input data t1,hr,Pam,d,model,fig
% - t1 - [C]   ambient atmospheric temperature
% - hr - [%]   relative humidity
% - Pam- [mbar]ambient atmospheric pressure
% - d  - [m]   distance source - receiver
% - model - [1 ili 2] model 1-ISO 2-ANSI
% - fig  - [1 ili 2] figure 1-No 2-Yes
%% f [Hz] frequency band, octave bands and one third octave
bands - function for generating the frequency band
[f,fc1,fup1,flo1,k1,fc2,fup2,flo2,k2] = frek();
%% Check input data
%% t1 [C] ambient atmospheric temperature
if (nargin < 1 || isempty(f) || ~isnumeric(f)); t1=20; end;
%% hr [%] relative humidity
if (nargin < 1 || isempty(f) || ~isnumeric(f)); hr = 1:1:100; end;
%% Pam [mbar] ambient atmospheric pressur
if (nargin < 1 || isempty(f) || ~isnumeric(f)); Pam = 1013.25;
end;
%% d [m] distance source - receiver
if (nargin < 1 || isempty(f) || ~isnumeric(f)); d = 200; end;
%% model [1 ili 2] 1-ISO 2-ANSI
if (nargin < 1 || isempty(f) || ~isnumeric(f)); model = 1; end;
%% fig [1 ili 2] figure 1-No 2-Yes
if (nargin < 1 || isempty(f) || ~isnumeric(f));fig = 2; end;
%% END INPUT DATA---OUTPUT DATA Aatm I f
% Aatm    [dB] - Atmospheric attenuation
% f          [Hz] - frequency band
%% FIGURE INPUT DATA ---------------------------------
   if fig == 1
   elseif fig == 2
figure(2);  plot(fc1,ones(1,k1),'rv','MarkerFaceColor',[1,0,0]);
hold on; plot(flo1,ones(1,k1),'bo'); plot(fup1,ones(1,k1),'b+');
leg1=legend(num2str(fc1,5),num2str(flo1,5),num2str(fup1,5))
; title(leg1,'Central frequency with lower and upper limits for
octave bands [Hz]');
      for i11 = 1:k1
line([flo1(i11) flo1(i11)],[0 1]);line([fup1(i11) fup1(i11)],[0
1]); end;
title ('Frequency subbands - ground attenuation'); xlabel
('Frequency subbands f [Hz]'); grid on; axis([0-max(fup1)/10
max(fup1)+ max(fup1)/10 0 2]);
figure(3);  plot(fc2,ones(1,k2),'rv','MarkerFaceColor',[1,0,0]);
hold on; plot(flo2,ones(1,k2),'bo'); plot(fup2,ones(1,k2),'b+');
leg1=legend(num2str(fc2,5),num2str(flo2,5),num2str(fup2,5))
; title(leg1,'Central frequency with lower and upper limits for
one third octave bands [Hz]');
      for i11 = 1:k2
line([flo2(i11) flo2(i11)],[0 1]);line([fup2(i11) fup2(i11)],[0
1]); end; title ('Frequency subbands - atmospheric
attenuation'); xlabel('Frequency subbands f [Hz]');grid on;
```

```
 axis([0-max(fup2)/10 max(fup2)+ max(fup2)/10 0 2]); end
%% END FIGURE INPUT DATA --------------------------
%% ATMOSPHERIC ATTENUATION
[alpha_iso,alpha_ansi,Prm] = alpha1(f,t1,hr,Pam);
% alfa_iso,alfa_ansi [dB/m] - atmospheric attenuation coefficients
Aatm_iso = alpha_iso*d;   % dB - atmospheric atten. on d [m]
Aatm_ansi=alpha_ansi*d;%dB - atmospheric atten. on d [m]
   if model == 1
      Aatm = Aatm_iso;  % ISO
   elseif model == 2
      Aatm = Aatm_ansi; % ANSI
   else
      disp('ERROR input data - "model"')
   end
   if fig == 1
   elseif fig == 2
      figure(4) % x f[Hz], y hr [%], z alpha [dB/m]
      hold on
      [~,hr_a] = size(hr); % hr matrix or one value
      if hr_a > 1   % hr matrix
mesh(f,hr,alpha_iso);hold    on;view(50,43);xlabel('f    [Hz]');
xlim([0 12000]); ylabel('hr [%]'); zlabel('\alpha [dB/m]');
         if Pam==Prm
title({'Atmospheric attenuation coefficient \alpha [dB/m]',['T = ',num2str(t1), char(176),'C;',' Pam = Prm = ',num2str(Pam), ' mbar']});
         else
title({'Atmospheric attenuation coefficient \alpha [dB/m]',['T = ',num2str(t1), char(176),'C;',' Pam= ',num2str(Pam), ' mbar',' Prm= ',num2str(Prm), ' mbar']});
         end; grid on;
      else    % hr one value
plot(f,alpha_iso); hold on; plot(f,alpha_ansi,'-.'); xlabel('f [Hz]'); ylabel('ISO \alpha [dB/m] and ANSI \alpha [dB/m]');
legend('\alpha ISO','\alpha ANSI');
         if Pam==Prm
title({'Atmospheric attenuation coefficient \alpha [dB/m]',['T = ',num2str(t1), char(176),'C;','hr = ',num2str(hr),' %',' Pam = Prm = ',num2str(Pam), ' mbar']});
         else
title({'Atmospheric attenuation coefficient \alpha [dB/m]',['T = ',num2str(t1), char(176),'C;','hr = ',num2str(hr),' %',' Pam= ',num2str(Pam), ' mbar',' Prm= ',num2str(Prm), ' mbar']});
         end; grid on; end;
      figure(5) % x f[Hz], y hr [%], z alpha [dB/m]
hold on; [~,hr_a] = size(hr); % hr matrix or one value
      if hr_a > 1   % hr matrix
mesh(f,hr,Aatm_iso, 'edgecolor','g'); hold on;
mesh(f,hr,Aatm_ansi,'edgecolor','r');  view(50,30);  xlabel('f [Hz]'); ylabel('hr [%]'); zlabel('\alpha [dB/m]'); legend('ISO: \alpha [dB]','ANSI: \alpha [dB]');
         if Pam==Prm
title({'Atmospheric attenuation coefficient \alpha [dB]',['d = ',num2str(d),' m;','T = ',num2str(t1), char(176),'C;',' Pam = Prm = ',num2str(Pam), ' mbar']});
         else
title({'Atmospheric attenuation coefficient \alpha [dB]',['d = ',num2str(d),' m;','T = ',num2str(t1), char(176),'C;',' Pam= ',num2str(Pam), ' mbar',' Prm= ',num2str(Prm), ' mbar']});
         end; grid on;
         else
plot(f,Aatm_iso); hold on; plot(f,Aatm_ansi,'-.'); xlabel('f [Hz]'); ylabel('ISO \alpha [dB] and ANSI \alpha [dB]');
legend('\alpha ISO','\alpha ANSI');
         if Pam==Prm
title({'Atmospheric attenuation coefficient \alpha [dB]',['d = ',num2str(d),' m;','T = ',num2str(t1), char(176),'C;','hr = ',num2str(hr),' %',' Pam = Prm = ',num2str(Pam), ' mbar']});
         else
title({'Atmospheric attenuation coefficient \alpha [dB]',['d = ',num2str(d),' m;','T = ',num2str(t1), char(176),'C;','hr = ',num2str(hr),' %',' Pam= ',num2str(Pam), ' mbar',' Prm= ',num2str(Prm), ' mbar']}); end; grid on; end; end;
   if fig == 1
   elseif fig == 2
   figure(12) % x f[Hz], y hr [%], z Aatm [dB]
   hold on; [~,hr_a] = size(hr); % hr matrix or one value
      if hr_a > 1   % hr matrix
mesh(f,hr,Aatm); view(43,41); xlabel('f [Hz]'); ylabel('hr [%]'); zlabel('A [dB]')
         if Pam==Prm
title({'Atmospheric attenuation Aatm [dB]',['d = ',num2str(d),' m; ','T = ',num2str(t1), char(176),'C; ',' Pam = Prm = ',num2str(Pam), ' mbar;']});
         else
title({'Atmospheric attenuation Aatm [dB]',['d = ',num2str(d),' m; ','T = ',num2str(t1), char(176),'C; ',' Pam= ',num2str(Pam), ' mbar; ',' Prm= ',num2str(Prm), ' mbar;']});
         end; grid on
      else
         plot(f,Aatm); xlabel('f [Hz]'); ylabel('Aatm [dB]');
         if Pam==Prm
title({'Atmospheric attenuation Aatm [dB]',['d = ',num2str(d),' m; ','T = ',num2str(t1), char(176),'C; ','hr = ',num2str(hr),' %; ',' Pam = Prm = ',num2str(Pam), ' mbar;']});
         else
title({'Atmospheric attenuation Aatm [dB]',['d = ',num2str(d),' m; ','T = ',num2str(t1), char(176),'C; ','hr = ',num2str(hr),' %; ',' Pam= ',num2str(Pam), ' mbar; ',' Prm= ',num2str(Prm), ' mbar;']});
         end; grid on; end; end;
   %% END ATMOSPHERIC ATTENUATION
   %% Function frek and alpha1
function [f,fc1,fup1,flo1,k1,fc2,fup2,flo2,k2] = frek()
   %% Function for generating the frequency bands
   % Octave bands (8 subbands) i 1/3 octave bands (24 subbands): Range 44 Hz to 11.314 kHz
   %% function call
   % 1. [f,fc1,fup1,flo1,k1,fc2,fup2,flo2,k2]=frek()
   % 2. press run
%% OUTPUT DATA 9 (f,fc1,fup1,flo1,k1,fc2,fup2,flo2, k2)
% f [Hz] - frequency band (all in one: octave bands and 1/3 octave bands)
```

```
% fc1 [Hz] - central freq. for all subbands (octave bands)
% fup1 [Hz] - upper limit freq. of subbands (octave bands)
% flo1 [Hz] - lower limit freq. of submerged (octave bands)
% k1 - num. of central freq. all subbands (octave bands) 8
% fc2 [Hz] - central freq. for all subbands 1/3 octave bands)
% fup2 [Hz] - upper limit freq. of subbands (1/3 octave bands)
% flo2 [Hz] - lower limit freq. of subbands (1/3 octave bands)
% k2 – num. of central freq. for all subbands (1/3 octave bands) 24
%% Calculate Octave Bands
fc1 = 10^3 * (2 .^(-4:3)); % 62.5 Hz -8 kHz ISO
fd1 = 2^(1/2); fup1 = ceil(fc1 * fd1); flo1 = fix(fc1 / fd1);
[~,k1] = size(fc1);
% Calculate Third Octave Bands
fc2  = 10^3 * (2 .^ ((-13:10)/3));
fd2 = 2^(1/6); fup2 = ceil(fc2 * fd2); flo2 = fix(fc2 / fd2);
[~,k2] = size(fc2); f = flo2(1):1:fup2(end);end
function [alpha_iso,alpha_ansi,Prm]=alpha1(f,t1,hr,Pam)
%% Atmospheric attenuation coefficients alpha_iso,
alpha_ansi [dB/m]
% ANSI S1.26-1995: Method for Calculation of the
Absorption of Sound by the Atmosphere
%% Function call
%1. [alpha_iso,alpha_ansi,Prm]=alpha1(100,20,50,1013.25)
%    [alpha_iso,alfa_bass,Prm]=alpha1(f,t1,hr,Pam)
% 2. press run
%% List of constants
Tr = 293.15;  % [K],    reference value
T01 = 273.16; % [K], 273.15+0.01,triple-point isotherm temp.
Pr = 101.325; % [kPa],  reference value
Prm = Pr*10;  % [mbar], reference value
%% Check input data
if (nargin < 1 || isempty(f) || ~isnumeric(f)); f=100; end;
if (nargin < 2 || isempty(t1) || ~isnumeric(t1)); t1=20; end;
if (nargin < 3 || isempty(hr) || ~isnumeric(hr)); hr=50; end;
if (nargin < 4 || isempty(Pam) || ~isnumeric(Pam)); Pam =
1013.25; end;
   %% Input data (f,t1,hr,Pa)
   % f    [Hz]          - frequency band
   % t1   [C]           - ambijentalna atmosferska temperatur
   T = 273.15+t1;%  [K], - C -> K
   % hr   [%]           - relative humidity
   % Pa   [mbar]        - ambient atmospheric pressur
   Pa = Pam/10;  % [kPa] - mbar -> kPa
   %% OUTPUT DATA 3 (alpha_iso,alpha_ansi,Prm)
% alpha_iso [dB/m] - Atmospheric atten. coeffic. in air ISO
% alpha_ansi[dB/m]- Atmospheric atten. coeffic. in air ANSI
% Prm [mbar]       - atmospheric pressur reference value
   %% ISO
Psat=Pr*10^(-6.8346*((T01/T)^1.261)+4.6151);[k10,~]=size
(hr.'); h=zeros(1,k10);fro=zeros(1,k10);frn = zeros(1,k10);
   for i10 = 1:k10
h(1,i10) = hr(i10)*(Psat/Pa); fro(1,i10)=(Pa/Pr)*(24+40400 *
h(1,i10) *(0.02+h(1,i10))/(0.391+h(1,i10))); frn(1,i10) =
(Pa/Pr)*((T/Tr)^(-1/2))*(9+280*h(1,i10)*exp(-4.17*
((T/Tr)^(-1/3)-1)));end;
[k20,~]=size(f.');[k3,~]= size(hr.');alpha_iso = zeros(k3,k20);
   for i20 = 1:k20
      for i10 = 1:k3
alpha_iso(i10,i20) = 8.686*(f(i20)^2)*((1.84*(10^-11) *
(Pa/Pr) ^(-1)*(T/Tr)^(1/2))+(T/Tr)^(-5/2)*(0.01275 * (exp(-
2239.1/T))*(fro(1,i10))/(((fro(1,i10))^2)+(f(i20)^2)))+
0.1068*(exp(-3352/T))*(frn(1,i10)/(((frn(1,i10))^2)+
(f(i20)^2)))); % dB/m
      end;end
   %% ANSI
Psat_b=Pr*10^(10.79586*(1-T01/T)-5.02808*log10(T/T01)+
1.50474*10^(-4)*(1-10^(-8.29692*(T/T01-1)))+4.2873*10^(-
3)*(-1+10^(4.76955*(1-T01/T)))-2.2195983);
   for i10 = 1:k10
h_b(1,i10)  =  hr(i10)*(Psat_b/Pa);fro_b(1,i10)=(Pa/Pr)*(24+
40400*h_b(1,i10)*(0.02+ h_b(1,i10))/(0.391+h_b(1,i10)));
frn_b(1,i10)  =  (Pa/Pr)*((T/Tr)^(-1/2))*(9+280*h_b(1,i10)*
exp(-4.17*((T/Tr)^(-1/3)-1)));end;alpha_ansi=zeros (k3,k20);
   for i20 = 1:k20
      for i10 = 1:k3
alpha_ansi(i10,i20) = 8.686*(f(i20)^2)*((1.84*(10^-
11)*(Pa/Pr)^(-1)*(T/Tr)^(1/2))+(T/Tr)^(-5/2)*(0.01275 *
(exp(-2239.1/T))*(fro_b(1,i10))/(((fro_b(1,i10))^2) +(f(i20)
^2)))+0.1068*(exp(-3352/T))*(frn_b(1,i10)/(((frn_b
(1,i10))^2)+(f(i20)^2)))); % dB/m
end; end;end end
```

REFERENCES

[1] ANSI S1.26-1995 (ASA 113-1995)

[2] D. Heimann, R.Blumrich, “Meteorological Aspects of Sound Propagation Modeling over Irregular Terrain,” Proceeding of Sound propagation outdoors, ICA 2001. year

[3] A.I. Tarreroa, J. Gonzálezb, M. Machimbarrenab, M. Arenalb, “Temperature & Trees Influence on Propagation Outdoors,” Proceeding of Sound propagation outdoors, ICA 2001. year

[4] G.A. Daigle, “Effects of atmospheric turbulence on the interference of sound waves above a finite impedance boundary,” JASA 65(1), , pp 45-49, 1979. year.

[5] J.E.Piercy, T.Embleton, L.Sutherland: “Review of noise propagation in the atmosphere,” JASA 61(6), pp 1403-1418, 1977. year

[6] J.S.Lamancusa, “Outdoor Sound Propagation. Noise control – chapter 10,” Penn State University, USA, 2009. year

[7] T.F.W. Embleton, “Tutorial on sound propagation outdoors,” 80 Sheardown Drive, Box 786, Nobleton, Ontario L0G 1N0, Canada, 1996. year JASA. 100 (1), pp 31-48 July 1996. year

[8] Internacional Standard ISO 9613-2, December 1996. year

[9] M. Bérengier, B.Gauvreau, P.Blanc-Benon, D. Juvé, “Outdoor Sound Propagation: A Short Review on Analytical and Numerical Approaches,” Acta acustica united with Acustica, Vol. 89, 980 – 991, November 2003. year

The Numerical Study of Atmospheric Attenuation of Outdoor Sound Propagation

Milan Mišković
Miomir Mijić
Miljko Erić

Usage of Averaging in Generation of Room Energy Decay Curve

Miljan Miletić, Dejan Ćirić and Marko Janković

***Abstract*—Acoustical quality of rooms is typically estimated by measuring a set of room impulse responses (RIRs) and extracting acoustical parameters from them including reverberation time and early decay time. These parameters are commonly determined by the backward integration of a squared RIR also known as Schroeder integration. In this way, a curve called backward integrated impulse response or energy decay curve (EDC) is obtained. After introduction of the Schroeder integration several decades ago, it has become a predominant method for EDC generation. One of the main reasons lies in the fact that the Schroeder integration considerably reduces the curve fluctuations providing smooth EDC. However, a weak point of this integration is cumulative summing of background noise. As a consequence, the late part of noisy EDC is bent upward. In this paper, an alternative method for generating smooth room EDC based on averaging is analysed. The averaging is also able to reduce the curve fluctuations. Besides, the cumulative summing is not a problem in this case. The target is to generate the EDC that deviates from the reference one (noiseless EDC) as little as possible. The effects of changing the number of points used for averaging are investigated.**

***Index Terms*—Energy decay curve; room impulse response; Schroeder's backward integration; averaging; reverberation time.**

I. Introduction

IN room acoustics, the generated sound is a combination of the direct sound coming from the source and indirect sound consisting of reflections coming from surfaces and objects in the room. This combination of direct and indirect sound, which is affected by the volume, shape and surfaces of the room, plays a key role in forming the room acoustic quality [1]. It is worth mentioning that every surface and object reacts to sound in a different way - some of them transmit or absorb sound while the others reflect or diffuse sound.

Evaluation of the acoustic quality of a room is done by analyzing relevant acoustical parameters. They are commonly calculated from measured room impulse responses (RIRs). A RIR can be measured using various techniques including the swept sine technique [2]. An important part of processing in parameter calculation is generation of the energy decay curve (EDC). More than five decades ago, Schroeder proposed procedure for obtaining the EDC based on the backward integration of a squared RIR [3]. It is also known as backward integrated impulse response or Schroeder's curve [4]. The Schroeder's backward integration has become more and more popular, and nowadays it is a preferable method for the EDC generation. Some international standards like ISO 3382 [5] and ISO 18233 [6] deal with the EDC and the way how it should be generated.

The two main problems of the Schroeder's integration come from background noise and finite upper limit of integration [7,8]. When a noisy RIR is backward integrated, which is always the case in reality (in measured RIRs), the noise energy is cumulatively added causing the EDC to bend upward. Over the last several decades, various noise compensation methods have been proposed, like RIR truncation, subtraction of mean square value of noise and correction for truncation [8].

Here, the results of study dealing with an alternative method for generating the EDC are presented. This method applies averaging of a squared RIR given in decibels in order to provide a smooth EDC that deviates from the reference EDC obtained from the noiseless RIR as little as possible. The required processing (including the averaging) is applied in the software package Matlab. The effects of length of averaging interval (window), that is, number of points used for the averaging are investigated. The goal is to find an optimal interval for averaging. The analysis is done on both broadband synthesized RIRs and RIRs filtered in third-octave bands.

After the introductory part, theoretical basis of the EDC is given. It is followed by the method section where the EDC generation and error calculation is explained. The next section contains the main results of the above mentioned analysis, which is followed by the discussion and conclusions as well as several guidelines for the future work.

II. Energy Decay Curve (EDC)

Since reverberation time is one of important acoustical parameters of a room, it should be estimated as precisely as possible. The usual method for this estimation is based on the linear regression of the EDC, although an approach based on nonlinear regression is also presented [9]. Besides reverberation time, one more parameter (early decay time - EDT) is also estimated from the EDC. In that regards, the EDC should be generated in such a way to reduce the effects of all disturbances as much as possible. There are two regular options used for EDC generation. The first one is related to

Miljan Miletić is with the College of Applied Technical and Technological Sciences, Kosančićeva 36, 37000 Kruševac, Serbia (e-mail: miljan.miletic@vhts.edu.rs).

Dejan Ćirić is with the Faculty of Electronic Engineering, University of Niš, 14 Aleksandra Medvedeva, 18106 Niš, Serbia (e-mail: dejan.ciric@elfak.ni.ac.rs).

Marko Janković is with the Faculty of Electronic Engineering, University of Niš, 14 Aleksandra Medvedeva, 18106 Niš, Serbia (e-mail: marestudio2004@gmail.com).

logarithm of squared decay obtained by the interrupted noise method where white or pink noise is applied for excitation. The second one is the previously mentioned Schroeder's backward integration of a squared RIR. This integration in the discrete domain can be represented by the following equation:

$$d(t_k) = \sum_{r=t_k}^{t_\infty} h_n(r)^2 = \sum_{r=t_k}^{t_\infty} \left(h(r)+n(r)\right)^2 \qquad (1)$$

where $d(t_k)$ is the discrete time decay function, k and r are integers, t_k is a discrete time equal to $t_k = k\,\Delta t$, and $k = 0, 1, 2, \ldots$, while $\Delta t = 1/f_s$ and f_s is the sampling frequency. Noisy RIR h_n is given as a sum of noiseless RIR h and noise n. As a result of (1), the backward integrated impulse response or EDC is generated, which describes the decay of acoustic energy in the room.

Measured RIRs always contain certain background noise consisting of ambient noise and measurement equipment noise. Presence of noise seriously affects the generated EDC. Thus, the Schroeder's integration of noiseless (exponentially decaying) RIR is a straight line on the decibel scale. On the other hand, the integration of a noisy RIR leads to the EDC consisting of three parts: the main (reverberation) decay, the second decay with smaller slope caused by the cumulative summing of noise and the steep roll off caused by the finite upper limit of integration, see Fig. 1.

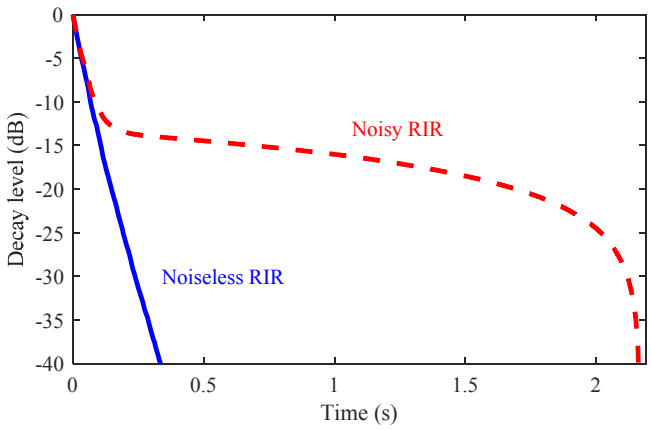


Fig. 1. EDCs of synthesized RIR with (noisy RIR) and without noise (noiseless RIR) obtained by Schroeder's backward integration.

III. EDC OBTAINED BY AVERAGING

The EDC can also be generated by logarithm of a squared RIR. Such an EDC (logarithmic squared RIR – LS RIR) has strong fluctuations in both parts – reverberation decay and noise, as presented in Fig. 2. Since data processing nowadays is much more powerful than it used to be, the linear regression can directly be applied even to this fluctuating EDC. However, the EDC fluctuations can lead to a larger deviation of the straight line approximation of the reverberation decay from true (target or reference) decay.

The fluctuations of LS RIR can be reduced by averaging. In this procedure, data points of the LS RIR within a current window are averaged to give a mean value. The current window is moved throughout the LS RIR from the starting point to the end point. The original LS RIR is then replaced by the mean values of the windows. Such a decay curve is called here averaged logarithmic squared RIR (ALS RIR).

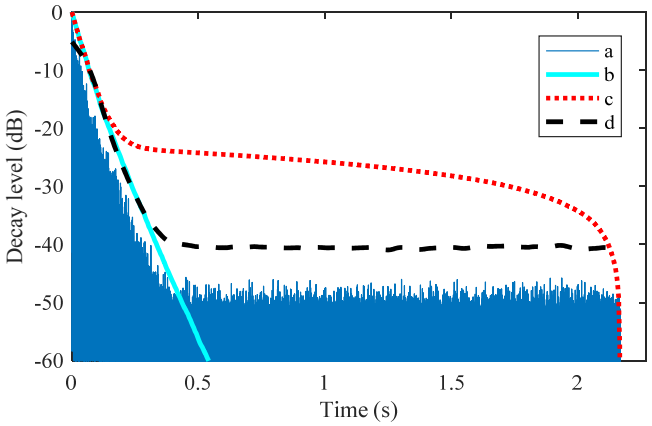


Fig. 2. Decay curves of synthesized noisy and noiseless RIR: (a) noisy LS RIR, (b) noiseless EDC obtained by Schroeder's integration, (c) noisy EDC obtained by Schroeder's integration and (d) averaged LS RIR (ALS RIR).

In this research, a basic moving average implemented in the software package Matlab is used for generation of ALS RIR. This averaging method replaces each data point with the average value of neighboring data points defined by the span or averaging window length (number of data points for mean value estimation) [10]. It is equivalent to low-pass filtering [11] where the result is defined by

$$y_a(i) = \frac{1}{2N+1}\left[y(i+N) + y(i+N-1) + \ldots + y(i-N)\right] \qquad (2)$$

where $y_a(i)$ represents the average (smoothed) value of the ith data point, and the span or averaging window is defined as $2N+1$. From (2), it is obvious that the averaged data point is located at the center of the window. Full averaging window spanning equally on both sides of the averaged data point can not be defined for the starting and ending data points. Illustration on how the averaging is done for the starting points in the case of the span equal to 7 is given in (3):

$$\begin{aligned} y_a(1) &= y(1) \\ y_a(2) &= \left[y(1)+y(2)+y(3)\right]/3 \\ y_a(3) &= \left[y(1)+y(2)+y(3)+y(4)+y(5)\right]/5 \\ y_a(4) &= \left[y(1)+y(2)+y(3)+y(4)+y(5)+y(6)+y(7)\right]/7 \\ &\ldots \end{aligned} \qquad (3)$$

Besides the EDCs of noiseless and noisy RIR obtained by the Schroeder's backward integration, Fig. 2 also shows the ALS RIR obtained by the moving average method. It can be seen that the noisy ALS RIR has different shape than the

noisy EDC. The effects of cumulative summing of noise and finite upper limit of integration are not present in the former curve, that is, in the ALS RIR.

IV. Method of Research

For the research purpose, noiseless RIRs are synthesized using the image source method (ISM) [12]. Noisy RIRs are generated by adding white or pink noise of different levels to the noiseless RIRs. Fig. 3 shows the waveforms of one of noisy and noiseless synthesized RIRs used in the analysis.

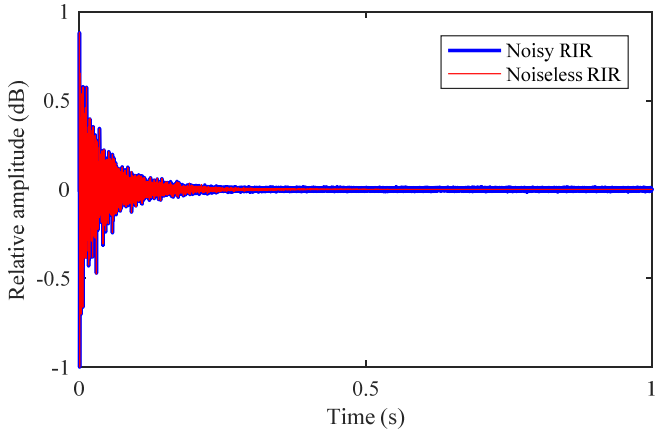


Fig. 3 Waveform of noisy and noiseless synthesized RIRs used in the analysis.

The question is how to evaluate the decay curves. A curve generated from a noiseless RIR by the Schroeder's backward integration is defined as the target or reference curve. Deviation of any considered decay curve from the reference one is calculated as a root mean square error (difference) given by

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \qquad (4)$$

where $\hat{y}_i$ are the data points of the reference curve, y_i are the data points of the considered decay curve, and n represents the number of points (time interval) where RMSE is calculated, called RMSE interval.

This time interval is determined in the following way: for every particular RIR and third-octave frequency band, the ALS RIR (generated using averaging window of length of 1000 data points) is applied for the RMSE interval calculation. The lower RMSE interval limit is set in the time point where the reference EDC has a decay level of -5 dB, this is t_{low}. The difference between the considered ALS RIR and reference EDC is positive going from the curve end. At the first cross of these curves, their difference changes the sign. This time point is set as an initial the end of the RMSE interval, that is, the initial t_{up}. Then, the decay level of the reference EDC is determined in the initial t_{up}. This decay level is then reduced by the margin, which can be a few decibels yielding L_{RMSE}. Here, the value of 2 dB is used for the margin. The final the end of the RMSE interval (the final t_{up}) is set at the time point where the reference EDC has the decay level equal to L_{RMSE}. Number of points present in the RMSE interval between t_{low} and t_{up} is equal to n. The described procedure for determination of RMSE interval is illustrated in Fig. 4. The determined RMSE interval is then used for various averaging window lengths applied to that particular RIR and in that particular frequency band.

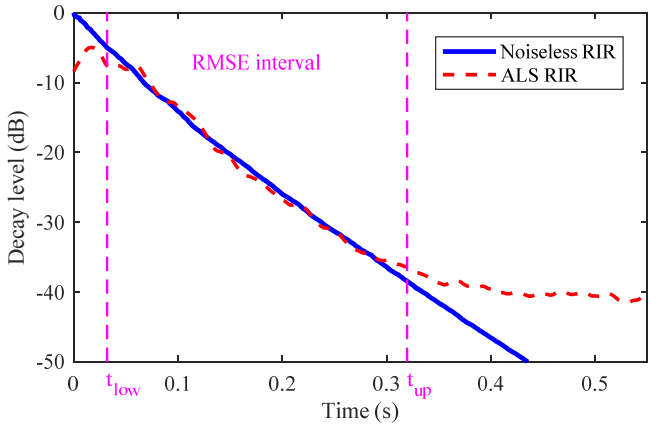


Fig. 4. Illustration of determination of RMSE interval where the RMSE between the considered (ALS RIR) and the reference EDC is calculated.

A number of data points that are averaged or averaging window leading to a single averaged value is considered to be an averaging parameter. The effects of changing this parameter are analyzed here. The analysis is carried out on both broadband RIRs and RIRs filtered in third-octave bands. The goal is to find an optimal averaging window length leading to the smallest deviation of this curve (ALS RIR) from the reference one, which is quantified by the RMSE.

During the analysis, besides the deviation of the considered decay curve from the reference one (RMSE), time necessary for performing the averaging is taken into account, too. Processing such as averaging was not possible some time ago due to limited processing power. Now, the processing power is considerably higher, and some new possibilities become available. The research was carried out on the PC computer under the Windows 7 Professional operating system. PC configuration includes Intel Core i7 870 CPU and 8GB of DDR3 ram memory.

The programming part of this research is realized using the software package Matlab. For that purpose, relevant Matlab scripts for the RIR synthesis, processing of RIRs in order to generate EDCs, evaluation of the decay curves and plotting the results are developed. Possibilities of generating an appropriately smooth decay curve by the averaging are first explored using the broadband RIRs. Then, the procedure is applied to the RIRs filtered in third-octave bands in the frequency range from 100 Hz to 8 kHz.

V. AVERAGING RESULTS

Although a number of RIRs are used for the analysis, the results are here illustrated using a representative example of a RIR. When the mean average method is applied to this broadband RIR, the generated ALS RIRs for 6 different averaging windows are presented in Fig. 5. These windows contain 100, 500, 1000, 2000, 5000 and 10000 data points. The same averaging windows are chosen for other figures given in this paper.

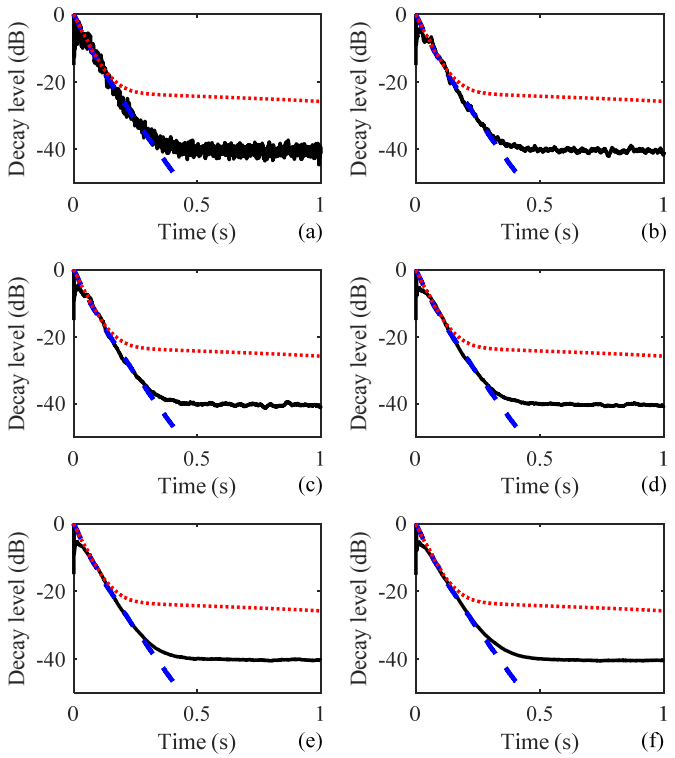


Fig. 5 Decay curves obtained from broadband synthesized RIR: reference EDC (- - -), noisy EDC (· · ·) and ALS RIR (—) generated by using averaging window of 100 (a), 500 (b), 1000 (c), 2000 (d), 5000 (e) and 10000 (f) data points.

Strong fluctuations of the LS RIR (see Fig. 2) are reduced by averaging in the ALS RIR. By increasing the averaging window length, the fluctuations become smaller and smaller. However, at the same time, due to specific averaging explained by (3), a certain disturbance appears at the beginning of the ALS RIR. This disturbance becomes more prominent with an increase of the averaging window length. For the averaging window of 100 data points, this effect is hardly visible.

When the moving average method is applied to the LS RIR obtained from the RIR filtered in third-octave bands at low frequencies, two above mentioned behaviors of reduction of fluctuations and appearance of disturbance in the ALS RIR beginning are also noticed, see Fig. 6. At low frequencies, even the reference EDC has certain fluctuations due to the nature of response decay in this frequency range. These fluctuations are not filtered out by the averaging. However, the reverberation and random fluctuations present in both the reverberation decay and noise are reduced by the averaging.

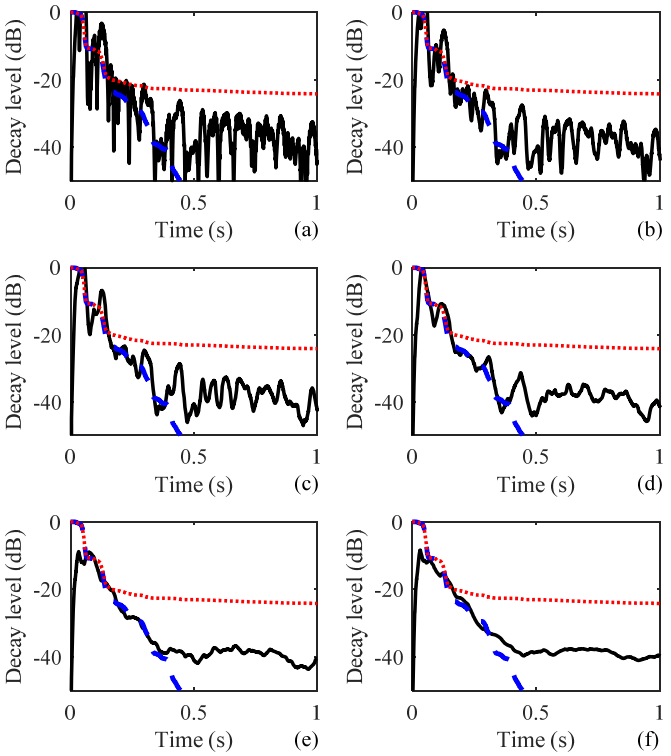


Fig. 6 Decay curves obtained from synthesized RIR filtered in the third-octave band at 200 Hz: reference EDC (- - -), noisy EDC (· · ·) and ALS RIR (—) generated by using averaging window of 100 (a), 500 (b), 1000 (c), 2000 (d), 5000 (e) and 10000 (f) data points.

Fluctuations of the LS RIR generated from the RIR filtered in third-octave bands at higher frequencies are reduced by the averaging even more efficiently. The results in one third-octave band are presented in Fig. 7. The situation regarding two noticed behaviors (fluctuation reduction and disturbance at the curve beginning with an increase of the averaging window length) is the same here.

Observing Figs. 5, 6 and 7, it can be concluded that the ALS RIRs coincide rather well with the reference EDCs in the RMSE interval. The deviations are somewhat larger at low frequency ranges where the ALS RIRs do not follow the reference EDCs completely. However, it seems that the decay rates of the ALS RIRs are very close to those of the EDCs. This will be quantified through the RMSE in the rest of the section.

Besides the disturbance of the ALS RIR at its beginning, the deviation of this curve from the reference EDC becomes larger with an increase of the averaging window in vicinity of the knee. This point represents the intersection of the reverberation decay and noise level. This is a consequence of taking more and more data points where the noise is dominant for averaging when a particular data point is close to the knee.

One more important observation from these figures is that the dynamic range of the ALS RIRs is considerably greater than the corresponding dynamic range of the noisy EDCs obtained by the Schroeder's backward integration. This is an

expected result since there is no cumulative summing of noise in the ALS RIRs.

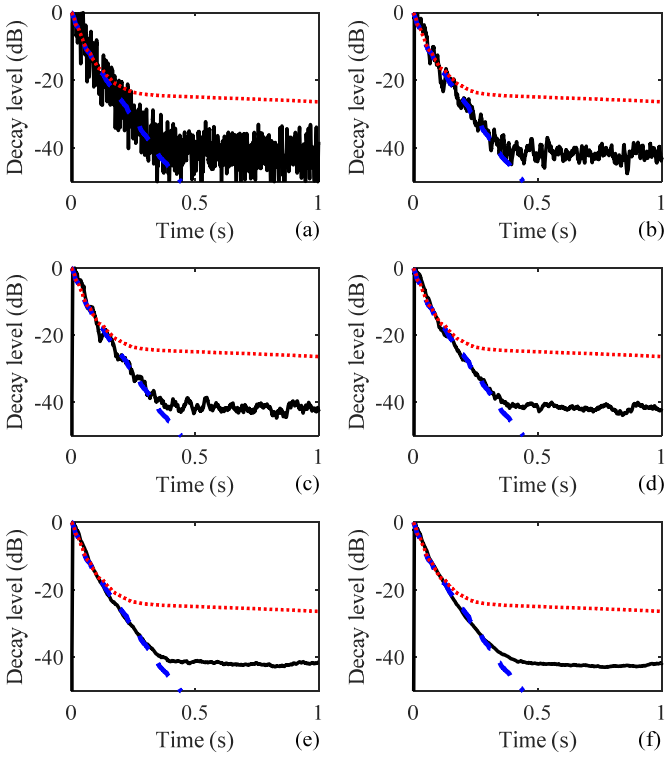


Fig. 7 Decay curves obtained from synthesized RIR filtered in the third-octave band at 2000 Hz: reference EDC (- - -), noisy EDC (· · ·) and ALS RIR (—) generated by using averaging window of 100 (a), 500 (b), 1000 (c), 2000 (d), 5000 (e) and 10000 (f) data points.

Deviation of an ALS RIR from the reference EDC is quantified by the RMSE calculated in the RMSE interval as described in the method section. This error for the representative example of RIR for both broadband response and response filtered in third-octave bands from 100 Hz to 5 kHz is given in Fig. 8. The first observation is that the RMSE is reduced by an increase of the averaging window length. This is the case in every frequency band, but also in the broadband RIR. The error reduction is the largest in the beginning of these plots, with initial increase of the averaging window length from several tens of data points to several hundreds of data points. However, after this initial reduction of the RMSE, there is a region of saturation of the RMSE reduction, and at the end of the plots the RMSE is even increased with an increase of the averaging window length. The saturation region is reached for shorter average window lengths at higher frequencies, see Fig. 8(c). The shape of presented RMSE at higher frequencies contains relatively wide plateau.

The previous observation indicates that there is an optimal length of the averaging window, that is, an optimal number of data points to be averaged. Now, the question is whether this optimal averaging window length is the same for all frequency bands or not. At low and mid frequencies, the optimal averaging window length is from 5000 to 10000 data points. With an increase in frequency, this optimal window length becomes a bit smaller, and it ranges from a few thousands to 5000 data points. For broadband RIR, optimal length of the averaging window is about 5000 data points. Taking all these data into account, it can be concluded that the optimal length of the averaging window is about 5000 data points in the considered situation.

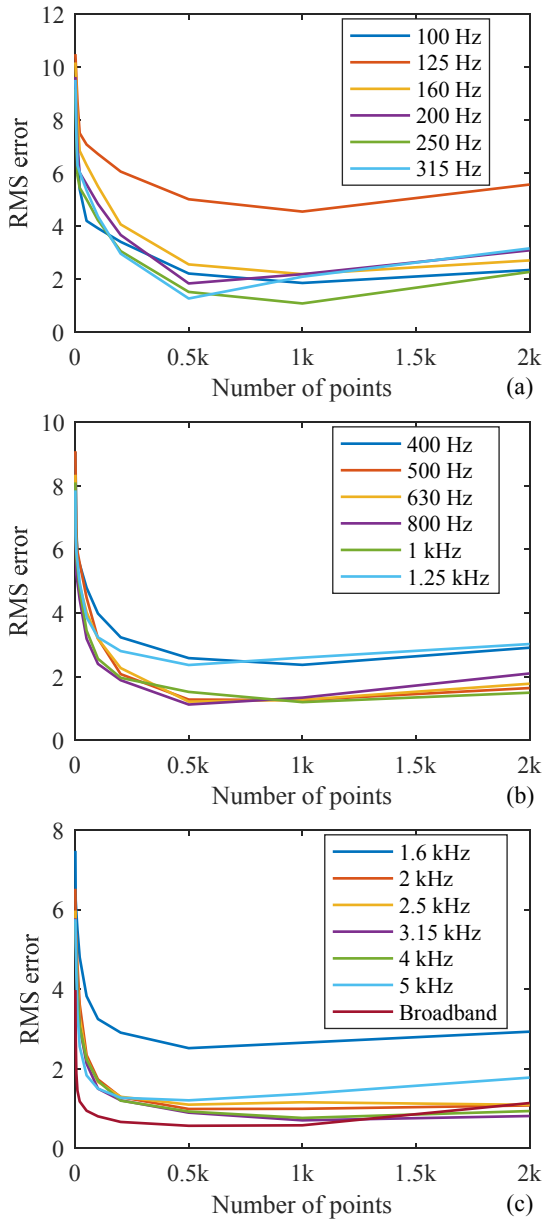


Fig. 8 RMSE calculated in the RMSE interval as RMS error between the considered ALS RIR and reference EDC for both broadband RIR and RIR filtered in third-octave bands.

Regarding the processing time necessary for moving average method, the longest time is required for the longest averaging window length of 20000 data points. For this window length, the averaging lasts about a few tens of seconds per frequency band. Having 18 frequency bands in overall, the processing for all bands will last about 5 minutes.

VI. Conclusion

An EDC generated by the Schroeder's backward integration exhibits drawbacks of cumulative summing of background noise and finite upper limit of integration of a noisy RIR. These drawbacks can be (fully or at least partly) overcome by generating the decay curve using an alternative method based on averaging of a logarithmic squared RIR (LS RIR). The averaging reduces the fluctuations of LS RIR, and at the same time, it enables generation of the decay curve of larger dynamic range than the one generated by the Schroeder's integration. This curve has a certain disturbance at its beginning since there is no full averaging window for data points close to the LS RIR start. Besides, the averaged LS RIR (ALS RIR) can be curved in vicinity of the knee due to effect of averaging noise points together with reverberation decay points when the averaging window is large enough.

By increasing the averaging window length (number of data points of the LS RIR to be averaged), fluctuations of the LS RIR are reduced to a higher extent. However, the disturbance effect at the decay curve (ALS RIR) beginning and curvature close to the knee indicate that the averaging window should not be too large. The investigation presented here shows that the optimal averaging window length is about 5000 data points for the analysed cases. The proposed procedure will be applied to the measured RIRs in the next phase of research.

Acknowledgment

This research is supported by the Ministry of Education, Science and Technological Development of Serbia through the project No. 36026.

References

[1] H. Kuttruff, *Room Acoustics, 4*th ed. New York, USA: Spon Press, 2001.

[2] S. Müller, P. Massarani "Transfer measurement with sweeps," *J. Audio Eng. Soc*, vol. *49,* no. *6,* pp. *443-471,* June 2001.

[3] M. R. Schroeder, "New method of measuring reverberation time," *J. Acoust. Soc. Am*, vol. *37,* no. *3,* pp. *409-412,* 1965.

[4] M. Guski, M. Vorländer, "Comparison of noise compensation methods for room acoustic impulse response evaluations," Acta Acust united Ac, vol. 100*,* no. *2,* pp. *320-327,* March 2014.

[5] *Acoustics - Measurement of Room Acoustic Parameters - Part 1: Performance Spaces*, ISO 3382-1, 2009.

[6] *Application of New Measurement Methods in Building and Room Acoustics*, ISO 18233, 2006.

[7] M. V. Janković, D. G. Ćirić, "Automated estimation of the truncation of room impulse response by applying a nonlinear decay model," *J. Acoust. Soc. Am*, vol. *139,* no. *3,* pp. *1047-1057,* March 2016.

[8] D. G. Ćirić, M. V. Janković, "Correction of room impulse response truncation based on a nonlinear decay model," *Appl Acoust*, vol. *132,* no. *3,* pp. *210-222,* March 2018.

[9] N. Xiang, "Evaluation of reverberation times using a nonlinear regression approach," *J. Acoust. Soc. Am*, vol. *98,* no. *4,* pp. *2112-2121,* 1995.

[10] Smooth response data. Mathworks. Available at: https://www.mathworks.com/help/curvefit/smooth.html.

[11] Filtering and smoothing data. Mathworks. Available at: https://www.mathworks.com/help/curvefit/smoothing-data.html

[12] E. A. Lehmann, A. M. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *J. Acoust. Soc. Am*, vol. *124,* no. *1,* pp. *269-277,* 2008..

Noise Control Solution of the HVAC System

Marko Ličanin, Darko Mihajlov, Momir Praščević, and Ana Đorđević

***Abstract*—Building construction trends in designing the modern buildings very often exploit the use of the centralized Heated Ventilated Air Conditioning (HVAC) units to maintain optimal water and air condition. These systems need fair amount of space and require significant airflow to operate properly. Being often placed at the open areas, where airflow is sufficient, they create the high noise levels that influence the living environment. This becomes a challenge for the noise control engineers to find the proper solutions for attenuating the generated noise, without enclosing the space around HVAC units. To account for the necessary airflow while still perform noise reduction, here, a simulation of the particular type of segmented barrier is done. Results of the simulation are, together with a noise measurements, then applied on the real case scenario of the HVAC units noise control. Results showed that the simulated barrier can be used as solution for this type of problems.**

***Index Terms*—Noise Control, Acoustic Barrier, COMSOL, HVAC, Absorption**

I. Introduction

Heated Ventilated Air Conditioning (HVAC) units are very often integral parts of modern buildings, especially those that have a community purposes (offices buildings, supermarkets, shopping malls, hospitals etc.). Those units are very convenient in the term of the efficiency, size, maintenance and heating/cooling power outputs, thus representing the logical choice instead of using many local wall mounted units. However, in many cases these types of machine systems are positioned on the rooftops of buildings, where the best cooling of the sub units, such as fans, pipes and compressors is provided. To perform sufficient cooling, air intake and exhaust should be given enough airflow [1]. Since the HVAC units often move a massive amount of air around them, they can become a noise sources with a significant acoustic power. This parameter is usually provided in the technical data-sheets of the units, where normal operation mode is very often used as reference by technicians who mount the units. As a result, noise reduction is in many cases neglected.

Working load of the HVAC systems is largely influenced by the nature of climate in which they operates. During the hot summers and cold winters HVACs put more power into providing desired temperatures, working beyond the normal operational mode. This translates into generation of a higher noise levels, thus grater disturbance on the environment. This is especially prominent during the summer, where people in residential areas and office buildings tend to open windows and balconies for fresh air. Noise problem then becomes on of the biggest environmental disturbance and quick, as well as effective, engineering solutions are sought. Unfortunately very fast solutions can do more harm than good, increasing the costs and creating additional problems.

Engineering challenge when performing the noise control of the HVAC systems is to provide the proper noise attenuation [2, 3] while at the same time allow sufficient airflow for units subsystems cooling. This paper tries to address one of the possible solutions to the problem, that can be applied in different situations. These solutions are studied through the project that has be successfully implemented for the central office building of one international bank in Belgrade. After identifying the problem at the site location, a series of simulation has been performed in COMSOL that exploits certain geometry of absorption panels. Based on the chosen geometry and measured noise spectrum, an optimal metal perforated panels has been designed and noise reduction calculated. Placement of the panels has been observed in the 1:1 built 3D model, and in the cooperation with a construction engineer, supporting structure designed. After the construction and panel mounting is performed, measurements have been retaken, and results of the practical application observed. Although, there is a lot of technical documentation behind the realized project, this paper tries to summarize most important aspects of this particular noise control design.

II. Analysis of the Problem

The initial step of the project was the existing condition evaluation at the site. Machine elements are located in the support building, next to the office building of the international bank. Three HVAC units (chillers) are located next to each other in the technical area enclosed with three walls covered with metal sheets, and one partition consisting of the metal blinds. Here is no ceiling above and this area is considered as a rooftop. These three units create excessive amount of noise which creates disturbance in the office building across the small parking access street. Sketchup representation of the described situation is presented in figure 1. The evaluation step included initial measurements of the noise level that has been taken at several positions at the site using Norsonic Nor 140 sound analyzer. Three measurement positions have been chosen at 1m distance from the each of

Marko Ličanin, University of Niš, Faculty of occupational safety, Čarnojevića 10A, 18000 Niš, Serbia (e-mail: marko.licanin@ znrfak.ni.ac.rs).

Darko Mihajlov, University of Niš, Faculty of occupational safety, Čarnojevića 10A, 18000 Niš, Serbia (e-mail: darko.mihajlov@ znrfak.ni.ac.rs).

Momir Praščević, University of Niš, Faculty of occupational safety, Čarnojevića 10A, 18000 Niš, Serbia (e-mail: momir.prascevic@ znrfak.ni.ac.rs).

Ana Đorđević, University of Niš, Faculty of Electronic engineering, Aleksandra Medvedeva 14, 18000 Niš, Serbia (e-mail: ana.djordjevic@ elfak.rs).

the noise sources (1, 2 and 3 on Fig. 1), while three positions has been considered on the noise path near the affected building (4, 5, 6 on Fig. 1).

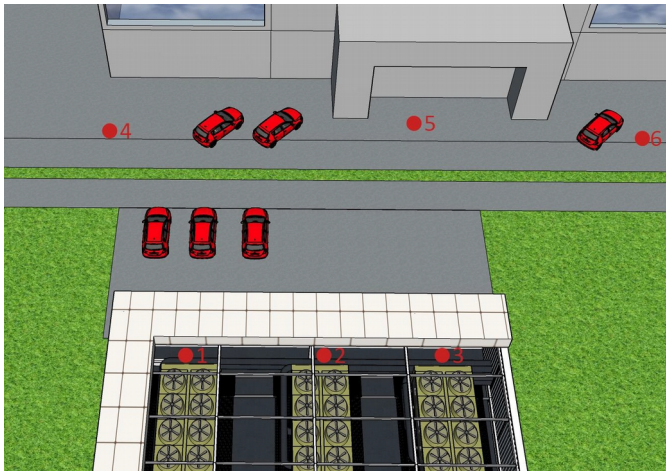


Fig. 1. Sketchup representation of the noise sources, affected office building and noise level measurement positions during the initial evaluation

Measurement interval has been set to 5 min, and measurements have been taken three times at each chosen position. HVAC units can be set in different operation modes and under the different output loads. For the measurement conditions, all of three chillers has been set to cooling and output load increased significantly beyond usual operation. This was done to account for the stationary noise that will be produced by the noise source during the hot summer days. To determine the equivalent noise disturbance, measurements would be done to account the all operational mode, and measurement time increased to at least 15 min [4]. However, for the purpose of establishing base line before noise control application, and evaluate relative noise level reduction after the noise control application, this measurement method was sufficient. Noise levels of measurement position 2 and 5 are presented in the Fig. 2 (top). Resolution of the observed noise levels is in the third octave bands.

In front of the units, measured noise level was similar for positions 1, 2 and 3, while for positions 4, 5 and 6 there was slight different in the spectrum due to the noise traffic. Even, though the street that separates the bank support facility and affected office building is somewhat secluded from the main road, noise influence from the parking zone was present. For this reason, background noise on positions 4, 5 and 6 has been also measured. However, the total broad spectrum noise level was not as different, as the variations in the spectrum. In front of the units (positions 1, 2 and 3), measurement showed between 79.8 dB and 81.8 dB, while at the street noise was 64.3 and 65.2 dB (positions 4, 5 and 6). When chiller units was turned off, background traffic noise levels were between 54.6 dB and 55.3 dB (positions 4, 5 and 6). After presenting the results to the investor, decision is made that noise control should provide at least of 8 dB of noise attenuation. The reason is that Belgrade is not yet acoustically zoned, but this is a process that will probably be executed in the near future. If the city area at the position of the bank office, area can become denoted as residential. In this case, all off the noise sources must not exceed 55 dB during the day and 45 during the night. By reducing more than 8 dB of noise, problem will be solved even by the most strict noise level regulations [5].

After the noise level evaluation, using the construction blueprint of the technical room, entire object has been made in 3D in 1:1 ratio so different engineering solutions can be visually exploited.

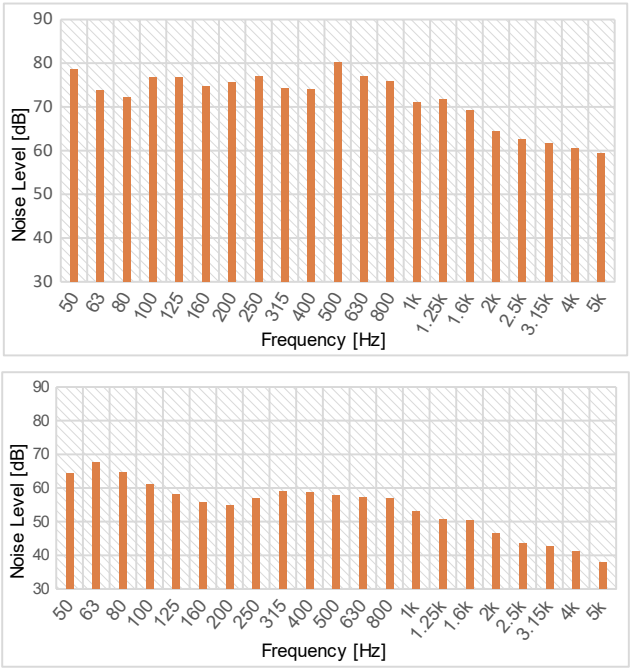


Fig. 2. Noise levels in third octave band, averaged per each band for three repeated measurements at position 2 (Top) and position 5 (bottom)

The engineering solution for this particular problem has been thoroughly observed from many different aspects. Air flow necessary for the proper operation of the HVAC subsystems has proved to be one of the major challenges. At one point, before the start of the project, chiller maintenance engineer was impatient and wanted a quick solution of the noise problem. As a result, hard styrodur barrier was installed which redirected the noise to the upper floors of the affected building and increased the temperature around the chillers. This influenced their operational stability. Eventually, barrier has been removed which created significant waste of time and effort, creating unnecessary costs.

III. DESIGN OF THE NOISE CONTROL SYSTEM

To solve the problem of the needed air transparency, investigation of the multi segment barrier has been considered. Traditional single segment barrier has proved to be of no use in this kind of situation. It has been clear from the start that some portion of the noise has to be redirected to a longer path, which will naturally attenuate the sound energy. This has been exploited, as the location of another office building where redirection was intended lies at much longer air distance from the HVAC units. At the entrance of this building noise level has been dominant by traffic along nearby busy street.

A. *Simulation*

For the purpose of the investigation, a simulation of the noise propagation has been performed in COMSOL Multiphysics. Designed segmented barrier that was tested in the simulation is composed of absorption metal perforated panels angled at the 25 degrees from the barrier axis. Fig. 3 shows the simulation results in the case where there was no barrier, and the case where barrier is included. Results are presented at frequencies of 100 Hz (Fig. 3 - top), 500 Hz (Fig. 3 - middle) and 1 kHz (Fig. 3 – bottom). As it can be seen, when barrier is placed, more noise energy is contained inside the technical room where chillers are placed. This is especially observable at middle to high frequency. Beside presented results, simulation has been done in much higher resolution of frequencies, and only some of them are presented here. It is important to note that during the simulation, no absorption on the segmented barrier has been set and only noise propagation is observed.

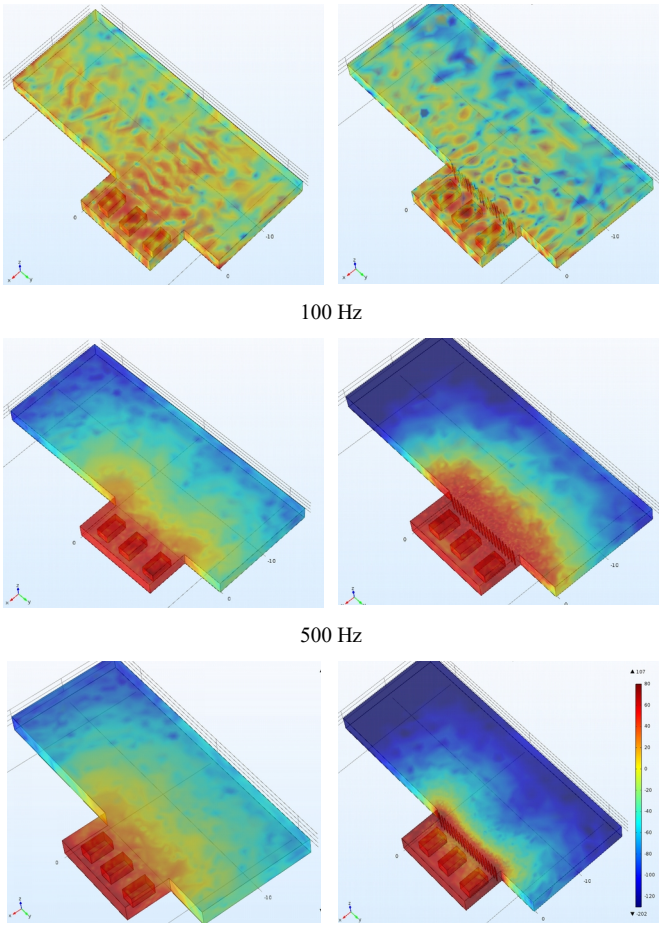


Fig. 3. Simulation results the sound propagation for different frequencies, without barrier (left) and with barrier (right)

Beside observation of the sound propagation, another simulation has been preformed using Raytrace method. As mention before, some of the noise has to redirected to the longer path and simulation has been used for evaluation of the geometry of the segmented barrier. In Fig. 4, results has been presented in a way that left portion of the figures represent the case without segmented barrier (barrier influence is disabled), while on the right side situation with barrier is shown (barrier is enabled). Ray trace metod works by releasing the vectors from the noise source in all aof the directions. By following the paths of these vectors, one can observe geometrical propagation of the sound at the chose n frequency. In presented simulation, a 10 000 rays ahas been released and the reflection influence of barrier redirection of sound observed. Main goal is to evaluate the number of rays, as well as the ray intensity that reaches the noise receiver after a certain points in time. Simulation length is 120 ms, with a 4 steps 30 ms each. In Fig. 4, first row is a situation after the first step, where each subsequent row represents the another step. As it can be observed, placement of the segmented barrier (Fig. 4 - right) truly redirects the sound to a longer path.

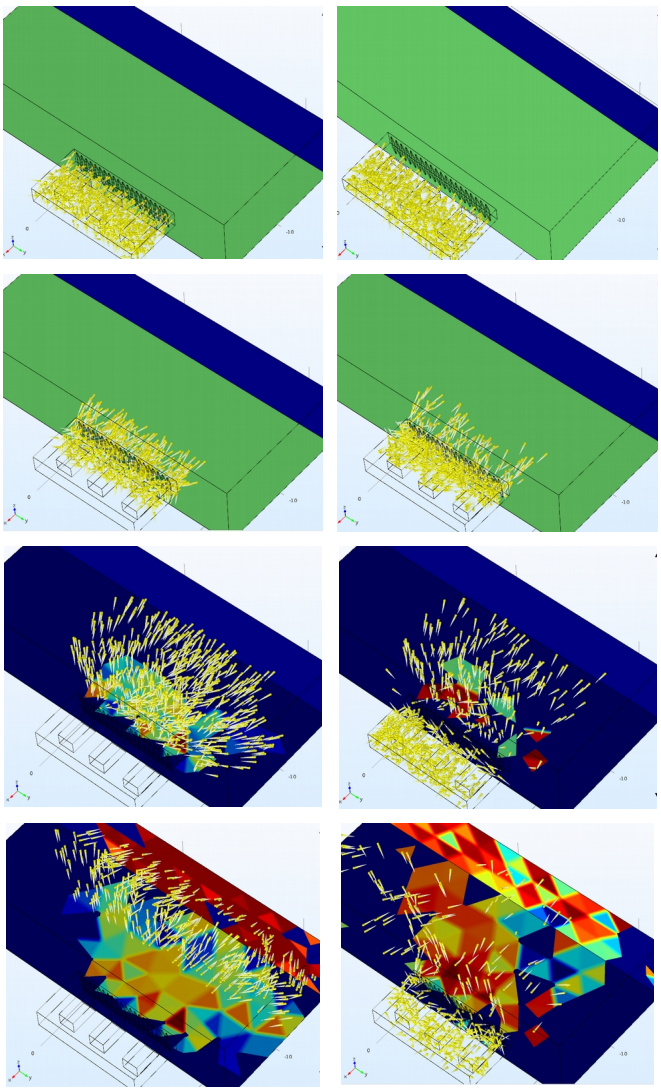


Fig. 4. Simulation results the segmented barrier noise redirection in the case where barrier is disabled in the model (left) and enabled (right). First step (0 ms), second step (60 ms), third step (90 ms) and fourth step (120 ms) shows the flow of the simulation.

B. *Segmented barrier*

Simulation results has confirmed that the assumed geometry of the segmented barrier can be used for solving the problem of the HVAC noise. To maximize the effectiveness

of the barrier, metal perforated panels has to be designed as well. The process of calculating the absorption coefficient of the perforated panels is well known in literature, and it not a subject of this paper. This process accounts for the front perforated panel thickness, hole repeat distance, hole radius, cavity depth, absorber thickness, air space thickness and flow resistivity[2, 3, 5]. Absorption can be calculated for each third octave band, and frequency dependent curve fitted to match the noise spectra of the HVAC, obtained by measurements in the initial phase of the project. Cross section of the designed metal perforated absorption panel (top), and its absorption coefficient (bottom) are presented in Fig. 5. Panel is filled with the mineral wool of 30 kg/m density.

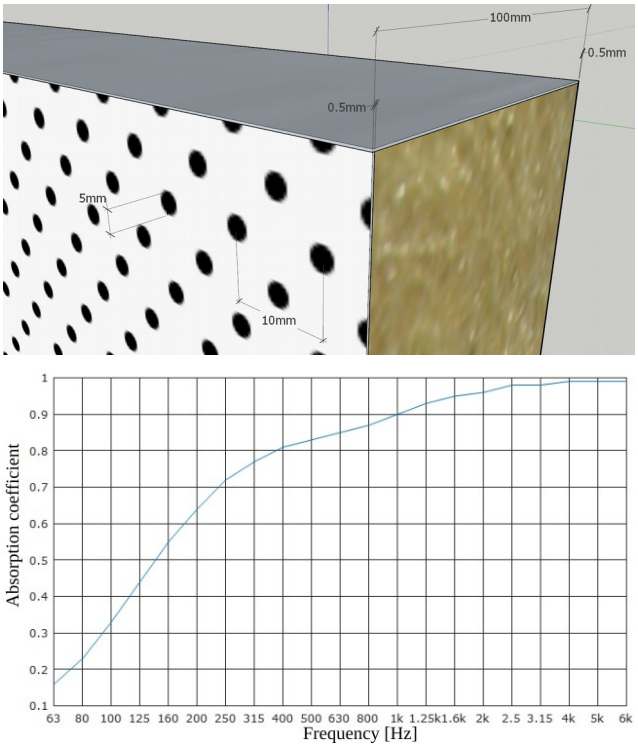


Fig. 5. Designed metal perforated panel cross-section (top) and the absorption coefficient curve (bottom)

Barrier is composed by connecting multiple panels using the metal construction material to form the rigid cage that will provide the sufficient static to the weight of the panels. In the design process, another segmented barrier has been consider. To ensure that noise generated by the fans of the chiller units will not reach the upper floors of the affected building, this additional barrier has been included on the top of the existing metal support beam structure on the site. Beside the absorption effect, it will also redirect the noise towards the sky. Sketchup representation of both of these barriers is presented in Fig. 6. For the easier observation, the rooftop and the upper barrier is removed from the 3D model (Fig. 6 – top). Support structure where upper barrier is placed is also visible on the Fig. 6. This metal beam structure is responsible for the static of the entire roof segment. The construction consultant has calculated that even with a strong wind stress on the support beams there will be no disturbance of the structural integrity of this section. If this has not been the case, additional support structure would have to be built.

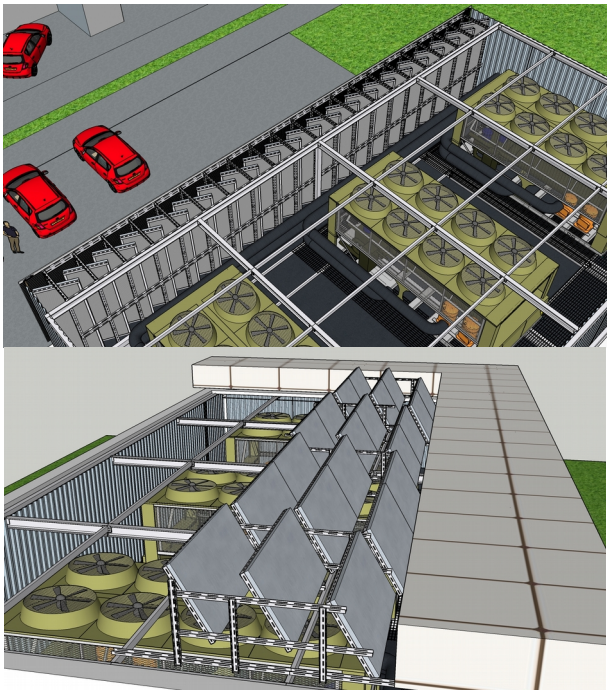


Fig. 6. Graphical representation in 3D of the lower barrier (top) and the barrier above the chiller (bottom)

IV. RESULTS

After the construction works and barriers mounting has been finished, validation process took place. This has been done by retaking all of the measurements in the same way as in the initial stage. Measurements has been performed in six positions with a 5 min averaging. For each position, noise level has been recorded three times and values averaged. HVAC units as a noise source has been set under the same working load as it was the case at the beginning of the process, thus the stationary noise conditions were established. At the positions 4, 5 and 6, background noise has been also measured, where chiller units were turned off. This has been done to determine the traffic noise level. If the noise levels of the background noise are similar to the noise levels when HVAC units are turned, at measurement positions 4, 5 and 6 (see Fig. 1), this would mean that the background noise is dominant. In this situation, for the project validation purpose, only measurement positions 1, 2 and 3, on the side of the lower barrier opposite to the noise sources positions, can be used to evaluate the effectiveness of the project. Results of the validation process can be observed in the Table I. In all of the measurement positions, noise level has been reduced more than 9 dB. Comparison of the background noise during the first and second measurements (measurement positions 4, 5 and 6), as well as the resulting noise level when units are working, is presented in Table II. We observe that background noise levels are consistent between measurements, which indicates that there was no significant change in the traffic density at this particular location during noise analysis. It is also very interesting to see that resulting noise levels after the application of the noise control are within the boundaries of the background noise as well. This is yet another proof that

project has been successful, as the noise level in the measurement positions will be masked by the background traffic noise. In the Fig. 7, several photos of the mounted barriers at the site are presented.

Fig.7. Documented photos during and after the mounting process

TABLE 1
MEASUREMENT RESULTS OF THE VALIDATION. COMPARISON BETWEEN INITIAL AND VALIDATION STAGE OF THE PROJECT

Meas. position	1.	2.	3.	4.	5.	6.
	Noise level [dB]					
Initial phase	81.8	79.8	80.1	64.4	65.2	64.3
Validation phase	71.9	70.1	70.2	54.4	54.8	54.3
Noise reduction	9.6	9.7	9.9	10.0	10.4	10.0

TABLE 2
COMPARISON OF THE BACKGROUND NOISE BETWEEN INITIAL STAGE, VALIDATION STAGE AND NOISE LEVEL WHEN NOISE SOURCE IS ACTIVE

Meas. position	4.	5.	6.
	Noise level [dB]		
Initial phase background noise	55.2	53.2	54.1
Background noise levels	53.8	54.8	54.2
Validation phase noise HVACs are active	54.4	54.8	54.3

V. CONCLUSION

When designing the noise control system around HVAC systems, there are many challenges that creates constraints and significantly influence the flow of the project. Here, a study case of one of the noise control projects is described from initial steps to the installation and finally validation. Measurements in the first stage has provided useful inputs on the noise level, spectrum of the noise and constraints related to the HVAC units. In the simulation stage, noise propagation and the behavior of the proposed segmented barrier has been observed. Gather data showed promising results that was used in the following stage of designing the physical barrier and absorption panel. After the material production and mounting stage, validation process has been performed by measuring noise levels in the same way as in the first stage. Results showed that total noise reduction achieved, at each of the measurement positions, is more than 9 dB. This exceeds the expectation of the project and proves the feasibility of this type of noise control application. Different types of segmented barrier geometry and configuration could provide even better results, which is a topic for the future research.

ACKNOWLEDGMENT

This paper is presented as a part of the projects "Development of the methodology and means for noise protection in urban areas" (No. TR - 037020). Authors would like to express gratitude to the Ministry of education, science and technological development of Republic of Serbia, for the financing support in these research.

REFERENCES

[1] R. McDowall, *Fundamentals of HVAC systems* 1st ed. Burlington, Great USA: Elsevier, 2007.
[2] F. Fahy, *Foundations of Engineering Acoustics* 2nd ed. London, Great Britain: Academin Press, 2003.
[3] D. A. Bies, C. H. Hansen, *Engineering noise control – Theory and Practice,* 3rd ed. London, Great Britain: Spon Press, 2003.
[4] M. Praščević , D. Mihajlov, D. Cvetković, "Permanent and semi-permanent noise monitoring," *Journal of Low Frequency Noise, Vibration and Active Control*, vol. *34,* no. *3,* pp. *251-268, Serbia*. March, 2015.
[5] G. Licitra, *Noise maps in the EU: Models and procedures* 1st ed. New York, USA: CRC Press, 2012.
[6] M. Norton, D. Karczub, *Fundamentals od Noise and Vibraion Analysis for Engineers 2nd* ed. New York, USA: Cambr. University Press, 2003.

Određivanje zavisnosti ostvarene vrednosti izolacione moći fasadnih pregrada od tipa izvora u urbanim sredinama

Miodrag Stanojević, Miloš Bjelić, Dragana Šumarac Pavlović, Miomir Mijić, Tatjana Miljković

***Apstrakt*— Oblik ugaone raspodele incidentne energije utiče na ostvarenu vrednost izolacione moći fasadne pregrade. U opštem slučaju oblik ove raspodele nije poznat. Primenom mikrofonskog niza i algoritama za prostorno-vremensku obradu signala moguće je eksperimentalno utvrditi funkciju gustine verovatnoće ugaone raspodele incidentne energije na fasadi zgrada. U urbanim uslovima postoji veliki broj izvora različitog tipa, spektralnog sadržaja buke koju emituje, snage itd. Zbog toga je uvedena pretpostavka da oblik ugaone raspodele, a samim tim i ostvarena izolaciona moć pregrade, zavisi od tipa izvora buke. U ovom radu prikazana je upotreba metodologije za eksperimentalno određivanje ugaone raspodele sa ciljem da se posmatraju razlike u ostvarenim vrednostima izolacione moći iste pregrade prilikom delovanja različitih tipova zvučnih izvora u urbanim uslovima. Ostvarena izolaciona moć izračunata je korišćenjem dobijenih funkcija gustine raspodele i pokazano je da se ona menja u vremenu i zavisi od trenutne strukture zvučnog polja. Takođe, izvršeno je poređenje ostvarenih vrednosti izolacione moći izračunatih za pojedinačne događaje i izolacione moći izračunate za duži vremenski period. Na taj način moguće je sagledati uticaje pojedinih tipova izvora na generalno stanje zvučne izolacije fasada.**

***Ključne reči*— lokalizacija, mikrofonski niz, proračun zvučne izolacije, saobraćajna buka, ugaona raspodela, urbani uslovi.**

I. UVOD

U literaturi je pokazano da vrednost zvučne izolacije, odnosno izolacione moći pregrade *R*, zavisi od frekvencije i ugaone raspodele incidentne energije [1-2]. Raspodela incidentne energije spoljašnje buke u opštem slučaju nije poznata. Uniformna ugaona raspodela je polazna pretpostavka u radovima u kojima se izračunava izolaciona moć pregrade. U literaturi se mogu pronaći merenja zvučne izolacije fasada korišćenjem standardizovanih tehnika merenja pomoću zvučnika i pomoću saobraćaja, i pokazano je da vrednosti $D_{2m,nT}$ mogu varirati i do 8 dB [3]. Izvesno je da ove varijacije potiču od razlike između incidentnih uglova realne saobraćajne buke i incidentnog ugla 45±5°, koji se koristi u tehnikama merenja sa zvučnikom [4]. Na osnovu toga zaključuje se da ugaona raspodela incidentne energije zavisi od konfiguracije terena u kom se analizirana zgrada nalazi. Konfiguraciju terena određuje širina ulice, broj saobraćajnih traka, prisutnost naspramnih zgrada, visina zgrada i sl.

Upotrebom optimizovanog neregularnog mikrofonskog niza i algoritama za prostorno-vremensku obradu signala moguće je odrediti nepoznate ugaone raspodele incidentne energije spoljašnje buke na fasadi zgrada u urbanim uslovima [5, 6]. Izvršeno je nekoliko merenja u urbanim uslovima kako bi se sagledale razlike u oblicima ugaonih raspodela za različite konfiguracije terena ispred ravni fasade [7]. Eksperimentalno određene ugaone raspodele incidentne energije spoljašnje buke i poznate vrednosti građevinskih parametara korišćene su za predikciju izolacione moći fasadnih pregrada [8, 9]. Pokazano je da vrednosti izolacione moći mogu varirati i do 10 dB za istu pregradu u različitim uslovima za pojedinačne frekvencijske opsege i oko 5 dB za merodavne vrednosti. Na osnovu oblika ugaonih raspodela definisane su kategorije konfiguracija terena u urbanim uslovima. Pokazano je da merna mesta koja pripadaju istoj kategoriji imaju sličan uticaj na izolaciona svojstva fasadnih pregrada [7].

U ovom radu analiziran je uticaj pojedinačnih izvora buke na ostvarene vrednosti izolacione moći pregrade. U urbanim uslovima postoji veliki broj izvora različitog oblika, veličine, snage i spektralnog sadržaja buke koji emituju. Zbog toga će se javiti razlike u oblicima funkcija gustina verovatnoće ugaonih raspodela incidentne energije na fasadnoj pregradi za različite tipove izvora. Različiti oblici ugaonih raspodela mogu dovesti do toga da ista fasadna pregrada ispoljava različita izolaciona svojstva u zavisnosti od toga koji izvori buke trenutno postoje ispred fasade. Za jednu situaciju iz [7] definisano je nekoliko tipova zvučnih izvora koji su snimani mikrofonskim nizom, a zatim je nad snimljenim signalima primenjen algoritam za prostorno-vremensku obradu signala pomoću koga je određen oblik ugaonih raspodela. Izvršen je proračun ostvarenih vrednosti izolacione moći za dva tipa materijala od kojih je izrađena pregrada, a zatim sprovedena analiza dobijenih rezultata.

Motivacija za ovo istraživanje je potreba da se smanji vreme potrebno za dobijanje ugaonih raspodela incidentne energije na mernom mestu. Da bi se predloženom metodom odredila ugaona raspodela potrebno je izvršiti snimanje buke mikrofonskim nizom u dužem intervalu vremena. Nakon toga potrebno je primeniti algoritme za obradu signala sa mikrofonskog niza kako bi se dobili oblici ugaonih raspodela. Vremenska i računska kompleksnost ovog

Miodrag Stanojević – Elektrotehnički fakultet, Univerzitet u Beogradu, Bulevar kralja Aleksandra 73, 11020 Beograd, Srbija (e-mail: miodragstanojevic@bitprojekt.co.rs).

Miloš Bjelić – Elektrotehnički fakultet, Univerzitet u Beogradu, Bulevar kralja Aleksandra 73, 11020 Beograd, Srbija (e-mail: bjelic@etf.rs).

Dragana Šumarac Pavlović – Elektrotehnički fakultet, Univerzitet u Beogradu, Bulevar kralja Aleksandra 73, 11020 Beograd, Srbija (e-mail: dsumarac@etf.rs).

Miomir Mijić – Elektrotehnički fakultet, Univerzitet u Beogradu, Bulevar kralja Aleksandra 73, 11020 Beograd, Srbija (e-mail: emijic@etf.rs).

Tatjana Miljković – Elektrotehnički fakultet, Univerzitet u Beogradu, Bulevar kralja Aleksandra 73, 11020 Beograd, Srbija (e-mail: tm@etf.rs).

postupka direktno je srazmerna trajanju snimljenog signala. Ideja je da se izvrši snimanje ključnih događaja koji se javljaju na toj lokaciji, da se izvrši određivanje oblika ugaone raspodele za te događaje i da se na osnovu toga odredi ukupan oblik raspodele, odnosno ispoljena vrednost izolacione moći pregrade. Time bi se značajno smanjilo vreme merenja ali i vreme obrade signala. Ovaj rad predstavlja uvodno istraživanje u kome bi se sagledao uticaj pojedinačnih događaja na oblik funkcije gustine verovatnoće ugaone raspodele incidentne energije. Poređenjem oblika raspodele za pojedinačne događaje i oblika dobijenih za duže vremenske intervale snimanja moglo bi se doći do veze koja bi omogućila smanjivanje vremena potrebnog za izračunavanje globalnog oblika raspodele.

Rad je organizovan kako sledi. U drugom poglavlju prikazani su mikrofonski niz, algoritam za prostorno-vremensku obradu signala sa mikrofonskog niza, postupak za izračunavanje ugaone raspodele i model za proračun izolacione moći pregrade. U trećem poglavlju prikazana je postavka eksperimenata. U četvrtom poglavlju prikazani su rezultati i njihova diskusija. U poslednjem poglavlju izneti su zaključci.

II. METODE

A. Mikrofonski niz

Mikrofonski niz se u ovom radu koristi kao merni alat za određivanje lokacije izvora u prostoru u odnosu na fasadnu pregradu. Korišćen je planarni mikrofonski niz sa 24 mikrofona. Na osnovu referentnog spektra buke u urbanim uslovima [10] i spektra dobijenog merenjem u urbanim uslovima [11] određen je frekvencijski opseg od interesa za analizu. Pozicije mikrofona u mikrofonskom nizu izabrane su tako da se u opsegu od 250 Hz do 2000 Hz dobiju što bolje karakteristike mikrofonskog niza. Širina glavnog loba određuje prostornu rezoluciju sa kojom je moguće odrediti pozicije zvučnih izvora, dok nivo bočnih lobova određuje prisustvo lažnih zvučnih izvora u rezultatima lokalizacije [7]. Postupak optimizacije i konstrukcije mikrofonskog niza dat je u [7, 13-14].

B. Algoritmi za prostorno-vremensku obradu signala

Određivanje ugaone raspodele incidentne energije bazira se na određivanju lokacije pojedinačnih zvučnih izvora pomoću mikrofonskog niza i algoritama za obradu signala sa mikrofonskog niza. Bazični algoritmi za obradu signala, kao što su *Delay and sum* algoritam (*DaS*) [15] i konvencionalni *beamforming* algoritam (CB) [16], nisu pogodni za korišćenje u ovoj aplikaciji zbog male dinamike koja se pomoću njih može ostvariti. U ovom radu za lokalizaciju izvora u prostoru korišćen je algoritam DAMAS2 [17]. Ovaj algoritam koristi rezultat dobijen pomoću CB algoritma, a zatim dekonvolucionim postupkom eliminiše uticaj mikrofonskog niza iz rezultata. Na taj način ostvaruje se veća dinamika, odnosno mogu se lokalizovati izvori koji su značajno tiši od dominantnih zvučnih izvora [18]. Kao rezultat algoritma dobija se „zvučna mapa“ na kojoj je vrednost zvučne energije predstavljena bojama.

C. Izračunavanje funkcije gustine verovatnoće ugaone raspodele incidentne energije na fasadi

Ugaonu raspodelu incidentne energije spoljašnje buke na fasadi potrebno je odrediti za uglove u odnosu na normalu, koji ne odgovaraju direktno vrednostima uglova korišćenim u algoritmima za prostorno-vremensku obradu signala. To znači da je potrebno izvršiti svođenje prostornih uglova azimuta i elevacije na uglove u odnosu na normalu. Uglovi se nalaze u opsegu od 0° do 90°, gde ugao od 0° odgovara normali na fasadu, a ugao 90° ima krak koji je paralelan sa fasadom. Postupak za svođenje prostorne raspodele energije, izračunate pomoću algoritama i mikrofonskog niza, na ugaonu raspodelu u odnosu na normalu opisan je u [5-7].

D. Proračun izolacione moći fasadne pregrade

Izolaciona moć monolitne fasadne pregrade može se izračunati na osnovu sledeće jednačine [19]:

$$R = -10\log(\tau), \tag{1}$$

gde τ predstavlja koeficijent transmisije analizirane pregrade. Koeficijent transmisije se izračunava na sledeći način [19]:

$$\tau(\omega) = \sum_{k=0}^{\pi/2} \frac{p(\theta_k)}{\left(1+\eta\frac{\omega m'}{2\rho c}\left(\frac{f}{f_c}\right)^2\cos\theta\sin^4\theta\right)^2 + \frac{\omega^2 m'^2}{4\rho^2 c^2}\cos^2\theta\left(1-\left(\frac{f}{f_c}\right)^2\sin^4\theta\right)^2}, \tag{2}$$

gde je:

- η faktor gubitaka u materijalu
- f frekvencija za koju se računa izolaciona moć
- ω kružna učestanost
- ρc specifična impedansa u vazduhu
- f_c karakteristična frekvencija materijala
- θ incidentni ugao
- m' površinska masa pregrade
- $p(\theta)$ funkcija gustine verovatnoće ugaone raspodele incidentne energije na površini pregrade

Poznavanjem građevinskih karakteristika pregrade i frekvencijski zavisnih funkcija gustine verovatnoće ugaone raspodele incidentne energije moguće je izračunati vrednosti izolacione moći.

III. POSTAVKA EKSPERIMENATA

U ovom poglavlju biće prikazana postavka eksperimenata koji su realizovani u urbanim uslovima pomoću mikrofonskog niza, kao i važni parametri algoritama za prostorno-vremensku obradu signala.

Sl. 1. Profil ulice u kojoj je vršeno merenje i pozicija mikrofonskog niza na fasadi (Ulica cara Nikolaja II)

Mikrofonski niz se postavlja direktno na fasadu zgrade koja je okrenuta ka ulici. Na Slici 1 prikazana je merna situacija u kojoj se naspram zgrade na koju je postavljen mikrofonski niz nalazi parking za automobile i park sa zelenilom. Ova ulica je dvosmerna i u njoj se odvija gust saobraćaj koga čine automobili, motocikli i vozila javnog gradskog prevoza (autobusi i trolejbusi).

Ukupno trajanje merenja iznosilo je 30 minuta i u tom periodu ulicom je prošao veliki broj vozila svih navedenih tipova. Za potrebe ovog rada izvršeno je izdvajanje tipskih događaja iz ukupnog signala sa svih mikrofona mikrofonskog niza. Prilikom izdvajanja tipskih događaja održani su vremenski odnosi između signala sa pojedinačnih mikrofona. Pod tipskim događajima se podrazumeva: prolazak automobila (u jednoj saobraćajnoj traci ili obe), prolazak trolejbusa (u jednoj saobraćajnoj traci ili obe) i tišina (nema prolazaka motornih vozila). Događaji su izabrani tako da predstavljaju najčešće izvore zvuka koji se mogu javiti na prikazanom mernom mestu. Poslednji tip izdvojenih događaja sadrži zvukove koji nisu nastali u blizini mernog mesta već potiču od udaljenih zvučnih izvora do kojih nema optičke vidljivosti. Za svaki tip događaja izdvojeno je nekoliko signala. Za akviziciju signala sa mikrofonskog niza korišćena je merna oprema opisana u [7, 13]. Mikrofonski niz je udaljen od ulice 3 m. Frekvencijski opseg u kom se vrši lokalizacija zvučnih izvora je [250 2000] Hz. Elevacioni ugao θ i azimutni ugao φ uzimaju vrednosti iz opsega [80 160]° i [-90 90], respektivno. Opseg vrednosti prostornih uglova izabran je u skladu sa time gde se realno mogu naći zvučni izvori na ulici i njenoj neposrednoj okolini. Trajanje bloka T nad kojim se vrši obrada signala sa mikrofonskog niza je 20 ms. Broj iteracija u dekonvolucionom algoritmu iznosi 2000.

IV. Rezultati i diskusija

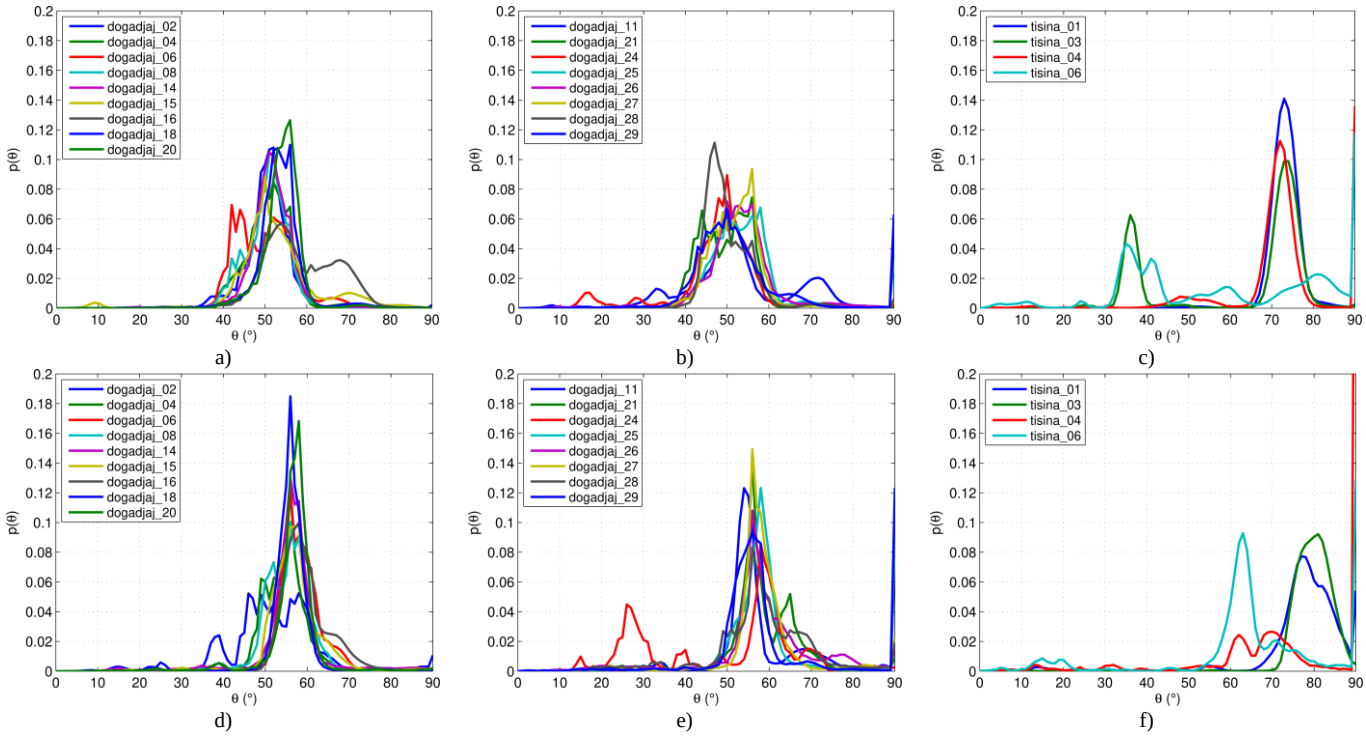
U ovom poglavlju prikazani su izgledi eksperimentalno određenih funkcija gustina verovatnoća ugaonih raspodela incidentne energije spoljašnje buke na fasadi zgrada. Raspodele su određene za različite događaje koji su se javili na mernoj lokaciji sa ciljem da se odredi razlika u ostvarenim vrednostima izolacione moći. Oblici raspodela izračunati su za 1/3 oktavne frekvencijske opsege, ali su zbog ograničenog obima rada prikazani samo neki frekvencijski opsezi. Zaključci koji su izvedeni za prikazane oblike važe i u ostalim opsezima. Na osnovu prikazanog modela za izračunavanje izolacione moći pregrade i eksperimentalno određenih ugaonih raspodela incidentne energije izračunate su ostvarene vrednosti izolacione moći fasadnih pregrada. U ovom poglavlju prikazane su vrednosti ostvarene izolacione moći fasadnih pregrada. Analizirane su fasadne pregrade izrađene od betona i stakla, kao dva najzastupljenija građevinska materijala u urbanim uslovima.

Na Slici 2 prikazani su izgledi funkcija gustina verovatnoće ugaonih raspodela incidentne energije na fasadi zgrade, prikazane na Slici 1. Izgledi raspodela dati su za tri tipa događaja i dva 1/3 oktavna frekvencijska opsega sa centralnim frekvencijama 1600 Hz i 2000 Hz. Analizirano je 9 događaja koji predstavljaju prolazak automobila, 8 događaja koji predstavljaju prolazak trolejbusa i 4 događaja koji predstavljaju ambijentalnu buku (tišinu). Ukupno trajanje snimaka prolaska automobila je 61 sekunda, prolaska trolejbusa 54 sekunde, a tišine 20 sekundi. Na Slikama 2 a) do c) uočavaju se određene razlike u oblicima raspodela za različite tipove događaja. Prolasci automobila saobraćajnicom ispred dovode do toga da se u ugaonoj raspodeli incidentne energije pojavljuje povećanje verovatnoće u oblasti prostornih uglova od 40° do 60°. Buka usled prolaska automobila nastaje zbog motora vozila, sistema za izduvne gasove i kontakta pneumatika automobila sa podlogom. Bliske pozicije dominantnih izvora prilikom prolaska automobila rezultovale su relativno malim opsegom prostornih uglova za koji se dobija povećana verovatnoća zvučne energije.

Trolejbusi kao drugi tip događaja koji se javljaju na ovom mernom mestu proizvode buku zbog kontakta pneumatika sa podlogom, elektromotora vozila kao i zbog kontaktne mreže preko koje se vrši napajanje vozila. Zbog veće prostorne raspodeljenosti izvora po površini vozila javlja se i širi opseg prostornih uglova u kojima se javlja povećanje verovatnoće nailaska zvučne energije. Prilikom prolaska trolejbusa taj opseg je širi nego prilikom prolaska automobila i iznosi od 30° do 70°. Posmatrajući treću vrstu tipskih događaja u kojoj nema prolaska vozila već samo postoji ambijentalna buka, primetno je da se oblast povećane verovatnoće zvučne energije razlikuje u velikoj meri u odnosu na dva prethodno pomenuta tipa događaja. Na Slici 2 c) se uočava više lokalnih minimuma i povećanje vrednosti verovatnoći u funkciji gustine raspodele za uglove bliske 90°. Povećanje verovatnoće u oblasti uglova bliskih 90° je značajno sa stanovišta vrednosti izolacione moći jer se za te incidentne uglove ostvaruje manja vrednost izolacione moći.

Na Slikama 2 d) do f) prikazane su funkcije gustine verovatnoće ugaone raspodele incidentne energije za tri tipa događaja i 1/3 oktavni frekvencijski opseg sa centralnom frekvencijom 2000 Hz. Razlike koje su postojale za prethodni frekvencijski opseg važe i u ovom slučaju. Izvesna odstupanja postoje u tome što su opsezi u kojima se javlja povećana verovatnoća zvučne energije pomereni za nekoliko stepeni ka većim vrednostima uglova. Na Slici 2 f) uočava se da postoji dominantna energija koja pogađa fasadu pod uglovima koji su bliski 90°, za razliku od slučaja prikazanog na Slici 2 c) kod koga je zvučna energija bila više distribuirana. Drugi i treći tip događaja, u odnosu na prvi događaj, imaju povećanu verovatnoću za prostorne uglove bliske 90°. To znači da će vrednost izolacione moći za ove događaje potencijalno biti manja za posmatrane frekvencijske opsege.

Izolaciona moć monolitne pregrade zavisi od funkcije gustine verovatnoće ugaone raspodele incidentne energije pa će razlike u oblicima ugaonih raspodela za analizirane događaje dovesti do toga da će ista fasadna pregrada ispoljiti različita izolaciona svojstva. Na Slici 3 punim crnim linijama prikazane su ostvarene vrednosti izolacionih moći fasadnih pregrada u slučaju kada je funkcija gustine verovatnoće incidentne energije određena na osnovu snimka od 30 minuta. Gornji red na Slici 3 predstavlja frekvencijsku zavisnost vrednosti izolacione moći betonske pregrade debljine 16 cm, dok donji red na istoj slici predstavlja frekvencijsku zavisnost vrednosti izolacione moći fasadne pregrade izrađene od stakla debljine 5 mm.

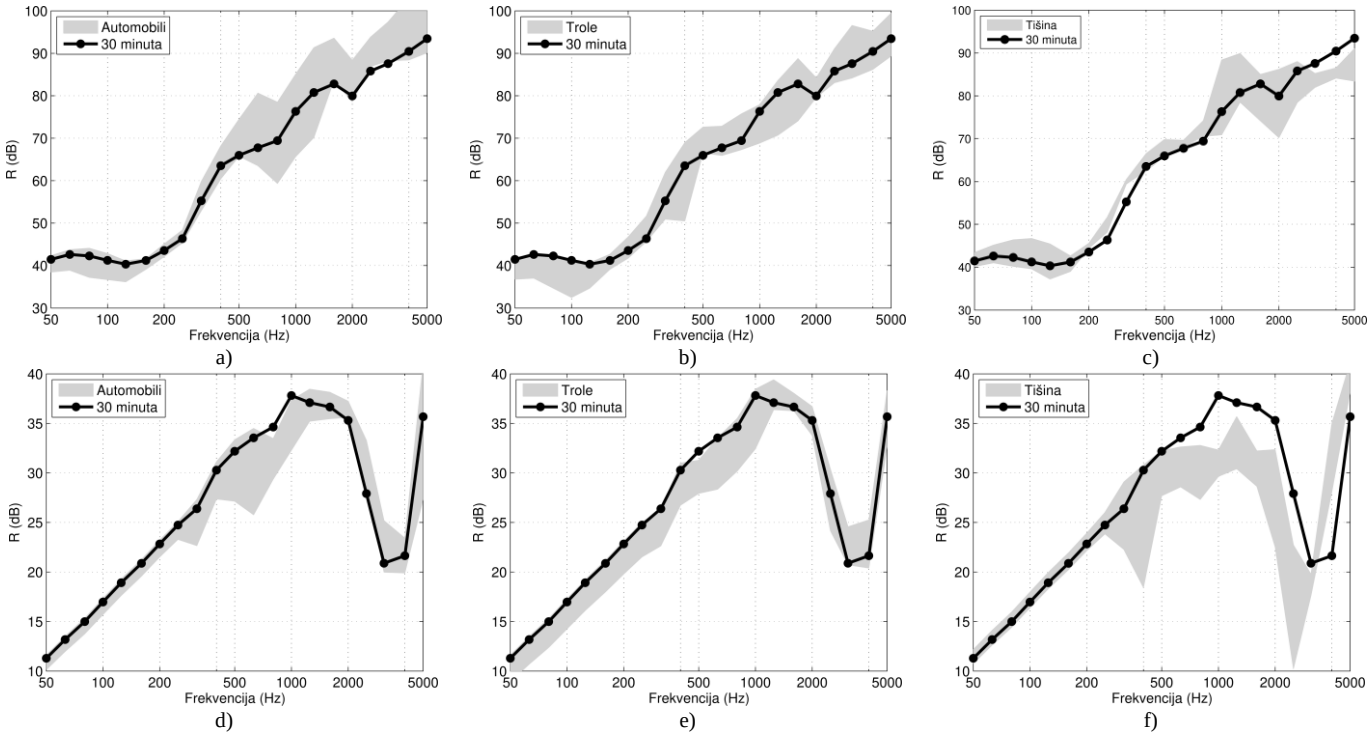


Sl. 2. Funkcije gustine verovatnoće ugaone raspodele incidentne energije spoljašnje buke za izdvojene događaje i 1/3 oktavne frekvencijske opsege: a) automobili 1600 Hz, b) trolejbusi 1600 Hz, c) tišina 1600 Hz, d) automobili 2000 Hz, e) trolejbusi 2000 Hz, f) tišina 2000 Hz

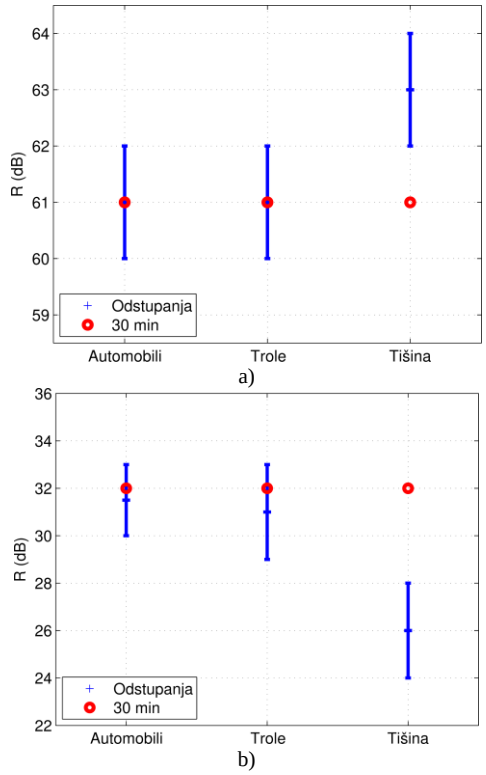
Građevinski podaci ovih prerada, potrebni za proračun izolacione moći, preuzeti su iz [20]. Na Slici 3 sivom bojom prikazana su odstupanja vrednosti izolacione moći za 1/3 oktavne frekvencijske opsege izračunate za nekoliko događaja istog tipa.

Na Slici 3 a) prikazana je frekvencijska zavisnost vrednosti izolacione moći betonske fasadne pregrade u slučaju prolaska automobila ispred mesta merenja. Odstupanja koja se javljaju u ostvarenim vrednostima izolacione moći za različite prolaske automobila iznose i do 20 dB za određene frekvencijske opsege. Najveća odstupanja javljaju se u opsegu od 500 Hz do 2000 Hz. Vrednosti izolacione moći betonske pregrade izračunate sa ugaonim raspodelama određenim za vremenski interval od 30 minuta nalaze se u gabaritima odstupanja vrednosti izolacione moći izračunatim za prolaske automobila. Za frekvencije veće od 2000 Hz vrednosti izolacione moći betonske pregrade izračunate na osnovu ukupnog trajanja snimka od 30 minuta nalaze se na donjoj granici gabarita odstupanja za ovaj tip događaja. Na Slici 3 b) prikazana je frekvencijska zavisnost odstupanja vrednosti izolacione moći betonske pregrade prilikom prolaska trolejbusa. Odstupanja su nešto manja u odnosu na slučaj prolaska automobila. Maksimalna odstupanja javljaju se u frekvencijskom opsegu od 500 Hz do 2000 Hz i iznose maksimalno 15 dB. Vrednosti izolacione moći izračunate za ukupno trajanje snimka, prikazane punom linijom, i u ovom slučaju nalaze se u granicama gabarita odstupanja. Na niskim frekvencijama, manjim od 500 Hz, vrednosti izolacione moći za snimak od 30 minuta nalaze se na gornjoj granici odstupanja za ovaj tip događaja. Odstupanja vrednosti izolacione moći za slučaj kada je analizirana ambijentalna tišina su manja u odnosu na dva prethodno analizirana tipa događaja. Sa Slike 3 c) može se videti da su odstupanja manja od 5 dB, izuzev nekoliko 1/3 oktavna frekvencijska opsega na srednjim frekvencijama. Vrednost izolacione moći određene na osnovu snimka od 30 minuta ne nalazi se unutar gabarita odstupanja za ovaj tip događaja za ceo frekvencijski opseg.

Na Slikama 3 d) do 3 f) prikazana su odstupanja vrednosti izolacione moći fasadne pregrade izrađene od stakla za tri analizirana tipa događaja, zajedno sa frekvencijskom zavisnošću vrednosti izolacione moći izračunate za ukupan signal, za isti tip pregrade. Najveća odstupanja za pojedinačne događaje javljaju se u slučaju kada nema prolaska vozila, odnosno za tip događaja nazvan tišina. Najveća odstupanja kod ovog tipa događaja javljaju se u opsegu od 2000 Hz do 2500 Hz, gde se nalazi i frekvencija incidencije staklene pregrade. U ovoj frekvencijskoj oblasti staklena pregrada ispoljava najmanje vrednosti izolacione moći. Za pomenuti opseg, funkcija gustine verovatnoće ima maksimum za uglove bliske 90° (Slika 2 f)), što doprinosi dodatnom smanjenju ostvarene izolacione moći analizirane pregrade. Sa Slike 3 f) može se uočiti da se ostvarene vrednosti izolacione moći staklene fasadne pregrade izračunate za ukupno trajanje signala ne nalaze u gabaritima odstupanja vrednosti izolacione moći izračunate za ambijentalnu tišinu. Rastojanje od gabarita u pojedinim frekvencijskim opsezima iznosi 10 dB. Za druga dva tipa analiziranih događaja odstupanja su približno ista, i manja u odnosu na slučaj kada je analizirana betonska pregrada. U oba slučaja izolaciona moć izračunata za ukupan snimak izlazi van granica gabarita odstupanja. Na Slici 4 plavom bojom prikazana su odstupanja jednobrojnih (merodavnih) vrednosti izolacionih moći [10] za pojedinačne događaje istog tipa i dve analizirane fasadne pregrade. Crvenom bojom prikazane su jednobrojne vrednosti izolacione moći izračunate za ceo snimak sa mikrofonskog niza. Vrednosti izolacione moći izračunate za ukupno trajanje snimka iznose 61 dB za betonsku pregradu, odnosno 32 dB za staklenu fasadnu pregradu.



Sl. 3. Frekvencijska zavisnost vrednosti izolacione moći različitih pregrada za izdvojene događaje: a) automobili beton, b) trole beton, c) tišina beton, d) automobili staklo, e) trole staklo, f) tišina staklo



Sl. 4. Odstupanja jednobrojnih vrednosti izolacione moći za izdvojene događaje i a) betonsku pregradu, b) staklenu pregradu

Sa Slike 4 se uočava da se jednobrojna vrednost izolacione moći izračunata na osnovu ukupnog signala nalazi u granicama odstupanja za događaje tipa prolaska automobila i događaje prolaska trolejbusa. Za slučaj betonske pregrade i dva navedena tipa događaja ova odstupanja iznose 2 dB, dok je odstupanje u slučaju staklene pregrade izračunate na osnovu snimaka prolaska trolejbusa 3 dB. U slučaju izračunavanja vrednosti izolacione moći na osnovu snimaka ambijentalne tišine, odstupanja su ista kao i u slučaju druga dva tipa događaja, ali se značajno razlikuju od vrednosti izolacione moći dobijene za ukupan signal. U slučaju betonske pregrade jednobrojna vrednost se nalazi u granicama od 62 dB do 64 dB za pojedinačne snimke tišine i ove vrednosti su veće od vrednosti dobijene za ukupno trajanje snimka. U slučaju kada se analizira pregrada od stakla jednobrojna vrednost se nalazi u granicama od 24 dB do 28 dB za pojedinačne snimke tišine i ove vrednosti su manje od vrednosti dobijene za ukupno trajanje snimka.

Prikazani rezultati ukazuju da se ispoljena vrednost izolacione moći fasadne pregrade u urbanim uslovima menja u vremenu i da zavisi od tipa događaja na mernom mestu u urbanim uslovima. Uticaj tipa događaja na vrednost izolacione moći posmatrane u dužem vremenskom intervalu zavisi od frekventnosti tog tipa događaja. Posmatrajući rezultate dobijene za snimke ambijentalne tišine zaključuje se da je doprinos ovog tipa događaja u ostvarenoj vrednosti izolacione moći za 30 minuta relativno mali, jer vrednost izolacione moći u ovom slučaju u većoj meri odstupa u odnosu na ukupnu. Broj prolazaka automobila i trolejbusa na ovoj mernoj lokaciji je veliki, pa je njihov uticaj na ostvarenu vrednost izolacione moći u intervalu od 30 minuta veliki.

V. ZAKLJUČAK

U ovom radu prikazana je upotreba metodologije za određivanje ugaone raspodele incidentne energije spoljašnje buke za karakteristične tipove izvora na jednoj mernoj lokaciji u urbanim uslovima. Pokazano je da ostvarena vrednost izolacione moći zavisi od tipa izvora jer postoji promena ugaone raspodele incidentne energije koja „napada“ fasadu. U urbanim uslovima postoje izvori različitih tipova (automobili, autobusi, trolejbusi itd.) što dovodi do promene strukture zvučnog polja. Promena strukture polja u vremenu ima za posledicu vremensku

zavisnost vrednosti izolacione moći. Uticaj određenih tipova izvora zavisi od konkretne konfiguracije terena u kome se nalazi fasadna pregrada ali i od veličine izvora, spektralnog sadržaja buke koju izvor emituje itd. Zbog toga se i za pojedinačne događaje istog tipa mogu javiti različite ostvarene vrednosti izolacione moći. U ovom radu pokazani su gabariti odstupanja. Za analiziranje zvučnog komfora u zgradama potrebno je generalno sagledati ostvarene vrednosti izolacione moći, što zahteva primenu metodologije za određivanje oblika raspodele nad signalima veće dužine, kako bi se obuhvatio veliki broj događaja svih tipova. Obrada i analiza takvih rezultata je vremenski i računski kompleksna. Ukoliko bi se definisali ključni događaji na posmatranoj mernoj lokaciji bilo bi dovoljno snimiti samo njih, što bi smanjilo vreme merenja, a samim tim i vreme potrebno za obradu signala. Međutim, na osnovu rezultata prikazanih u ovom radu zaključuje se da svi tipovi događaja nemaju isti uticaj na ostvarenu vrednost izolacione moći. To znači da se prostim sabiranjem uticaja malog broja pojedinačnih događaja neće dobiti rezultat kao kada se merenje izvrši u dužem vremenskom intervalu, npr. 30 minuta. Broj događaja određenog tipa, kao i ugaona raspodela incidentne energije koju oni stvaraju utiču na ukupan rezultat. Zbog toga je potrebno uvesti korekcione faktore koji bi uvažili uticaj pojedinih tipova događaja na generalnu sliku o stanju izolacione moći na analiziranoj lokaciji. Uvođenje korekcionih faktora u cilju smanjenja vremena potrebnog za globalno sagledavanje izolacione moći u urbanim uslovima biće tema budućih istraživanja.

ZAHVALNICA

Ovaj rad je realizovan u okviru projekta TR 36026 koga finansira Ministarstvo prosvete, nauke i tehnološkog razvoja Republike Srbije.

LITERATURA

[1] C.Brutel-Vuilmet, C.Guigou-Carter, M.Villot, „A Study of the Influence of Incidence Angle on Sound Reduction Index Using NAH-Phonoscopy. Acta Acustica United with Acustica, 2007;Vol. 93: 364–374.

[2] D. Šumarac Pavlović, F. Pantelić, S. Bojičić, M. Bjelić, „Airborne sound insulation of monolithic partition as a function of incidence angles“, Proc. Forum Acusticum, Krakow 2014.

[3] G.Vermeir, G.Geentjens, W.Bruyninckx, „Measurement and calculation experiences on façade sound insulation“, Proc INTER-NOISE 2004.

[4] ISO 140-5 „Acoustics - Measurement of sound insulation in buildings and of building elements – Part 5: Field measurements of airborne sound insulation of façade elements and façades“.

[5] M. Stanojević, M. Bjelić, D. Šumarac Pavlović, M. Mijić, Measurements of noise energy angular distribution at the building envelope using microphone arrays, Applied Acoustics, Vol 140, 283-287 (2018).

[6] M. Bjelić, M. Stanojević, D. Šumarac Pavlović, M. Mijić, „Određivanje uglova incidencije buke u urbanim sredinama“, ETRAN, Kladovo, jun 2017, Broj rada (zbornik radova CD): AK 1.1, ISBN: 978-86-7466-692-0.

[7] M. Bjelić, „Analiza ugaone raspodele incidentne energije spoljašnje buke primenom mikrofonskog niza“, Univerzitet u Beogradu, Elektrotehnički fakultet, Doktorska disertacija, jun 2018.

[8] M. Bjelić, M. Stanojević, D. Šumarac Pavlović, M. Mijić, T. Miljković, „Analiza ugaone raspodele incidentne energije spoljašnje buke u urbanim uslovima“, ETRAN, Palić, jun 2018, Zbornik radova 49-54, ISBN: 978-86-7466-752-1.

[9] M. Bjelić, „Analiza ugaone raspodele incidentne energije spoljašnje buke na fasadama zgrada u urbanim uslovima pomoću mikrofonskog niza“, 26th Telecommunications forum TELFOR 2018, Belgrade, November 2018, CD Proceedings paper No. 8.9., ISBN: 978-1-5386-7170-2.

[10] ISO 717-1:1996 „Acoustics – rating of sound insulation in buildings and of building elements – Part 1: Airborne sound insulation“.

[11] C. Buratti, E. Belloni, E. Moretti, **"**Façade noise abatement prediction: New spectrum adaptation terms measured in field in different road and railway traffic conditions“, Appl. Acoust. 2014;76:238–248.

[12] J. Hald and J. Christensen, „A novel beamformer array design for noise source location from intermediate measurement distances“, J. Acoust. Soc.Am. 112, 2448, DOI: 10.1121/1.4780077, (2002).

[13] M. Bjelić, M. Stanojević, D. Šumarac Pavlović, M. Mijić, „Dizajn mikrofonskog niza optimizovanog za monitoring saobraćajne buke“, ETRAN, Zlatibor, jun 2016, Broj rada (zbornik radova CD): AK 1.2, ISBN: 978-86-7466-618-0.

[14] M. Bjelić, M. Stanojević, D. Šumarac Pavlović, M. Mijić, „Microphone array geometry optimization for traffic noise analysis“, The Journal of the Acoustical Society of America, Vol 141(5), 3101-3104 (2017).

[15] U. Michel, "History of acoustic beamforming", Berlin , 2006. Berlin Beamforming Conference.

[16] T.F. Brooks, W.M. Humphreys, „A deconvolution approach for the mapping of acoustic sources (DAMAS) determined from phased microphone arrays“, *Journal of Sound and Vibration* 294.4, 856-879, 2006.

[17] R.P. Dougherty, „Extensions of DAMAS and Benefits and Limitations of Deconvolution in Beamforming“, *AIAA*, 2961.11, 2005.

[18] K. Ehrenfried, L. Koop, „A comparison of iterative deconvolution algorithms for the mapping of acoustic sources“, *AIAA journal*, 45.7:1584-1595, 2007.

[19] L. Beranek, "Noise Reduction. New York": McGraw-Hill Book Company, Inc., 1960.

[20] H. Kurtović, **"**Priručnik za proračun zvučne izolacije". Beograd : Elektrotehnički fakultet, Laboratorija za akustiku, 1994.

ABSTRACT

The shape of angular distribution of the incident energy influences in-situ values of the sound reduction index of façade elements. In general case, the shape of this distribution is unknown. Utilizing a microphone array and space-time signal processing algorithms it is possible to experimentally obtain a probability density function of the angular distribution of the incident energy on the building envelope. In urban environments there is a large number of sources which differ in type, spectral content they emit, power, etc. This leads to the assumption that the shape of angular distribution, and hence the value of in-situ sound reduction index will depend on the sound source type. This paper presents the methodology for experimental determination of noise angular distribution and its application in observing the differences of the in-situ sound reduction index of the same façade element, when it is exposed to different sound sources in urban environments. The in-situ sound reduction index is calculated using the experimentally obtained probability density functions and it is shown that it varies in time and depends on the momentary structure of the sound field. Furthermore, a comparison is presented of the in-situ values calculated for individual events and calculated for a longer time interval. In this way, it is possible to observe the influences of certain sound sources types on the façade sound insulation performance.

Estimation of the dependency of façade sound reduction in-situ values on the type of noise sources in urban environments

Miodrag Stanojević, Miloš Bjelić, Dragana Šumarac Pavlović, Miomir Mijić, Tatjana Miljković

Uticaj „tišine“ na zvučni komfor

Miomir Mijić, Dragana Šumarac Pavlović, Miloš Bjelić, Tatjana Miljković

***Apstrakt*— U novije građenim zgradama, stambenim i poslovnim, sve su češće pritužbe na nedostatak privatnosti govora kao bitnog elementa zvučnog komfora. Jedini numerički pokazatelj koji se danas koristi kao formalni indikator stanja je građevinska izolaciona moć pregrada između prostorija na osnovu koje se izvodi zaključak o postignutom akustičkom komforu. Međutim, iskustva iz prakse pokazuju da se pri istim vrednostima izolacione moći u zgradama mogu zateći različita stanja privatnosti. Zbog toga se u radu razmatraju uzroci te pojave i mogućnosti primene kompleksnijih numeričkih pokazatelja stanja privatnosti. Na osnovu toga se predlaže dopuna forme standardnih izveštaja o merenju zvučne izolacije za pouzdaniju procenu stanja privatnosti govora i mogući pravci delovanja u projektovanju i izgradnji zgrada.**

***Ključne reči*— ambijentalna buka, tišina, zvučna izolacija, zvučni komfor**

I. Uvod

Stanje koje bi se moglo nazvati „potpunom tišinom“ predstavlja jedan zanimljiv ideal o kome pričaju i o kome maštaju mnogi ljudi. Oni takvo akustičko stanje vide kao ambijent u kome će naći svoj mir, i u kome im nikakav zvuk neće skretati pažnju. Takve stavove često prati i interesovanje za načinima na koji se „potpuna tišina“ može postići u stanu, ili bar u jednoj sobi.

Sa druge strane, u današnje vreme čini se da je broj žalbi na neadekvatan zvučni komfor u stambenim zgradama u stalnom porastu. Neke od njih na kraju dovedu i do vrlo ozbiljnih sporova. U dugom vremenskom periodu žalbe su se pretežno odnosile na izvore buke u okruženju: na industriju, na tuđe klima uređaje, na saobraćaj, muziku iz ugostiteljskih objekata i slično. Takve žalbe adresirane su na vlasnike izvora buke sa zahtevima da se oni isključe ili utišaju. Problemi te vrste uglavnom se rešavaju pomoću inspekcijskih službi i komunalne policije, drugim rečima postoji nekakav pravni sistem kojim je to regulisano.

U novije vreme dešava se sve češćeg pomeranja fokusa stanarskog nezadovoljstva zvučnim komforom sa klasičnih izvora buke, pa se žalbe ljudi na njegov nedostatak sve ćešće odnose na okolnost da u svojoj sobi čuju razgovor iz susednih stanova. Pri tome je jasno da taj problem ima dva smera – sa druge strane zida podjednako će se čuti njihov glas. Razgovor iz komšiluka jeste ometajući, ali istovremeno i važan signal ljudima da je ugrožena njihova privanost [1]. U takvim okolnostima novo je to da su žalbe, umesto ka vlasnicima izvora buke, usmerene protiv građevinara i eventualno investitora koji su zgradu napravili. Zanimljivo je da su u tome vrlo retko žalbe usmerene na projektante zgrade, mada se pokazalo da su takvi problemi najčešće rezultat njihovih propusta.

Nisu samo stambene zgrade mesta gde se javljaju pritužbe na nedovoljan zvučni komfor, to jest privatnost. I u novim poslovnim zgradama postoje okolnosti, možda još češće nego u stambenim, u kojima se ljudi žale na nedostatak privatnosti ili na ometanja govorom. Sudeći po broju poziva u novije vreme koji dolaze iz raznih firmi sa molbom da se nekako rešava problem čujnosti razgovora iz njihovih „važnih“ prostorija, čini se da je danas privatnost razgovora važan, ali zanemaren aspekt poslovnih zgrada. Očigledno je da ima mnogo drugih faktora koji su u projektovanju važniji od pitanja privatnosti.

Nesumnjivo je da problem zvučnog komfora u stanovanju nije novi, ali izgleda da su ga neke okolnosti, bar u Srbiji, vremenom učinile izraženijim. Žalbe na nedovoljan zvučni komfor najšećće dolaze iz novih zgrada, stambenih i poslovnih. Reklo bi se da pojave ometanja iz okruženja ljudi više ne prihvataju kao neumitnost stanovanja ili kao deo poslovnog folklora. Jedno objašnjenje za porast takvih žalbi je da se prag tolerancije ljudi na ometanja vremenom spustio. Može se takođe pretpostaviti da u novim, luksuznim zgradama koje podrazumevaju i višu cenu prostora ljudi generalno imaju veća očekivanja, pa u tom smislu i veće zahteve u domenu zvučnog komfora. U Srbiji to znači da se u projektovanju zahtevi izolacije pregrada između stanova i pregrada oko sala za sastanke moraju preispitati u kontekstu promena u stavu prema zvučnom komforu.

Zvučni komfor u zgradama nominalno se rešava u fazi projektovanja zgrade kada se usvajaju pregrade sa adekvatnim zvučnoizolacionim svojstvima. Postoji regulativa koja definiše minimalne zahteve za izolacione moći zidova i tavanica kojima su okruženi boravišni prostori u zgradama [2]. Karakteristično za takav problem kada se javi je i u tome što je njegovo rešavanje, to jest popravljanje zvučne izolacije, vrlo kompleksno i po nekada nema elegantnog rešenja. Takođe postoje i okolnosti kada je to sasvim nemoguće.

Zbog toga je veoma važno da se razumeju kompleksni razlozi koji dovode do nezadovoljstva zvučnim komforom kako bi se na adekvatan način rešavali takvi problemi, presvega u projektovanju, ali i pri sanaciji nekog zatečenog stanja. U

Miomir Mijić – Elektrotehnički fakultet, Univerzitet u Beogradu, Bulevar Kralja Aleksandra 73, 11020 Beograd, Srbija (e-mail: emijic@etf.rs)

Dragana Šumarac Pavlović – Elektrotehnički fakultet, Univerzitet u Beogradu, Bulevar Kralja Aleksandra 73, 11020 Beograd, Srbija (e-mail: dsumarac@etf.rs)

Miloš Bjelić – Elektrotehnički fakultet, Univerzitet u Beogradu, Bulevar Kralja Aleksandra 73, 11020 Beograd, Srbija (e-mail: bjelic@etf.rs)

Tatjana Miljković – Elektrotehnički fakultet, Univerzitet u Beogradu, Bulevar Kralja Aleksandra 73, 11020 Beograd, Srbija (e-mail: tm@etf.rs)

ovom radu je razmatran uticaj jednog specifičnog faktora zvučnog komfora, a to je nivo ambijentale buke, odnosno tišine u sobama na osećaj privatnosti ljudi u zgradama. U tome postoji kompleksno pitanje načina da se takav uticaj na neki način kvantifikuje. U radu su diskutovane neke mogućnost za to kao i načini za prevazilaženje neadekvatnog stanja privatnosti.

II. FAKTORA KOJI ODREĐUJU PRIVATNOST U ZGRADAMA

Čujnost i razumljivost govora iz susednih prostorija u zgradama, a to znači iz stana u stan, iz kancelarije u kancelariju ili iz kancelarije u zajednički prostor, prevashodno je određena stanjem zvučne izolacije pregradne konstrukcije koja ih deli. To je tema koja se obrađuje u procesu projektovanja zgrada, pri čemu se ne polazi od apsolutnog zahteva da se svaka prekomerna buka u prostorijama učini nečujnom u okruženju. Normativi koji propisuju uslove za pregrade u zgradi, to jest njihova izolaciona svojstva, zasnivaju se na podacima o očekivanom nivou zvuka koji nastaje pri razgovoru ljudi u prostorijama i očekivanom nivoa ambijentalne buke. U stambenim zgradama to je u izvesnom smislu limitirano kućnim redom, a u poslovnim zgradama zahtevima tehnologije rada. Prema korišćenom normativu u Srbiji [2] očekivani ekvivalentni nivo zvuka u normalnim prostorijama u zgradama (takozvane boravišne prostorije) ne prelazi 70 dBA. Prostorije u kojima nivo zvuka premašuje tu vrednost spadaju u posebne kategorije označene kao „bučne“, ili „veoma bučne“. Njihova zvučna izolacija se u projektovanju posebno tretira, pa one nisu tema u razmatranju problema privatnosti.

Jedno ranije istraživanje je pokazalo kakvo je stanje nivoa zvuka u boravišnim prostorijama stanova, to jest koliko ozbiljna može biti buka u prosečnim domaćinstvima kada se u njima odvijaju svakodnevne aktivnosti [3]. Isto istraživanje je takođe dalo rezultate koji pokazuju stanje nivoa zvuka u stanovima pri ekscesnim dešavanjima kao što su proslave, slušanje glasne muzike i slično. U sklopu publikovanih rezultata tog istraživanja nalaze se i informacije o izmerenim nivoima zvuka pri razgovoru ljudi u stanovima. Oni pokazuju da se ekvivalentni nivo govora u sobama kreće u opsegu od 60 dBA do 73 dBA. To znači da su rezultati merenja potvrdili stav unet u regulativu da je očekivani ekvivalentni nivo buke u stanovima pri normalnim životnim aktivnostima do 70 dBA.

U poslovnim zgradama nivoi buke u prostorijama su po prirodi stvari u granicama kontrolisanim zahtevima rada. U literaturi postoje rezultati istraživanja koji pokazuju stanje nivoa zvuka pri razgovoru ljudi u salama za sastanke kakve se uobičajeno nalaze u poslovnim zgradama [4,5]. Na osnovu merenja u 79 sala zaključeno je da se ekvivalentni nivo govora u njima kreće oko 60 dBA, sa varijacijama u opsegu ±2 dB. Očigledno je da se u takvoj govornoj komunukaciji ljudi neguje izvesna kultura govora, pa se nivo glasa podeša da bi se govor razumeo u prostoriji, čak i kada se koristi sistem za doozvučavanje. Ovi rezultati pokazuju da su očekivani nivoi govora u stanovima viši od nivoa u konferencijskim salama. To je posledice pre svega manje zapremine boravišnih prostorija u stanovima u odnosu na konferencijske sale u poslovnim zgradama, kao i načina kako se govori u domaćoj atmosferi u odnosu na prostorije za zvanične sastanke.

Zahtevi zvučne izolacije definisani u normativima za zvučnu zaštitu u zgradama podešavani su sa idejom da se govor ljudi pri ovakvim nivoima zvuka ne razume, ili uopšte ne čuje sa suprotne strane zidova. Šta to sa aspekta razumljivosti govora može da znači u kvalitativnom smislu nalazi se, između ostalog, u tekstu predloga jednog novog ISO standarda (postupak za klasifikaciju zgrada) u pogledu postignute zvučne izolacije u njima [6]. U njemu se kao ilustracija navodi moguća čujnosti govora između susednih prostorija pri određenim vrednostima zvučne izolovanosti D_{nT}. Ova ilustracija je prikazana u Tabeli I. Vidi se, na primer, da je pri izolovanosti između prostorija 52 dB očekivano da glasan govor bude razumljiv, a normalan govor jedva razumljiv. Iz ovog prikaza sledi da bi govor postao teško razumljiv tek pri izolovanosti većoj od 56 dB. Vidi se razlika između pojmova razumljivosti i čujnosti govora, pri čemu čujnost podrazumeva da se prepoznaje zvuk govora, ali ne i šta se izgovara.

TABELA I
PREGLED ČUJNOSTI GOVORA PRI RAZLIČITIM STANJIMA IZOLACIJE

D_{nT}	**glasan govor**	**normalan govor**
52 dB	razumljiv	jedva razumljiv
56 dB	jedva razumljiv	teško razumljiv
60 dB	čujan	čujan
64 dB	jedva čujan	jedva čujan
68 dB	retko čujan	nečujan

Međutim, dosadašnja praksa autora ovog rada na sanaciji zvučne izolacije po zgradama pokazala je da izolovanost između prostorija ili izolaciona moć pojedinačnih pregrada ne određuju jednoznačno stanje razumljivosti govora, a to znači i stanje privatnost. Drugim rečima, stečeno iskustvo u izvesnom smislu relativizira opise iz Tabele 1. Opšta je činjenica u telekomunikacijama da odnos signal/šum određuje da li će neki signal biti detektabilan. Isto važi i za detekciju govora čulom sluha, pa je odnos signal/šum na mestu slušanja ono što određuje da li će se neki govor čuti i razumeti. Pri tome se pod signalom podrazumeva akustički govorni signal koji deluje na čulo sluha slušaoca, sa svim njegovim dinamičkim specifičnostima, a pod šumom se podrazumeva ambijentalna buka u okruženju tog slušaoca.

Jasno je da se zbog dinamičkih karakterstika signala govora o veličini odnosa signal/šum može govoriti samo statistički, ali se u prvoj aproksimaciji može problem posmatrati i kroz njihove ekvivalentne vrednosti, kako god da su izmerene. Nivo govornog signala na mestu njegovog nastanka u svakodnevnoj govornoj komunikaciji u stanovima i kancelarijama određen je ograničenom zvučnom snagom vokalnog trakta i naučenim načinom na koji ga ljudi uobičajeno koriste.

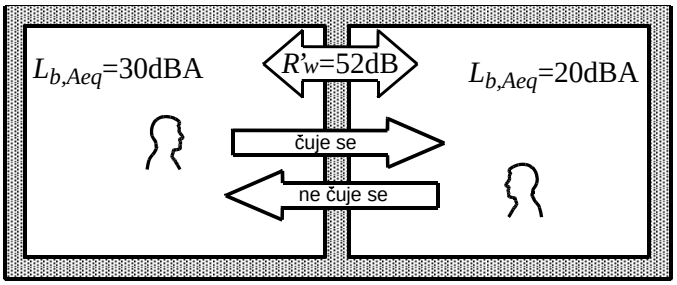
U literaturi je ukazano da odnos signal/šum ono što određuje čujnost i razumljivost govora [4]. Pokazano je da govor postaje nerazumljiv kada srednja vrednost odnosa signal/šum po opsezima 1/3 oktave približno opadne ispod −2 dB [4]. Pri tome se taj odnos posmatra u frekvencijskom opsegu od

značaja za govor, što je ograničeno na raspon od 160 Hz do 5 kHz. Tako je privatnost govora funkcija nivoa govora na mestu slušanja, što je samo po sebi jasno, ali i nivoa ambijetalne buke na tom istom mestu, što je skreće pažnju i na ambijent u kome se sluša kao faktor. Prema tome, i privatnost govora u zgradama zavisiće od ta dva faktora: zvučne izolacije i stanja „tišine“. To je objašnjenje zašto u istim uslovima zvučne izolacije i pri istom intenzitetu govora on može biti primetan u susednoj prostoriji, ali i ne mora – odnos signal/šum može biti različit.

U urbanoj sredini ambijentalna buka u zgradama potiče od više izvora: saobraćajna i druga buka koja dospeva kroz prozore i druge fasadne otvore, buka od raznih izvora u zgradi koja dospeva u sobe kroz konstrukciju zgrade i instalacije, kao i buka uređaja u prostorijama (u stanovima frižider, vodovod, klima uređaj, u poslovnim zgradama ventilacija i klimatizacija, računari, razna kancelarijska oprema, itd.).

III. Jedna praktična ilustracija problema privatnosti

Brojni su primeri koji dokazuju da privatnost zavisi od stanja ambijentalne buke u prostorijama. Jedan slučaj iz novije konsultantske prakse zabeležen u stambenoj zgradi prilikom merenja zvučne izolacije šematski je objašnjen na slici 1. Merenje stanja izolacije urađeno je po žalbi jednog stanara. Izmerena merodavna vrednost građevinske izolacione moći između dva stana bila je 52 dB, kao što je naznačeno na slici. To je vrednost propisana kao zahtev u normativu [2] koja je najšire primenjuje kao kriterijum pri projektovanju zgrada. To znači da stanje zvučne izolacije zadovoljava danas važeći normativ.



Sl. 1. Ilustracija jednog analiziranog slučaja koji je pokazao da vrednost zvučne izolacije ne utvrđuje jednoznačno stanje privatnosti u prostorijama.

Međutim, utvrđeno je da se stanje privatnosti u dve susedne prostorije doživljava različito, iako ih razdvaja ista pregrada sa nominalno zadovoljavajućom vrednošću građevinske izolacione moći. U jednom smeru utvrđeno je da se govor iz susedstva može čuti, što je i iniciralo žalbu stanara, dok u suprotnom smeru takva pojava nije konstatovana i stanari nisu imali primedbe na privatnost. Ova razlika u čujnosti govora naznačena je na slici. U analizi stanja razumljivosti govora koja je sprovedena u oba smera, to jest sa obe strane zida, korišćena je ista grupa ljudi kao izvor govora i isti slušaoci sa druge strane, tako da je eliminisan mogući uticaj individualnih razlika govornika i slučalaca.

Da bi se objasnio ovaj fenomen, pored zvučne izolacije, izvršeno je merenje ekvivalentne vrednosti nivoa ambijentalne buke u obe prostorije. Rezultati su takođe prikazani u slici. U prostoriji u kojoj nije konstatovano da se čuje govor iz susednog stana ekvivalentna vrednost ambijentane buke je bila oko 30 dBA. Upravo ova vrednost je u domaćim normativima navedena kao maksimalno odzvoljeni nivo buke tokom noći u boravišnim prostorijama stambenih zgrada. Ovakvo stanje buke je dominantno posledica zatečenog stanja izolacione moći prozora i raznih izvora zvuka u stanu. U prostoriji susednog stana u kojoj se registruje govor iz susedstva izmerena vrednost ekvivalentnog nivoa ambijentalne je bila samo oko 20 dBA. Utvrđeno je da su u tom stanu stavljeni novi fasadni prozori velike izolacione moći i svi kućni uređaji su novi, sa niskim nivoom buke koju stvaraju. Iako je stanje zvučne izolacije jedinstveno, ta razlika u nivou ambijentalne učinila je da odnos signal/šum u dve prostorije bude različit, pa je zbog toga različito i stanje privatnosti, to jest stanje čujnosti i razumljivost govora.

Pokazani primer jasno pokazuje da stanje privatnosti kao elementa zvučnog komfora ne određuje samo vrednost izolacione moći građevinskih pregrada koje okružuju prostorije, već i stanje ambijentalne buke u okolini potencijalnih slušaoca. I pored toga u dve faze provere stanja – u projektovanju i pri merenju zvučne izolacije za tehnički prijem zgrade – proračunava se i meri samo zvučna izolacija. To naravno da ima opravdanja jer se time proverava kalitet projekta i same gradnje. Ali, da bi se dobila pouzdanija slika o zvučnom komforu u zgradi analiza stanja se mora proširiti sa podacima o nivou ambijentalne buke u prostorijama. Tek tada se mogu izvoriti kvalitetni zaključci o stanju zvučnog komfora u zgradi i procenjivati realno stanje privatnosti.

Standardno testiranje akustičkog kvaliteta novoizgrađenih zgrada danas obuhvata isključivo merenje merodavne vrednosti građevinske izolacione moći R'_w, pa su za nekakvu ocenu stanja privatnosti na raspolaganju samo vrednosti tog parametra. Činjenica je da ambijentalna buka nije predmet merenja u izgrađenim zgradama, osim kada se pojave žalbe, a one su u praksi gotovo isključitvo inicirane nekom mašinskom opremom postavljenom u blizini. Činjenica je takođe da u ovom trenutku ne postoje podaci o prosečnom stanju ambijentalne buke u stanovima. Tema za buduće istraživanje je i stanje ambijentalne buke u stanovima tokom tihih perioda kao što je noć, kada se problem privatnosti dodatno intenzivira. Zbog toga se ne mogu izvoditi relevantni zaključci o očekivanoj privatnosti u zgradama čak i kada postoje podaci o zvučnoj izolaciji.

IV. Pregled mogućih pristupa kompleksnijem kvantifikovanju privatnosti u zgradama

Opisani primer iz prakse pokazao je nedostatak koji se javlja kada se značenje pokazatelja zvučne izolacije u zgradama ekstrapolira i na indikaciju privatnosti i čujnosti govora. To ukazuje na potrebu da se skup indikatora akustičkog kvaliteta zgrada na neki način dopuni i drugim podacima koji će zaokružiti sliku o stanju privatnosti. U literaturi se mogu naći neka rešenja za kvantifikovanje stanja poverljivosti govora na

način koji je kompleksniji od jednostavnog navođenja podatka o vrednosti izolovanosti D_{nT} ili merodavne vrednosti građevinske izolacione moći R'_w. Jedan mogući pristup bio bi pokušaj da se ono što postoji u tom domenu u literaturi prilagodi specifičnoj nameni ocene akustičkog kvaliteta zgrada.

Jedno moguće rešenje pronađeno je u američkom standardu ASTM E2638 [7]. Postupak koji je u njemu izložen zasnovan je na radovima Bredlija i saradnika [8,9]. Taj standard definiše parametar nazvan Speech Privacy Class (SPC). To je kvantitativna mera stanja privatnosti govora u prostoriji sa aspekta potencijalnog slušaoca koji se nalazi izvan nje. Parametar SPC se u standardu definiše kao mera bezbednosti i poverljivosti govora u salama za sastanke. Iako s takvom specifičnom namenom, on ima potencijal da se primenjuje i za ocene u stambenim zgradama. Vrednost SPC je definisana kao zbir dva faktora koji određuju čujnost govora između susednih prostorija: nenormalizovane zvučne izolovanosti D između njih i nivoa buke L_b u tački gde se može nalaziti potencijalni slušalac. Podrazumeva se da su obe veličine u decibelima.

Da bi se došlo do vrednost SPC sve relevantne veličine se mere u opsezima 1/3 oktave, pa se i izolovanost računa po tim opsezima:

$$D_i = L_{pred,i} - L_{prij,i} \tag{1}$$

gde je $L_{pred,i}$ nivo zvuka u predajnoj prostoriji u i-tom opsegu 1/3 oktave, a $L_{prij,i}$ nivo zvuka u prijemnoj prostoriji u istom frekvencijskom opsegu. Obe veličine se u standardu posmatraju u opsegu relevantnom za govor, a to je od 160 Hz do 5000 Hz, što je ukupno 16 opsega, to jest 16 vrednosti D_i. Za potrebe izračunavanja vrednost SPC ukupna izolovanost se definiše kao neponderisana srednja vrednost po svim opsezima 1/3 oktave:

$$D = \sum_{i=0}^{16} D_i/16 \tag{2}$$

gde je i oznaka opsega 1/3 oktave. Na isti način se izračunava i nivo buke u prijemnoj prostoriji:

$$L_b = \sum_{i=0}^{16} L_{b,i}/16 \tag{3}$$

S obzirom da je standard ASTM E2638 namenjen oceni poverljivosti razgovora i curenja akustičkog signala izvan nekog prostora, nivo zvuka na prijemnoj strani L_{prij} i ambijentalna buka L_b mere se u tački koja je na 25 cm od pregradnog zida prema predajnoj prostoriji i na visini između 1,2 m i 2 m.

Prethodno pomenuta istraživanja nivoa zvuka govora u salama za sastanke i u stanovima pokazala su statističke osobine ovih zvukova i limite u kojoma se kreću. To je omogućilo da se u nalaženju odnosa signal/šum pri slušanju govora iz susednog stana nivo signala posmatra kroz vrednost izolovanosti D. U skladu s tim u standardu ASTM E2638 definisan je parametar Speech Privacy Class (SPC) kao:

$$\text{SPC} = D + L_b \tag{4}$$

Kao što se vidi iz izraza (4) ovako koncipiran parametar predstavlja zbir dva nezavisna faktora koji određuju privatnost. Jedan član, izolovanost D, pokazuje stanje zvučne izolacije, ali indirektno i nivo govornog signala. Drugi član, nivo ambijentalne buke, definiše efekat maskiranja zvukovima iz okruženja. Sabrane, ove dve veličine pokazuju ukupni efekat na stanje privatnosti govora. Što je veća vrednost SPC, manji je odnos signal/šum na mestu slušanja.

U radu Bredlija [4], kao i u prilogu standarda ASTM E2638 [7] opisana su očekivana stanja privatnosti govora za različite vrednosti parametra SPC. U Tabeli 2 predstavljeni su ti opisi da bi se stekao uvid u značenje njegovih numeričkih vrednosti. Opisi definišu verovatna stanja razumljivosti i čujnosti govora. Iz prikazanih opisa može se zaključiti da tek pri vrednostima SPC preko 70 privatnost postaje relativno zadovoljavajuća, ali ne i apsolutna, to jest maksimalno moguća. Jasno je da se zbog svoje dinamike u okviru dužeg govora, ali i dinamike unutar svake rečenice, nivo govora može posmatrati samo statistički preko dugovremene efektivne vrednosti. U tom smislu tumačenje značaja pojedinih vrednosti SPC iz Tabele II može samo da definišu najverovatiju učestanost pojavljivanja manje ili više čujnih segmenata.

TABELA II
OPISI STANJA PRIVATNOSTI ZA RAZLIČITE VREDNOSTI PARAMETRA SPC

SPC	opis stanja
< 60	govor je skoro uvek čujan i često se razume
60-65	zvuk govora uglavom čujan, kratke izgovorene fraze se povremeno razumeju
65-70	zvuk govora često čujan, kratke izgovorene fraze su ređe razumljive
70-75	zvuk govora je retko čujan, govor suštinski nerazumljiv (kratke fraze čujne najviše jednom u 15 minuta)
75-80	zvuk govora retko čujan, govor nerazumljiv
80-85	zvuk govora eventualno čujan na svakih 15 minuta, govor nerazumljiv
> 85	zvuk govora neprimetan

U delu sveta gde se primenjuje ISO standardizacija uvrđivanje faktora koji određuju zvučni komfor zasnovano je na merenju građevinske izolacione moći prema standardu ISO 16283-1 [10], i ekvivalentnog nivoa ambijentalne buke L_{Aeq} izraženog u dBA prema standardu ISO 1996-1 [11]. Rezultat merenja zvučne izolacije iskazuje se merodavnom vrednošću izolacione moći R'_w koja se izračunava prema standardu IS0 717-l [12]. Autori ovog rada u jednom svom ranijem radu predložili su jednu modifikaciju u definiciji SPC da bi se bazirala na vrednostima iskazanim kao rezultati merih postupaka prema ISO standardizaciji primenjivanoj u Evropi [13]. Predložen je ekvivalent SPC nazvan Indeks privatnosti govora (IPG) koji je definisan kao:

$$IPG = R'_w + L_{A,eq} \quad (2)$$

U prethodnom istraživanju autora ovog rada eksperimentalno je pokazano da vrednosti IPG sa dovoljnom tačnošću estimira vrednost SPC, a to znači da se u istom smislu može korisiti i za ocenu stanja privatnosti govora između prostorija koristeći standardne pokazatelje defnisane ISO standardima koji se rutinski mere [13].

Zanimljivo je da je značaj nivoa ambijentalne buke u zaštiti privatnosti prepoznat i u standardnim postupcima za kvantifikovanje stanja privatnosti u *open space* kancelarijama. U standardu koji definiše relevantne parametre za ocenu kvaliteta ovih specifičnih prostora predviđeno je i merenje ekvivalentnog nivoa ambijentalne buke [14]. U proceduri opisanoj u tom standardu uticaj buke se posmatra posredno, i to kroz opadanje vrednosti indeksa prenosa govora sa rastojanjem. Posmatra se rastojanje od govornika na kome se govorni signal utopi u postojeću ambijentalnu buku i tako praktično nestane razumljivost govora. U mnogim okolnostima se u poslovnim prostorijama za te namene koruste pomoćna sredstva za kontrolu nivoa buke u prostorijama.

Mada je SPC izvorno definisan kao parametar koji treba da pokazuje mogućnost prisluškivanja razgovora u okruženju sala za sastanke poslovnih zgrada, ideja sadržana u tom konceptu može se iskorititi za ocenu zvučnog komfora u drugim okolnostima, pa i u stambenim zgradama. Parametar IPG kao njegov pandan može se direktno koristiti za izvođenje zaključaka o stanju privatnosti, jer je u saglasju sa postojećim metodama ocenjivanja zgrada.

V. DISKUSIJA O MOGUĆIM PRISTUPIMA U PRAKSI

U prostorijama savremenih stanova i poslovnih zgrada ideal „tišine“ o kome maštaju mnogi ljudi danas je skoro pa dostignut, čak i u urbanoj sredini, ako se pod time podrazumeva da je u njima vrlo nizak ekvivalentni nivo buke, praktično ispod praga primetnosti čulom sluha. Do toga je došlo zahvaljujući opštoj težnji u građevinarstvu ka višem standardu ambijenata za stanovanje i rad. U korenu toga su i savremeni zahtevi termičke izolacije i energetske efikasnosti, proizvodnja sve tiših uređaja koji se koriste u stanovima i kancelarijama, pa i veći prostorni standardi zbog kojih se u zgradama u proseku broj kvadratnih metara prostora po osobi izgleda vremenom povećao. Sve je to na svoj način doprinelo toliko željenom stanju „tišine“ u zgradama.

Međutim, u praksi svakodnevnog života pokazalo se da je ideal „potpune tišine“ postao uzrok specifičnog problema koji se najpribližnije može označiti kao „prekomerna tišina“. Takvo stanje kome se težilo neočekivano je otvorilo „pandorinu kutiju“ neželjenih zvukova iz okruženja, a među njima je i razgovor suseda. Naime, u takvim uslovima nestaje blagotvorno zaštitno dejstvo umerene ambijentalne buke koja maskira mnogo toga što bi bilo ometajuće, iako ljudi toga nisu svesni jer je ne primećuju i ne skreće im pažnju.

U praksi projektovanja poslovnih prostorija odavno je postalo standard da se problem „prekomerne tišine“ rešava dodatnim sistemima za maskiranje pomoću nekog generisanog šuma. Negde je to prosto šum ventilacije ili sistema za klimatizaciju. Kada to nije dovoljno, pribegava se veštačkim sredstvima. Istorijski gledano, počelo je sa fontanama u „open space“ kancelarijama, a danas na tržištu postoje brojna rešenja namenskih audio sistema za maskiranje. Postoji ISO standard koji definiše parametre za ocenu postignute privatnosti govora u takvim kancelarijskim prostorima, među njima i za kvantifikovanje postignutog efekta maskiranja govora [14]. U proceduri njihovog merenja vrši se i određivanje ekvivalentnog nivoa buke $L_{A,eq}$.

Međutim, u praksi projektovanja i izgradnje stambenih zgrada taj aspekt je i dalje potpuno zanemaren jer je nepoznat kako projektantima, tako i stanarima. U njima postoje dva moguća pravca delovanja da se spreči „prekomerna tišina“ i tako obezbedi privatnost govora. Prvi je da se zanemari činjenica o ekstremno niskim nivoima ambijentalne buke koji postoje u zgradi, ali da se zahtevi u odnosu na izolaciona svojstva pregrada značajno povećaju kako bi se obezbedio očekivani zvučni komfor. To bi zahtevalo da zidovi između stanova imaju građevinsku izolacionu moć i do 60 dB, a ne minimalnih 52 dB kao što je zapisano u domaćoj regulativi. U zemljama Evrope gde postoji klasifikacija zgrada u pogledu akustičkog kvaliteta vrednosti izolacione moći preko 60 dB još ranije je propisana za najvišu klasu. Međutim, realno je da takav zahtev u praksi veoma komplikuje izbor materijalizacije pregradnih zidova, često zahteva i poseban pristup u rešavanju konstrukcije zgrade, a svakako poskupljuje gradnju. Ono što je možda najveći problem u tome je što nužno zahteva izvestan otklon od ustaljene graditeljske prakse u Srbiji, kako projektantske, tako i izvođačke.

Drugi pravac je da se i u stanovima, kao što se radi u poslovnim prostorijama, deluje u domenu ambijentalne buke, to jest da se na neki način spreči pojava „prekomerne tišine“. To se može postići inovativnim rešenjima u domenu zvučne izolacije. To bi značilo da se na neki način omogući doziranje prodora spoljašnje buke, a ne samo njeno maksimalno potiskivanje. Međutim, u praksi to može biti delikatno za realizaciju jer podrazumeva usložnjavanje standardnih građevinskih elemenata (na primer prozora). Znatno jednostavniji put sprečavanja „prekomerne tišine“ je da se kontrola nivoa ambijentalne buke realizuje nekim pomoćnim, „negrađevinskim“ sredstvima. Naravno, u granicama nivoa unutar kojih ne privlači pažnju ljudi i sa pažljivo odabranim karakterom emitovanog signala.

Kroz primer parametra Speech Privacy Class (SPC), odnosno novopredloženog Indeksa privatnosti govora (IPG), pokazan je način na koji se može kvantifikovati stanje čujnosti i razumljivosti govora. Ovo bi trebalo da inicira neke praktične korekcije standardnih procedura akustičkih merenja u zgradama kako bi korisnici prostora imali jasniju indikaciju postignutog stanja.

Prvi korak ka tome moglo bi biti proširenje formulara za prikaz rezultata merenja zvučne izolacije sa poljem u koji bi se unosili dodatni podaci o izmerenoj vrednosti ekvivalntnog nivoa buke u prostorijama između kojih je izvršeno merenje (prredajna i prijemna). Za realizaciju takve ideje potrebno je preciznije definisati uslove za takvo merenje da bi dobijeni

rezultat bio merodavan za ocenu stanja pri realnoj upotrebi prostorija. Ovakav pristup je posebno svrsishodan u slučajevima kada se merenje vrši u postojećim zgradama povodom žalbi stanara.

Drugi korak bi bio traganje za načinima sanacije kada stanje privatnosti ne zadovoljava korisnike zgrade. Do sada je to uvek vodilo ka nekakvim korekcijama zvučne izolacije, ali je to po pravilu destruktivno za postojeći enterijer. Ova analiza je pokazala da se rešenja mogu tražiti i u domenu kontrole ambijentalne buke unutar boravišnih prostorija stanova. Time se otvara široko polje za istraživanje koje uključuje razvoj tehnoloških ili građevinskih sredstava, ali takođe i istraživanje u oblasti koja je u literaturi nazvana *soundscape*. Ta tema za prostore stanova još nije obrađivana.

VI. ZAKLJUČAK

Primer iz prakse opisan na početku i definicija SPC u američkom standardu na direktan način su pokazali značaj“tišine“ u prostorijama za stanje privatnosti govora. Iz svega prikazanog jasno proizilazi da prekomerna, bolje rečeno „nekontrolisana“ tišina, ako nije praćena odgovarajućim merama zvučne izolacije, otvara potencijalni problem nedovoljne privatnosti govora. Zanimljivo je da stanje prekomerne tišine u zgradama nastaje kao sekundarna posledica čitavog niza zahteva i procesa u savremenom građevinarstvu, a ne kao neki unapred postavljeni cilj kome se svesno težilo.

U radu je pokazano da podatak o građevinskoj izolacionoj moći iskazan u projektu ili izmeren u zgradi nije dovoljan da se proceni stanje privatnosti govora između stanova ili kancelarijskih prostorija. Zbog toga je analizirana moguća upotreba parametra koji u svojoj definiciji integrišu doprinos zvučne izolacije i ambijentalne buke. Predloženo je da se formular izveštaja o merenju vrednosti građevinske izolacione moći dopuni podacima o ekvivalentnom nivou ambijentalne buke u prijemnoj i predajnoj prostoriji. To bi omogućilo da se za svaku okolnosti merenja procenjuje i Indeks privatnosti govora, a iz toga i stanje privatnosti koje korisnici zgrade mogu očekivati.

Najzad, rad aktuelizuje pitanje metodologije za kontrolu nivoa ambijentalne buke u prostorijama kada pregradni elementi zgrade učine da je njen nivo u prostorijama nepotrebno nizak. Načini za to mogu ići u pravcu maskiranja elektroakustičkim sredstvima ili posebnim konstrukcijama drugih elemenata zgrade, kao na primer prozora, koji će na neki način omogućiti doziranje „tišine“.

ZAHVALNICA

Ovaj rad je napravljen kao deo istraživanja u okviru projekta broj TR36026 koga finansira Ministarstvo prosvete, nauke i tehnološkog razvoja Republike Srbije.

LITERATURA

[1] SRPS ISO 6242-3 „Visokogradnja. Izražavanje zahteva korisnika, Deo 3: Akustički zahtevi“, 1997.
[2] SRPS U.J6.201 „Akustika u zgradarstvu – Tehnički uslovi za projektovanje i građenje zgrada“
[3] M.Adnađević, M. Mijić, D. Sumarac Pavlović, D. Mašović, ″Noise in dwellings generated in normal home activities – general approach″, Forum Acusticum, Aalburg, 2011, Proceedings, 1335-1340
[4] J.Bradley, B. Gover, „Speech Privacy Class for Rating the Speech Privacy of Meeting Rooms“, Canadian Acoustics, Vol. 36 No. 3 (2008) 22-23
[5] J.Bradley, B.Gover, „Speech levels in meeting rooms and the probability of speech privacy problems“, The Journal of the Acoustical Society of America 127, 815 (2010)
[6] ISO/TC 43/SC 2 N 1218, TU0901 Proposal CS for NWIP 2013-11-19 − “Acoustic classification scheme for dwellings”
[7] ASTM E2638 – 10, „Standard Test Method for Objective Measurement of the Speech Privacy Provided by a Closed Room“, 2017.
[8] B.Gover, J.Bradley, „Measures for assessing architectural speech security (privacy) of closed offices and meeting rooms“ The Journal of the Acoustical Society of America 116, 3480 (2004)
[9] J.Bradley, and B.Gover, „Speech levels in meeting rooms and the probability of speech privacy problems“, The Journal of the Acoustical Society of America **127**, 815 (2010)
[10] ISO 16283-1Acoustics — Field measurement of sound insulation in buildings and of building elements — Part 1: Airborne sound insulation
[11] ISO 1996-1 Acoustics — Description, measurement and assessment of environmental noise — Part 1: Basic quantities and assessment procedures
[12] ISO 717-l Acoustics - Rating of sound insulation in buildings and of building elements - Part 1: Airborne sound insulation
[13] M.Bjelić, T.Miljković, D.Šumarac Pavlović M.Mijić, „Speech Privacy Class in ISO standardisation’s world“, paper in preparation for publishing
[14] ISO 3382-3 Acoustics — Measurement of room acoustic parameters — Part 3: Open plan offices

ABSTRACT

In newly built buildings, both residential and business, there are increasing of complaints about a lack of speech privacy as an essential element of acoustic comfort. The only numerical indicator that is used today as a formal information about acoustic comfort is the sound reduction index of the partitions between the dwellings. Any conclusion about the achieved acoustic comfort in the building can be based only on such an information. However, some recent experience shows that with same value of sound reduction index different privacy can be identified. The paper discusses the causes of such occurrence, as well as the possibilities of applying more complex numerical indicators if the speech privacy. On that basis, the paper proposes an upgrading of the standard form for sound insulation measurement report for more accurate information about speech privacy, as well as the possible solution in the design and construction of buildings.

Influence of “silence” on acoustic comfort

Miomir Mijić, Dragana Šumarac Pavlović, Miloš Bjelić, Tatjana Miljković

Platforma za realizaciju napredne akustičke kamere

Iva Salom, Vladimir Čelebić, Vladimir Ćatić, Jovana Novaković, Bratislav Planić, Veljko Janić, Marko Ralić i Dejan Todorović

***Apstrakt*—U ovom radu prikazana je jedna realizacija napredne akustičke kamere, koja obuhvata specijalno projektovane module sa digitalnim MEMS mikrofonima, platformu za akviziciju do 32 signala i skladištenje podataka, dodatne module, koji se mogu priključiti u zavisnosti od primene akustičke kamere (GPS, meteorološki moduli) i video kameru. *Beamforming* algoritam realizovan je u programskom paketu MATLAB i kao rezultat postprocesiranja dobija se akustička mapa snimanog područja sa zadatom rezolucijom. Realizovan sistem pokazao se kao pouzdana platforma za realizaciju različitih mikrofonskih nizova sa dodatnim opcijama u zavisnosti od konkretne primene.**

***Ključne reči*—akustička kamera; mikrofonski niz; beamforming; MEMS mikrofon; FPGA; DSP.**

I. Uvod

Akustička kamera je jedna od mogućih primena mikrofonskog niza u kombinaciji sa video kamerom. Vizuelizacija zvučnog polja vrši se primenom određenog algoritma za prostorno-vremensku obradu signala sa mikrofonskog niza, i preklapanjem dobijene akustičke mape snimanog područja sa snimkom sa video kamere.

Od kako se početkom dvehiljaditih godina pojavila prva akustička kamera na tržištu, kreće nagli razvoj ove tehnologije, tako da danas akustička kamera predstavlja moderan inženjerski alat koji se, kroz identifikaciju i utvrđivanje položaja izvora zvuka, kao i kvantifikovanje i analizu pojedinačnih izvora zvuka, sve više koristi za različite namene: za određivanje i karakterizaciju izvora buke, u analizi akustike prostorija, za ispitivanje zvučne izolacije, za utvrđivanje kvarova u industrijskim postrojenjima (detekcijom/praćenjem poremećaja zvučnog polja), prilikom dizajniranja i testiranja vozila (auto-industrija, avio-industrija), u robotskim sistemima itd. [1]-[12]. Kao posledica toga, danas je na tržištu dostupan veliki broj akustičkih kamera različitih karakteristika i realizacija, u zavisnosti od primene [13]-[17].

Performanse akustičke kamere zavise od nekoliko parametara [2]: frekvencijski opseg - donju graničnu frekvenciju načelno određuje najveća dimenzija mikrofonskog niza, dok gornju graničnu frekvenciju određuje minimalno rastojanje između mikrofona, kao i frekvencija odabiranja audio signala; prostorna rezolucija predstavlja mogućnost razdvajanja dva izora zvuka na malom rastojanju i zavisi od frekvencije odabiranja audio signala, kao i udaljenosti izvora od mikrofonskog niza; selektivnost zavisi od širine glavnog luka i potiskivanja bočnih lukova, što je u direktnoj vezi sa konfiguracijom mikrofonskog niza (prostornim rasporedom mikrofona), tj. njegove karakteristike usmerensti (*beampattern*); pojava lažnih izvora zavisi od potiskivanja bočnih lukova, što takođe zavisi od karakteristike usmerenosti mikrofonskog niza; dinamički opseg zavisi od broja i osetljivosti mikrofona u mikrofonskom nizu.

Nakon akvizicije audio signala mikrofonskim nizom vrši se obrada signala (*array processing*). Svi ovi algoritmi mogu se definisati kao prostorno filtriranje (*beamforming*), odnosno izdvajanje akustičkog signala iz određenog pravca. *Beamforming* algoritmi se koriste za određivanje pravca nailska zvuka (*Direction of Arrival* - DoA) [1]-[3]. Postoji veliki broj *beamforming* algoritama koji se razlikuju po svojim karakteristikama pa je za konkretnu primenu potrebno izabrati odgovarajući algoritam (preciznost, složenost u pogledu zahtevanih resursa, brzina, potreban broj signala itd.) [1]-[3],[18],[19].

Ukoliko se javi potreba za specifičnom primenom akustičke kamere, kao što je na primer snimanje buke iz vazduha, kada je akustičku kameru neophodno montirati na bespilotnu letilicu pa mora zadovoljavati određene zahteve za malom težinom, autonomnim napajanjem, uzimanjem u obzir dodatnih parametara, kao što su meteorološki i slično, često gotova rešenja nisu pogodna za upotrebu. Sa druge strane, može se javiti i potreba za prilagođenjem postojećeg ili razvojem novog algoritma. U takvim slučajevima pristupa se razvoju sopstvenog rešenja.

U ovom radu prikazano je jedno razvijeno rešenje akustičke kamere bazirano na *Xilinx* Zynq-7000 platformi [20], koje predstavlja unapređenu verziju akustičke kamere opisane u [21]. Zynq-7000 platforma je tipa AP SoC (*All Programmable System on Chip*), koja uključuje procesor (PS – *Processing System*) i FPGA (*Field Programmable Gate Array*) programabilnu logiku (PL – *Programmable Logic*). Ovakva konfiguracija posebno je pogodna za realizaciju akustičke kamere jer omogućava optimalno iskorišćenje resursa kroz implementaciju odgovarajućih funkcija sistema na odgovarajućem podsistemu [22]. Ovde se pre svega misli na mogućnosti paralelnog procesiranja FPGA logike i implementaciju akvizicije i obrade

Iva Salom – Institut Mihajlo Pupin, Univerzitet u Beogradu, Volgina 15, 11060 Beograd, Srbija (e-mail: iva.salom@pupin.rs).
Vladimir Čelebić – Institut Mihajlo Pupin, Univerzitet u Beogradu, Volgina 15, 11060 Beograd, Srbija (e-mail: vladimir.celebic@pupin.rs).
Vladimir Ćatić – Institut Mihajlo Pupin, Univerzitet u Beogradu, Volgina 15, 11060 Beograd, Srbija (e-mail: vladimir.catic@pupin.rs).
Jovana Novaković – Institut Mihajlo Pupin, Univerzitet u Beogradu, Volgina 15, 11060 Beograd, Srbija (e-mail: jovana.novakovic@pupin.rs).
Bratislav Planić – Institut Mihajlo Pupin, Univerzitet u Beogradu, Volgina 15, 11060 Beograd, Srbija (e-mail: bratislav.planic@pupin.rs).
Veljko Janić – Institut Mihajlo Pupin, Univerzitet u Beogradu, Volgina 15, 11060 Beograd, Srbija (e-mail: veljko.janic@pupin.rs).
Marko Ralić – Institut Mihajlo Pupin, Univerzitet u Beogradu, Volgina 15, 11060 Beograd, Srbija (e-mail: marko.ralic@pupin.rs).
Dejan Todoorović – Dirigent Acoustics, Mažuranićeva 29, 11000 Beograd, Srbija (e-mail: dejan.todorovic@dirigent-acoustics.rs).

velikog broja ulaznih signala u realnom vremenu, zbog čega FPGA tehnologija poslednjih godina preuzima primat u realizaciji sistema sa mikrofonskim nizovima [23].

Pored platforme za akviziciju i skladištenje podataka na memorijski medij (SD karticu) u realnom vremenu, koja omogućava akviziciju sa maksimalno 32 mikrofona, sistem uključuje i specijalno projektovane module sa digitalnim MEMS mikrofonima, dodatne module koji se mogu priključiti u zavisnosti od primene (*Global Positioning System* – GPS modul, meteorološki moduli - temperatura, vazdušni pritisak, vlažnost) i video kameru, kao nezavisan uređaj sa sopstvenim napajanjem i memorijom.

Prilikom projektovanja najpre je bilo neophodno ispitati karakteristike sistema. U tu svrhu realizovana je simulacija u softverskom paketu MATLAB sa implementiranim *delay-and-sum beamforming* algoritmom [1]-[3]. Pomoću simulacije, u skladu sa mogućnostima primenjene platforme i zahtevima primene, izabrana je realizovana konfiguracija mikrofonskog niza [24]. Posebna pažnja posvećena je mehaničkoj realizaciji nosača akustičke kamere sa zadatim karakteristikama, kao i izboru komponenata da bi se zadovoljio uslov za malom težinom.

Program za obradu snimljenih podataka zajedno sa korisničkim interfejsom realizovan je u softverskom paketu MATLAB na osnovu programa za simulaciju, uz preklapanje dobijenih akustičkih mapa sa snimcima sa video kamere i kombinovanjem sa podacima sa drugih modula. U radu su data poređenja simulacije sa realnim snimcima na jednm primeru karakterističnog signala.

II. Softverska simulacija i izbor parametara

Osnovni zadatak realizovane MATLAB simulacije je da što preciznije prikaže ponašanje mikrofonskog niza sa mogućnošću podešavanja svih parametara, kako bi se izabrala optimalna konfiguracija za konkretnu primenu. Na Sl. 1 prikazan je korisnički interfejs realizovane simulacije, na kojoj se uočavaju parametri simulacije koji se mogu podešavati, kao i rezultat simulacije.

Konfiguracija mikrofonskog niza podrazumeva zadavanje broja i prostornog rasporeda mikrofona. Položaj mikrofona se zadaje izborom standardnih konfiguracija (zvezda, krug i sl.) sa ravnomerno raspoređenim mikrofonima (izbor do 4 kružnice sa nezavisnim izborom poluprečnika, broja mikrofona na krugu i ugaone rotacije u odnosu na *x* osu). Pored planarne konfiguracije mikrofonskog niza dodatno je moguće realizovati prostornu konfiguraciju zadavanjem *z* pomeraja svakog od krugova (paralelno *xy* ravni) ili definisati položaj mikrofona na sferi zadatog poluprečnika. Pri izboru predefinisanih konfiguracija smatra se da je centar mikrofonskog niza u koordinatnom početku.

Karakteristika mikrofona - mikrofoni imaju podrazumevanu omnidirekcionu karakteristiku usmerenosti, ali se može zadati proizvoljna karakteristika usmerenosti. U simulaciji se može uključiti slučajano pojačanje ulaznih signala koji odgovaraju pojedinim mikrofonima do ±2 dB što odgovara odstupanju u kalibraciji realnih mikrofona.

Izvori zvuka zadaju se polarnim koordinatama, koeficijentom slabljenja i odgovarajućim *.wav* fajlom. Proizvoljan *.wav* fajl može biti izabran kao izvor zvuka. Generisana je baza audio fajlova koji se najčešće koriste u simulacijama (*Dirac*-ov impuls, sinusni tonovi različitih frekvencija, beli i roze šum itd.). U trenutnoj realizaciji mogu se uneti do dva izvora zvuka.

Izbor algoritma – u trenutnoj realizaciji implementiran je *delay-and-sum beamforming* algoritam. Kašnjenja signala na mikrofonima, nastala kao posledica prostiranja akustičkog talasa, predstavljaju izvor informacija za *beamforming* algoritam.

Prostorni ugao proračuna, koji se koristi kao jedinica rezolucije, podešava se u stepenima i predstavlja vertikalni i horizontalni ugaoni krak.

Filtriranje ulaznih signala standardnim tercnim i oktavnim filtrima u opsegu od 200 Hz do 20 kHz omogućava ispitivanje osetljivosti mikrofonskog niza u različitim opsezima učestanosti.

Prikaz rezultata simulacije predstavlja normalizovani nivo zvuka u zavisnosti od prostornog ugla mapiran na ravan korišćenjem azimutne ekvidistantne projekcije, obrađen prema zadatom opsegu prikazivanja i preklopljen sa zadatom statičnom fotografijom scene, koja uzima u obzir zadati ugao snimanja video kamere postavljene u centru akustičke kamere.

Imajući u vidu primenu projektovane akustičke kamere određen je frekvencijski opseg od 250 Hz do 9 kHz, i s obzirom da realizovana platforma podržava akviziciju sa maksimalno 32 mikrofona vršene su simulacije sa nekoliko različitih konfiguracija mikrofonskog niza od 32 mikrofona sa različitim izvorima zvuka, koje su poređene po pitanju različitih parametara, a pre svega po selektivnosti i potiskivanju signala u odnosu na posmatrani pravac. Da bi se obezbedilo veće potiskivanje zvuka sa zadnje strane akustičke kamere izabrana je prostorna konfiguracija.

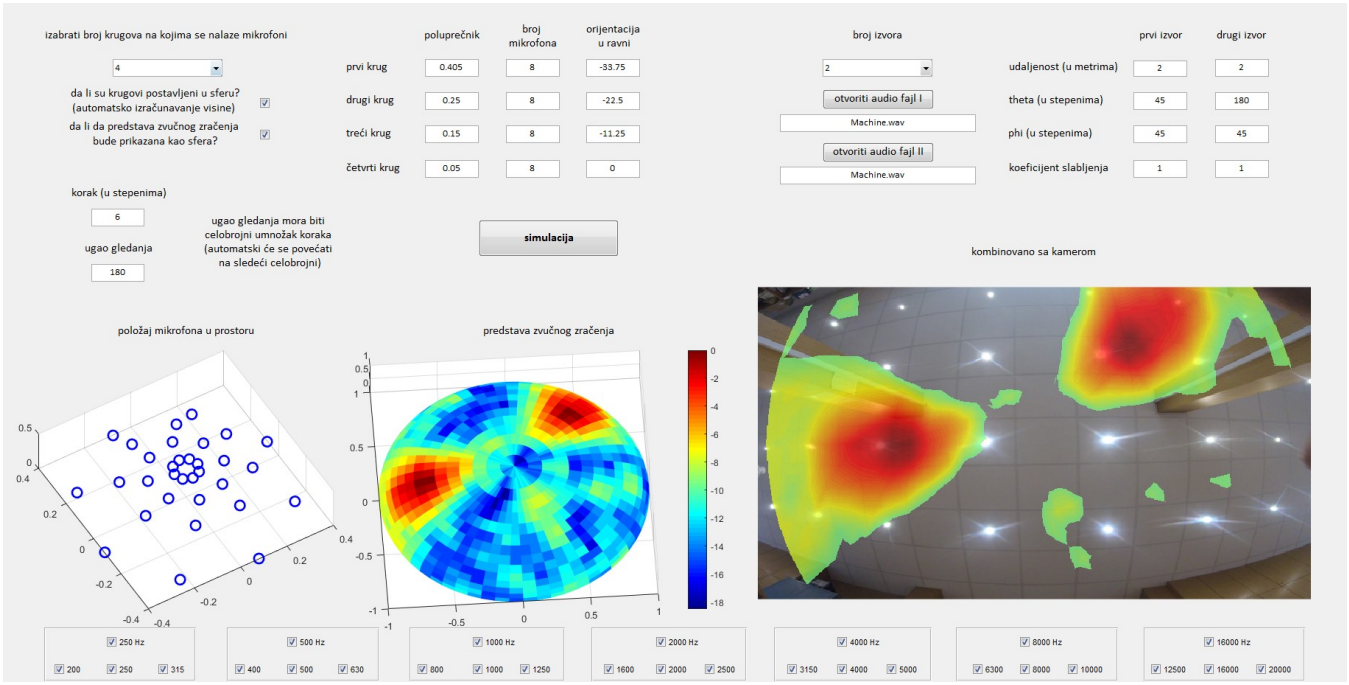
Konfiguracija mikrofonskog niza izabrana za realizaciju sastoji se od 4 kružnice sa po 8 ravnomerno raspoređenih mikrofona, pri čemu se kružnice nalaze na sferi poluprečnika 40 cm, a poluprečnici kružnica su 5 cm, 15 cm, 25 cm i 40 cm. Pri tome su mikrofoni na različitim kružnicama međusobno rotirani za po 11.25 stepeni. Izabrana konfiguracija mikrofonskog niza može se uočiti na prozoru korisničkog interfejsa simulacije, prikazanom na Sl. 1.

III. Opis sistema

Razvijena akustička kamera, prikazana na Sl. 2, obuhvata 1) senzorski blok (mikrofonski niz), 2) platformu za akviziciju, obradu i skladištenje podataka, 3) GPS modul, 4) modul za prikupljanje meteoroloških podataka (temperatura, vlažnost vazduha i pritisak), 5) video kameru, 6) napajanje sistema, 7) mehanički nosač. Blok za obradu podataka (*postprocessing*) i korisnički interfejs predstavlja PC računar sa odgovarajućim programima.

A. Senzorski blok

Osnovna uloga senzorskog bloka je odabiranje akustičkog signala i AD konverzija. Mikrofonski niz se sastoji od 32 mikrofona koji su postavljeni u odgovarajuću konfiguraciju. Za konkretnu primenu zahteva se da mikrofoni imaju omnidirekcionu karakteristiku usmerenosti, širok frekvencijski odziv, zadovoljavajuću osetljivost, kao i zadovoljavajući odnos signal/šum. Zbog svojih malih dimenzija i otpornosti na smetnje, naročito zbog relativno dugačkih linija od samog mikrofona do modula za akviziciju, korišćeni su digitalni MEMS mikrofoni [25]. Odabrani su omnidirekcioni mikrofoni



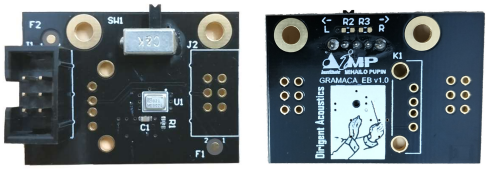
Sl. 1. Korisnički interfejs MATLAB simulacije sa rezultatom za odabranu konfiguraciju mikrofonskog niza i zvuk motora u opsegu 1.5 kHz – 2.5 kHz.

ADMP621 proizvođača *TDK InvenSense* [26] zbog prihvatljivog frekvencijskog odziva u rasponu frekvencija od 100 Hz do 16 kHz, dobrog odnosa signal-šum od 65 dBA i osetljivosti od -46 dBFS. Interesantna je činjenica da su se konkretni mikrofoni pojavili na tržištu pre više od pet godina, a pri tome i dalje sa datim karakteristikama predstavljaju najbolji izbor za primenu u ovakvom sistemu.

Sl. 2. Realizovana akustička kamera.

Odabrani digitalni MEMS mikrofoni predstavljaju integrisano rešenje koje obuhvata akustički pretvarač, predpojačavač audio signala i sigma-delta konvertor unutar jednog integrisanog kola. Male dimenzije pakovanja digitalnih MEMS mikrofona omogućavaju lako postavljanje komponenti na elektronsku štampanu ploču pomoću mašina za automatsko ulaganje komponenata. Digitalni interfejs ovakvih mikrofona omogućava njihovo direktno povezivanje sa digitalnim integrisanim kolima, kao što su FPGA programabilna integrisana kola, bez potrebe za dodatnim komponentama poput analogno-digitalnog konvertora koji bi bio neophodan pri upotrebi analognih mikrofona. Mikrofoni ovakvog tipa zahtevaju, pored veza sa napajanjem i masom sistema, jedino ulazni signal takta frekvencije između 1 MHz i 3 MHz. Svaki od mikrofona povezan je preko četiri signala tj. signala takta, mase, napajanja i izlaznog signala podataka. Izlazi signala podataka grupe koju čine dva mikrofona multipleksirani su impulsno-širinskom modulacijom, PDM (*Pulse Density Modulation*). Na ovaj način mikrofoni iz iste grupe dele izlazni signal podataka i sinhronišu se na isti ulazni signal takta. Ovakva raspodela konekcija rezultuje smanjenjem fizičkog broja veza sa FPGA programabilnom logikom na polovinu od ukupnog broja implementiranih digitalnih MEMS mikrofona.

Digitalni MEMS mikrofoni, svaki ponaosob, asemblirani su na jednostavnu elektronsku štampanu ploču, označenu kao EB (*Extension Board*). EB ploča, pored MEMS mikrofona i pasivnih komponenti sadrži i prekidač, kojim se podešava na koju ivicu takta se vrši odabiranje podataka sa MEMS mikrofona iz mikrofonskog para koji su povezani na zajedničke linije. Svaka od ploča realizovana je na takav način da se lako može povezati sa flet kablom za ostvarivanje fizičke veze preko odgovarajućeg konektora. Ovakvom realizacijom omogućena je laka zamena akustičkih senzora u slučaju pojedinačnih otkaza, kao i jednostavne izmene u konfiguraciji senzorskog niza. Elektronska štampana ploča EB sa digitalnim MEMS mikrofonom prikazana je na Sl. 3.



Sl. 3. Štampana ploča EB sa digitalnim MEMS mikrofonom.

B. Platforma za akviziciju, obradu i skladištenje podataka

U platformi za akviziciju, obradu i skladištenje podataka vrši se:

- prikupljanje više digitalnih akustičkih signala u PDM formatu,
- konverzija digitalnih akustičkih signala u zahtevani oblik (PDM signal u PCM signal zahtevane frekvencije odabiranja i bitske rezolucije),
- povezivanje sa GPS modulom,
- povezivanje sa modulom za prikupljanje meteoroloških parametara,
- skladištenje audio zapisa sa svih mikrofona na memorijskom modulu (SD kartica), kao i podataka sa GPS i meteoroloških modula.

S obzirom na to da optimalno rešenje problema akvizicije i obrade većeg broja akustičkih signala u realnom vremenu zahteva paralelno procesiranje signala svakog od senzora, za realizaciju je odabrana platforma bazirana na FPGA programabilnoj logici [10]-[12],[27]-[31]. Pored mogućnosti paralelne obrade velike količine podataka dodatne prednosti programabilne logike ogledaju se u povećanoj pouzdanosti, vremenu izvršenja i maloj potrošnji. Za realizaciju funkcionalnog prototipa iskorišćeno je gotovo rešenje - razvojna ploča *Digilent* Arty Z7-10 [32], bazirana na AP SoC čipu *Xilinx* Zynq-7000 [20], koji uključuje dvojezgrani ARM Cortex A-9 procesor (PS) i FPGA programabilnu logiku iz serije Xilinx-7 (PL). Glavna prednost ove platforme ne ogleda se samo u ovim njenim pojedinačnim delovima (PS i PL), već i u mogućnosti njihovog zajedničkog funkcionisanja u okviru jednog sistema, uz brzu komunikaciju, ostvarenu pre svega njihovim povezivanjem preko AXI (*Advanced eXtensible Interface*) interfejsa. Akvizicija podataka i deo obrade vrše se u programabilnoj logici (PL), dok se skladištenje podataka vrši na jednom od procesora (PS). Razvojna ploča *Digilent* Arty Z7-10 odabrana je jer omogućava povezivanje velikog broja digitalnih I/O signala, kao i dodatnih modula preko Pmod konektora, sadrži interfejs za povezivanje sa mikro SD karticom, a pri tome se odlikuje optimizacijom potrošnje, cene i veličine.

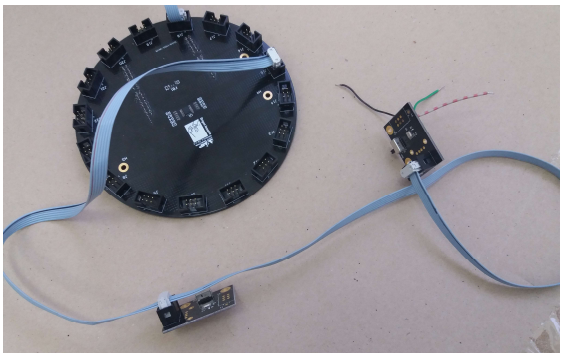
Za povezivanje senzorskog bloka i platforme za akviziciju, obradu i skladištenje podataka projektovana je centralna štampana ploča, označena kao CB (*Central Board*), koja pored potrebnih pasivnih komponenata, sadrži konektor za povezivanje sa *Digilent* Arty Z7-10 razvojnom pločom, sa jedne strane, i 16 odgovarajućih konektora za povezivanje sa EB pločama preko flet kablova zadatih dužina i sa odgovarajućim konektorima, sa druge strane. Na svaki flet kabl povezane su po 2 EB ploče. Na Sl. 4 prikazan je način povezivanja pomenutih komponenata.

C. GPS modul

GPS modulom se preko UART interfejsa prihvataju podaci u obliku informacija o trenutnom položaju sistema (geografske širine i dužine, i nadmorske visine) i tačnog vremena. Izabran je *Digilent* Pmod GPS [33], sa preciznošću geografskih koordinata od 3 m koja zadovoljava potrebe razvijenog sistema.

D. Modul za prikupljanje meteoroloških podataka

Modul za prikupljanje meteoroloških podataka sastoji se iz *Digilent* modula Pmod HYGRO [34] (temperaturni i senzor vlažnosti vazduha) i Pmod NAV [35] (senzor atmosferskog pritiska) sa kojih se očitavaju potrebni parametri.



Sl. 4. Povezivanje CB ploče sa komponentama sistema.

E. Video kamera

Video kamera je deo sistema akustičke kamere koji predstavlja nezavisan uređaj sa sopstvenim napajanjem i memorijom.

F. Napajanje

Za napajanje sistema izabrano je autonomno baterijsko napajanje velikog kapaciteta i malih dimenzija i težine, u vidu eksterne baterije za punjenje mobilnih telefona sa USB mikro konektorom. Namenjeno je za napajanje same akustičke kamere, kao i određenih modula koji se povezuju na platformu. Cilj je da se potrošnja sistema svede na minimum, tako da se obezbedi kontinualno snimanje (od bar 1 h), nakon čega je moguće napuniti ili zameniti bateriju.

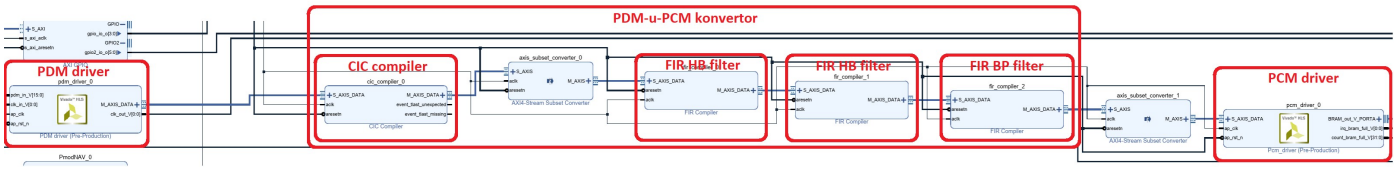
G. Mehanički nosač

Mehanički nosač morao je da zadovolji uslove za malom težinom, kao i za jednostavnim montiranjem ostalih komponenata sistema, a pre svega mikrofona na tačno zadatim pozicijama. Mehanički nosač je projektovan i izdeljen na segmente u softverskom paketu *Solid Works*, a zatim je odštampan pomoću 3D štampača. Masa nosača je oko 0.5 kg.

H. Blok za obradu podataka i korisnički interfejs

Blok za obradu podataka (*postprocessing*) i korisnički interfejs predstavlja PC računar sa specijalno napisanim programom za izdvajanje audio podataka iz vremenskog multipleksa, kao i softverskim paketom MATLAB, u kome je realizovan *delay-and-sum beamforming* algoritam za obradu podataka sa mikrofonskog niza. Iako je među različitim algoritmima najjednostvniji za realizaciju i hardverski najzahtevniji u smislu zahtevanih kapaciteta, pokazao se kao veoma dobro rešenje za veliki broj primena, pa i za konkretnu realizaciju akustičke kamere. Dodatno, moguće je implementirati i druge algoritme, izvršiti poređenje između različitih algoritama, i izabrati odgovarajući algoritam za određenu primenu. U sam algoritam je moguće uključiti podatke dobijene sa dodatnih modula, kao i uključiti napredni algoritnmi za potiskivanje neželjene buke, kao što je buka propelera bespilotne letilice.

Obrada audio zapisa i primena algoritma sprovodi se na delovima signala trajanja 125 ms, nakon čega se vrši preklapanje dobijenih akustičkih mapa sa snimcima dobijenim sa video kamere. Sinhronizacija audio i video zapisa realizovana je na osnovu tačnog vremenskog trenutka impulsne pobude (pljesak) u oba zapisa i skraćivanjem zapisa kod koga je ovaj događaj detektovan kasnije. Na ovaj način je dobijena sinhronizacija sa tačnošću od oko 250 ms.



Sl. 5. Deo realizacije PL (*Xilinx* Vivado Design Suite): PDM-u-PCM konvertor.

IV. SOFTVERSKA REALIZACIJA

FPGA sistemi najčešće se dizajniraju primenom jezika kao što je VHDL (*Very high-speed integrated circuit Hardware Description Language*). Za realizaciju dela sistema u FPGA komponenti (PL) korišćeno je softversko razvojno okruženje *Xilinx* Vivado Design Suite, koje, između ostalog, pruža mogućnost dizajniranja pojedinih delova sistema u vidu IP (*Intellectual Property*) blokova pomoću Vivado HLS softverskog paketa za projektovanje na osnovu alogoritama pisanih u nekom od viših programskih jezika, kao što je C/C++. Za realizaciju PS dela sistema korišćeno je *Xilinx* SDK razvojno okruženje za aplikativni razvoj.

Analizom zahteva sistema za konkretnu primenu određene su potrebne karakteristike izlaznih audio fajlova: frekvencija odabiranja 32 kHz sa rezolucijom od 24 bita.

A. Realiacija PL dela sistema

U okviru PL dela sistema realizovana je akvizicija podataka sa mikrofona u PDM formatu i konverzija u PCM format. Deo realizacije PL dela sistema, realizovanog u *Xilinx* Vivado Design Suite okruženju prikazan je na Sl. 5. Za manipulisanje podacima koji se vode na blok u kome se vrši konverzija iz PDM formata u PCM format, kao i izlaznim podacima iz tog bloka, kreirani su IP blokovi **PDM drajver** i **PCM drajver** u Vivado HLS softverskom paketu. Blok **PDM-u-PCM konvertor** dobijen je kaskadnim povezivanjem postojećih filtarskih blokova (*Xilinx* CIC Compiler v4.0 i FIR Compiler v7.2).

S obzirom na to da performanse jednog ovakvog sistema (pre svega u smislu ispravne akvizicije podataka u realnom vremenu sa zadatog broja ulaznih signala) zavise od načina na koji se pristupa memoriji, veoma je važno da sistem bude strukturiran tako da se omogući što veći protok podataka uz korišćenje minimalne količine resursa. U ovoj realizaciji zadovoljavajuć kvalitet postignut je korišćenjem DMA (*Direct Memory Access*) kontrolera.

1) PDM drajver

Na digitalne MEMS mikrofone dovodi se signal takta frekvencije 3.072 MHz, kojom se vrši odabiranje PDM signala na izlazu iz mikrofona. Akvizicija se vrši paralelno sa svih kanala na uzlaznu/silaznu ivicu signala takta. Jednobitni PDM podaci šalju se na izlaz bloka u vidu dvobitne predstave u komplementu dvojke, prema zahtevu formata ulaza u naredni blok (CIC (*Cascaded Integrator Comb*) filtar). Podaci se šalju u strimu, u TDM (*Time-Division Multiplexing*) formatu.

2) PDM-u-PCM konvertor

U okviru bloka **PDM-u-PCM konvertor** vrši se konverzija signala: ulazni jednobitni PDM signal frekvencije odabiranja 3.072 MHz konvertuje se u 24-bitni 32 kHz PCM signal pa je faktor decimacije 96. Ova konverzija realizovana je u tri koraka: 1) decimacija korišćenjem CIC filtra; 2) filtriranje korišćenjem dva 2:1 HB (*Half Band*) filtra; 3) filtriranje FIR (*Finite Impulse Response*) filtrom propusnikom opsega [36]-[38]. S obzirom na to da je sigma-delta modulator u MEMS mikrofonu četvrtog reda, CIC decimator je petog reda, sa faktorom decimacije 24, dok su faktori decimacije HB filtara 2.

3) PCM drajver

Ovaj blok realizovan je tako da prima 24-bitne PCM podatke sa više kanala u strimu i slaže ih u vremenski multipleks, kako bi se podaci skladištili na SD kartici u tom obliku.

B. Realiacija PS dela sistema

Za realizaciju PS dela sistema korišćeno je *Xilinx* SDK razvojno okruženje za aplikativni razvoj. Ono je kompatibilno sa *Xilinx* Vivado Design Suite okruženjem iz kog je eksportovana konfiguracija dizajnirane hardverske platforme, kako bi na njoj mogao da se razvija softver. Softver se bazira na obradi prekida generisanih po skupljanju određene količine podataka u PCM drajveru. Ovi podaci se dalje smeštaju u bafer i skladište na SD karticu zajedno sa podacima sa dodatnih modula (GPS, meteorološki moduli) u fajlovima određene veličine i formata. Na PC računaru na kome se vrši obrada ovih podataka napisan je program za izdvajanje pojedinačnih audio zapisa sa svakog od mikrofona, kao i podataka sa dodatnih modula.

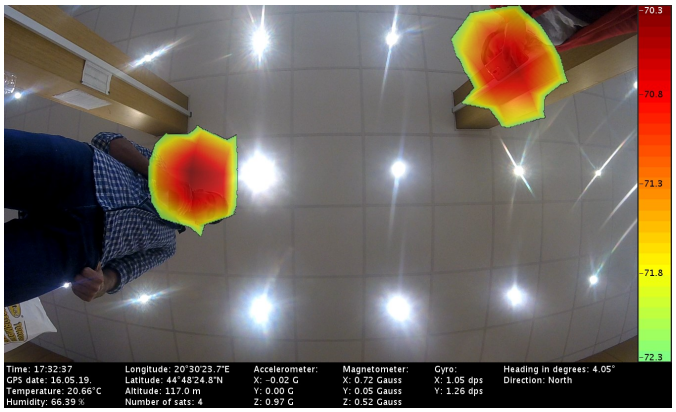
V. REZULTATI

Verifikacija rada realizovanog sistema izvršena je snimanjem realnih signala i poređenjem dobijenih rezultata sa rezultatima simulacije. Na Sl. 6 prikazan je jedan frejm dobijen u programu za obradu na osnovu snimljenih signala sa realizovanom akustičkom kamerom. Tačkasti izvori su zvučnici mobilnog telefona, koji se nalaze na rastojanju od oko 2 m od centra mikrofonskog niza i sa kojih je emitovan zvuk motora, filtriran u opsegu od 1.5 kHz do 2.5 kHz. Na slici se mogu uočiti podaci dobijeni sa GPS i meteoroloških modula. Na Sl. 1 prikazani su rezultati simulacije za istu konfiguraciju i izvor zvuka. Sa slika se može uočiti dobro poklapanje rezultata simulacije i merenja. Generalno, snimanjem različitih izvora zvuka i analizom dobijenih rezultata potvrđen je zadovoljavajuć rad projektovanog sistema.

VI. ZAKLJUČAK

U ovom radu prikazana je jedna platforma za realizaciju napredne akustičke kamere bazirana na *Xilinx* Zynq-7000 AP SoC čipu. Pokazano je da je platforma pogodna za implementaciju dodatnih mogućnosti, kao što su dodavanje odgovarajućih modula, implementiranje i testiranje različitih *beamforming* algoritama, a zbog svoje fleksibilnosti pored konkretne primene može se primenjivati u svim sistemima gde se zahteva akvizicija sa većeg broja senzora. Realizovana akustička kamera ispunila je i zahtev za što manjem težinom (ukupna težina kamere sa svim komponentama je 1.3 kg).

Dodatno, moguća je brza i jeftina proizvodnja, i povezivanje većeg broja ovakvih platformi. Autori nastavljaju rad na poboljšanu karakteristika realizovane platforme u smeru dodatnog povećanja broja ulaznih signala, implementiranja novih algoritama, a pre svega na proširenju mogućnosti same platforme, kao što je realizacija *beamforming* algoritma u realnom vremenu na jednom od procesora sistema, koji je u trenutnoj realizaciji neaktivan.



Sl. 6. Rezutat merenja sistemom – zvuk motora u opsegu 1.5 kHz – 2.5 kHz.

ZAHVALNICA

Ovaj rad je delimično finansiran od strane Inovacionog Fonda u okviru programa saradnje nauke i privrede (projekat ID 50038), kao i od strane Ministarstva prosvete, nauke i tehnološkog razvoja Republike Srbije (projekat TR32038).

LITERATURA

[1] J. Benesty, J. Chen, Y. Huang, "Microphone array signal processing," Springer, 2008.
[2] H. L. Van Trees, "Optimum array processing – Part IV of detection, estimation and modulation theory," John Wiley & Sons Inc, 2002.
[3] M. Brandstein, D. Ward, editors, "Microphone Arrays, Signal Processing Techniques and Applications," New York : Springer-Verlag, 2001.
[4] D. De Vries, E. M. Hulsebos, "Parameterization and Reproduction of Concert Hall Acoustics Measured with a Circular Microphone Array," Audio Engineering Society 112th Convention, Paper No. 5579, Munich, Germany, May 10-13, 2002.
[5] D. Khaykin, B. Rafaely, "Acoustic analysis by spherical microphone array processing of room impulse responses, "The Journal of the Acoustical Society of America. Vol. 132(1), pp. 261–270, 2012.
[6] U. Michel, B. Barsikow, P. Böhning, "Localisation of sound sources on moving vehicles with phased microphone arrays," InterNoise, Proc.Conference pp. 4069–4075, Prague, Czech Republic, August 22-25, 2004.
[7] M. Orman, C. Pinto, "Usage of acoustic camera for condition monitoring of electric motors," TENCON 2013 - IEEE Region 10 Annual International Conference, Proceedings, Oct. 22-25, 2013
[8] M. Mijić, D. Mašović, D. Šumarac Pavlović, M. Adnađević, „Funkcionalni model planarnog mikrofonskog niza s jeftinim komercijalnim mikrofonima," 19. TELFOR 2011, Zbornik radova str. 1040-1043, Srbija, Beograd, 22 - 24. novembar 2011.
[9] M. Bjelić, M. Stanojević, D. Šumarac Pavlović, M. Mijić, „Detekcija slabih tačaka u zvučnoj izolaciji primenom mikrofonskih nizova," 59. ETRAN, Srebrno jezero, jun 2015, Zbornik radova CD: AK 2.3
[10] F. Perrodin, J. Nikolic, J. Busset, R. Siegwart, "Design and calibration of large microphone arrays for robotic applications," IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2012), pp. 4596-4601, Vilamoura, 7-12 Oct. 2012.
[11] B. Zimmermann, C. Studer, "FPGA-based real-time acoustic camera prototype," Proceedings of 2010 IEEE International Symposium on Circuits and Systems (ISCAS), Paris, France, May 30-June 2 2010.
[12] B. da Silva, L. Segers, A. Braeken, A. Touhafi, "Runtime Reconfigurable Beamforming Architecture for Real-Time Sound-Source Localization," 26th International Conference on Field-Programmable Logic and Applications, Lausanne, Switzerland, August 29-September 2nd 2016.
[13] www.acoustic-camera.com
[14] www.lmsintl.com
[15] www.bksv.com
[16] www.norsonic.com
[17] www.microflown.com
[18] M. M. Erić, "Some Research Challenges of Acoustic Camera," 19th TELFOR 2011, Conf. Proc. pp. 1036-1039, Serbia, Belgrade, November 22-24, 2011.
[19] R.A. Mucci, "A comparison of efficient beamforming algorithms," IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 32, No. 3, pp. 438-557, 1984.
[20] Xilinx: Zynq-7000 All Programmable SoC Technical Reference Manual, UG585 (v1.12.1), December 6, 2017, http://www.xilinx.com
[21] J. Novaković, I. Salom, V. Čelebić, B. Planić, V. Ćatić, V. Janić, D. Todorović, „Realizacija akustičke kamerena platformi Zynq-7000," 62. ETRAN, Palić, jun 2018, Zbornik radova CD: AK 1.1
[22] L. H. Crockett, R. A. Elliot, M. A. Enderwitz, R. W. Stewart, *The Zynq Book: Embedded Processing with the Arm Cortex-A9 on the Xilinx Zynq-7000 All Programmable Soc*, United Kingdom, Strathclyde Academic Media, 2014.
[23] B. da Silva, A. Braeken, A. Touhafi, "FPGA-Based Architectures for Acoustic Beamforming with Microphone Arrays: Trends, Challenges and Research Opportunities," Computers No. 7, Vol. 41, August 2018
[24] V. Ćatić, I. Salom, V. Čelebić, D. Todorović, "MATLAB Simulation of an Acoustic Camera using Delay-and-Sum Method," Audio Engineering Society 140th Convention, Novi Sad, Serbia, May 2017.
[25] P. Von Pflug, D. Krischker, "Aspects of the use of mems microphones in phased array systems," InterNoise, Proc.Conference pp. 5983–5993, Hong Kong, August 27-30, 2017.
[26] InvenSnese: ADMP621 datasheet — Wide Dynamic Range Microphone with PDM Digital Output, Technical Report DS-ADMP621-00, InvenSense Inc. San Jose, CA, USA, 2013.
[27] P. Graham, B. Nelson, "FPGA-Based Sonar Processing," Proceedings of the Sixth ACM/SIGDA International Symposium on FieldProgrammable Gate Arrays (FPGA '98), pp. 201-208, 1998.
[28] H. Ye, J. Whittington, I. Himawan, T. Kleinschmidt, M. Mason, "FPGA implementation of dual-microphone delay-and-sum beamforming for incar speech enhancement and recognition," AutoCRC Conference 2009, Conference Proceedings, Melbourne, Convention and Exhibition Centre, Melbourne, Victoria, 5 March 2009.
[29] I. Salom, V. Čelebić, M. Milanović, D. Todorović, J. Prezelj, "An implementation of beamforming algorithm on FPGA platform with digital microphone array," Audio Engineering Society 138th Convention, Warsaw, Poland, May 7-10, 2015.
[30] B. G. Tomov, J. A. Jensen, "A new architecture for a single-chip multichannel beamformer based on a standard FPGA," IEEE Ultrasonics Symposium, Vol. 2, pp. 1529-1533, Atlanta, GA, 07 Oct – 10 Oct 2001.
[31] D. Harders, "Development and Implementation of a FPGA based digital beamformer for an ultrasonic imaging system," thesis for the Degree of Bachelor of Digital Systems with Honours, School of Computer Science and Software Engineering at Monash University, November 2003.
[32] Digilent: Arty Z7 Reference Manual, https://reference.digilentinc.com/reference/programmable-logic/arty-z7/reference-manual
[33] https://reference.digilentinc.com/reference/pmod/pmodgps/reference-manual
[34] https://reference.digilentinc.com/reference/pmod/pmodhygro/reference-manual
[35] https://reference.digilentinc.com/reference/pmod/pmodnav/reference-manual
[36] P.A. Uchagaonkar, S.A. Shinde, V.V. Patil, R.K. Kamat, "FPGA based sigma – Delta analogue to digital converter design," International Journal of Electronics and Computer Science Engineering 1(2), pp. 508-513, 2012.
[37] N. Hegde, "Seamlessly Interfacing MEMS Microphones with Blackfin Processors," Analog Devices, Engineer-to-Engineer Note EE-350, Rev 1 – August 3, 2010.
[38] Lj. Milić, "Multirate Filtering for Digital Signal Processing: MATLAB Applications," Information Science Reference, New York, 2009.

ABSTRACT

This paper presents a realization of an advanced acoustic camera, consisting of digital MEMS microphone modules, a platform for acquisition and storing of 32 input signals, additional modules (such as GPS and meteorological modules), and a video camera. Beamforming algorithm is implemented in MATLAB which results in an acoustic map of the analyzed space with a given resolution. The system can be used for realization of various types of microphone arrays with advanced characteristics depending on the application.

Platform for Realization of Advanced Acoustic Camera

Iva Salom, Vladimir Čelebić, Vladimir Ćatić,
Jovana Novaković, Bratislav Planić,
Veljko Janić, Marko Ralić, Dejan Todorović

Pozicioniranje mikrofona prilikom snimanja audio karakteristika motora putničkih vozila

Marko Milivojčević, Filip Pantelić, Dejan Ćirić

***Apstrakt*— U ovom radu je izvršena analiza audio zapisa dobijenih snimanjem zvuka u oblasti ispod motornog prostora putničkih vozila pokretanih motorima sa unutrašnjim sagorevanjem. Snimanje je ralizovano za tipove vozila sa longitudinalno i transverzalno postavljenim motorom u devet tačaka na površini sa izraženom refleksijom. Prilikom snimanja korišćena su dva tipa mikrofona, merni mikrofon sa omnidirekcionom karakteristikom i pzm (boundary) mikrofon sa kardioidnom karakteristikom. Izvršena analiza ima za cilj da ukaže na mogućnost upotrebe pzm mikrofona za snimanje zvučnih karakteristika motora sa unutrašnjim sagorevanjem prelaskom vozila iznad mikrofona koji su svojom kontrukcijom daleko pogodniji za postavljanje na tlo, kao i da prikaže uticaj relativnog položaja mikrofona u odnosu na izvor zvuka na analizirane karakteristike audio zapisa.**

***Ključne reči*—Akvizicija zvuka, pozicioniranje mikrofona; pzm (boundary) mikrofoni; motori vozila sa unutrašnjim sagorevanjem; spektralna analiza; karakteristike mikrofona.**

I. Uvod

Prilikom frekvencijske analize zvuka generisanog radom motora vozila sa unutrašnjim sagorevanjem sa ciljem određivanja pogonskog goriva realizovane u radu [1], utvrđeno je da se audio zapis najpogodniji za analizu dobija pozicioniranjem mikrofona ispod motornog prostora vozila. Kako bi se kreirala banka audio zapisa potrebnih za mašinsko učenje sistema automatske identifikacije pogonskog goriva motora sa unutrašnjim sagorevanjem na osnovu zvuka potrebno je prikupiti veliki broj audio uzoraka. Za ovu akviziciju, praktično bi bilo najjednostavnije da se vozilo kratak vremenski period u radu pozicionira iznad mikrofona. Kako bi tako ostvarena merenja bila relevantna za dalju analizu, u ovom radu je izvršena analiza uticaja položaja mikrofona u oblasti ispod motornog prostora vozila. Dodatno je izvršena komparacija pzm mikrofona, koji je fizički, zbog svog oblika, najpogodniji za ovu vrstu akvizicije zvučnih zapisa, sa klasičnim mernim mikrofonom kako bi se procenila eventualna degradacija spektra identičnog uzorka zvuka.

U radu je opisan postupak pozicioniranja dva tipa mikrofona u oblasti ispod motornog prostora, neposredno iznad podloge, za dva vozila sa različitom konfiguracijom motora. Cilj je uočiti razlike u prostiranju zvuka koje bi mogle da utiču na kvalitet i karakteristike audio zapisa. U realnim uslovima, kada se prikupljaju uzorci zvuka velikog broja vozila, nije moguće uvek imati identičan položaj mikrofona u odnosu na izvor zvuka. Zbog toga je u ovom radu ispitan položaj mikrofona prilikom kreiranja audio zapisa koji je bio promenljiv. Drugim rečima, mikrofoni su postavljani u fiksne tačke ravnomerno geometrijski raspoređene u oblasti ispod motornog prostora. Simulacija realnih uslova prikupljanja velikog broja uzoraka je podrazumevala rad vozila u režimu praznog hoda, odnosno neopterećenog motora sa stabilnim brojem obrtaja radilice. Izvršena analiza je bazirana na nivou zvučnog polja (zvuka) i spektralnim karakteristikama snimljenih audio sekvenci. Pretpostavka je da su te karakteristike prvi preduslov da se merni mikrofon zameni pzm mikrofonom.

Rezultati komparacije zvučnih zapisa ukazuju da je moguće realizovati sistem za prikupljanje ovog tipa audio zapisa sa pzm mikrofonom i da njegov položaj ispod motornog prostora ne utiče na karakteristike bitne za dalju analizu.

II. Pregled literature

Pregledom literature može se zaključiti da akvizicija zvuka motora vozila sa unutrašnjim sagorevanjem i analiza karakteristika nije široko zastupljena tema. Većina prezentovanih radova se odnosi na snmanje vibro-akustičkih karakteristika, što je razumljivo, obzirom da je cilj uglavnom dijagnostifikovanje neispavnosti u radu motora sa unutrašnjim sagorevanjem [2][3][4]. Međutim, kod ovih primena se pozicioniranje mikrofona vrši u samom motornom prostoru, što u ovom radu nije od interesa. Pokušaji dijagnostifikovanja anomalija u radu motora sa unutrašnjim sagorevanjem rezultovalo je i određenim komercijalnim rešenjima poput pametnog mikrofonskog niza proizvedenog od strane velikog proizvođača automobilskih komponenti Magneti Marelli [5][6] koji se u motornom prostoru pozicionira čvrstom vezom u oblasti usisnog voda.

Klasifikacija vrste vozila na osnovu akustičkih signala (generisanog zvuka) je realizovana u okviru analize buke u saobraćaju sa fokusom na tip vozila (teretno, putničko ili motocikl) [7][8][9], bez osvrta na pogonsko gorivo datih vozila. Ukoliko bi cilj klasifikacije bio selekcija vozila pokretanih motorima sa unutrašnjim sagorevanjem na osnovu pogonskog goriva, kao što je opisano u radu [1] nijedan od navedenih načina pozicioniranja mikrofona nije u potpunosti pogodan, posebno ako se uzme u obzir evidentni napredak u zvučnoj izolaciji proizvođača vozila.

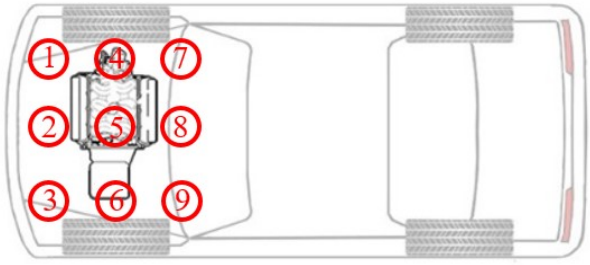
Marko Milivojčević – Visoka škola elektrotehnike i računarstva strukovnih studija u Beogradu, Vojvode Stepe 283, 11000 Beograd, Srbija (e-mail: markom@viser.edu.rs)

Filip Pantelić – Visoka škola elektrotehnike i računarstva strukovnih studija u Beogradu, Vojvode Stepe 283, 11000 Beograd, Srbija (e-mail: filipp@viser.edu.rs)

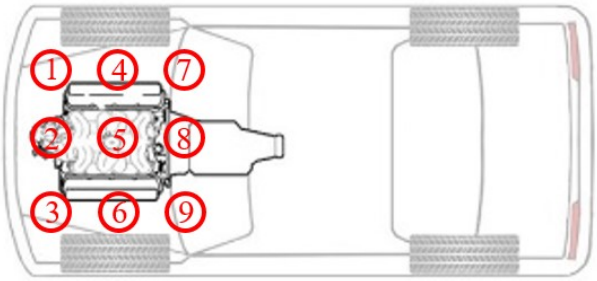
Dejan Ćirić – Elektronski fakultet u Nišu, Univerzitet u Nišu, Aleksandra Medvedeva 14, 18000 Niš, Srbija (e-mail: dejan.ciric@elfak.ni.ac.rs).

III. POZICIONIRANJE MIKROFONA I KARAKTERISTIKE OPREME

Za prikupljanje audio zapisa u oblasti ispod motornog prostora, korišćena su dva vozila, jedno vozilo je imalo uzdužno (longitudinalno) postavljen motor, dok je drugo vozilo imalo poprečno (transverzalno) postavljen motor: Razlog ovakvog izbora vozila je da bi se uvideo mogući uticaj položaja motora na dobijene rezultate. Vozila su pokretana motorima sa unutrašnjim sagorevanjem koji koriste benzin kao pogonsko gorivo i identičan sistem ubrizgavanja, kako bi zvučni zapisi bili što približnijih karakteristika. Za svako od vozila je merni, a zatim i pzm mikrofon postavljan u devet tačaka. Tačke u kojima su prikupljani audio uzorci su izabrane tako da pokrivaju celokupan motorni prostor, pri čemu je oblast ispod motornog prostora procenjena na dimenzije "1.2 m × 1.2 m", a zatim izdeljena na zone kvadratnog oblika i veličine "40 cm × 40 cm", merne tačke su zatim pozicionirane u centrima kvadrata. Tačka "1" se nalazi u centru prvog kvadrata, odnoso na koordinatama "x=20 cm, y=20 cm" posmatrano u odnosu na prednju stranu vozila. Položaj mernih tačaka je prikazan na Sl. 1 i 2.



Sl. 1. Položaj i raspored mernih tačaka za vozilo sa poprečno postavljenim motorom.

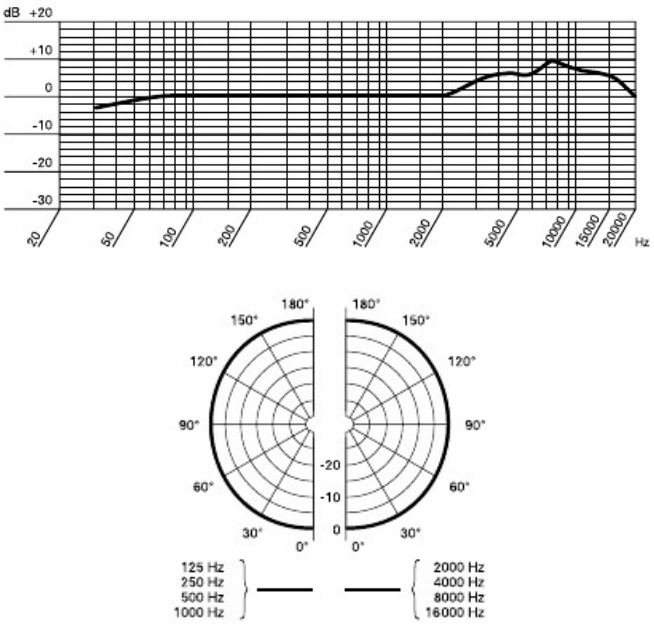


Sl. 2. Položaj i raspored mernih tačaka za vozilo sa uzdužno postavljenim motorom.

Merenje je vršeno u zatvorenoj prostoriji koja nije potpuno zvučno izolovana od okolnih uticaja. U svakoj od mernih tačaka snimani su audio uzorci trajanja 5 s. Uzimajući u obzir da motor sa unutrašnjim sagorevanjem i četiri radna takta u praznom hodu radi na približno 700 obrtaja radilice u minutu, i da je za četiri radna ciklusa sva četiri cilindra potrebno dva obrtaja radilice u sekundi, može se zaključiti da je potrebno približno 0.2 s kako bi motor prošao kroz sve radne cikluse u ovom režimu [10]. Izabrano vreme trajanja audio uzoraka je produženo zbog nesavršenosti mernog okruženja i mogućnosti usrednjavanja i lakše obrade u slučaju da se na snimku pojavi neki neželjeni zvuk.

Oprema korišćena za merenje se sastojala iz dva seta. Jedan set je bio vezan za merni mikrofon, dok je drugi bio praktičnije realizovan sa pzm mikrofonom. Set sa mernim mikrofon se sastojao od mernog mikrofona DBX RTA-M i audio interfejsa Behringer UMC404HD. Set sa pzm mikrofonom se sastojao od mikrofona AKG C562BL i ručnog snimača Zoom H4Nsp. Pzm mikrofon karakteriše robusna konstrukcija, mala visina od svega 9 mm [11] pogodna za postavljanje ispod vozila (Sl. 3), hemisferna karakteristika usmerenosti i frekvencijska karakteristika data na Sl. 4. Snimanje je obavljeno sa frekvencijom odabiranja fs=44100 Hz i dinamičkom rezolucijom od 16 bita.

Sl. 3. Izgled korišćenog pzm mikrofona AKG C562BL.



Sl. 4. Frekvencijska karakteristika mikrofona AKG C562BL [11].

Karakteristika nivoa zvučnog polja (nivoa zvuka) snimaka je računata u okviru softvera SoundForge, dok je frekvencijska analiza rađena u Matlab-u.

IV. KARAKTERISTIKE SNIMLJENIH ZVUKOVA MOTORA

Kao početni korak u analizi zabeleženih audio uzoraka izvršeno je izračunavanje srednje kvadratne vrednosti (RMS-a) za svaki od snimaka. Postupak analize dobijenih vrednosti je podrazmevao četiri slučaja: slučaj uzdužno postavljenog motora snimanog mernim mikrofonom, slučaj uzdužno postavljenog motora snimanog pzm mikrofonom, slučaj

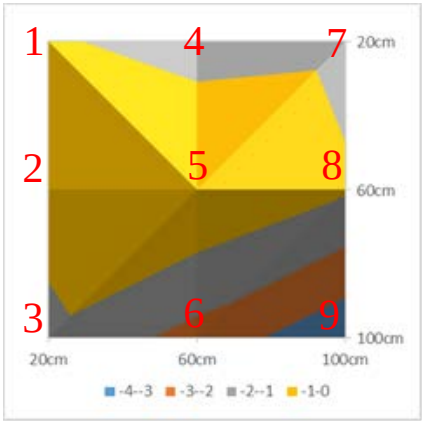
poprečno postavljenog motora snimanog mernim mikrofonom i slučaj poprečno postavljenog motora snimanog pzm mikrofonom. Analiza je sprovedena za režim rada motora u praznom hodu, odnosno neopterećenog motora. Izračunati nivoi RMS-a u odnosu na tip vozila, tip mikrofona i položaj mikrofona su dati u Tabeli I.

TABELA I
IZRAČUNATI NIVOI RMS U ZAVISNOSTI OD TIPA VOZILA I TIPA MIKROFONA

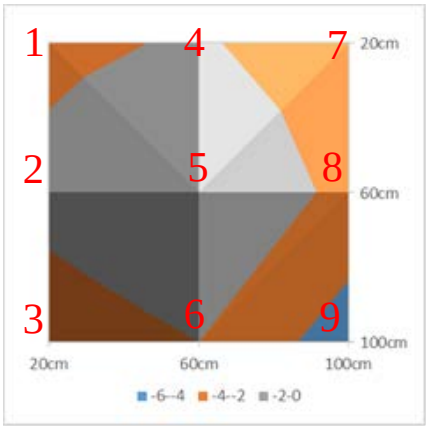
	Poprečno postavljen motor	Uzdužno postavljen motor
Tip mikrofona	PZM	
Merna pozicija 1	-23.95 dB	-20.94 dB
Merna pozicija 2	-23.04 dB	-22.93 dB
Merna pozicija 3	-25.01 dB	-22.73 dB
Merna pozicija 4	-22.61 dB	-19.56 dB
Merna pozicija 5	-21.34 dB	-19.37 dB
Merna pozicija 6	-23.89 dB	-20.62 dB
Merna pozicija 7	-24.52 dB	-20.73 dB
Merna pozicija 8	-23.34 dB	-19.10 dB
Merna pozicija 9	-26.37 dB	-20.13 dB
Tip mikrofona	Merni mikrofon	
Merna pozicija 1	-19.968 dB	-26.017 dB
Merna pozicija 2	-19.792 dB	-27.044 dB
Merna pozicija 3	-20.521 dB	-27.701 dB
Merna pozicija 4	-20.455 dB	-25.127 dB
Merna pozicija 5	-19.075 dB	-25.484 dB
Merna pozicija 6	-21.428 dB	-26.016 dB
Merna pozicija 7	-20.321 dB	-26.214 dB
Merna pozicija 8	-19.963 dB	-25.098 dB
Merna pozicija 9	-22.866 dB	-25.907 dB

Kako su se zbog razlike u konstrukciji vozila, apsolutni nivoi zvuka dva različita vozila razlikovali, prikazani rezultati su predstavljeni kao relativni nivoi - relativni u odnosu na tačku sa najvišim nivoom RMS-a. Na Sl. 5 i 6 su prikazana odstupanja RMS-a u odnosu na tačku sa najvišim nivoom RMS-a za vozilo sa poprečno postavljenim motorom u slučaju da su uzorci snimljeni mernim, odnosno pzm mikrofonom, respektivno. Na Sl. 5, 6, 8 i 9 su varijacije nivoa izračunatih RMS vrednosti izražene u dB prikazane različitim bojama koje su označene na legendi ispod svakog grafika.

U slučaju vozila sa poprečno postavljenim motorom može se zaključiti da je najviši nivo RMS zabeležen u mernoj tački "5", odnosno njegov nivo je -19.075 dB za merni i -21.34 dB za pzm mikrofon. Relativna odstupanja su za merni mikrofon u granicama od 0 dB do -3.791 dB, dok su za pzm mikrofon nešto veća i kreću se u granicama od 0 dB do -5.03 dB. Na Sl. 5. se posebno u mernoj tački "1" može uočiti uticaj usmerenosti pzm mikrofona koja utiče na niži nivo primljenog signala u odnosu na merni mikrofon.

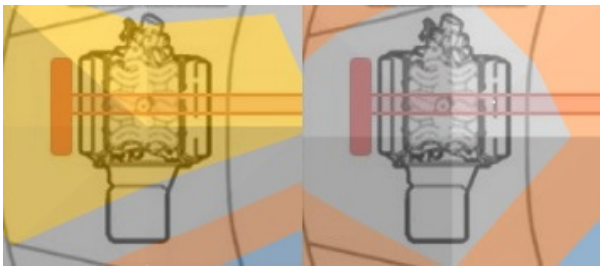


Sl. 5. Prikaz odstupanja RMS vrednosti nivoa zvuka u dB za vozilo sa poprečno postavljenim motorom snimano mernim mikrofonom.



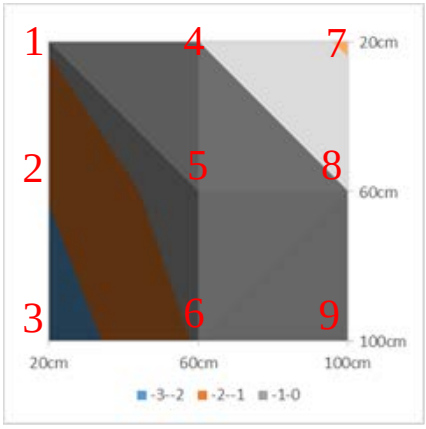
Sl. 6. Prikaz odstupanja RMS vrednosti nivoa zvuka u dB za vozilo sa poprečno postavljenim motorom snimano pzm mikrofonom.

Iako formirano polje RMS vrednosti nivoa zvuka nema velikih varijacija, ipak se može uočiti izvestan karakterističan oblik višeg nivoa koji se poklapa sa konstrukcijom i položajem izduvnog sistema vozila (označenog crvenom bojom), što je prikazano na Sl. 7. Varijacije u nivou RMS-a su na Sl. 7 i 10 prikazane različitim bojama oslanjajući se na grafike 5 i 6 odnosno 8 i 9.

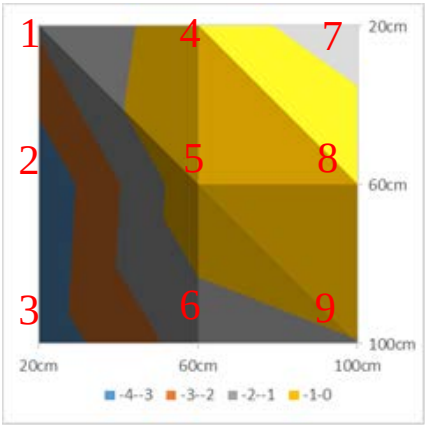


Sl. 7. Uporedni prikaz RMS nivoa kod vozila sa poprečno postavljenim motorom u odnosu na položaj izduvnog sistema.

Na Sl. 8 i 9 su prikazana odstupanja RMS vrednosti nivoa zvuka u odnosu na tačku sa najvišim RMS nivoom za vozilo sa uzdužno postavljenim motorom u slučaju da su uzorci snimljeni mernim odnosno pzm mikrofonom, respektivno.



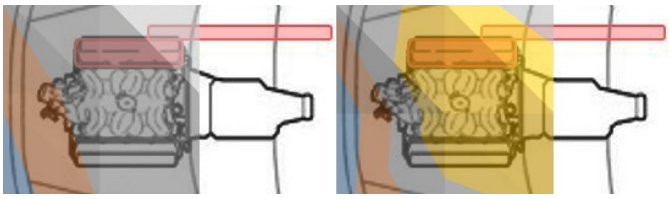
Sl. 8. Prikaz odstupanja RMS vrednosti nivoa zvuka u dB za vozilo sa uzdužno postavljenim motorom snimano mernim mikrofonom.



Sl. 9. Prikaz odstupanja RMS vrednosti nivoa zvuka u dB za vozilo sa uzdužno postavljenim motorom snimano pzm mikrofonom.

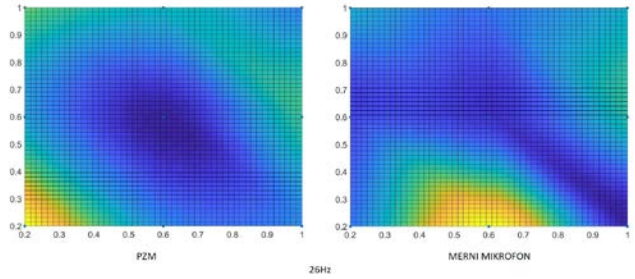
U slučaju vozila sa uzdužno postavljenim motorom može se zaključiti da je najviši RMS nivo zvuka takođe zabeležen u mernoj tački “5”, odnosno njegova vrednost je -25.484 dB za merni i -19.37 dB za pzm mikrofon. Relativna odstupanja su za merni mikrofon u granicama od 0 dB do -2.603 dB dok su za pzm mikrofon nešto veća i kreću se u granicama od 0 dB do -3.83 dB.

Kod vozila sa uzdužno postavljenim motorom formirano zvučno polje takođe nema velikih varijacija u RMS nivou, ali se može uočiti izvestan karakterističan oblik višeg nivoa koji se poklapa sa konstrukcom i položajem izduvnog sistema vozila (označenog crvenom bojom), što je prikazano na Sl. 10.

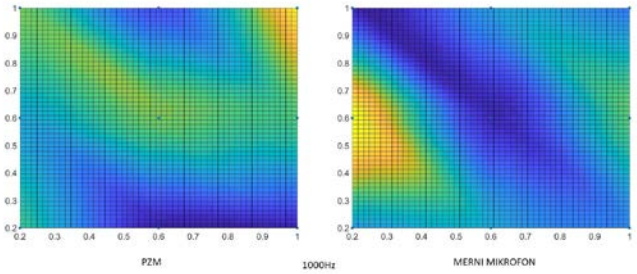


Sl. 10. Uporedni prikaz RMS nivoa kod vozila sa uzdužno postavljenim motorom u odnosu na položaj izduvnog sistema.

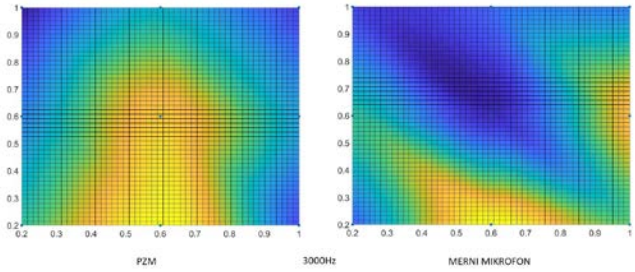
Osvrtom na rad [1] i imajući u vidu zaključak da bi pri klasifikaciji pogonskog goriva motora sa unutrašnjim sagorevanjem posebno bio značajan deo spektra u opsegu 3-4 kHz, izvršena je analiza nivoa zvučnog polja na odgovarajućim frekvencijama, kao i oblika spektra u ovom opsegu kako bi se uvideo uticaj tipa mikrofona. Na Sl. 11 do 15 dat je prikaz nivoa zvučnog polja samo za vozilo sa uzdužno postavljenim motorom, pri čemu su svetlijom bojom prikazani viši nivoi zvučnog polja.



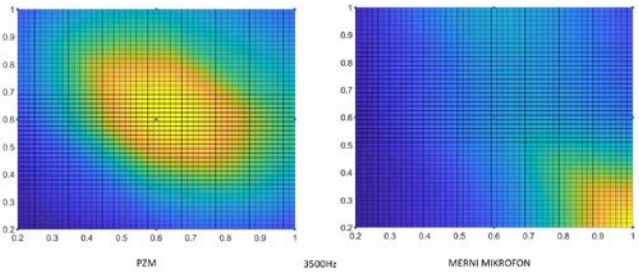
Sl. 11. Nivo zvučnog polja za vozilo sa uzdužno postavljenim motorom na učestanosti od 26 Hz



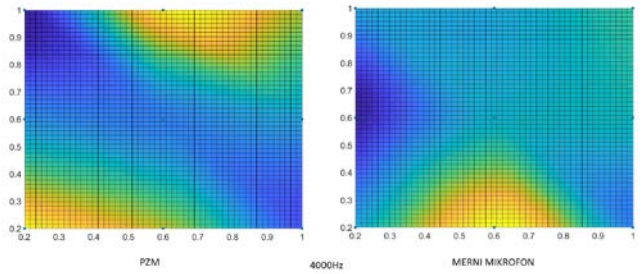
Sl. 12. Nivo zvučnog polja za vozilo sa uzdužno postavljenim motorom na učestanosti od 1000 Hz.



Sl. 13. Nivo zvučnog polja za vozilo sa uzdužno postavljenim motorom na učestanosti od 3000 Hz.



Sl. 14. Nivo zvučnog polja za vozilo sa uzdužno postavljenim motorom na učestanosti od 3500 Hz.



Sl. 15. Nivo zvučnog polja za vozilo sa uzdužno postavljenim motorom na učestanosti od 4000 Hz.

Posmatrajući uporedne karakteristike nivoa zvučnog polja snimane pomoću dva mikrofona sa Sl. 11 do 15 može se uočiti da postoje određene razlike između ovih karakteristika. One su pre svega posledica različitog frekvencijskog odziva i karakteristike usmerenosti mikrofona. Pored toga, može se reći da je oblik zvučnog polja formiran na osnovu snimaka u identičnim uslovima za pzm mikrofon uniformniji, pre svega u centralnom delu analiziranog motornog prostora. Na ovaj način su uzorci snimljeni pzm mikrofonom pogodniji za analizu u slučaju kada nije moguće predvideti tačnu poziciju izvora zvuka u odnosu na mikrofon.

V. ZAKLJUČAK

Posmatrajući rezultate merenja sa aspekta RMS vrednosti nivoa zvuka može se zaključiti da, iako pzm mikrofon ima usmereniju karakteristiku, to ne rezultira velikim varijacijama u nivou zvuka, a posebno ne na karakterističan oblik formiranog polja. Ukoliko se uzme u obzir njegova izuzetno mala visina od 9 mm, ovaj mikrofon je moguće koristiti umesto mernog mikrofona u cilju prikupljanja velikog broja uzoraka audio zapisa motora sa unutrašnjim sagorevanjem. Male varijacije u RMS nivoima zvuka, kao i relativno uniformna raspodela zvučnog polja ispod centralnog dela motornog prostora u frekvencijskom opsegu od interesa ukazuju da položaj mikrofona u oblasti ispod motornog prostora nema veliki uticaj na parametre zvuka. Ovo pruža mogućnost fiksiranja pzm mikrofona na podlogu iznad koje prelaze i kratko se zadržavaju vozila (naplatna rampa, ulazna rampa parkinga, benzinska pumpa) radi snimanja zvuka motora. Dalji rad podrazumeva izdvajanje novog seta parametara od interesa kao i analiza uticaja tipa mikrofona na karakteristike istih.

LITERATURA

[1] M. Milivojčević, F. Pantelić, D. Ćirić, "Comparison of frequency characheristic of sound generated by internal combustion engines depending on fuel," Proc. 26th Noise and Vibration, Niš, Serbia, pp. 115-120, 6-7 December 2018.

[2] S. K. Yadav, P. K. Kalra, "Automatic fault diagnosis of internal combustion engine based on spectrogram and artificial neural network," Proceedings of 10th WSEAS (Int. Conference on Robotics, Control and Manufacturing Technology), Hangzhou, China, 11-13 April, 2010.

[3] N. Cavina, A. Businaro, N. Rojo, M. De Cesare, L. Paiano, A. Cerofolini, "Combustion and intake/exhaust systems diagnosis based on acoustic emissions of a GDI TC engine," Proc. 71st Conf. Italian Thermal Machines Engineering Association, ATI2016, 14-16 September 2016.

[4] J. Yao, Y. Xiang, S. Qian, S. Wang, "Radiation noise separation of internal combustion engine based on Gammatone – Robust ICA Method," *Shock and Vibration*, Article ID 7565041, Vol. 2017, 2017.

[5] M. Turqueti, E. Oruklu, J. Saniie, "Smart acoustic sensor array (SASA) system for real-time sound processing applications," *Smart Sensors and Mems*, pp. 492-517, 2014.

[6] https://www.magnetimarelli.com/business_areas/powertrain/competences/acoustics, accessed April 2019.

[7] M.A. Sobreira-Seoane, A. Rodriguez Molares, J. L. Alba, "Automatic classification of traffic noise," Proc. Acoustics '08, Paris, France, pp. 6221-6226, 29th June- 04th July 2008.

[8] K. Haddad, W. Song, X. Valero, "Environmental sound classification in realistic situations," Proc. Forum Acusticum, Krakow, Poland, 7-12 September 2014.

[9] Karol J. Piczak, "Environmental sound classification with convolutional neural networks," Proc. IEEE Int Workshop on Machine Learning for Signal Processing, Boston, USA, 17–20 September 2015.

[10] S.Raynor, R.H.Haas, "Natural frequencies of multicylinder engines", *International Journal of Mechanical Sciences*, vol. 5, no. 1, pp. 69-75, January–February 1963.

[11] *AKG C562BL Boundry Microphone Specification Sheet*, AKG.

ABSTRACT

In this paper, analysis of audio samples obtained by recording of sound below the compartment of internal combustion engines of passenger vehicles is carried out. Sound is recorded using two vehicle types (with longitudinally and transversally set engine) in nine points located on a reflective surface. Recordings are made by two microphones – by an omni directional measurement microphone and by a pzm (boundary) microphone with cardioid polar pattern. The performed analysis has a goal to show the potentials for usage of the pzm microphone for recording of sound of internal combustion engine when a passenger car is passed above the microphone. This microphone has more convenient construction for mounting on a hard floor. Another goal of the presented analysis is to show the influence of the microphone relative position in reference to the sound source on the audio characteristics of recorded sound.

Positioning of Microphone while Recording the Audio Characteristics of Passenger Vehicle Engines

Marko Milivojčević, Filip Pantelić, Dejan Ćirić

Krive opadanja dobijene u reverberacionoj komori pri merenju koeficijenta apsorpcije

Dejan Ćirić, Kristian Jambrošić, Nikola Stojković

***Apstrakt*—Jedan od klasičnih pristupa za merenje koeficijenta apsorpcije test uzorka je baziran na merenju vremena reverberacije prazne reverberacione komore i vremena reverberacije kada se test uzorak nalazi u komori. Dimenzije i osnovne karakteristike reverberacione komore su definisane u relevantnim standardima. Kada se ispune određeni uslovi, naročito kada je zapremina komore manja od preporučenih vrednosti, mogu se javiti neželjeni efekti sopstvenih modova prostorije na niskim frekvencijama. Uticaj ovih efekata pri merenju koeficijenta apsoprcije test uzoraka na krive opadanja zvuka je analiziran u radu. Merenja impulsnih odziva su izvršena u reverberacionoj komori zapremine oko 65 m^3. U opsezima terci ispod 160 Hz se mogu javiti specifična višestruka (najčešće dvostruka) opadanja zvuka koja daju konkavni oblik krivim opadanja. U ovakvim slučajevima se postavlja pitanje na koji način izračunati vreme reverberacije.**

***Ključne reči*—Krive opadanja; reverberaciona komora; koeficijent apsorpcije; sopstveni modovi prostorije.**

I. UVOD

KOEFICIJENT apsorpcije se meri primenom metoda impedansne cevi ili metoda reverberacione komore (prostorije) koji predstavlja standardizovani metod merenja [1]. Drugo pomenuti metod pretpostavlja da u ovakvoj prostoriji postoji difuzno zvučno polje [2]. Tokom godina primene bilo je dosta oprečnih mišljenja, pogotovu o preciznosti ovog metoda na niskim frekvencijama. Jedan od razloga može biti odstupanje od pretpostavljenog difuznog zvučnog polja, koje nije jednostavno realizovati u praksi.

Analiza kapaciteta reverberacione komore da aproksimira difuzno zvučno polje se bazira na raznim deskriptorima. U ove deskriptore spadaju granična frekvencija, broj modova, prostorna uniformnost reveberantnog zvučnog polja, zakrivljenost krivih opadanja zvuka i preciznost merenog vremena reverberacije [2].

U reverberacionim komorama se može očekivati da krive opadanja zvuka prikazane na logaritamskoj skali imaju oblik prave linije. Zbog toga primena klasičnog načina izračunavanja vremena reverberacije linearnom regresijom koristeći opseg od 20 ili 30 dB deluje sasvim prihvatljivo. Međutim, postoje slučajevi kada krive opadanja zvuka odstupaju od prave linije. Ove zakrivljenosti krivih opadanja se lakše detektuju kada su krive dobijene *Schroeder*-ovim integraljenjem unazad [3] nego kada su dobijene metodom prekinutog šuma (eng. *interrupted noise method*) [4].

Ovaj rad se bavi analizom krivih opadanja zvuka dobijenih pri merenju koeficijenta apsorpcije u reverberacionoj komori čija je zapremina manja od 150 m^3 koja se preporučuje u standardu ISO 354 [4]. Fokus je na krivama opadanja zvuka koje nemaju pravolinijski oblik. Ovakve krive pokazuju ili dvostruko opadanje ili jaku zakrivljenost tako da je njihov oblik konkavan. Ova pojava je dominantna na niskim frekvencijama, tipično ispod 160 Hz. Analizirani su rezultati merenja prazne reverberacione komore, kao i komore sa test uzorcima. Pored toga, merenja u praznoj prostoriji su obavljena sa i bez difuzora.

II. KRIVE SA VIŠESTRUKIM OPADANJEM

Višestruka opadanja zvuka su uobičajena pojava kada postoji više spojenih prostorija (eng. *coupled spaces*). Procena parametara opadanja u ovakvim slučajevima se često svodi na njihovu aproksimaciju sumom više eksponencijalnih opadanja [5,6]. Međutim, višestruka opadanja se mogu javiti i u prostorijama gde nema spojenih prostora već se radi o jedinstvenom celovitom prostoru [7]. U takvim slučajevima, sopstveni modovi prostorije dovode do toga da je opadanje zvuka zbir više eksponencijalnih opadanja umesto jednog jedinstvenog [8]. Na taj način se umesto pravolinijske krive opadanja dobija kriva koja je zakrivljena. Postoje i slučajevi kada se na krivoj opadanja mogu jasno primetiti višestruka pravolinijska opadanja sa različitim strminama.

Pomenuta višestruka opadanja se mogu javiti kako na niskim frekvencijama, tako i na srednjim i visokim frekvencijama. Pod niskim frekvencijama se često podrazumeva opseg ispod *Schroeder*-ove frekvencije [2,9,10]. Karakteristično za ovaj frekvencijski opseg je da postoji manje sopstvenih modova tako da dolazi i do njihovog manjeg preklapanja. Zakrivljenost na niskim frekvencijama je posledica promenljivih vremena opadanja aksijalnih, tangencijalnih i bočnih (eng. *oblique*) modova [11]. Kako je vreme reverberacije veće kod aksijalnih modova, a mnogo manje kod tangencijalnih i bočnih modova [12], onda će se kao rezultat javiti zakrivljenost krivih opadanja. Praktično se krive opadanja svakog od modova kombinuju u krivu koja predstavlja globalno opadanje zvuka. Ova kriva onda nije prava linija, i izbor estimacije vremena opadanja (reverberacije) dobija još više na značaju.

Kada je mikrofon postavljen u ugao prostorije, izmereni

Dejan Ćirić – Elektronski fakultet u Nišu, Univerzitet u Nišu, Aleksandra Medvedeva 14, 18000 Niš, Serbia (e-mail: dejan.ciric@elfak.ni.ac.rs).

Kristian Jambrošić – Fakultet elektrotehnike i računarstva, Sveučilište u Zagrebu, Unska 3, HR-10000 Zagreb, Hrvatska (e-mail: Kristian.Jambrosic@fer.hr).

Nikola Stojković – Elektronski fakultet u Nišu, Univerzitet u Nišu, Aleksandra Medvedeva 14, 18000 Niš, Serbia (e-mail: stojkovicnikola@elfak.rs).

nivo zvučnog pritiska je veći u odnosu na srednju vrednost u prostoriji usled efekata interferencije. Ovo povećanje je 3, 6 i 9 dB za aksijalne, tangencijalne i bočne modove, respektivno [12]. Na sličan način se javljaju efekti interferencije i kada se zvučnik postavi u ugao. Ovakvo ponašanje ima jak uticaj na balans između modova.

Na srednjim i visokim frekvencijama situacija je nešto drugačija. Ovde neravnomerna raspodela apsorpcionih površina može dovesti do višestrukog (više-eksponencijalnog) opadanja [13], odnosno opadanja koje odstupa od Sabinovog zakona [14].

Značajno je istaći da zakrivljenost krive opadanja samo po sebi ne mora da znači da zvučno polje nije difuzno [8]. Međutim, zakrivljenost krivih opadanja može biti ozbiljan problem. Tako, pri merenjima koeficijenta apsorpcije u različitim laboratorijama, jedan od osnovnih uzroka razlika u rezultatima može biti različit pristup tretiranju višestrukih opadanja.

U literaturi je prezentovano da kod krivih sa višestrukim opadanjem početak krivih opadanja sadrži srednju vrednost svih modova u prostoriji [8]. Za razliku od toga, poslednji deo krivih opadanja sadrži modove koji sporije opadaju. U referenci 15 je predloženo da je moguće koristiti početni deo krive opadanja kako bi se izračunao koeficijent apsorpcije test uzorka. Eksperimenti u skaliranom modelu reverberacione komore su pokazali da se bolji rezultati za određivanje koeficijenta apsorpcije dobijaju kada se koristi početni deo opadanja umesto vremena reverberacije dobijenog na osnovu opadanja od 20 ili 30 dB [8]. Na ovaj način se dobijaju manje vrednosti vremena reverberacije nego što bi se po teoriji očekivalo, ali je to bolje rešenje nego koristiti opadanje sa manjom strminom. Prezentovani su i pristupi koji nisu favorizovali karakterizaciju opadanja zvuka u reverberacionoj komori koristeći samo jednu vrednost [16].

III. Granična frekvencija reverberacione komore

Kod prostorija malih dimenzija, dominantan uticaj na odziv ima modalno ponašanje, odnosno akustika malih prosorija se često karakteriše neregularnom raspodelom zvuka na niskim frekvencijama [12]. Reverberacione komore spadaju u prostorije specijalne namene, čija se konstrukcija razlikuje od običnih prostorija. Međutim, neregularna raspodela zvuka na niskim frekvencijama se može javiti i u ovim komorama.

Jedan od značajnih parametara kojim se opisuje reverberaciona komora je granična frekvencija koja pokazuje donju graničnu frekvenciju difuznosti zvučnog polja. Dve vrste definicija se koriste za određivanje donje granične frekvencije: jedna je bazirana na konceptu faktora modalnog preklapanja dok je druga bazirana na broju modova u određenom frekvencijskom opsegu [2]. Iz prvo pomenute defincije prositiče tzv. *Schroeder*-ova frekvencija koja odgovara faktoru modalnog preklapanja vrednosti 3 [9], koja je prilično restriktivna sa tačke gledišta primene [17]. *Schroeder*-ova frekvencija deli frekvencijski odziv prostorije na nisko frekventni (ispod *Schroeder*-ove frekvencije) gde dominiraju modovi prostorije, i visoko frekventni (iznad *Schroeder*-ove frekvencije) gde dominira gusto modalno preklapanje sa statističkim (Gausovim) svojstvima [10]. *Schroeder*-ova frekvencija je definisana izrazom

$$f_c = 2000 \cdot (T/V)^{0.5} \tag{1}$$

gde je T vreme reverberacije, a V zapremina prostorije. Za konkretnu reverberacionu komoru Elektronskog fakulteta u Nišu (u kojoj su izvršena merenja) zapremine 65.05 m^3, ukoliko se za vreme reverberacije uzme vrednost od 5 s, dobija se *Schroeder*-ova frekvencija 554 Hz. Međutim, u mnogim praktičnim merenjima nisu pronađeni dokazi da se nisko frekventni region prostirao na frekvencije veće od polovine *Schroeder*-ove frekvencije [10]. U skladu sa tim je predložena i modifikovana granična frekvencija koja je približno 0.45 vrednosti *Schroeder*-ove frekvencije [10]. U slučaju reverberacione komore Elektronskog fakulteta u Nišu, poslednji izraz bi dao graničnu frekvenciju od 250 Hz.

Drugo pomenuta definicija je statistički bazirana i često se koristi. Granična frekvencija dobijena na ovaj način odgovara modalnom broju (broju modova) 20 u bilo kom frekvencijskom opsegu. Za jedan tercni opseg, ova granična vrednost (f_c) se dobija pomoću izraza

$$f_c = \frac{343}{\sqrt[3]{V/4}} \tag{2}$$

U konkretnom slučaju komore zapremine 65.05 m^3, granična frekvencija iznosi oko 135 Hz. Za komore zapremine od 150 m^3, granična frekvencija prema (2) ima vrednost oko 102 Hz, dok za komore zapremine 200 m^3, granična frekvencija iznosi 93 Hz.

Stepen linearnosti krivih opadanja može biti dobar indikator difuznosti zvučnog polja [2]. Da bi se kvantifikovao stepen nelinearnosti krive opadanja ili drugim rečima njene zakrivljenosti, koristi se koeficijent korelacije (r) između krive opadanja i aproksimacione prave koja na najbolji način reprezentuje datu krivu opadanja [11]. Efektivna mera zakrivljenosti se može dobiti i pomoću odstupanja od idealne korelacije [11], što se predstavlja jednačinom

$$\kappa = 1000\left(1 - abs(r)\right) \tag{3}$$

IV. Merenja i obrada odziva

U okviru COST akcije 15125 realizuje se *round robin* eksperiment koji se bavi merenjem koeficijenta apsorpcije 3 unapred pripremljena test uzorka. Akcenat je na uticajima različitih reverberacionih komora, opreme, obrade i timova koji vrše merenja, sa posebnim osvrtom na rezultate na niskim frekvencijama. U tom kontekstu, izvršena su merenja koefcijenta apsorpcije 3 test uzorka u prethodno pomenutoj reveberacionoj komori Elektronskog fakulteta u Nišu. Prostorija je nepravilnog oblika, nema paralelnih zidova, najveća stranica osnove je 4.08 m, dok je najmanja stranica osnove 3.67 m. Najveća visina je 4.33 m, dok je najmanja

visina 3.87 cm. Svi zidovi su refleksioni, vrata su dvostruka, a unutrašnja površina vrata je prekrivena pločicama.

Za namenu povećanja difuznosti zvuka, mogu se koristiti difuzori koji se postavljaju da vise sa tavanice. Postoji 5 ovakvih difuzora površine od 0.8 do 2 m^2, videti Sl. 1, što je u skladu sa standardom ISO 354 [4]. Raspored difuzora je slučajan. Merenja u praznoj komori su obavljena sa i bez difuzora, dok su sva merenja sa test uzorcima obavljena kada su difuzori bili postavljeni.

Sl. 1. Reverberaciona komora sa difuzorima.

Izračunavanja koeficijenta apsorpcije u reverberacionoj komori su bazirana na izmerenim vrednostima vremena reverberacije prazne komore i kada je u nju postavljen test uzorak. Za ova merenja su primenjene dve tehnike, *swept sine* tehnika [18] i tehnika prekinutog šuma. Kod prvo pomenute tehnike je kao pobuda korišćen *swept sine* signal dužine 30 s, frekvencijskog opsega od 20 Hz do 11 kHz koji je ponovljen 2 puta sa pauzom od 30 s. U ovom radu će biti prikazani samo rezultati dobijeni *swept sine* tehnikom.

Pobudni signal je reprodukovan neusmerenim zvučnim izvorom u obliku sfere koji sadrži 12 pobudnih jedinica u rasporedu dodekedra, videti Sl. 2. Odzivi su snimani mernim mikrofonom Bruel & Kjaer, type 4144. Merna oprema je još sadržala pojačavač audio signala, eksternu zvučnu karticu, mikrofonsko napajanje i laptop računar sa koga je vršena reprodukcija pobude i snimanje odziva.

Sl. 2. Postavka za merenja sa prikazom zvučnog izvora i mikrofona sa test uzorkom u reverberacionoj komori.

U uputstvima za *round robin* test je već definisano da se za merenja koristi 12 regularnih kombinacija položaja zvučnog izvora i mikrofona i dodatna 2 položaja mikrofona u uglovima komore. Dakle, ukupno se koristi 14 kombinacija položaja izvora i mikrofona za svaki test uzorak. Pri tome, izabrane su 4 pozicije za zvučni izvor i za svaki od njih po 3 pozicije za merni mikrofon koje su zadovoljile uslove definisane standardom ISO 354. Za jedan položaj zvučnog izvora, definisane su i 2 dodatne pozicije mikrofona u uglovima komore. Za svaku od pozicija izvora i mikrofona je definisana visina izvora, odnosno mikrofona.

Isti položaji izvora i mikrofona se koriste i za merenja u praznoj komori i u komori sa test uzorcima. Za svaki test uzorak uključujući praznu komoru i komoru sa uzorkom, setovi merenja se vrše sa dva ponavljanja. Sa svim neophodnim pripremama, merenja su trajala nešto manje od 3 nedelje.

Na osnovu izmerenih odziva se najpre izdvajaju impulsni odzivi prostorije. Ovi odzivi se zatim filtriraju u opsezima terci, a zatim se primenom *Schoreder*-ovog integraljenja unazad dobijaju krive opadanja. Dalje sledi obrada radi izračunavanja koeficijenata apsorpcije koja nije predmet ovog rada. Obrada je obavljena u programskom paketu Matlab. Ovde će biti prikazani samo rezultati koji se odnose na dobijene krive opadanja.

V. Analiza krivih opadanja

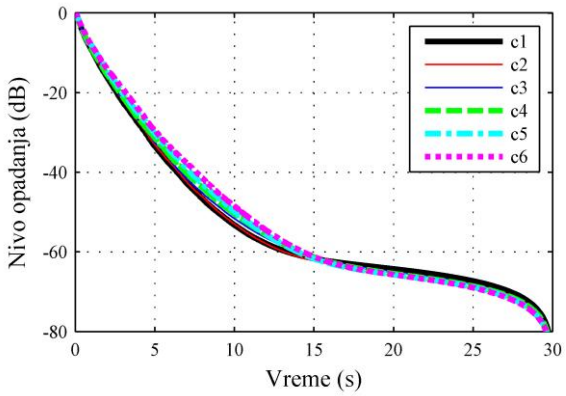
A. Uticaj difuzora

Najpre je anliziran uticaj difuzora na krive opadanja izmerene u praznoj reverberacionoj komori. Za tu namenu se porede krive opadanja dobijene sa i bez difuzora. Kao što je prethodno pomenuto, posebna pažnja se posvećuje krivama u opsezima terci na niskim i srednjim frekvencijama.

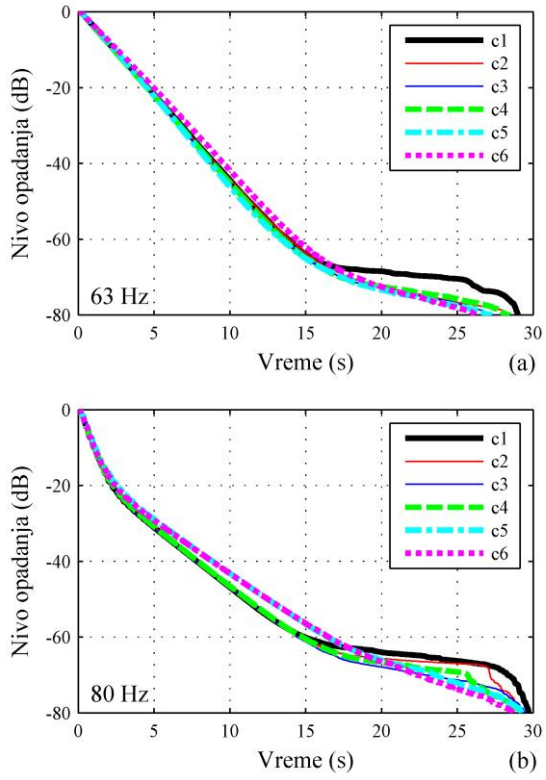
Krive opadanja dobijene *Schoreder*-ovim integraljenjem unazad širokopojasnih impulsnih odziva su prikazane na Sl. 3. Dato je 6 krivih opadanja koje su generisane na osnovu odziva izmerenih kada je u reverberacionoj komori bilo postavljeno svih 5 difuzora, kao i odziva izmerenih kada je u komori bilo 4 difuzora, zatim 3, 2 i 1 difuzor, kao i odziva izmerenih u praznoj komori bez difuzora. Uticaj difuzora je vidljiv kod prikazanih širokopojasnih krivih opadanja. Sve krive pokazuju određenu zakrivljenost gde je jasno vidljiv konkavni oblik. Strmine reverberacionog opadanja nisu potpuno identične kod ovih krivih. Prelaz od krive bez difuzora do krive sa svih 5 difuzora je postepen.

Krive opadanja dobijene na isti način kako je opisano za širokopojasne odzive ali za odzive filtrirane u opsezima terci na 63 Hz i 80 Hz su prikazane na Sl. 4(a) i (b), respektivno. Izabrana su dva tercna opsega koji ilustruju dve različite situacije prisutne na niskim frekvencijama. Sl 4(a) prikazuje slučaj kada je uticaj difuzora na krive opadanja manji, i gde gotovo da ne postoji zakrivljenost krivih opadanja, odnosno njihov konkavni oblik. Za razliku od toga, Sl. 4(b) prikazuje situaciju gde postoji određeni uticaj difuzora i gde je prisutno dvostruko opadanje kod krivih opadanja. Prvo opadanje ima veću striminu opadanja i manji dinamički opseg, dok drugo

opadanje ima manju strminu i veći dinamički opseg. Generalno se može reći da je uticaj difuzora najmanji na najnižim frekvencijama, i da se on povećava kako se povećava frekvencija. Pri tome, postoje tercni opsezi gde je uticaj difuzora veći, i oni gde je uticaj difuzora manji.

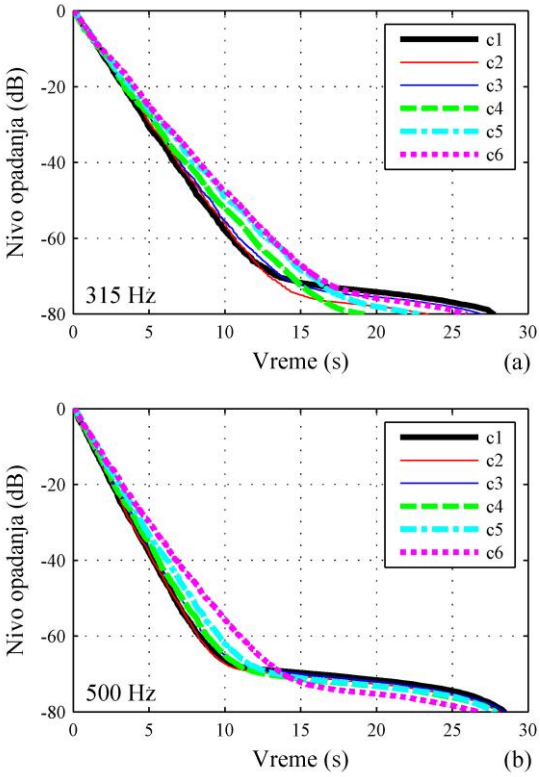


Sl. 3. Krive opadanja dobijene *Schroeder*-ovim integraljenjem unazad širokopojasnih impulsnih odziva izmerenih za jednu kombinaciju položaja zvučnog izvora i mikrofona: sa svih 5 difuzora (c1), sa 4 difuzora (c2), sa 3 difuzora (c3), sa 2 difuzora (c4), sa 1 difuzorom (c5) i bez difuzora (c6).



Sl. 4. Krive opadanja dobijene *Schroeder*-ovim integraljenjem unazad impulsnih odziva filtriranih u opsegu terci na 63 Hz (a) i na 80 Hz (b) izmerenih za jednu kombinaciju položaja zvučnog izvora i mikrofona: sa svih 5 difuzora (c1), sa 4 difuzora (c2), sa 3 difuzora (c3), sa 2 difuzora (c4), sa 1 difuzorom (c5) i bez difuzora (c6).

Može se reći da je uticaj difuzora najveći na srednjim frekvencijama, ili preciznije u tercnim opsezima na nekoliko stotina herca. Ilustracija ovakave situacije je prikazana na Sl. 5. U oba prikazana tercna opsega se može videti da se sa povećanjem broja difuzora prisutnih u reverberacionoj komori smanjuje zakrivljenost krivih opadanja. Drugim rečima rečeno, smanjuje se prisustvo dvostrukog opadanja tako što se strmina drugog (dela) opadanja povećava i ona postaje gotovo identična kao strmina prvog (dela) opadanja. Najveća razlika između krivih opadanja je ona koja se javlja u komori bez difuzora i u komori sa 3, 4 ili 5 difuzora. Pri tome, krive opadanja dobijene za prisustvo 3, 4 i 5 difuzora su veoma slične jedna drugoj.

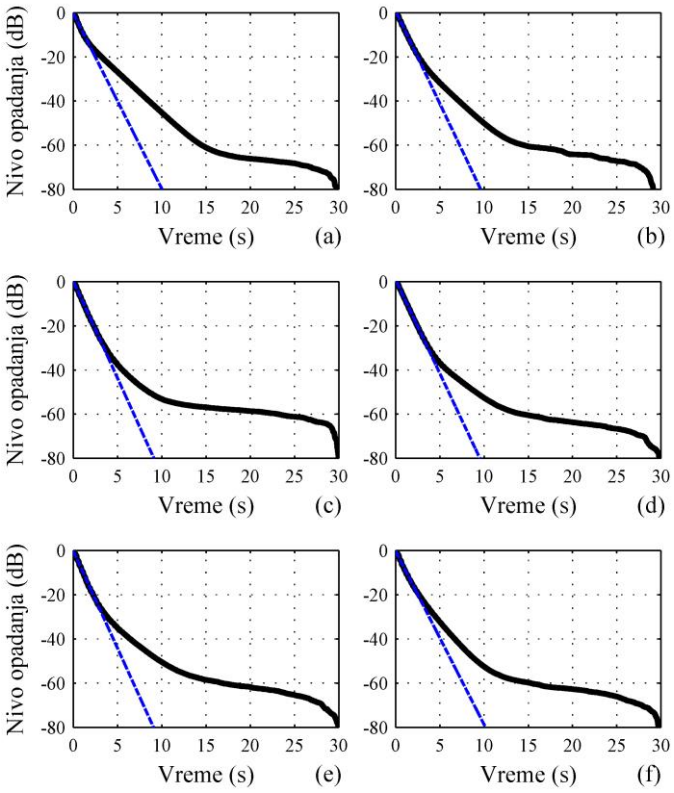


Sl. 5. Krive opadanja dobijene *Schroeder*-ovim integraljenjem unazad impulsnih odziva filtriranih u opsegu terci na 315 Hz (a) i na 500 Hz (b) izmerenih za jednu kombinaciju položaja zvučnog izvora i mikrofona: sa svih 5 difuzora (c1), sa 4 difuzora (c2), sa 3 difuzora (c3), sa 2 difuzora (c4), sa 1 difuzorom (c5) i bez difuzora (c6).

B. Krive opadanja prazne komore i sa test uzorkom

Rezultati prikazani u ovoj pod-sekciji se odnose na merenja sa svim difuzorima u komori. I ovde je posebna pažnja posvećena krivama opadanja na niskim i srednjim frekvencijama. Na niskim frekvencijama, postoje tercni opsezi gde krive opadanja imaju potpuno regularan (gotovo pravolinijski) oblik. Takvi slučajevi nisu analizirani ovde. Sa druge strane, postoje tercni opsezi gde krive opadanja pokazuju zakrivljenost, i karakteristični primeri su prikazani u nastavku. Jedan od njih se dobija u tercnom opsegu na 80 Hz kada se merenja vrše u praznoj reverberacionoj komori (bez test uzorka), Sl. 6. Na ovoj frekvenciji, za većinu kombinacija položaja izvora i mikrofona se dobija određena (manja ili veća) zakrivljenost krivih opadanja. Najčešće je to oblik dvostrukog (eksponencijalnog) opadanja poput onog na Sl.

6(b) ili 6(e). U određenim slučajevima, ovo dvosktruko opadanje nije očigledno, već kriva opadanja u delu reverberacionog opadanja ima konkavan oblik.

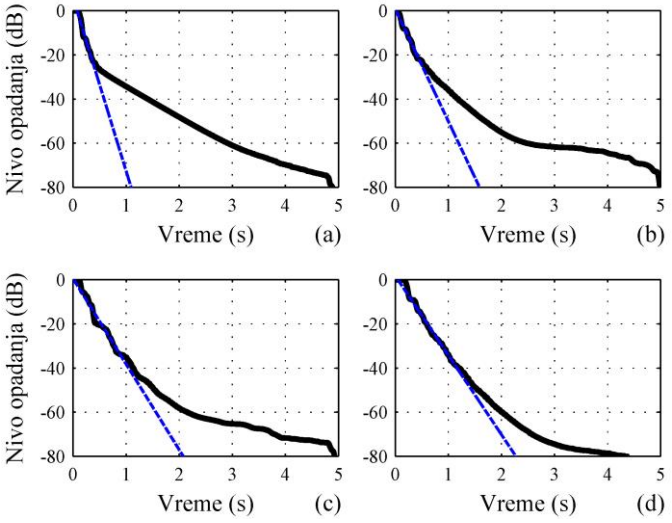


Sl. 6. Krive opadanja (—) dobijene *Schroeder*-ovim integraljenjem unazad 6 impulsnih odziva reverberacione komore bez test uzorka filtriranih u opsegu terce na 80 Hz za različite kombinacije položaja zvučnog izvora i mikrofona, kao i aproksimacione prave (- - -) dobijene linearnom regresijom početnog dela opadanja.

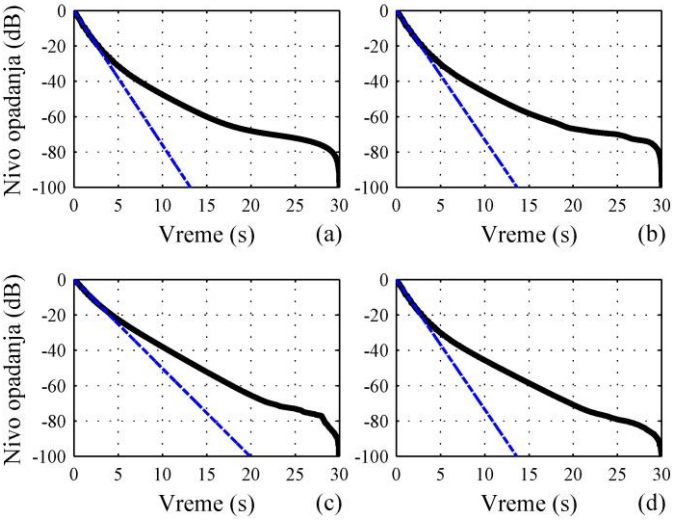
Situacija u ovom tercnom opsegu je nešto drugačija kada se bilo koji od merenih test uzoraka postavi u reverberacionu komoru. Krive opadanja tada su uglavnom regularne, imaju pravolinijski oblik. U jednom manjem broju slučajeva, javlja se oblik krive gde najpre dolazi do naglog opadanja nivoa zvuka, a zatim se uspostavlja relativno stabilan režim konstantne strmine opadanja, videti Sl. 7.

Pored karakterističnog dvostrukog opadanja dobijenog u praznoj reverberacionoj komori (Sl. 6), u ovakvim mernim uslovima se dobijaju i krive opadanja koje imaju dosta veliku zakrivljenost koja se ogleda u konkavnom obliku krive, ali bez vidljivog dvostrukog opadanja. Ovakva situacija se najčešće javlja u tercnom opsegu na 125 Hz, kao što je prikazano na Sl. 8. I u ovom tercnom opsegu postoje kombinacije položaja zvučnog izvora i mikrofona koje daju dvostruko opadanje, ali jedan manji broj kombinacija koji rezultira regularnim krivama opadanja.

Vredi istaći da se zakrivljenosti krivih opadanja prazne reverberacione komore ulgavnom javljaju do frekvencije 160 Hz. Međutim, u situacijama kada se merenja obavljaju sa test uzorkom u komori, onda se na srednjim frekvencijama (od 315 Hz do 630 Hz ili evenutalno do 800 Hz) može javiti blaga zakrivljenost krivih opadanja tako da kriva ima konstantno konkavan oblik u delu reverberacionog opadanja. Jaka zakrivljenost sa izraženim dvostrukim opadanjem na 400 Hz je prikazana na Sl. 9. Razlog zakrivljenosti na ovim frekvencijama je uglavnom neravnomerna raspodela apsorpcije. Naime, samo je pod u ovom slučaju apsorpciona površina, dok su sve ostale površine refleksione.



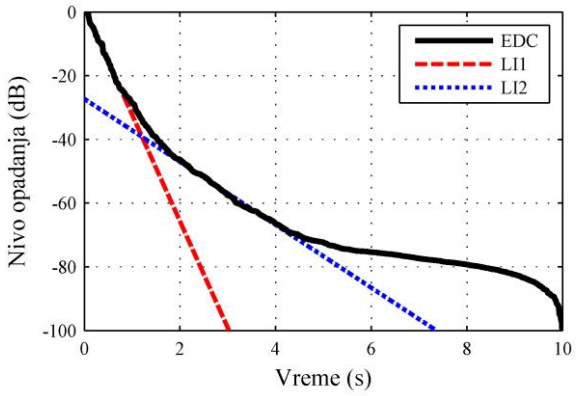
Sl. 7. Krive opadanja (—) dobijene *Schroeder*-ovim integraljenjem unazad 4 impulsna odziva reverberacione komore sa test uzorcima filtriranih u opsegu terce na 80 Hz za različite kombinacije položaja zvučnog izvora i mikrofona, kao i aproksimacione prave (- - -) dobijene linearnom regresijom početnog dela opadanja.



Sl. 8. Krive opadanja (—) dobijene *Schroeder*-ovim integraljenjem unazad 4 impulsna odziva reverberacione komore bez test uzorka filtriranih u opsegu terce na 125 Hz za različite kombinacije položaja zvučnog izvora i mikrofona, kao i aproksimacione prave (- - -) dobijene linearnom regresijom početnog dela opadanja.

Kada kriva opadanja pokazuje zakrivljenost poput one prikazane na Sl. 9, postavlja se pitanje na koji način izračunati vreme reverberacije. Da li je bolje koristiti početno opadanje (prikazano isprekidanom linijom na Sl. 9) ili kasno opadanje

(prikazano linijom sa tačkicama na Sl. 9).



Sl. 9. Kriva opadanja (EDC) dobijena *Schroeder*-ovim integraljenjem unazad impulsnog odziva reverberacione komore sa test uzorkom filtriranog u opsegu terce na 400 Hz, kao i aproksimaciona prava (LI1) dobijena linearnom regresijom početnog dela opadanja, kao i drugog dela opadanja (LI2).

VI. Zaključak

Krive opadanja zvuka dobijene na osnovu izmerenih impulsnih odziva prostorije mogu imati regularan pravolinijski oblik na logaritamskoj skali u decibelima. Međutim, odstupanja od ovakvog ponašanja se mogu javiti pre svega u manjim prostorijama. Ovakva odstupanja su karakteristična za niže i eventualno srednje frekvencije. Na niskim frekvencijama, dominantan uzrok mogu biti različite brzine opadanja zvuka aksijalnih, tangencijalnih i bočnih modova, dok dominantan uzrok na srednjim frekvencijama može biti neravnomerna raspodela apsorpcije. Kao posledica pomenutog ponašanja, javljaju se zakrivljenosti krivih opadanja, odnosno višestruka (eksponencijalna) opadanja.

Zakrivljenosti krivih opadanja na niskim i srednjim frekvencijama u reverberacionoj komori relativno malih dimenzija je analizirano u ovom radu. Pomenute neregularnosti (zakrivljenosti) se mogu delimično korigovati u reverberacionoj komori postavljanjem difuzora. Pri tome, efekat je izraženiji na srednjim nego na niskim frekvencijama.

I pored primene difuzora, u konkretnoj reverberacionoj komori se dobijaju krive opadanja sa dvostrukim opadanjem i konkavnim oblikom na niskim frekvencijama u praznoj reverberacionoj komori, kao i na srednjim frekvencijama kada se u komori nalazi test uzorak. Ostaje otvoreno pitanje na koji način vršiti izračunavanje vremena reverberacije kada krive opadanja pokazuju pomenute zakrivljenosti.

Zahvalnica

Ovo istraživanje je podržano od strane Ministarstva prosvete, nauke i tehnološkog razvoja Republike Srbije kroz projekat 36026.

Literatura

[1] E. J. Carlisle, R. J. Hooker, “Small chamber reverberant absorption measurement,” Proc. Acoustics 2004, Gold Coast, Austrlia, 3-5 November, 2004.

[2] M. M. Hasan, M. Hodgson, “Effectiveness of reverberation room design: Room size and shape and effect on measurement accuracy,” Proc. 22nd ICA (International Congress on Acoustics), Buenos Aires, Argentina, 5-9 September, 2016.

[3] M. R. Schroeder, “New method of measuring reverberation time,” *J. Acoust. Soc. Am*, vol. *37*, no. *3*, pp. *409-412*, 1965.

[4] *Acoustics - Measurement of Sound Absorption in a Reverberation Room*, ISO 354, 2003.

[5] N. Xiang, Y. Jing, A. C. Bockmann, “Investigation of acoustically coupled enclosures using a diffusion- equation model,” *J. Acoust. Soc. Am*, vol. *126*, no. *3*, pp. *1187-1198*, June 2009.

[6] N. Xiang, P. Coggans, T. Jasa, P. Robinson, “Bayesian characterization of multiple-slope energy decays in coupled-volume systems,” *J. Acoust. Soc. Am*, vol. *129*, no. *2*, pp. *741-752*, Feb 2011.

[7] H. Kuttruff, *Room Acoustics*, 6th ed. New York, USA: Spon Press, 2016.

[8] J. Balint, F. Muralter, M. Nolan, C-H Jeong, “Energy decay curves in reverberation chambers and the influence of scattering objects on the absorption coefficient of a sample,” Proc. Euronoise 2018, Crete, Greece, pp. 2025-2030, 27-31 Mai, 2018.

[9] M. Schroeder, “The “Schroeder frequency” revisited,” *J. Acoust. Soc. Am*, vol. *99*, no. *5*, pp. *3240-3241*, Mai 1996.

[10] M. Skålevik, “Schroeder frequency revisited,” Proc. Forum Acusticum 2011, Aalborg, Denmark, 27 June -1 July, 2011.

[11] K. Bodlund, “Monotonic curvature of low frequency decay records in reverberation chambers,” *J. Sound Vib*, vol. *73*, no. *1*, pp. *19-29*, Nov 1980.

[12] J. Rindel, “Modal energy analysis of nearly rectangular rooms at low frequencies,” *Acta Acust un Acust*, vol. *101*, no. *6*, pp. *1211-1221*, Nov-Dec. 2015.

[13] E. Nilsson, “Decay processes in rooms with non-diffuse sound fields Part I: Ceiling treatment with absorbing material,” *Building Acoustics*, vol. *11*, no. *1*, pp. *39-60*, March 2004.

[14] N. G. Kanev, “Reverberation in a trapezoidal room,” *Acoustical Physics*, vol. *59*, no. *5*, pp. *559-564*, 2013.

[15] H. Kuttruff, “Eigenschaften und Auswertung von Nachhallkurven,” *Acustica*, vol. *8*, pp. *273-280*, 1958

[16] F. V. Hunt, L. Beranek, D. Y. Maa, “Analysis of sound decay in rectangular rooms,” *J. Acoust. Soc. Am*, vol. *11*, pp. *80-94*, 1939.

[17] R. Ramakrishnan, A. Grewal, “Reverberation rooms and spatial uniformity,” *Proc Canadian Acoust*, vol. *36*, no. *3*, pp. *28-29*, 2008.

[18] S. Müller, P. Massarani “Transfer measurement with sweeps,” *J. Audio Eng. Soc*, vol. *49*, no. *6*, pp. *443-471*, June 2001.

Abstract

One of classical approaches for measurements of absorption coefficient of test sample is based on measurements of reverberation time of empty reverberation room and reverberation time of the same room when the test sample is in the room. Dimensions and important characteristics of a reverberation room are defined in the relevant standards. When certain conditions are fulfilled, especially when the reverberation room volume is smaller than recommended values, some disturbing effects of room modes can appear at lower frequencies. Influence of these effects in measurements of absorption coefficient of test samples on energy decay curves is analysed in the paper. Impulse response measurements are carried out in the reverberation room of volume of about 65 m^3. In third-octave bans below 160 Hz, specific multi-exponential (often double) sound decays yielding concave shape of decay curves are present. In such cases, there is a question how to estimate the reverberation time.

Energy Decay Curves Obtained in Reverberation Chamber During Measurements of Absorption Coefficient

Dejan Ćirić, Kristian Jambrošić, Nikola Stojković

Upotreba različitih obeležja za prepoznavanje drvenih duvačkih instrumenata korišćenjem neuralnih mreža

Tatjana Miljković, Miloš Bjelić, Dragana Šumarac Pavlović, Goran Kvaščev

***Apstrakt*— U procesima automatskog prepoznavanja instrumenata koriste se različita obeležja zasnovana na analizi spektralnog sadržaja audio zapisa. U ovom radu prikazan je hroma profil tonova kao novo obeležje, izdvojeno iz audio zapisa muzičkog sadržaja, na osnovu kojeg se vrši prepoznavanje duvačkih instrumenata pomoću neuralne mreže. Kako bi hroma profil predstavljao validno obeležje izvršena je komparativna analiza sa MFCC koeficijentima. Duvački instrumenti koji su bili od interesa u ovom istraživanju su: klarinet, flauta i oboa. Metodologija karakterizacije duvačkih instrumenata bazirana na hroma profilu tonova pokazala je očuvanje osnovnih razlika u karakteristikama instrumenata. Hroma profil predstavlja prikaz relativnih odnosa energije na pojedinim tonovima unutar oktave. Posmatrani su hroma profili tonova računati nad celim signalom i po prozorima signala. Isti principi računanja obeležja primenjena su i na MFCC koeficijente. Korišćenjem hroma profila tonova i MFCC koeficijenata kao ulaznih parametara neuralne mreže ostvareni su visoki procenti prepoznavanja instrumenata.**

***Ključne reči*— audio obeležja; hroma profil; MFCC; muzički instrumenti; neuralna mreža; prepoznavanje.**

I. Uvod

Digitalna revolucija u distribuciji muzike podstakla je ogroman interes za pronalaženje različitih načina na koje se informacione tehnologije mogu primeniti na ovakvu vrstu sadržaja. Stoga, sve veća količina audio i muzičkih signala ukazuje na potrebu za inteligentnim pretraživanjem, prikupljanjem i obradom muzičkog sadržaja pomoću automatizovanih metoda [1]. Takođe, razvitkom tehnologije javila se potreba za različitim multimedijalnim aplikacijama koje su imale za cilj da izvršavaju automatsku transkripciju nota, klasifikaciju muzičkih žanrova, identifikaciju pevača kao i prepoznavanje muzičkih instrumenata. Stečena znanja o osobinama i građi zvuka stvorila su osnov za dalja istraživanja audio zapisa muzičkog sadržaja, ali takođe i za određivanje atributa zvuka muzičkih instrumenata, pomoću savremenih softverskih alata. Kao reference pri istraživanju muzičkih signala najčešće se uzimaju u obzir nekoliko osnovnih parametara koji ih opisuju i to: visina tona, melodija, harmonija i ritam [2].

Iako postoje raznovrsni muzički sadržaji njihova osnova je ista i čine je tonovi. Ton predstavlja zvuk koji ima izraženu tonsku visinu. Spektar tona je diskretan i sastavljen od harmonijskog niza frekvencija koji se u muzičkoj literaturi naziva alikvotni niz [2]. Harmonijski niz frekvencija kao takav se prema osnovnom tonu odnosi kao 1:2:3:4:*n*, gde je *n* određena granična frekvencija. Iako je intervalski odnos harmonika prema osnovnom tonu uvek jednak, kod različitih muzičkih instrumenata razlikuje se relativna jačina pojedinih alikvota. Razlike u relativnoj jačini pojedinih alikvota formiraju anvelopu spektra tona koja predstavlja jedan od faktora kojim se definiše boja tona.

Proces percepcije tonske visine kod čoveka formira se na logaritamskoj frekvencijskoj osi. Proces koji se odvija u čulu sluha dovodi do preslikavanja harmonijskog niza frekvencija na logaritamsku frekvencijsku osu. Pri preslikavanju spektralnih komponenti tona na hroma krug pozicije nekih komponenti odgovaraju određenim tonovima unutar jedne oktave. Analizom pozicija harmonika određenog tona na hroma krugu nakon preslikavanja pokazano je da se ukupna energija tona dominantno nalazi na osnovnoj frekvenciji tona i na frekvencijama koje odgovaraju intervalima velike terce i čiste kvinte [3]. Raspodela energije jednog tona unutar hroma kruga predstavlja pogodan podatak za prepoznavanje različitih karakteristika kojima se mogu opisati muzički signali. Takođe, zbog izražene harmonijske strukture muzički ton se može jednostavno matematički modelovati. U literaturi je pokazano da takav model predstavlja jedan od glavnih ulaznih parametara algoritama za automatsko prepoznavanje atributa muzičkih instrumenata [4].

U cilju automatskog prepoznavanja muzičkih instrumenata razvijen je veliki broj algoritama. Iako je broj algoritama značajan, velika većina se zasniva na izdvajanju karakterističnih obeležja. Prvobitna istraživanja u pravcu izdvajanja obeležja muzičkih audio zapisa zasnivala su se na tehnikama izdvajanja govornih obeležja kao što su: LPC (*Linear Prediction Coefficients*) i MFCC (*Mel-Frequency Cepstral Coefficients*) koeficijenti [5]. Marques i Moreno su pomoću ovih obeležja uspeli da ostvare klasifikaciju od oko 70% muzičkih audio zapisa [6]. Eronen i Klapuri, razvili su sistem u kojem izdvajaju preko 20 različitih obeležja [7]. Koristili su kako spektralna tako i vremenska obeležja, među

Tatjana Miljković – Elektrotehnički fakultet, Univerzitet u Beogradu, Bulevar kralja Aleksandra 73, 11020 Beograd, Srbija (e-mail: tm@etf.rs).

Miloš Bjelić – Elektrotehnički fakultet, Univerzitet u Beogradu, Bulevar kralja Aleksandra 73, 11020 Beograd, Srbija (e-mail: bjelic@etf.rs).

Dragana Šumarac Pavlović – Elektrotehnički fakultet, Univerzitet u Beogradu, Bulevar kralja Aleksandra 73, 11020 Beograd, Srbija (e-mail: dsumarac@etf.rs).

Goran Kvaščev – Elektrotehnički fakultet, Univerzitet u Beogradu, Bulevar kralja Aleksandra 73, 11020 Beograd, Srbija (e-mail: kvascev@etf.rs).

kojima su različite varijacije MFCC koeficijenata, spektralni centroid, "udarna"(*attack*) i "opadajuća"(*decay*) faza tona i druge. Uprkos većem broju kako vremenskih tako i spektralnih obeležja rezultati klasifikacije instrumenata iznosili su oko 80% uspešnosti [8].

Kako bi se povećao procenat uspešnog prepoznavanja instrumenata izvršeno je istraživanje sa ciljem pronalaska novih obeležja muzičkih signala koji bi se koristili kao ulazni parametri veštačkih neuralnih mreža. U ovom radu prikazana je metodologija za analizu audio sadržaja muzičkih signala koja se naziva hromatogram. Kada se govori o hromatogramu kao alatu za analizu muzičkih signala podrazumeva se da je na osnovu hromatograma moguće izvršiti analizu raspodele energije u vremenu datog audio zapisa, kao i ustanovljavanje energetskog bilansa na očekivanim parcijalama analiziranih tonova. Na osnovu hromatograma muzičkih signala dobijeni su hroma profili datih signala, gde hroma profil predstavlja prikaz raspodele energije signala u vremenu. Analizom hroma profila pomenutih instrumenata uočeni su različiti bilansi enegrije za različite instrumente. Na osnovu uočenih razlika u hroma profilima instrumenata pokazano je da hroma profil predstavlja adekvatno obeležje jednog instrumenta, kao i povoljan ulazni parametar veštačke neuralne mreže. Kako bi se verifikovala tvrdnja da hroma profil predstavlja adekvatno obeležje jednog instrumenta, u ovom radu izvršena je komparativna analiza prepoznavanja drvenih duvačkih instrumenata pomoću neuralnih mreža upotrebom kako hroma profila, tako i MFCC koeficijenata. Cilj ovog rada je da pokaže da hroma profil predstavlja adekvatno obeležje, poput MFCC koeficijenata, na osnovu kog se može izvršiti prepoznavanje drvenih duvačkih instrumenata pomoću veštačke neuralne mreže.

Rad je organizovan kako sledi. U drugom poglavlju prikazani su hromatogram, hroma profil i kepstralna obeležja kao metodologije za analizu audio signala. U narednom poglavlju prikazani su eksperimentalni rezultati i diskusija dobijenih rezultata. Na kraju dat je zaključak o mogućnosti korišćenja hroma profila i MFCC koeficijenata kao ulaznih parametara neuralnih mreža na osnovu kojih se dobijaju visoki procenti prepoznavanja drvenih duvačkih instrumenata.

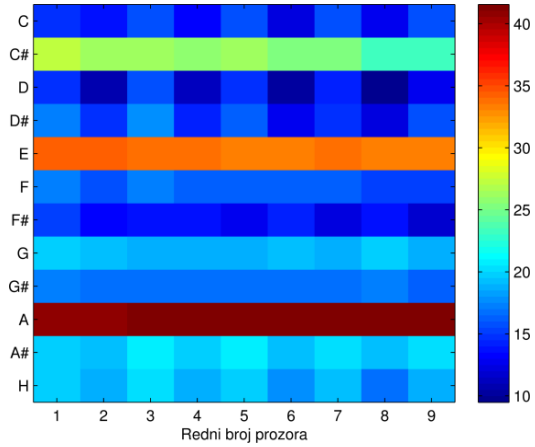
II. METODOLOGIJA

A. Hromatogram i Hroma profil

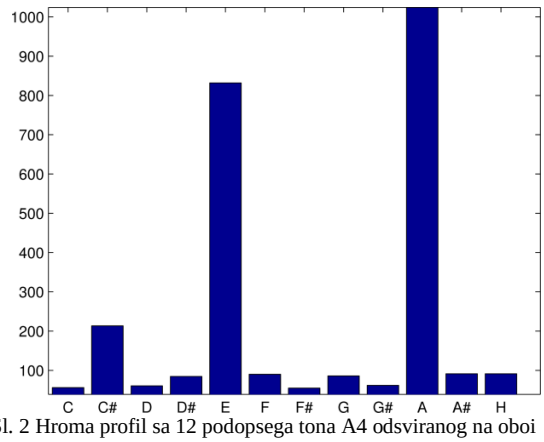
Pri automatskom prepoznavanju sadržaja audio signala koriste se različita obeležja zasnovana na analizi spektralnog sadržaja. Jedan od alata na osnovu kojeg se izdvajaju obeležja spektralnog sadržaja signala je hromatogram. Hromatogram predstavlja vremensko-frekvencijski prikaz signala gde je kompletan spektar signala transponovan na jednu izabranu oktavu koja je izdeljena na proizvoljan broj podopsega.

Izračunavanje hroma karakteristika jednog muzičkog signala se svodi na množenje hroma matrice i spektrograma datog signala, čime se dobija nova matrica koja se naziva hromatogram [8]. Spektrogram signala predstavlja matricu $\boldsymbol{S}$ dimenzija NxM dobijenu pomoću kratkovremene Furijeove transformacije (*STFT*). Redovi matrice $\boldsymbol{S}$ odgovaraju svakoj od N frekvencija i imaju indekse k, $k\epsilon[0, N\text{-}1]$. Broj kolona matrice $\boldsymbol{S}$ odgovara broju prozora u vremenu M. Hroma matrica $\boldsymbol{C}$ je dimenzija KxN, gde K predstavlja broj opsega u okviru jedne oktave [9]. Vrednost parametra K je proizvoljna, ali se uglavnom uzima vrednost 12, gde 12 predstavlja broj polutonova koji čine temperovanu skalu. Na Slici 1 prikazan je izgled hromatograma sa 12 podopsega za ton A4 odsviran na oboi.

Pored hromatograma, na osnovu kog se može posmatrati raspodela energije signala unutar oktave, postoji još jedan alat za analizu raspodele energije i naziva se hroma profil. Hroma profil predstavlja jedan vid hromatograma. Sumiranjem hromatograma za sve vremenske prozore audio signala dobija se hroma profil datog audio signala. Računanjem hroma profila na ovaj način ne dolazi do značajnijeg gubitka informacija o raspodeli energije posmatranog signala. Potreba za pojavom hroma profila audio signala javila se usled želje za preglednijim prikazom relativnih odnosa energije na pojedinim tonovima unutar oktave. Na Slici 2 prikazan je hroma profil sa 12 podopsega tona A4 odsviranog na oboi.



Sl. 1. Izgled hromatograma za ton A4 odsviran na oboi sa 12 podopsega



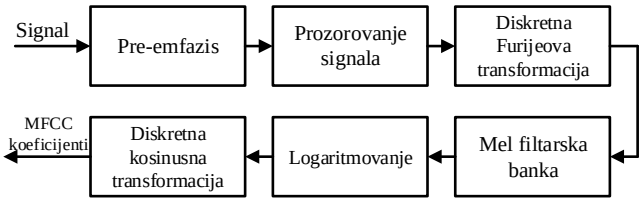
Sl. 2 Hroma profil sa 12 podopsega tona A4 odsviranog na oboi

B. Kepstralna obeležja

U domenu audio signala, kepstralna obeležja su prvobitno korišćena za analizu govora. Danas se kepstralna obeležja primenjuju u svim poljima pretraživanja audio signala kao što su govor, muzika i analiza zvukova okoline [10-12]. Kepstralna obeležja predstavljaju frekvencijski prikaz logaritmovanog amplitudskog spektra signala. Takođe,

omogućavaju primenu euklidske metrike, kao meru distanci zbog njihove ortogonalne prirode koja olakšava poređenje na osnovu sličnosti [13].

Na Slici 3 prikazana je blok šema postupka za određivanje MFCC (*Mel-frequency Cepstral Coefficients*) koeficijenata. Pre izdvajanja segmenata signala, odnosno prozorovanja, vrši se korekcija signala upotrebom pre-emfazisa. Zatim se vrši prozorovanje signala nekim od prozora, kao što su Hanov i Hammingov. Dužina prozora, odnosno trajanje segmenta signala koji se analizira zavisi od konkretne aplikacije i obično iznosi do 100 ms. Nakon toga se za svaki prozor izračunava DFT. Spektar signala u svakom prozoru se propušta kroz mel filtarsku banku koja se sastoji od filtara trouganog oblika [14]. Broj filtara zavisi od broja MFCC koeficijenata koji se želi odrediti. Zatim se na signale na izlazu iz svakog filtra primenjuje logaritmovanje i izračunavanje DFT. Na taj način se za svaki prozor dobija niz koeficijenata. Broj koeficijenata zavisi od tipa signala i vrste analize, ali se uglavnom uzima prvih 12 koeficijenata.



Sl. 3. Blok šema procedure za određivanje kepstralnih koeficijenata

C. Neuralna mreža

Veštačke neuralne mreže (*Artificial Neural Networks*) predstavljaju grupu računarskih modela. Izvorno, neuralne mreže inspirisane su biološkim strukturama, posebno strukturom mozga zbog njegove sposobnosti da rešava kompleksne probleme [15]. Veštačke neuralne mreže dobile su ime zahvaljujući njihovoj strukturi koju čini pregršt procesorskih jedinica, koje su poznate kao neuroni organizovani po slojevima. Svaki od slojeva povezan je sa prethodnim i sledećim slojem i na taj način formira se mreža sačinjena od povezanih neurona. Postoje tri tipa slojeva neuralne mreže: ulazni, skriveni i izlazni. Ulazni sloj predstavlja vezu sa polaznim parametrima neuralne mreže. Skriveni slojevi su zaduženi za transformaciju ulaznih podataka u prihvatljive informacije za izlazni sloj. U zavisnosti od broja skrivenih slojeva može se meriti sofisticiranost informacija koje su predviđene za izlazni sloj [16].

III. EKSPERIMENTALNI REZULTATI I DISKUSIJA

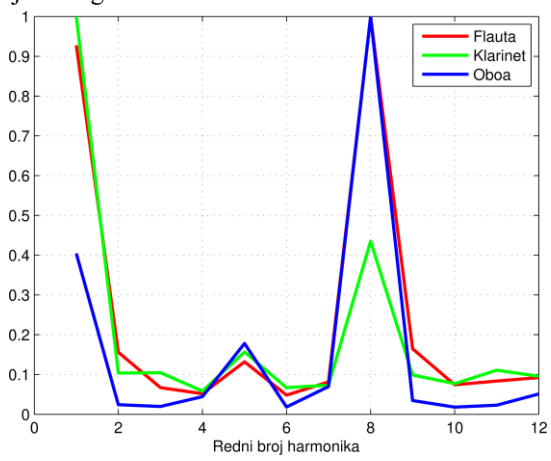
A. Baza snimaka

Problemi koji se rešavaju pomoću veštačkih neuralnih mreža zahtevaju podatke na kojima će se vršiti testiranje i verifikacija same mreže. U ovom radu korišćena je baza snimaka preuzeta sa internet stranice Britanske filharmonije [17]. Audio snimci koji čine bazu odsvirani su na: klarinetu, flauti i oboi. Snimke po sadržaju čine tonovi hromatske lestvice celokupnog registra posmatranih instrumenata. Bazu čini ukupno 2047 snimaka, od toga je 763 snimaka tonova flaute, 746 tonova klarineta i 538 tonova oboe. Snimci koji se nalaze u bazi su izolovani, tj. spadaju u monofone audio zapise, te tako predstavljaju idealan slučaj za prepoznavanje i izdvajanje obeležja audio zapisa muzičkog sadržaja. Takođe, odlikuju se različitim trajanjem: 1, 0.5 i 0.25 sekundi. Pored toga, audio zapisi su odsvirani u četiri različita dinamička manira: pianissimo, piano, mezzo-forte i forte.

B. Hroma profil i MFCC koeficijenti drvenih duvačkih instrumenata

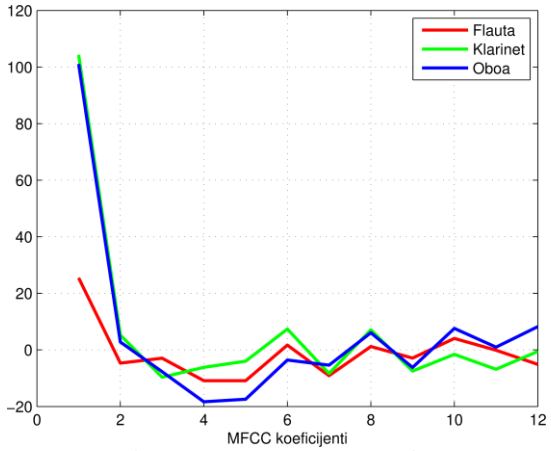
Kako se neuralnoj mreži prosleđuju ulazni parametri potrebno je izvršiti pretprocesiranje podataka, odnosno izdvojiti obeležja iz audio signala. U ovom radu kao obeležja drvenih duvačkih instrumenata od interesa izabrani su: MFCC koeficijenti i hroma profil.

Kako bi se izračunao hroma profil za pojedine tonove potrebno je prvo odrediti hromatogram datih tonova. Za izračunavanje spektrograma tonova instrumenata korišćen je Hammingov prozor dužine 2048 odbiraka sa 50% preklapanja. Zatim je množenjem spektrograma sa hroma matricom, spektrogram signala transponovan na oktavu izdeljenu na 12 podopsega i time je dobijen hromatogram. Sumiranjem hromatograma za sve vremenske prozore signala dobijen je hroma profil. Ovako opisan proračun podrazumeva da je sa 12 vrednosti koje odgovaraju hroma profilu opisan ceo signal, odnosno posmatrani ton. Te u tom slučaju ulazni parametri neuralne mreže predstavljaju 12 vrednosti hroma profila tona. Na Slici 4 prikazani su hroma profili sa 12 podopsega za ton A4 odsviran na flauti, klarinetu i oboi. Sa slike se može uočiti da za sva tri instrumenta postoji nagomilavanje energije na prvom harmoniku, koji odgovara osnovnom tonu i na petom i osmom harmoniku, koji odgovaraju intervalima velike terce i čiste kvinte. Takva raspodela energije se objašnjava prirodom tona, jer je spektar tona sačinjen od harmonijskog niza frekvencija. Takođe, na osnovu slike se može zaključiti da postoje vidljive razlike u raspodeli energije za posmatrana tri instrumenta, što predstavlja dobar preduslov da hroma profil predstavlja obeležje nekog instrumenta.



Sl. 4. Hroma profili sa 12 podopsega tona A4 odsviranog na flauti, klarinetu i oboi

Radi komparativne analize broj koeficijenata koji su korišćeni je 12. Za izračunavanje koeficijenata korišćen je Hammingov prozor sa 50% preklapanja. Vrednost pre-emfazis koeficijenta iznosi 0.97. Korišćen je frekvencijski opseg od 200 Hz do 20 kHz, kako bi se pokrile sve frekvencije koje posmatrani instrumenti mogu da proizvedu i 30 filtara u mel frekvencijskoj banci. Takođe, kao i u slučaju računanja hroma profila, ovaj slučaj proračuna podrazumeva da je ceo signal opisan samo sa prvih 12 MFCC koeficijenata. Na Slici 5 prikazano je 12 MFCC koeficijenata za ton A4 odsviran na flauti, klarinetu i oboi. Sa slike se može uočiti da postoje izvesne razlike u vrednostima MFCC koeficijenata za posmatrane instrumente. Varijacije u vrednostima MFCC koeficijenata ukazuju da se na osnovu njih mogu okarakterisati instrumenti.

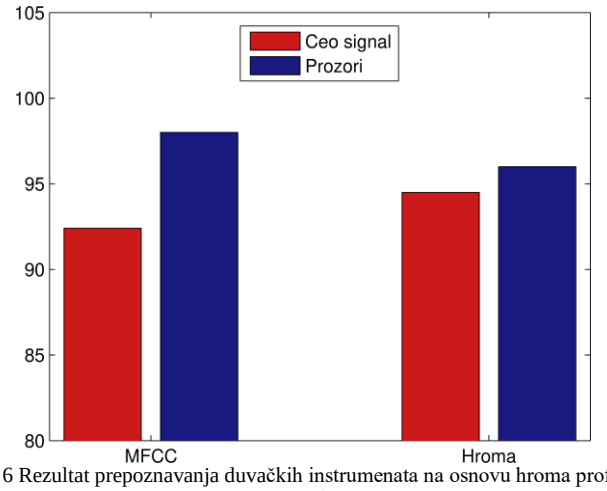


Sl. 5. 12 MFCC koeficijenata za ton A4 odsviran na flauti, klarinetu i oboi

Kada se govori o obeležjima koja predstavljaju signal uglavnom se ona izračunavaju na manjim delovima signala, odnosno prozorima. Kako bi se uočio uticaj različitog broja obeležja jednog signala, izvršeno je računanje kako hroma profila tako i MFCC koeficijenata za svaki prozor posmatranog signala. Da bi se uspešno sprovela komparativna analiza hroma profila i MFCC koeficijenata kao obeležja signala, potrebno je izvršiti prozorovanje signala ili na prozore jednake dužine trajanja ili na jednaki broj prozora. Kako bazu snimaka čine snimci različite dužine trajanja nije bilo moguće prozorovati signale tako da svi prozori budu jednake dužine trajanja. Zbog toga je u daljem toku analize uzeto u obzir prozorovanje signala na jednak broj prozora. Svaki signal iz baze snimaka podeljen je na 19 prozora, te se za svaki prozor računa po 12 vrednosti koje predstavljaju hroma profil datog segmenta i 12 MFCC koeficijenata. Što dalje implicira, da se jedan signal opisuje sa ukupno 228 vrednosti u oba slučaja. Te tada ulaz neuralne mreže čini 228 vrednosti koje predstavljaju obeležja celog signala računata po prozorima.

C. *Rezultati prepoznavanja*

Kao klasifikator instrumenata korišćena je MLP (*Multilayer Perceptron*) *feed forward* neuralna mreža sa propagacijom unazad. Struktura mreže se sastoji od ulaznog sloja, koji je povezan sa početnim parametrima neuralne mreže, odnosno sa hroma profilima instrumenata i MFCC koeficijentima. U zavisnosti od toga da li su kao ulazni parametri korišćeni hroma profili tonova ili MFCC koeficijenti računati nad celim signalnom ili po prozorima, broj ulaznih parametara neuralne mreže bio je 12, odnosno 228. Broj neurona koji je korišćen u skrivenom sloju je 40. Izlaz neuralne mreže predstavlja sloj od tri neurona. Takođe, neuralna mreža je potpuno povezana, odnosno svaki neuron je povezan sa svakim neuronom iz sledećeg sloja bilo direktno ili indirektno. Skup podataka koji se koristi za prepoznavanje podeljen je na podskupove za trening, validaciju i test. Podskup za trening čini 70% skupa podataka, dok su podskupovi za validaciju i test jednaki i iznose 15% skupa podataka. U procesu treniranja neuralne mreže korišćen je skalirani konjugovani gradijentni algoritam (*Scaled Conjugate Gradient Algorithm*). Kao aktivaciona funkcija neurona korišćena je tangent sigmoid funkcija (*Tansig function*) [18]. Pri treniranju mreže nije korišćena regularizacija.

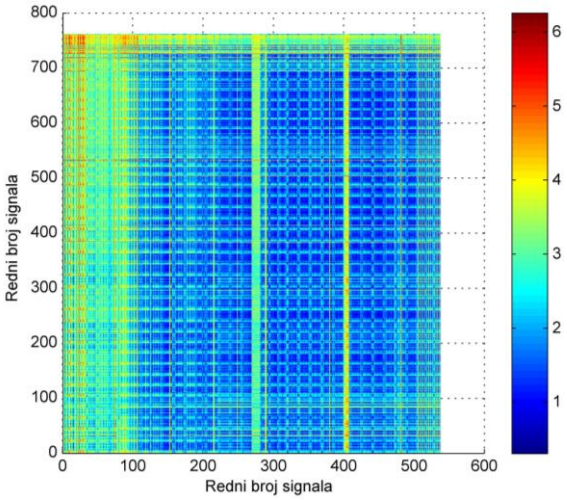


Sl. 6 Rezultat prepoznavanja duvačkih instrumenata na osnovu hroma profila i MFCC koeficijenata

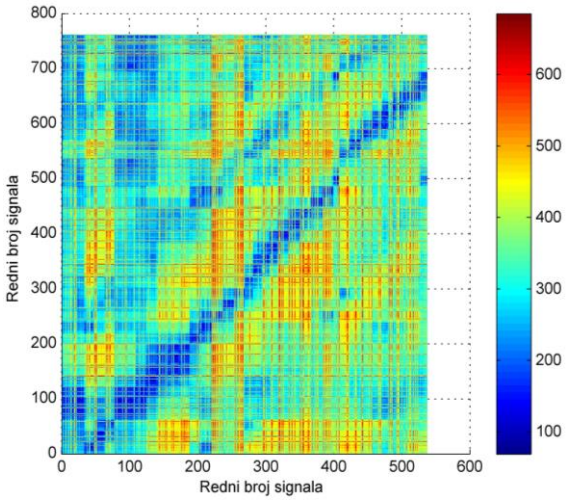
Na Slici 6 prikazan je rezultat prepoznavanja duvačkih instrumenata na osnovu hroma profila i MFCC koeficijenata koji su predstavljali ulazne parametre neuralne mreže. Pri prepoznavanju uzeta su u obzir dva slučaja i to: kada je ceo signal predstavljen sa 12 koeficijenata i kada su izračunati koeficijenti za svaki prozor signala. Kada su kao ulazni parametri neuralne mreže uzeti MFCC koeficijenti, tačnije njih 12, procenat uspešnog prepoznavanja instrumenta iznosi 92%. Dok za slučaj kada su ulazni parametri neuralne mreže predstavljali hroma profili tonova sa 12 podopsega procenat uspešnog prepoznavanja iznosio je 95%. Ako su obeležja signala računata po prozorima, odnosno ukoliko je broj ulaznih parametara neuralne mreže bio 228, procenat prepoznavanja u slučaju hroma profila iznosi 96%. Ukoliko je ulaz neuralne mreže bio niz od 228 vrednosti, koje predstavljaju MFCC koeficijente računate za svaki prozor signala, procenat prepoznavanja instrumenata iznosi 98%.

Kako bi se utvrdio razlog za ostvarene veće procente uspešnog prepoznavanja instrumenata za slučaj kada su

korišćena obeležja računata po prozorima signala formirane su matrice sličnosti. Prikazane su matrice sličnosti, gde se jedna dimenzija matrice odnosi na ukupan broj snimaka flaute, a druga na ukupan broj snimaka oboe. Pri računanju matrice snimci su sortirani onim redosledom koji zauzimaju i u hromatskoj lestvici. Takođe, sortiranje snimaka je izvršeno i prema oktavi kojoj snimci pripadaju i to u rastućem poretku, od treće do šeste oktave. Na Slici 7 prikazana je matrica sličnosti flauta-oboa, koja je računata na osnovu hroma profila tonova flaute i oboe. Na osnovu slike se može zaključiti da postoje izvesne razlike medju tonovima instrumenata flaute i oboe, naročito u nižim oktavama. Upravo zbog postojećih razlika ostvareni su visoki procenti prepoznavanja neuralne mreže. Na Slici 8 prikazana je matrica sličnosti flauta-oboa, koja je računata na osnovu MFCC koeficijenata tonova flaute i oboe. Takođe, sa slike se uočava da postoje razlike među tonovima posmatranih instrumenata. Činjenica da su razlike prikazane matricom sličnosti za slučaj MFCC koeficijenata za dva reda veličine veće u odnosu na razlike za slučaj hroma profila, potvrđuje rezultat gde su ostvareni veći procenti prepoznavanja za slučaj kada su ulazni parametri neuralne mreže bili MFCC koeficijenti signala računati po prozorima.



Sl. 7. Matrica sličnosti flauta-oboa hroma



Sl. 8. Matrica sličnosti flauta-oboa MFCC

Iako je procenat prepoznavanja instrumenata za slučaj kada su ulazni parametri predstavljali obeležja računata po prozorima signala veći u odnosu kada su ulaz činila obeležja koja su karakterisala ceo signal, takvi parametri zahtevaju više vremena za obradu. Stoga, u pogledu vremenske kompleksnosti neuralna mreža kod koje su ulazni parametri predstavljali 12 koeficijenata ima prednost za korišćenje u aplikacijama koje rade u realnom vremenu. Time se postiže ušteda u vremenu na štetu nešto nižem ostvarenom procentu uspešnog prepoznavanja instrumenata.

IV. ZAKLJUČAK

U ovom radu prikazana je upotreba različitih obeležja za prepoznavanje drvenih duvačkih instrumenata korišćenjem neuralnih mreža. Obeležja koja su uzeta u razmatranje su: hroma profil i MFCC koeficijenti. Pomenuta dva obeležja su uzeta u razmatranje, kako bi se hroma profil prikazao kao novo obeležje audio zapisa, čija bi se verifikacija utvrdila poređenjem sa u literaturi poznatim MFCC koeficijentima. Duvački instrumenti koji su prepoznavani su: klarinet, flauta i oboa. Ovi instrumenti odabrani su tako da pokrivaju širok opseg varijacija u pogledu načina pobuđivanja, oblika cevi i strukture rezonantnih modova. Prikazane razlike između instrumenata pružaju dobru osnovu za upotrebu hroma profila kao obeležja na osnovu kojih neuralna mreža vrši prepoznavanje instrumenta. Ostvareni su visoki procenti prepoznavanja instrumenata za slučaj kada su kao ulazni parametri neuralne mreže uzeti hroma profili i MFCC koeficijenti, računati kako za ceo signal tako i po prozorima signala. Iako prikazana neuralna mreža ostvaruje bolji učinak u prepoznavanju instrumenata za slučaj kada su ulazni parametri predstavljali obeležja računata po prozorima signala, uočeno je da takvi parametri zahtevaju više vremena za njihovu obradu. Te u smislu vremenske kompleksnosti rada neuralne mreže, kao i pretprocesiranja samih signala, povoljnije je kao ulazne parametre korsititi obeležja računata za ceo signal, odnosno uzeti u obzir samo 12 vrednosti. U tom slučaju procenti prepoznavanja instrumenata na osnovu hroma profila iznosi 95%, dok na osnovu MFCC koeficijenata prepoznato je 92% instrumenata. Cilj rada bio je ispitati mogućnost upotrebe hroma profila tonova duvačkih instrumenata kao obeležja audio zapisa muzičkog sadržaja na osnovu kog je moguće izvršiti prepoznavanje instrumenta. Pokazano je kako kroz ostvarene visoke procente prepoznavanja duvačkih instrumenata, tako i kroz komparativnu analizu sa već verifikovanim obeležjem da je hroma profil tonova adekvatno obeležje koje se može koristiti u daljim istraživanjima. Tako bi se u nekim od budućih istraživanja uzeli u obzir hroma profili ne samo monofonih nego i polifonih audio zapisa.

LITERATURA

[1] Lerch, Alexander. *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics*. s.1.: John Wiley & Sons, 2012. 1118393503.

[2] R. N. Shepard, “Circularity in judgements of relative pitch”, J. Acoust. Soc. Amer., vol. 36, pp. 2346–2353, 1964.
[3] Meinard Mller, “Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications”, 1st edition, Springer Publishing Company, Incorporated, 2015, ISBN:3319219448.
[4] D. L. Ishwar, K. Sethi, N. Dimitrova, T. McGeec, “Classification of general audio data for content-based retrieval”, Pattern Recognition Letters, vol. 22(5), pp. 533-544, 2001.
[5] J. Marques, P.J. Moreno, “A study of musical instrument classification usinggaussian mixture models and support vector machines”, Cambridge Research LaboratoryTechnical Report Series CRL,4, 1999.
[6] J. Marques, P.J. Moreno, “A study of musical instrument classification usinggaussian mixture models and support vector machines”, Cambridge Research LaboratoryTechnical Report Series CRL,4, 1999.
[7] A. Eronen, A. Klapuri, , “Musical instrument recognition using cepstral coefficientsand temporal features”, InICASSP'00, Vol. 2, pp. II753-II756, 2000.
[8] N. Segal, “Automatic Musical Instrument Recognition Using Convolutional Neural Networks”, MSc in Digital Signal Processing Project Report, 2016.
[9] M. Slavković-Ivić, D. Šumarac Pavlović, „Analiza muzičkih tonova pomoću hromatograma”, ETRAN, Zlatibor, jun 2016, Zbornik radova, ISBN: 978-86-7466-618-0.
[10] P. Gaunard, G.C. Mubikangiey, C. Couvreur, V. Fontaine, “Automatic classification of environmental noise events byhidden Markov models”, Appl Acoust, 54(3), 187–206, 1998
[11] M. Cowling, R. Sitte, “Comparison of techniques for environmental sound recognition”, Pattern Recognit Lett, 24(15), 2895–907, 2003.
[12] Y. Zhang, T. Ogata, S. Nishide, T. Takahashi, H.G. Okuno, “Classification of known and unknown environmental sounds based on self-organized space using a recurrent neural network”, Adv Rob, 25(17), 2127-41, 2011.
[13] Dalibor Mitrović, Matthias Zeppelzauer, Christian Breiteneder, “Features for Content-Based Audio Retrieval”, Advances in Computers vol.78, ISSN: 0065-2458.
[14] P.N. Garner, “Cepstral normalisation and the signal tonoise ratio spectrum in automatic speech recognition”, Speech Commun, 53(8), 991–1001, 2011.
[15] I.A. Basheer, M. Hajmeer, “Artificial neural networks: fundamentals, computing, design, and application”, Journal od microbiological methods, 43(1), pp.3-31, 2000.
[16] S. Haykin, “Neural Network A Competetive Foundation”, 2th ed.Prentice Hall International, 1990.
[17] http://www.philharmonia.co.uk/explore/sound_samples/clarinet
[18] S. Masood, S. Gupta and S. Khan, "Novel approach for musical instrument identification using neural network,", Annual IEEE India Conference (INDICON), New Delhi, pp. 1-5, doi: 10.1109/INDICON.2015.7443497, 2015.

ABSTRACT

In the processes of automatic recognition of the content of audio signals, various features based on the analysis of spectral content are used. In this paper are presented the chroma tones profile as a new feature separated from music content on the basis of which recognition of wind instruments by neural network is performed. In order to represent chroma profile as valid feature comparative analysis with MFCC coefficients has been done. Wind instruments that were of interest to this research are: clarinet, flute and oboe. The methodology for the characterization of wind instruments based on chroma tone profiles has shown the preservation of the basic differences in the characteristics of the instruments. The chroma profile represents relative energy ratios on individual tones within octave. Chroma profile of tones were been observed, which were calculated over complete signal and per windows of signal. The same principles of calculating the features were applied to the MFCC coefficients. Using chroma profile of tones and MFCC coefficients as the input parameters of the neural network, high percentages of instrument recognition were achieved.

Use of different features for recognizing wooden wind instruments using neural networks

Tatjana Miljković, Miloš Bjelić, Dragana Šumarac Pavlović, Goran Kvaščev

Akustički prenos podataka baziran na OFDM tehnici

Tatjana Miljković, Miloš Bjelić, Miljko Erić

***Apstrakt*— Prenos podataka korišćenjem audio signala (*data over sound*) u poslednjih nekoliko godina se koristi sve više. Iako koncept nije nov, njegov rast podstaknut je velikim brojem IoT uređaja i naprednih tehnika za obradu signala. Ova tehnologija može biti alternativa drugim tehnologijama za prenos podataka na malom rastojanju. Prednosti akustičkog (zvučnog) prenosa podataka u odnosu na radio tehnologije je mogućnost upotrebe na mestima na kojima je prenos radio talasa onemogućen. Prednost je i to što većina uređaja poseduje zvučnik i mikrofon i ne zahteva se upotreba dodatnih hardverskih komponenti. Ograničenje akustičkog prenosa podataka može biti manji protok u odnosu na druge tehnologije. U ovom radu prikazana je upotreba akustičkog signala za prenos podataka u vazduhu u čujnom delu audio opsega. Akustički prenos podataka realizovan je korišćenjem OFDM tehnike prenosa akustičkog signala. Mogućnost prenosa je ilustrovana rezultatima simulacije i realnih eksperimenata, bez tehnika zaštitnog kodovanja. Uvedena je tehnika za kompenzaciju vremenskog kašnjenja za necelobrojne vrednosti u broju odbiraka. Za uvedene parametre predajnika ostvaren je bitski protok od 3.2 Kbit/s. Dobijeni rezultati pokazali su mogućnost upotrebe ovakvog sistema za prenos informacija na malim rastojanjima u realnim uslovima.**

***Ključne reči*— akustički prenos, akustički OFDM, OFDM, prenos podataka, QAM.**

I. Uvod

Prenos podataka bežičnim putem koristi se jedan vek i kroz istoriju je doživeo velike promene. Današnji sistemi za bežični prenos podataka koristeći radio tehnologije mogu postići velike protoke u gotovo svakoj tački na planeti. Personalni uređaji koji se svakodnevno koriste poseduju mogućnosti za prenos podataka na velikom rastojanju korišćenjem javnih mobilnih mreža. Međutim, u poslednjih nekoliko decenija trend je da uređaji međusobno komuniciraju na manjim rastojanjima, formirajući lokalne mreže, uz vrlo malu potrošnju energije. Takvi uređaji predstavljaju osnovu za Internet of Things - IoT platforme. Na malim rastojanjima između predajnika i prijemnika mogu se koristiti i akustički signali za prenos informacija [1]. Prenos podataka korišćenjem akustičkog (zvučnog) signala kao nosioca poznat je već dugo u literaturi [2]. Pri akustičkom prenosu podataka zbog malog frekvencijskog opsega uređaja za reprodukciju i prijem signala ostvaruju se značajno manji protoci podataka u odnosu na radio prenos (reda nekoliko kb/s). Akustički prenos podataka se tradicionalno koristi za podvodne komunikacije [3]. U ovoj komunikaciji uslovi propagacije su vrlo nepovoljni za akustički signal pa se koriste razne vrste ekvalizacije kanala [4, 5]. Zbog ograničenog frekvencijskog opsega akustičkog kanala upotrebljavaju se tehnike za efikasnije korišćenje spektra [6]. Tokom prethodnih godina u ovu svrhu dominantno je korišćena OFDM (*Orthogonal Frequency-Division Multiplex*) tehnika [7]. OFDM predstavlja poznatu tehniku prenosa sa više podnosilaca [8]. Najveća prednost OFDM sistema u odnosu na sisteme sa jednim nosiocem je mogućnost za prilagođavanje lošim uslovima za prenos kroz kanal, kao što su slabljenje, uskopojasne smetnje i frekvencijski selektivni feding [9]. Upotrebom MIMO (*Multiple Input Multiple Output*) sistema u akustičkoj komunikaciji može se ostvariti dodatno povećanje protoka [10, 11]. Korišćenjem ovih tehnika uz OFDM ostvareni su protoci od par desetina Kbit/s, uz relativno malu verovatnoću greške.

U ovom radu analizirana je mogućnost upotrebe akustičkog signala za prenos podataka u vazduhu. Za prenos podataka korišćena je OFDM tehnika, odnosno akustički ekvivalent OFDM signala. Akustički prenos podataka realizovan je u čujnom audio opsegu (do 4.096 kHz). Kao podaci za prenos korišćena je *grayscale* slika. Realizovano je nekoliko simulacija i eksperimenata sa ciljem da se ispita kvalitet prenosa podataka kroz akustički kanal bez korišćenja tehnika zaštitnog kodovanja i bez korišćenja cikličnog prefiksa (CP). Pre slanja akustičkog signala u kom su kodovani podaci vršeno je slanje preambule sa sekvencom koju karakteriše ravan spektar. Preambula je korišćena na prijemnoj strani za određivanje početka korisnog signala metodom kroskorelacije. Preambula je korišćena i za ekvalizaciju kanala u eksperimentima sa realnim prenosom. U radu je prikazana i tehnika pomoću koje se vrši kompenzacija kašnjenja primljenog signala koje u opštem slučaju ne mora biti jednako celom broju odbiraka. Metoda je verifikovana u simulacijama akustičkog prenosa između zvučnika i mikrofonskog niza. U realizovanim eksperimentima zvučnik i mikrofon nalazili su se na rastojanju oko 2 metra u anehoičnim uslovima, što u literaturi predstavlja relativno veliko rastojanje između predajnika i prijemnika. Primenom ekvalizacije kanala postignut je relativno dobar kvalitet rekonstruisane slike. Ovaj rad predstavlja uvodno istraživanje čiji je cilj bio ispitivanje mogućnosti realizacije sistema za prenos podataka kroz akustički kanal u realnim uslovima u vazduhu.

Rad je organizovan kako sledi. U drugom poglavlju prikazan je postupak generisanja akustičkog OFDM signala. U trećem poglavlju prikazana je postavka realizovanih eksperimenata i simulacija. U četvrtom poglavlju prikazani su najvažniji rezultati u ovom istraživanju, kao i njihova

Tatjana Miljković – Elektrotehnički fakultet, Univerzitet u Beogradu, Bulevar kralja Aleksandra 73, 11020 Beograd, Srbija (e-mail: tm@etf.rs).

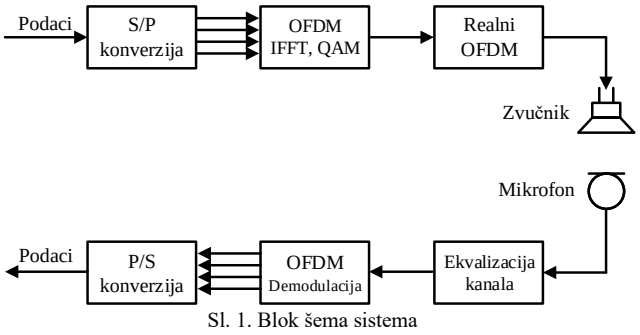
Miloš Bjelić – Elektrotehnički fakultet, Univerzitet u Beogradu, Bulevar kralja Aleksandra 73, 11020 Beograd, Srbija (e-mail: bjelic@etf.rs).

Miljko Erić – Elektrotehnički fakultet, Univerzitet u Beogradu, Bulevar kralja Aleksandra 73, 11020 Beograd, Srbija (e-mail: miljko.eric@etf.rs).

diskusija. U petom poglavlju izneti su zaključci.

II. METODOLOGIJA

U ovom poglavlju prikazan je način generisanja akustičkog OFDM signala na osnovu korisničkih podataka koje je potrebno preneti. Prikazani su parametri sistema na predajnoj strani kao i način kvantifikacije kvaliteta prenosa podataka. Na Slici 1 prikazana je principijelna blok šema sistema za prenos podataka korišćenjem akustičkog signala.



Sl. 1. Blok šema sistema

Kao informacija koja se prenosi pomoću prikazanog sistema u ovom radu korišćena je fotografija (slika). Slika je odabrana zbog toga što se nakon prenosa i demodulacije akustičkog signala na prijemnoj strani pored objektivne mere može i vizuelno izvršiti procena kvaliteta prenosa informacije.

Slika koju je potrebno preneti akustičkim signalom mapirana je sa 512 OFDM podnosioca. Polovina podnosilaca, odnosno 256 podnosioca, predstavljaju realan deo spektra, a ostalih 256 podnosioca predstavljaju konjugovano kompleksne parove realnog dela spektra. Odabrano je da frekvencija odabiranja iznosi 8192 Hz, pa se na širini spektra od 4096 Hz nalazi 256 OFDM podnosilaca. Širina svakog podkanala iznosi 16 Hz. Prvih 32 podnosioca (koji se nalaze u spektru na frekvencijama od 0 do 512 Hz), kao i poslednjih 24 podnosioca (koji se nalaze u spektru na frekvencijama od 3712 do 4096 Hz) se ne koriste. Prvih 32 podnosilaca se ne koriste zbog mogućih ograničenja zvučnika za reprodukciju niskih frekvencija. Na 20-om podnosiocu, što odgovara frekvenciji od 320 Hz, nalazi se prostoperiodični signal. Ovaj signal će se u daljim istraživanjima koristiti za rešenje problema razdešenosti takta na prijemnoj i predajnoj strani. Preostalih 200 podnosioca korišćeni su za mapiranje slike. Mapiranje slike je izvršeno 256 QAM modulacijom podnosilaca. OFDM signal je kompleksan i kao takav ne može se direktno reprodukovati preko zvučnika. Zbog toga je potrebno oformiti realni OFDM signal. Signalu sa prethodno generisanim podnosiocima potrebno je odrediti simetrične konjugovano kompleksne parove, a zatim izvršiti inverznu Furijeovu transformaciju. Dobijeni signal je realan i moguće ga je reprodukovati direktno preko zvučnika.

Zbog slabljenja prilikom prenosa akustičkog signala, na prijemnoj strani pre demodulacije primljenih OFDM simbola potrebno je izvršiti ekvalizaciju kanala. Zbog toga se pre slanja OFDM simbola koji prenosi informaciju šalje preambula od 3 OFDM simbola sa poznatom sekvencom sa ravnim amplitudskim spektrom [12], ukupne dužine 1536 odbiraka. Preambula se koristi i za vremensku sinhronizaciju, odnosno za detekciju početka OFDM simbola sa rezolucijom od 1 perioda samplovanja signala.

Kako bi se kvantifikovao kvalitet prenosa podataka pomoću akustičkog signala potrebno je uvesti objektivnu meru sličnosti poslate i primljene informacije. Pošto je slika informacija koja se prenosi kao mera razlike korišćena je kumulativna vrednost greške koja se računa na sledeći način:

$$greška = \sum_{i=1}^{m}\sum_{j=1}^{n}\left|I_{Tx,i,j} - I_{Rx,i,j}\right|, \quad (1)$$

gde su I_{Tx} i I_{Rx} poslati i rekonstruisani pikseli slike, respektivno. Broj vrsta piksela u slici iznosi *m*, a broj kolona piksela slike iznosi *n*. Na osnovu prethodnog izraza zaključuje se da kumulativna funkcija greške predstavlja sumu razlike svih piksela i da nema jedinicu.

III. POSTAVKA EKSPERIMENATA I SIMULACIJA

U ovom poglavlju biće prikazana postavka simulacije akustičkog prenosa, kao i eksperimenata koji su realizovani. Na Slici 2 prikazana je *grayscale* fotografija koja je korišćena kao informacija koja se prenosi akustičkim signalom u realizovanim simulacijama i eksperimentima. Broj bita za kodovanje vrednosti piksela je 8, pa se vrednosti piksela nalaze se u opsegu 0 do 255.

Sl. 2. Slika koja se koristi za prenos

Kao polazna tačka u ovom istraživanju korišćene su simulacije akustičkog prenosa informacije između dve tačke u prostoru. Na osnovu slike koju je potrebno preneti generisan je realni OFDM signal koji se reprodukuje sa zvučnika. Simulacija prostiranja akustičkih talasa zasniva se na množenju svake spektralne komponente signala sa odgovarajućim *steering* vektorom.

Takođe, važno je definisati temperaturu ambijenta zbog zavisnosti brzine prostiranja zvuka od temperature. Kao prijemna tačka u prostoru uzeta je pozicija jednog od mikrofona iz mikrofonskog niza [13]. Izabrano je da talasni front nailazi na mikrofonski niz pod uglom određenim azimutnim uglom 5° i elevacionim uglom 90°.

Nakon simulacije izvršeni su i realni eksperimenti. Eksperimenti su realizovani u Laboratoriji za Akustiku Elektrotehničkog fakulteta u Beogradu. Na Slici 3 prikazana je postavka eksperimenta. Kao prijemnik akustičkog signala korišćen je jedan od mikrofona iz mikrofonskog niza, kao i u simulacijama. Za akviziciju signala sa mikrofona korišćen je *Pulse* sistem [14]. Akustički OFDM signal reprodukovan je sa monitorskog zvučnika [15]. Rastojanje između

mikrofona i zvučnika iznosilo je oko 2 metra. Početak signala na prijemu detektovan je korišćenjem autokorelacije preambule koja se šalje pre OFDM signala.

Sl. 3. Eksperimentalna postavka

IV. REZULTATI I DISKUSIJA

U ovom poglavlju prikazani su rezultati simulacija i eksperimenata prenosa informacija korišćenjem akustičkog OFDM signala. Parametri simulacije odabrani su tako da odgovaraju eksperimentalnoj postavci prikazanoj na Slici 3. Prikazani rezultati predstavljaju izgled prenete fotografije nakon demodulacije i obrade signala na prijemnoj strani.

A. Rezultati dobijeni na osnovu simulacija

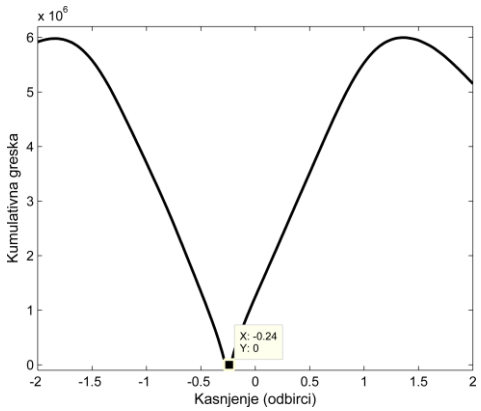
Množenjem spektralnih komponenti generisanog akustičkog OFDM signala sa *steering* vektorom dobijeni su signali na mikrofonskom nizu. Za potrebe ovog rada koristi se primljeni signal sa jednog mikrofona. Korišćenjem kroskorelacije određen je početak korisnog signala sa tačnošću od jednog vremenskog odbirka, a zatim izvršen postupak demodulacije primljenog OFDM signala i rekonstrukcija slike. Na Slici 4 prikazana je rekonstruisana slika dobijena na osnovu simulacije akustičkog prenosa.

Sl. 4. Primljena slika

Poredeći rezultat sa Slike 4 sa originalnom fotografijom (prikazanoj na Slici 2) evidentno je da postoje razlike. Na osnovu Slike 4 se prepoznaje sadržaj prenete fotografije, ali je oštrina fotografije narušena. Razlike se najlakše uočavaju na svetlijim delovima slike. Vrednost kumulativne greške, izračunate na osnovu izraza (1), iznosi 1245514. Prosečna vrednost razlika piksela poslate i primljene slike iznosi 10. Prosečna vrednost razlika piksela poslate i primljene slike računa se kao količnik kumulativne greške i ukupnog broja piksela.

U realizovanoj simulaciji nije modelovan realan akustički kanal, već samo kašnjenje usled prostiranja, pa je za očekivati da je rekonstruisana fotografija identična kao poslata. Korišćenjem kroskorelacije kao metode za detekciju početka korisnog dela signala može se postići vremenska tačnost do nivoa jednog odbirka. U realizovanoj simulaciji zbog izabranih pozicija zvučnika i mikrofona i zadate frekvencije odabiranja može se dogoditi da vremenski trenutak u kom počinje korisni signal ne odgovara celobrojnom broju odbiraka. Zbog toga se može desiti da je narušena fazna karakteristika primljenog signala. OFDM prenos je osetljiv na promenu faznog stava, pa je pretpostavljeno da se zbog toga javlja razlika između primljene i poslate slike.

Kako bi se ispitala navedena pretpostavka izvršena je izmena u simulaciji akustičkog prenosa. Nakon detekcije početnog vremenskog trenutka korisnog signala metodom kroskorelacije, izvršeno je dodatno vremensko kašnjenje. Pošto je potrebno zakasniti signal za necelobrojnu vrednost odbiraka τ, kašnjenje je izvršeno u frekvencijskom domenu. Za zadatu poziciju mikrofona u prostoru potrebno je odrediti vrednost τ koja će dovesti do idealne rekonstrukcije prenete fotografije. Vrednost τ određena je iterativnim postupkom tako što su pretpostavljane vrednosti vremenskog kašnjenja u opsegu [-2/fs 2/fs], a nakon toga izvršeno kašnjenje signala i demodulacija tako dobijenog OFDM signala. Za svaku vrednost pretpostavljenog vremenskog kašnjenja računata je vrednost kumulativne greške. Korak „pretrage“ iznosio je $f_s/100$. Nakon izračunavanja vrednosti greške za sve vrednosti vremenskog kašnjenja dobijena je zavisnost prikazana na Slici 5.



Sl. 5. Zavisnost vrednosti greške od vrednosti vremenskog kašnjenja

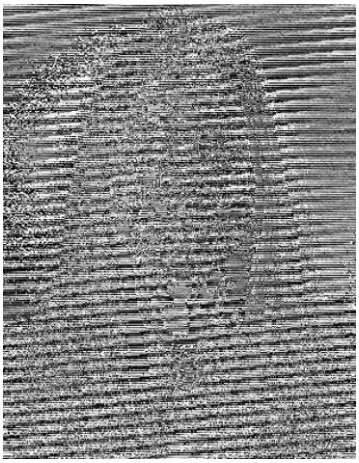
Sl. 6. Rekonstruisana fotografija sa unetim kašnjenjenjem

Na Slici 5 se vidi da je minimalna vrednost greške postignuta za vrednost vremenskog kašnjenja τ=0.24 odbirka. Vrednost greške u ovom slučaju iznosi 0, što znači da je rekonstruisana slika identična kao poslata. U simulacijama nije dodavan šum. Na Slici 6 prikazana je rekonstruisana slika za pronađenu vrednost necelobrojnog vremenskog kašnjenja u odbircima.

Dodatni postupak vremenskog kašnjenja ponovljen je i za ostale mikrofone iz mikrofonskog niza. Za svaki mikrofon je pronađena optimalna vrednost dodatnog kašnjenja i u svim slučajevima dobijena je idealno rekonstruisana fotografija. Na osnovu prethodno prikazanih rezultata zaključuje se da je za idealnu rekonstrukciju signala potrebno izvršiti dodatno vremensko kašnjenje, kako bi se očuvao početni fazni stav primljenog akustičkog OFDM signala. Ova procedura primenjena je u realizovanim eksperimentima, prikazanim u nastavku ovog poglavlja.

B. Rezultati dobijeni na osnovu eksperimenata

Na Slici 7 prikazana je fotografija dobijena demodulacijom akustičkog OFDM signala sa jednog mikrofona iz mikrofonskog niza u eksperimentu čija je eksperimentalna postavka prikazana na Slici 3. Na osnovu rekonstruisane slike ne može se zaključiti šta je njen sadržaj. Vrednost kumulativne greške iznosi 4831274. Prosečna razlika po pikselu između poslate i primljene slike iznosi 40. Akustički kanal kroz koji je vršen prenos nije idealan, odnosno postoji odstupanje od ravne amplitudske i linearne fazne karakteristike. Da bi se uspešno rekonstruisala informacija preneta kroz akustički kanal potrebno je izvršiti ekvalizaciju kanala. Ekvalizacija kanala izvršena je korišćenjem preambule koja se šalje pre generisanog OFDM signala. Preambula se sastoji od tri OFDM simbola sa poznatom sekvencom koje imaju ravan amplitudski spektar.



Sl. 7. Rekonstruisana fotografija bez korišćenja ekvalizacije kanala

Nakon primenjene ekvalizacije kanala izvršeno je određivanje početka korisnog dela signala, vremensko necelobrojno kašnjenje i demodulacija signala. Na Slici 8 prikazana je rekonstruisana slika. Sadržaj slike se za razliku od situacije bez ekvalizacije sada može prepoznati. Postoje izvesne razlike u odnosu na originalnu sliku koje se pre svega odnose na oštrinu slike. Vrednost kumulativne greške u ovom slučaju iznosi 1300189. Prosečna vrednost greške po pikselu je 10.

Sl. 8. Rekonstruisana fotografija sa korišćenom ekvalizacijom kanala

Fotografija preneta kroz realan akustički kanal korišćenjem OFDM signala u ovom eksperimentu nije identična kao poslata fotografija. Eksperiment je realizovan u anehoičnoj prostoriji pa je bilo za očekivati idealnu rekonstrukciju poslate fotografije. Jedan od razloga koji je mogao uticati da ne dođe do idealne rekonstrukcije je razlika u radnim frekvencijama odabiranja hardverskih sistema na predaji i prijemu. Svođenje na istu frekvenciju odabiranja izvršeno je decimacijom [16]. Početni rezultati dobijeni u ovom radu mogu biti unapređeni pre svega korišćenjem sistema koji rade sa istim taktom. Dalja unapređenja sistema za akustički prenos podataka baziran na OFDM tehnici mogu biti i uvođenje zaštitnih vremenskih intervala i cikličnog prefiksa kod OFDM simbola. Takođe, uvođenjem tehnika zaštitnog kodovanja signala generisanog na predajnoj strani može dovesti do smanjivanja kumulativne vrenosti greške. Bitski protok ostvaren pri ovakvom prenosu iznosi 3.2 kbit/s. Vrednost protoka bi se značajno uvećala korišćenjem veće frekvencije odabiranja, odnosno veće širine kanala za prenos. U ovom radu širina kanala iznosila je 4096 Hz. Dodatno povećanje protoka moglo bi se postići i povećavanjem reda QAM modulacije koja se korisiti za generisanje OFDM signala.

V. ZAKLJUČAK

U ovom radu prikazan je postupak akustičkog prenosa podataka korišćenjem OFDM tehnike. Prikazan je način generisanja akustičkog OFDM signala koji se može reprodukovati u akustičkom domenu. Realizovani eksperimenti i simulacije pokazali su da je moguće izvršiti prenos podataka korišćenjem ove tehnike, pri čemu je ostvaren protok od 3.2 kbit/s. Simulacijom akustičkog prenosa pokazano je da je neophodno izvršiti kompenzaciju vremenskog kašnjenja signala na prijemu, pre nego što se izrvši demodulacija OFDM signala. Vrednost kašnjenja ne mora biti jednaka celom broju odbiraka, pa je zbog toga kašnjenje signala izvršeno u frekvencijskom domenu. Vrednost kašnjenja određena je iterativnim postupkom. Pokazano je da se nakon ovog postupka preneti podaci mogu idealno rekonstruisati. Postupak određivanja tačne vrednosti kašnjenja potrebno je odrediti samo jednom za trenutnu poziciju mikrofona u prostoru. To znači da bi se pre slanja korisnih podataka mogla poslati referentna slika, koju bi unapred znao prijemnik, kako bi se odredila vrednost kašnjenja potrebna za idealnu rekonsturukciju. Realizovani

eksperimenti pokazali su neophodnost upotrebe ekvalizacije kanala kako bi se podaci mogli uspešno rekonstruisati. Rekonstruisani podaci na prijemnoj strani nisu identični kao poslati, ali pokazuju da je ovakav način prenosa moguć. Da bi se ovakav prenos mogao koristiti u realnim aplikacijama potrebno je koristiti još neka predprocesiranja signala. Korišćenjem cikličnog prefiksa i zaštitnog intervala kod OFDM simbola, kao i tehnika zaštitnog kodovanja mogla bi se smanjiti verovatnoća greške u prenosu. Buduća istraživanja ići će u pravcu uvođenja navedenih predobrada ali i varijacije širine kanala za prenos i reda QAM modulacije, sa ciljem da se ostvari što veći protok i da se smanji verovatnoća greške pri prenosu.

ZAHVALNICA

Ovaj rad je realizovan u okviru projekta TR 36026 koga finansira Ministarstvo prosvete, nauke i tehnološkog razvoja Republike Srbije. Autori se zahvaljuju profesorki Jeleni Ćertić na korisnim savetima i sugestijama.

LITERATURA

[1] K. Marneweck, J. Nesfield, A. Mehrabi, D. Jones, „ Why data-over-sound is an integral part of any IoT engineer's toolbox: Chirp + Arm = frictionless low power connectivity", White Paper, 2019.

[2] D. Wax, „MFSK-The Basis for Robust Acoustical Communications", OCEANS 81, Boston, MA, 1981, pp. 61-66. doi: 10.1109/OCEANS.1981.1151680.

[3] X. Han, J. Yin, G. Yu, P. Du, „Experimental demonstration of single carrier underwater acoustic communication using a vector sensor", Applied Acoustics, Volume 98, 2015, Pages 1-5, ISSN 0003-682X, https://doi.org/10.1016/j.apacoust.2015.03.019.

[4] Y. Xiao, F. Yin, „Blind equalization based on RLS algorithm using adaptive forgetting factor for underwater acoustic channel", China Ocean Eng 2014;28(3):401–8.

[5] J. Nelson, A. Singer, U. Madhow, C. McGahey, "BAD: Bidirectional arbitrated decision-feedback equalization," IEEE Trans. Commun. 53, 214–218 (2005).

[6] H. Esmaiel, D. Jiang, „Review Article: Multicarrier Communication for Underwater Acoustic Channel", Int. J. Communications, Network and System Sciences, 2013. 6: p. 361-376.

[7] C. Polprasert, A. Ritcey James, M. Stojanovic, „Capacity of OFDM systems over fading underwater acoustic channels", J Ocean Eng, 2011;36(4):514–24.

[8] B. Le Floch, M. Alard and C. Berrou, "Coded orthogonal frequency division multiplex [TV broadcasting]," in Proceedings of the IEEE, vol. 83, no. 6, pp. 982-996, June 1995., doi: 10.1109/5.387096.

[9] S. B. Weinstein, "The history of orthogonal frequency-division multiplexing [History of Communications]," in IEEE Communications Magazine, vol. 47, no. 11, pp. 26-35, November 2009., doi: 10.1109/MCOM.2009.5307460.

[10] J. V. Candy, A. J. Poggio, D. H. Chambers, B. L. Guidry, C. L. Robbins, and C. A. Kent, "Multichannel time-reversal processing for acoustic communications in a highly reverberant environment," J. Acoust. Soc. Am. 118, 2339–2354, 2005.

[11] G. Qiao, Z. Babar, L. Ma, S. Liu, J. Wu, „MIMO-OFDM underwater acoustic communication systems—A review", Physical Communication, Volume 23, 2017, Pages 56-64, ISSN 1874-4907, https://doi.org/10.1016/j.phycom.2017.02.007.

[12] N Suehiro, M Hatory, Modulatable orthogonal sequences and their application to SSMA systems. IEEE Trans. Inf. Theory. 34(1), 93–100 (1988)

[13] Tehnička dokumentacija proizvođača, dostupno na: http://www.bksv.com/products/transducers/acoustic/acoustical-arrays, pristupano 5.5.2019.

[14] Tehnička dokumentacija proizvođača, dostupno na: http://www.bksv.com/Products/frontends/lanxi.aspx, pristupano 5.5.2019.

[15] Tehnička dokumentacija proizvođača, dostupno na: http://www2.jblpro.com/catalog/General/Product.aspx?PId=24&MId=7 , pristupano 5.5.2019.

[16] Lj. Milić, „Multirate Filtering for Digital Signal Processing", New York: Hershey. p. 192. ISBN 978-1-60566-178-0, 2009.

ABSTRACT

Over the past few years data transfer using audio is more used. While the concept is not new, its current growth is being fuelled by the rise of smartphones, IoT devices and improved signal-processing technology. This technology can be an alternative to other data transmission. Advantages of acoustic (sound) data transmission in relation to radio technologies is the possibility of use in places where radio transmission is disabled. The advantage is that most devices have a speaker and a microphone and no additional hardware components are required. The limitation of acoustic data transmission may be a smaller flow than other technologies. This paper presents the use of an acoustic signal for data transmission in the air in the audio band. Acoustic data transmission is realized using an acoustic OFDM signal. The transmission was realized in simulations and real experiments without encoding techniques. A time delay compensation technique has been introduced that can take non-integer values in the number of samples. For the introduced transmission parameters, a bit rate of 3.2 kbit/s has been achieved. The results obtained showed the possibility of using such a system for the information transmission at small distances in realistic terms.

Data transmission based on acoustic OFDM technique

Tatjana Miljković, Miloš Bjelić, Miljko Erić

Surface Plasmon Polariton-like Propagation Induced by Structural Dispersion of Substrate Integrated Waveguide and Its Application in Microwave Circuits and Sensing

Vesna Crnojević-Bengin, Norbert Čeljuska, Žarko Šakotić, Mihailo Drljača, Goran Kitić, Vasa Radonić, Nikolina Janković
BioSense Institute, University of Novi Sad

***Abstract*— In this paper, we present microwave components and sensors based on surface plasmon polariton-like (SPP-like) propagation induced by structural dispersion of substrate integrated waveguide (SIW). We develop a theoretical framework upon which the novel microwave diplexer, dual-band filters, and sensor are designed and fabricated. Basic building block is a multilayer SIW, where layers represent sub-SIWs with different dielectric filling materials and/or different geometric parameters. The layers are designed to have effective permittivities of opposite signs in certain frequency ranges, which enables SPP-like propagation to occur at their interfaces. A detailed theoretical and numerical analysis together with numerical optimization has been performed to design the novel components, which were afterwards fabricated using different techniques. Excellent filtering operation has been confirmed with the designed diplexer and dual-band bandpass filters, while extreme sensitivity has been demonstrated with the proposed sensor. These results show the great potential of SPP-like phenomenon for design of microwave components and sensors.**

***Index Terms*— Dual-band bandpass filter, surface plasmon polariton (SPP), diplexer, sensor**

I. Introduction

The coupling between free electrons and photons at the interface of metal and dielectric medium enables a peculiar phenomenon – surface plasmon polaritons (SPP) [1] that is characterized by a strong field confinement, which has been widely applied in optical communications, imaging, and sensors [2-4]. Due to the advantageous properties of SPPs and the fact that SPPs naturally occur at optical frequencies, there have been proposed different concepts to engineer SPP phenomenon in the frequency ranges other than optical, including microwaves and terahertz. One of the concepts that effectively mimics SPPs are spoof or designer surface plasmon polaritons, which have been widely used in different microwave circuits including transmission lines, filters, antennas, couplers, splitters, absorbers, and circulators [5-21]. Unlike genuine SPPs, spoof SPPs are supported by specifically designed structures, predominantly by grooved strips whose geometrical properties are used to tailor the behavior of spoof SPPs.

Recently, a novel "natural" SPP-like concept in microwave regime has been proposed, which is based on the exploitation of the well-known structural dispersion of the electromagnetic modes in parallel-plate waveguide structure filled only with materials with positive permittivity [22]. Namely, if two waveguides with different cut-off frequencies are coupled, then the two exhibit effective permittivities of opposite signs in a certain frequency range, which opens up a possibility for SPP-like behavior to occur at their interface.

In this paper, we demonstrate how this promising concept can be employed to design novel circuits and sensors that operate at microwave frequencies [23-26]. Namely, it is well-known that the rapid growth of communication systems imposes the requirements for high-performance, low-cost, low-profile components that operate at two or more non-harmonically related frequencies. On the other hand, microwave sensors have attracted a considerable attention since they can provide non-invasive, label-free, as well as real-time detection and thus they have been utilized in various applications including dielectric constant sensing [27-36], food quality control, and concentration measurements of liquid solutions. Nevertheless, microwave sensors very often suffer from low resolution and sensitivity, which limits their applicability when a minute change has to be detected, which is the case in engine oil quality and fuel blends detection [37-41], or vegetable oils quality analysis and adulteration detection [41-43].

To address these requirements, we design SPP-like propagation using substrate integrated waveguide (SIW) that is comprised of several layers, i.e. sub-SIW structures, each having particular width and dielectric filling material, and consequently particular cut-off frequency. Since the layers are designed to have effective permittivities of opposite signs in certain frequency ranges, SPP-like propagation occurs at their interface. Besides slow-wave behavior and field confinement, SPP-like propagation also provides a transmission zero in the spectral response, and enables a clear separation of a passband and stopband in the spectrum paving a way towards good filtering operation. On the other hand, strong field enhancement at the interface and sensitivity of the transmission zero to the surrounding dielectric environment

represents a promising sensing scheme.

This concept is the underlying idea of the operating principle of the circuits and sensor that will be presented here – diplexer consisting of two SPP-based filters, three dual-band filters, and a dielectric constant sensor. In the following, we will present a theoretical background of the SPP-like phenomenon, its realization in SIW structure, as well as filter and sensor operation principles. The numerical analysis and results will be provided together with fabrication procedure and measurement results.

II. Theoretical Background

To present the underlying idea of the circuits and sensor, we first consider the basic building block - SIW consisting of two sub-SIW sections, whose layout and corresponding geometric and material parameters are shown in Fig. 1. In general case, the widths and dielectric filling materials are different for each sub-SIW.

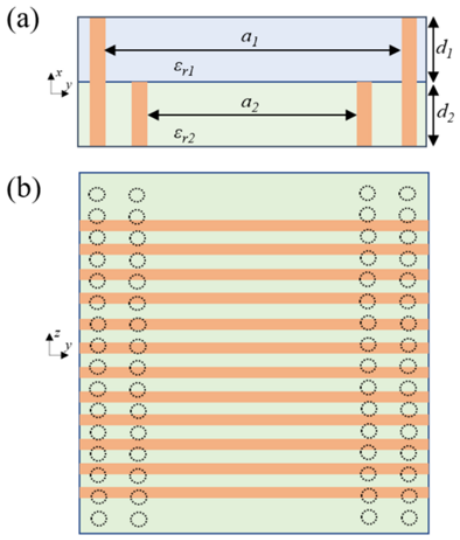


Fig. 1. (a) Cross section of a SIW composed of two sub-SIW sections, (b) top view of the interface of two sub-SIW sections.

Substrate integrated waveguide (SIW) represents a guiding structure whose dominant propagating mode is TE_{10}, whilst TM modes are not supported. The cut-off frequency of the TE_{10} mode is given as:

$$f_{cTE10} = \frac{c}{2a\sqrt{\varepsilon_r}}, \tag{1}$$

where ω is angular frequency, c is speed of light, a is SIW's width, and ε_r is the dielectric constant of the filling material. In other terms, we can say that SIW is characterized by the effective permittivity:

$$\varepsilon_e = \varepsilon_r - \left(\frac{c}{2af}\right)^2, \tag{2}$$

where f is frequency, and it can be seen that the effective permittivity is negative below and positive above the cut-off frequency.

SPP-like propagation can be achieved if two SIWs with different cut-off frequencies, i.e. with different effective permittivities in a certain frequency range, are coupled. This can be realized using the configuration shown in Fig. 1, so that the two SIWs represents sub-SIWs comprising the major SIW. Since each sub-SIW is bounded by the top and bottom metal plates, the interface metal plate is replaced by an array of wires, which provides coupling of the fields in the sub-SIWs, as well as accumulation of electric charges and consequently the condition that the normal components of the electric field at the interface are opposite to each other.

To further analyze the properties of the proposed structure, the dispersion relation is derived under approximation that a structure is a 2D partially-filled waveguide [25]:

$$\sqrt{\beta^2 + \frac{1}{4a_2^2} - \varepsilon_{r2}k_o^2}\left(\varepsilon_{r1} - \frac{1}{4a_1^2k_o^2}\right) \tanh\left(d_2\sqrt{\beta^2 + \frac{1}{4a_2^2} - \varepsilon_{r2}k_o^2}\right) = -\sqrt{\beta^2 + \frac{1}{4a_1^2} - \varepsilon_{r1}k_o^2}\left(\varepsilon_{r2} - \frac{1}{4a_2^2k_o^2}\right) \tanh\left(d_1\sqrt{\beta^2 + \frac{1}{4a_1^2} - \varepsilon_{r1}k_o^2}\right), \tag{3}$$

where β is the propagation constant in the z direction, k_o is the wavenumber in vacuum, a_i is the width and d_i is the height of the SIWs as shown in Fig 1 (a), where i takes values 1 and 2.

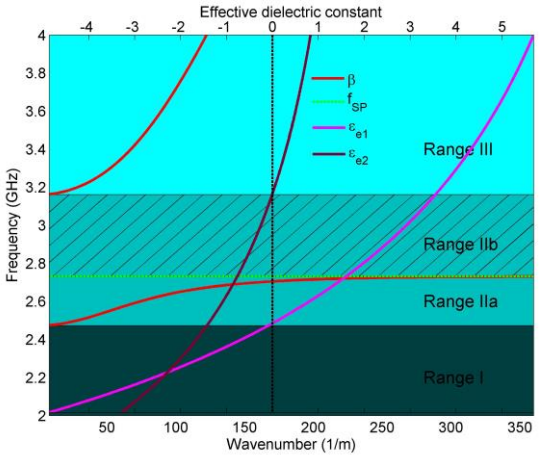


Fig. 2. Dispersion relation and effective permittivities of the sub-SIWs. The geometrical parameters of the structure are the following: a1 = 34.8 mm, a2 = 21.7 mm, ε1 = 2.2 and ε2 = 9.2.

Fig. 2 show the dispersion relation and characteristic frequency ranges for the two-layer SIW, for arbitrarily chosen values of the characteristic parameters – dielectric constants and sub-SIW's width. However, we consider that the characteristic values are such that the cut-off frequency of the upper layer is lower than that of the bottom layer. One can note three frequency ranges – in the first one both effective permittivities are negative, in the second one they are of opposite signs, whilst in the third frequency range both effective permittivities are positive. It is clear that the first range entirely suppresses propagation, whilst in the third one the conventional propagation occurs.

For the SPP-like propagation, the second frequency range is of major importance and it can be divided into two sub-regions, IIa and IIb, as shown in Fig. 2. Namely, in the region IIa, the absolute value of the negative effective permittivity is greater than that of the positive effective permittivity, causing SPP-like propagation, which is confirmed by the corresponding dispersion curve. The same curve goes to the infinity at the frequency at which the permittivities have the same values but opposite signs, causing a transmission zero to occur in the spectrum. In the region IIb, the negative permittivity has lower absolute value in comparison to the positive permittivity, and thus no propagation can occur, which is confirmed by the fact that dispersion curve does not exist in that region. In summary, owing to the nature of the SPP-like phenomenon, the two-layer structure provides stopbands in the I and IIb regions, and consequently the passband in the frequency range IIa. Since the widths and dielectric filling materials are different in general case, this indicates that transmission zero and consequently passband can be positioned arbitrarily.

The previous analysis represents the underlying idea of the formation of SPP-like propagation, and it can be further expanded to three-layer structure or adjusted to the case of half-mode SIW. As such, the analysis is the initial step in design of the microwave circuits and sensors, as it will be demonstrated.

III. MICROWAVE SINGLE-BAND FILTERS AND DIPLEXER

The SPP-like behavior in two-layer SIW can be used to realize single-band bandpass filters, since the two-layer structure can provide a passband owing to the transmission zero that stems from the SPP-like propagation. A single-band filter can be realized as a microstrip tapered two-layer SIW structure where two sub-SIWs have different filling dielectric materials and widths, Fig. 3 (a) and (b). To demonstrate this possibility, we have designed two single-band filters that operate at 2.4 and 2.6 GHz, respectively.

They have been designed according to the theoretical analysis in the previous section and in that sense, opposite signs of effective permittivity have been achieved relying on both different sub-SIWs' width and their filling materials. Dielectric materials that are used are Rogers RT 5800 (ε_{r1}=2.2, $tan\delta_1$=0.0009, d_1=0.51 mm) and TMM 10 (ε_{r2}=9.2, $tan\delta_2$=0.002, d_2=1.27 mm). The dimensions have been optimized using CST software package and their final values in millimeters are: a_{11}=34.8, a_{12}=21.7, a_{21}=32, a_{22}=20, L=40, L_T=5, W_{T1}=5.22, W_{T2}=4.8, w_{50}=3.4, w_w=0.4, d_w=0.4, n_w=31, d_{via}=0.5, p_{via}=0.25, where a_{11}/a_{21} and a_{12}/a_{22} are the widths of the upper and lower sub-SIW structures for bandpass filter operating at 2.4 and 2.6 GHz, respectively. The filters are fed by the tapered microstrip lines, which have also been optimized to achieve good impedance matching.

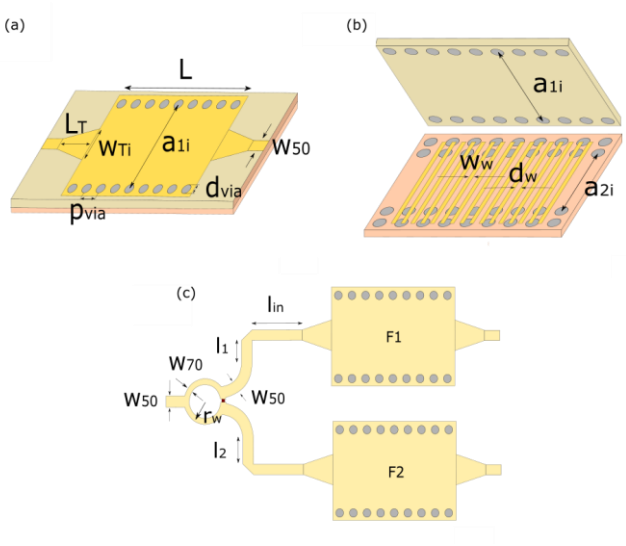


Fig. 3. (a) Two layer SIW with a microstrip transition (b) SPP interface (c) Final layout of the diplexer.

Furthermore, the single-band filters can be used to design a diplexer and to that end, we have designed a diplexer operating at 2.4/2.6 GHz that consists of the two single-band bandpass filters incorporated into diplexer using a convectional Wilkinson power divider, Fig. 3 (c) [26].

It should be noted that for the SIW bandpass filter that operates at 2.4 GHz, it is important not only to position the central frequency at 2.4 GHz, but also to position the transmission zero at 2.6 GHz, i.e. at the higher operating frequency of the diplexer, to achieve high selectivity as well as stopband suppression. To that end, the cut-off frequency of the upper sub-SIW structure is chosen to be at 2.35 GHz, whilst the cut-off frequency of the bottom sub-SIW structure should be placed above 2.8 GHz. Thus, the transmission zero and the stopband is positioned in the frequency range 2.5-2.65 GHz, which coincides with the higher operating band. As for the filter with the central frequency of 2.6 GHz, the cut-off frequency of the upper sub-SIW structure is placed at 2.55 GHz, whilst the cut-off frequency of the bottom sub-SIW structure is placed above 3 GHz, so its transmission zero is placed at 2.8 GHz, i.e. closely to the passband.

The simulated responses of the individual filters are shown in Fig 4 and they have the following characteristics - their 3-dB bandwidths are 8.3 and 7.69%, the insertion losses -0.735 and -0.78 dB, and the return losses are below -35 and -30 dB, respectively. The passbands are characterized by the excellent selectivity which is primarily due to transmission zeros that

occur at 2.59 GHz and 2.78 GHz.

The final response of the diplexer is also shown in Fig. 4 (b) and (c) and it is in a good agreement with the responses of the individual filters. The central operating frequencies are positioned at 2.4 and 2.63 GHz, and their 3-dB bandwidths are 10.4 and 7.98%, respectively. The insertion loss is -3.21 dB and 3.41 dB, whilst the return losses are 12 and 26 dB, respectively. The passbands are characterized by the excellent selectivity which is primarily due to transmission zeros that occur at surface-plasmon frequencies at 2.6 GHz and 2.8 GHz respectively. The isolation between outputs ports is below -20dB in the frequency range of interest. Transmission zero of the filter at 2.4 GHz is located at 2.6 GHz and ensures high stopband suppression below -26 dB.

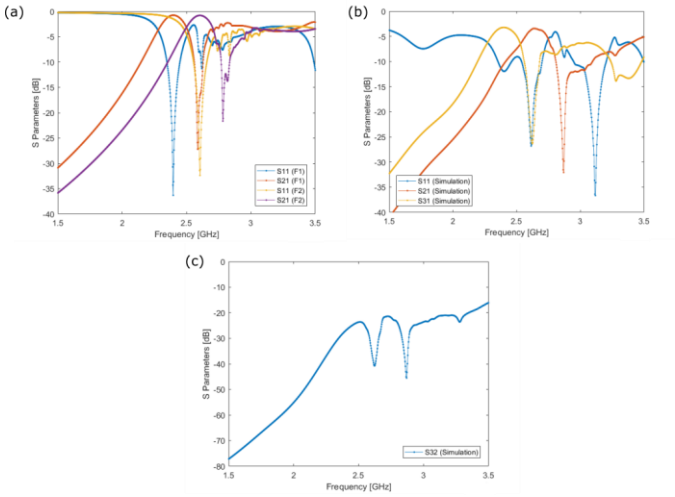


Fig. 4. Simulated (a) individual filter responses, (b) diplexer response and (c) its isolation

IV. Multi-Band Microwave Filters

The concept of microwave filters based on SPP-like propagation can be further extended to multi-band filters using multilayer SIW structures. Introduction of multiple sub-SIWs with different cut-off frequencies leads to the introduction of multiple transmission zeros and passbands in the spectrum. We note that the position of transmission zeros, and consequently passbands and stopbands can be chosen arbitrarily, since they depend on geometrical and material parameters of the individual sub-SIW structures. This property, together with SIW wave guiding capabilities, meets the requirement for high performance, low-cost, multiband microwave filters with non-harmonically related passbands.

To demonstrate this potential, we have designed two dual-band bandpass filters that are based on three-layer SIW structure, where each layer represents a sub-SIW structure with specific width and filling material, Fig 5. Since the sub-SIWs' width and dielectric constants are mechanisms for independent control of the passbands, theoretically there are six degrees of freedom in filter design. However, it should be noted that the top sub-SIW should have the lowest cut-of frequency and the bottom one the highest cut-of frequency or vice versa, to achieve dual-band operation. Otherwise, the two SPP-like propagation would overlap and only one passband would be formed. Also, to further simplify the procedure and simultaneously keep the design freedom it is judicious to use either the same width or the same dielectric constant for the three sub-SIWs.

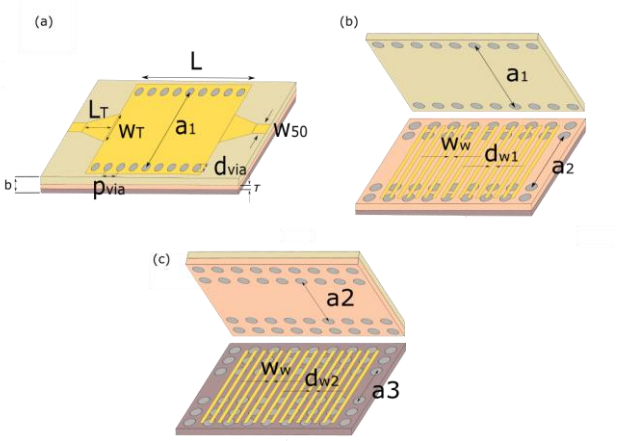


Fig. 5. (a) Layout of the proposed multi-band filters (a) top view (b) first SPP interface (c) second SPP interface

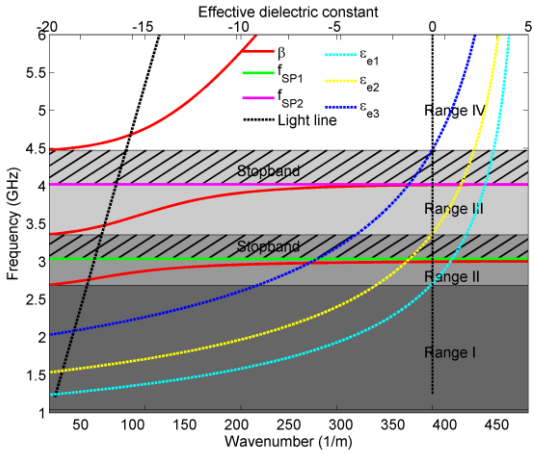


Fig. 6. (a) Dispersion relation of the multiband filter (b) effective permittivities of the three sub-SIWs. The geometrical parameters of the structure are the following: a_1=25 mm, a_2=20 mm, a_3=15 mm, ε_1=9.8, ε_2=4.5, ε_3=2.2, d=τ=1 mm and b=3 mm.

To demonstrate the operating principle, we extend the theoretical analysis given in the Section II, to the three-layer structure and obtain the following dispersion relation:

$$\frac{k_{y3}}{\varepsilon_{e3}}\tanh\left(k_{y3}d\right)=-\frac{k_{y2}}{\varepsilon_{e2}}\coth\left(k_{y2}\frac{\tau}{2}+\psi\right),$$
$$-\frac{k_{y1}}{\varepsilon_{e1}}\tanh\left(k_{y1}(b-d-\tau)\right)=-\frac{k_{y2}}{\varepsilon_{e2}}\coth\left(k_{y2}\frac{\tau}{2}-\psi\right), \quad (4)$$
$$k_{yi}=\sqrt{\beta^2-k_o^2\varepsilon_{e3}},$$

where β is the propagation constant in the z direction, k_0 is the wavenumber in vacuum, k_{yi} is the wavenumber in the y

direction, ε_{ei} effective dielectric constant, whilst i take the values 1, 2, and 3, denoting the top, middle, and bottom layer, respectively, and where d, τ, and b are the geometrical parameters of the structure as shown in Fig. 5(a).

Figure 6 shows the dispersion relations along with effective permittivities of each sub-SIW, where sub-SIWs have the same geometrical parameters and different dielectric constants. Similar to Figure 2 which shows a specific spectral region that enables passband in the spectrum, here we have two such spectral regions that enable the existence of the separate SPPs and consequently two passbands.

Following this analysis, we first propose a filter that comprises three sub-SIWs of the same width filled with different dielectric materials [23]. The dielectric constant of the middle layer and its width have been chosen to position its cut-off frequency in the range 2.7–3.2 GHz to be able to position the passbands around 2.4 and 3.5 GHz. Afterwards, the dielectric constants of the top and bottom sub-SIWs have been chosen to obtain the passbands at the desired frequencies.

The three layers are realized using the following dielectric substrates: Rogers TMM10i with relative permittivity ε_{r1}=9.8, dielectric loss $tan\delta_1$=0.002, and thickness $t_1=b-d-\tau$=1.27mm, Neltec NH9450 with relative permittivity ε_{r2}=4.5, dielectric loss $tan\delta_2$=0.0027, and thickness τ=0.768mm, and Rogers RT5880 with relative permittivity ε_r=2.2, dielectric loss $tan\delta$=0.0009, and thickness d=0.51mm.

The final geometrical parameters of the filter in millimeters are the following: a=22, L=40, L_T=12.5, W_{in}=4.5, W_t=5, w_w=0.2, d_{w1}=1, n_{w1}=33, d_{w2}=0.4, n_{w2}=65, d_{via}=0.8, p_{via}=1.1, where dvia represents the diameter of the via, pvia the spacing between the vias, whilst n_{w1} and n_{w2} represent the number of wires at the interface between neighboring materials.

To demonstrate the second tuning mechanism, we proposed a filter that comprises three sub-SIWs with different widths and same dielectric filling materials. To that end, the dielectric substrate Neltec NH9450 with relative permittivity ε_{r2}=4.5, dielectric loss $tan\delta_2$=0.0027, and thickness τ=0.768 mm, has been used. The widths a_1, a_2, and a_3 have been determined to achieve the passbands around the frequencies 4.7 and 5.5 GHz. The final geometrical parameters in millimeters of the filter are following: a_1=16, a_2=13.8, a_3=12.4, L=25, L_T=5, W_{in}=4.5, W_T=9, w_w=0.25, d_{w1}=0.25, n_{w1}=45, d_{w2}=0.375, n_{w2}=30, d_{via}=0.5, p_{via}=0.8.

In both filters, sub-SIWs are coupled through the array of wires, and the arrays together with microstrip feeding lines have been optimized to achieve good impedance matching. More detail on this can be found in [23].

The fabricated filters are shown in Fig. 7, while Fig 8. shows the simulated and measured response of the filters.

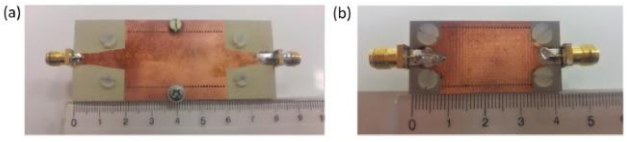


Fig. 7. Fabricated dual-band filters (a) First dual-band filter (b) Second dual-band.

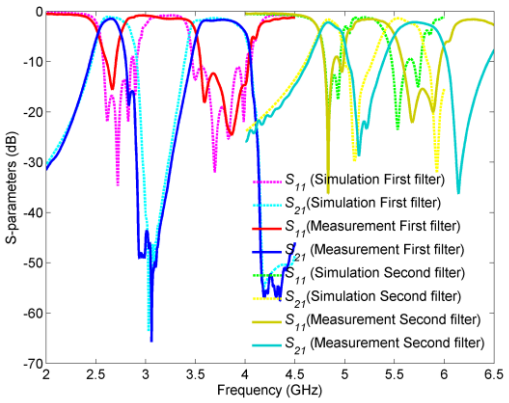


Fig. 8. Comparison of simulated and measured response for First- and Second dual-band filter.

The first filter exhibit two passbands at 2.65 and 3.75 GHz, with insertion loss of 1.47 and 1.69 dB, and the 3-dB bandwidths of 8.7% and 13.3%, respectively. The central frequencies of the second filter are 4.8 and 5.7 GHz, their insertion losses 2.22 and 2.17 dB, and the 3-dB bandwidths 5.2% and 8.2%, respectively. Both filters are characterized by good in-band and out-of-band performance as well as by excellent selectivity owing to the transmission zeros. The difference in the insertion losses between the first and the second filter can be attributed to the fact that the first filter comprises three different substrates whose dielectric losses are $tan\delta_1$=0.002, $tan\delta_2$=0.0027, and $tan\delta_3$=0.0009, respectively, whilst in the second proposed filter only the substrate with $tan\delta_2$=0.0027 has been used. Therefore, the losses in dielectric are more pronounced in the second proposed filter, which results in higher insertion losses.

The proposed filters were fabricated using standard printed circuit board (PCB) technology and each layer was fabricated separately and afterwards assembled into the final structure using screws, which inevitably causes small air gaps between the layers. Since SPP-like propagation is confined to the interface between layers, air gaps could influence the response introducing a small shift in the frequency spectrum. Their thickness cannot be measured precisely, but can be approximated by the thickness of the wires, i.e. copper thickness. In addition, imperfections in soldering of the connectors and employed three port measuring method could be a reason for higher return losses in the fabricated structure.

In order to overcome technology related limitations, we also proposed LTCC dual-band microwave filter based on “natural” SPP concept. LTCC enables accurate, robust, and low-cost production of multilayer circuits that can be easily integrated. In other words, employment of LTCC technology prevents occurrence of air gaps and misalignment in the structure, which provides very reliable fabrication of demanding multilayer structures. Since the filter is fabricated using LTCC technology, all layers have the same filling dielectric material, and the cut-off frequencies of the sub-SIWs are controlled exclusively by their widths.

The green tapes that are used in the LTCC process as the

dielectric material, have the relative permittivity 5.6, dielectric loss 0.001, and thickness 0.24 mm. The final geometrical parameters of the filter in millimeters are the following: a_1=16, a_2=13.8, a_3=12.5, L=40, L_T=5, W_{in}=4.5, W_T=5, w_w=0.2, d_{w1}=1, n_{w1}=33, d_{w2}=0.4, n_{w2}=65, d_{via}=0.8, p_{via}=1.1, where dvia represents the diameter of the via, pvia the spacing between the vias, whilst n_{w1} and n_{w2} represent the number of wires at the interface between neighboring materials.

During the lamination and co-firing process of LTCC process, slight shrinking of the structure has arisen due to inhomogeneity of the proposed structure, which is investigated using computed tomography, Fig. 9 (b).

The simulated and measured results are shown in Fig. 10. The simulated and measured responses show the same trend and *a* slight discrepancy is due to the difference in the simulated and fabricated dimension of a_1, and manufacturer tolerance in terms of dielectric constant.

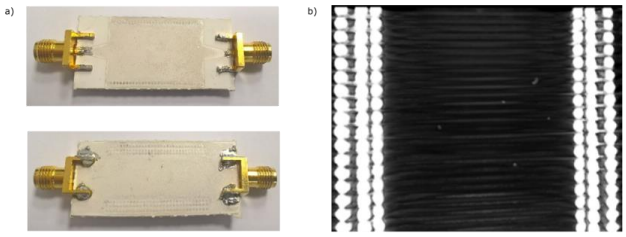


Fig. 9. a) Top and bottom view of the fabricated LTCC filter, b) inner structure of the proposed filter captured using CT tomography.

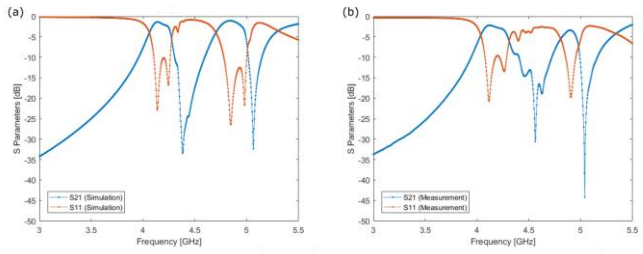


Fig. 10. (a) Simulated and (b) measured S parameters of the proposed dual-band bandpass filter.

V. MICROWAVE SINGLE-BAND FILTERS AND DIPLEXER

Relying on the previous discussions, we have proposed a microwave dielectric constant sensor [25], which is however based on half-mode (HM) SIW structure. Although SIW can provide lower radiation losses in comparison to HMSIW, HMSIW has comparable performance in the frequency range of interest, sufficiently good to support propagation and the proposed sensing concept. However, unlike the cumbersome structures in conventional SIW, HMSIW provides significantly smaller structure.

The layout of the proposed sensor is shown in Fig. 11. The HMSIW is comprised of two parts filled with a dielectric substrate material, where the bottom part also features a reservoir that hosts the sensed liquid analyte. The two parts are coupled through an array of wires, and since they differ in the dielectric constant, SPP occurs causing a sharp transmission zero in HMSIW response, which is very sensitive to small changes of the dielectric constant of the analyte. In that manner, minute changes in dielectric constant can be detected, which is not the case with the majority of the microwave dielectric constant sensors.

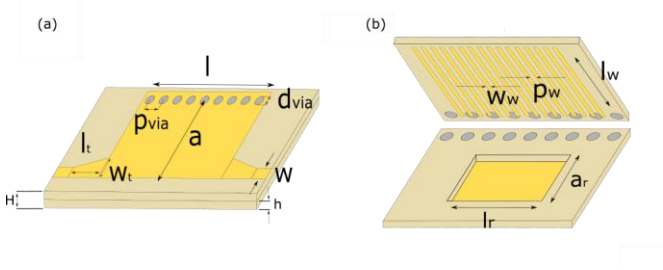


Fig. 11. Layout of the sensor based on HMSIW.

Fig. 12 shows HMSIW numerical transmission response for different materials in the reservoir, whose real part of the dielectric constant ranges from 2.4 to 3.8. The responses have been obtained using finite element method in CST solver, and the geometrical parameters are chosen to provide the response of interest in the range from 2 to 3 GHz. Obtained geometrical parameters in millimeters are: a=15, a_r=12.1, l=40, l_r=20, H=1.27, h=0.635, w=1.9, w_t=5, l_t=30, l_w=12.7, w_w=0.1, p_w=0.5, d_{via}=0.8, p_{via}=1.1. The dielectric substrate was Rogers TMM6 with relative permittivity ε_r=6, dielectric loss $tan\delta$=0.0023, and thickness t=0.635 mm, Also, the width of the microstrip lines that feed HMSIW have been optimized to achieve good impedance matching and low-loss response.

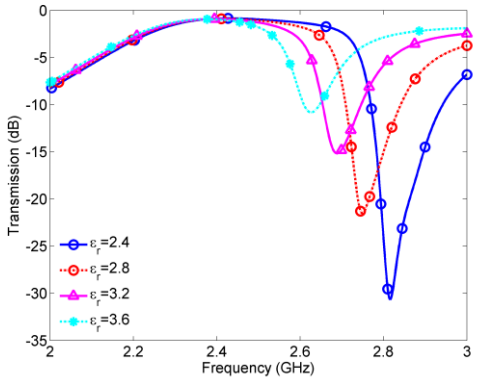


Fig. 12. HMSIW numerical transmission response for different materials in the reservoir.

One can notice that a cut-off frequency of the overall structure is the same for all responses – 2.2 GHz. However, the responses differ in the position of the sharp transmission zero. Namely, when the absolute values of the two dielectric constants are equal, propagation is not allowed and a sharp transmission zero occurs in HMSIW response. Using the calculation method proposed in [6], both real and imaginary parts of the unknown dielectric constant can be calculated:

$$\mathrm{Re}(\varepsilon_{e2}) = -\mathrm{Re}(\varepsilon_{e1}),$$
$$\mathrm{Im}(\varepsilon_{e2}) = \mathrm{Im}\left(\varepsilon_{r2}\left(1-\frac{\omega_2^2}{\omega(\omega+i\omega\gamma_2)}\right)\right) = \left|\mathrm{Re}(\varepsilon_{e2})\right|, \quad (5)$$

where ε_{ei} and ε_{ri} represent complex effective permittivities and real parts of the dielectric constant, ω_{pi} and γ_i are the plasma frequency and the damping rate of the plasma oscillations, where index i denotes the substrate (1) and sensing analyte (2). Q_d is the Q-factor of the plasma resonance and Υ is a derived parameter. A more detailed derivation can be found in [6].

To confirm the predicted sensor operation, we have fabricated the sensor using standard PCB technology, Fig. 13, and tested for different binary toluene/methanol mixtures. In the pure toluene whose dielectric constant is equal to 2.4-i*0.11, different amounts of methanol were added using micropipette to achieve the real parts of the dielectric constant values in the range from 2.4 to 3.8. Consequently, due to the losses in methanol, the range of the imaginary part of the dielectric constant in the mixtures was from 0.1 to 0.5. The exact value of the dielectric constant of the mixtures have been determined using Kraszewski formula.

Fig. 14 show the measured responses of the sensor for different binary toluene/methanol mixtures and they exhibit the same trend as the simulated ones. The dielectric constants of the analytes have been determined using the spectral positions of the transmission zeros in the measured responses and above mentioned calculation method. The obtained values have been compared to those calculated using Kraszewski formula and this is shown in Fig. 15.

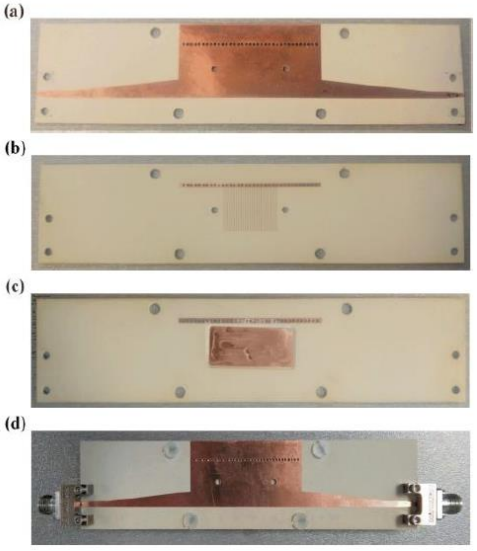


Fig. 13. Photographs of the fabricated sensor (a) top layer (b) bottom face of the top layer with wires (c) drilled tank in the bottom layer (d) assembled sensor.

One can note an excellent agreement between the two sets of the values for both real and imaginary part of the dielectric constant - the relative error is lower than 4% and 9%, respectively, for the whole set of measurement

The proposed sensor exhibits excellent sensitivity, which is significantly higher than the sensitivity of other recently proposed sensors [25]. Also, it should be noted that a particular strength of the proposed sensor is the detection of the real part of the dielectric constant in a very small range, whilst the majority of the published sensor are aimed at the detection of a wide range of the dielectric constant. We note here that the range of the real part of the dielectric constant values detectable by the proposed sensor can be widened using substrate with higher dielectric constant of HMSIW.

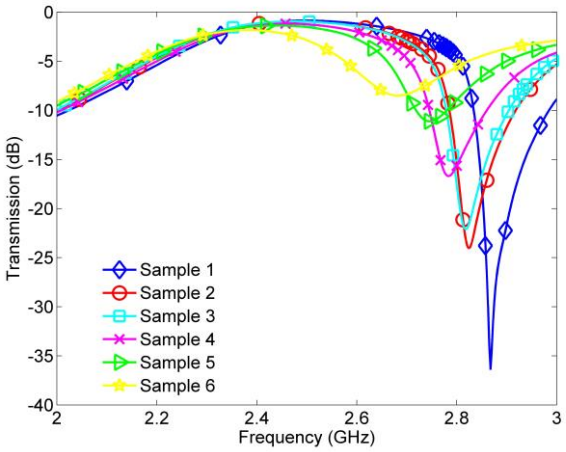


Fig. 14. Measured response of the fabricated sensor

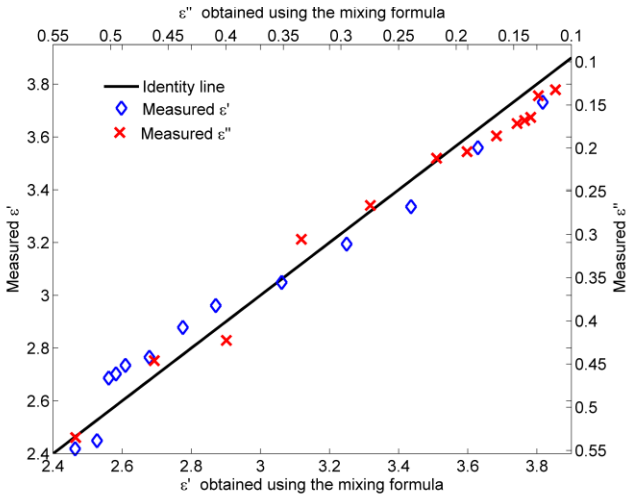


Fig. 15. Comparison of dielectric constants obtained using the mixing formula and spectral positions of the transmission zeros in the measured responses: real part and imaginary part.

These results confirm that the proposed sensor has a great potential for highly sensitive dielectric constant detection in liquid analytes in microwave regime, which can be applied in quality analysis of engine and edible oils. Moreover, the sensor represents a compact and low-cost solution since it is fabricated in low-profile configuration using PCB technology.

VI. Conclusions

In this paper, we have demonstrated the concept of "natural" SPP-like propagation and its applicability for

microwave components and sensors. We shown how SPP-like propagation can be realized in SIW structure and afterwards be applied to design single-band and dual-band filters, diplexer and dielectric constant sensor. Basic building block is a multilayer SIW, where layers represent sub-SIWs with different dielectric filling materials and/or different geometric parameters. The layers are designed to have effective permittivities of opposite signs in certain frequency ranges, which enables SPP-like propagation to occur at their interfaces. A detailed theoretical and numerical analysis together with numerical optimization was shown and afterwards used to design the novel components. Excellent filtering operation has been confirmed with the designed diplexer and dual-band bandpass filters, while extreme sensitivity has been demonstrated with the proposed sensor.

REFERENCES

[1] S. Maier, *Plasmonics: fundamentals and applications*, 1th ed., New York, USA: SSBM, 2007.

[2] Y. Li, *Plasmonic optics: Theory and applications*, SPIE Press, 2017.

[3] J. Wang, W. Lin, E. Cao, X. Xu, W. Liang and X. Zhang, "Surface Plasmon Resonance Sensors on Raman and Fluorescence Spectroscopy," Sensors, *17, 12, 2719*, 2017.

[4] L. Xiangang and Y. Lianshan, "Surface Plasmon Polaritons and Its Applications," IEEE Photon. J., *4, 2, 590-595*, 2012.

[5] J. Pendry, "Mimicking Surface Plasmons with Structured Surfaces," Science, *305, 847–848*, 2004.

[6] R. Sambles, A. Hibbins and M. Lockyear, "Manipulating Microwaves with 'Spoof ' Surface Plasmons," SPIE Newsroom , 2009.

[7] J. Zhu, S. Liao, S. Li and Q Xue, "Half-spaced Substrate Integrated Spoof Surface Plasmon Polaritons Based Transmission Line," Sci. Rep., *7, 8013*, 2017.

[8] D. Zhang, K. Zhang, Q. Wu, X. Ding, and X. Sha, :High-efficiency surface plasmonic polariton waveguides with enhanced lowfrequency performance in microwave frequencies," Opt. Exp., *25, 2121*, 2017.

[9] L. Tian, et al., "Compact spoof surface plasmon polaritons waveguide drilled with L-shaped grooves," Opt. Exp., *24, 28693*, 2016.

[10] L. Ye, et al., "Strongly confined spoof surface plasmonp waveguiding enabled by planar staggered plasmonic waveguides," Sci. Rep., *6, 38528*, 2016.

[11] C. Chen, " A new kind of spoof surface plasmon polaritons structure with periodic loading of T-shape grooves," AIP Adv., *6, 105003*, 2016.

[12] A. Kianinejad, Z. Chen, and C. Qiu, "Low-loss spoof surface plasmon slow-wave transmission lines with compact transition and high isolation," IEEE Trans. Microwave Theory Tech., *64, 10*, 3078–*3086*, 2016.

[13] Z. Li, et al., "High-contrast gratings based spoof surface plasmons," Sci. Rep., *6, 21199*, 2016.

[14] X. Liu, L. Zhu, Q. Wu and Y. Feng, "Highly-confined and low-loss spoof surface plasmon polaritons structure with periodic loading of trapezoidal grooves," AIP Adv., *5, 077123*, 2015.

[15] B. Xu, et al., "Tunable band-notched coplanar waveguide based on localized spoof surface plasmons," Opt. Lett., *40, 4683*, 2015.

[16] J. Wu, et al., "Bandpass filter based on low frequency spoof surface plasmon polaritons," Electron. Lett., *48, 269*, 2012.

[17] D. Guan, et al., "A wide-angle and circularly polarized beam-scanning antenna based on microstrip spoof surface plasmon polariton transmission line," IEEE Antennas Wireless Propag. Lett., *16, 2538–2541*, 2017.

[18] Y. Meng, et al., "Broadband spoof surface plasmon polaritons coupler based on dispersion engineering of metamaterials," Appl. Phys. Lett., *111, 151904*, 2017.

[19] X. Gao, et al., "Ultra-wideband surface plasmonic Y-splitter," Opt. Exp. *23, 23270*, 2015.

[20] S. Li, et al., "Hybrid metamaterial device with wideband absorption and multiband transmission based on spoof surface plasmon polaritons and perfect absorber," Appl. Phys. Lett., *106, 181103*, 2015.

[21] T. Qiu, J. Wang, Y. Li, J. Wang and S. Qu, "Broadband circulator based on spoof surface plasmon polaritons," J. Phys. D. Appl. Phys., *49, 355002*, 2016.

[22] C. Della Giovampaola, N. Engheta, "Plasmonics without negative dielectrics", Physical Review B, *93, 19*, 2016.

[23] N. Cselyuszka, Z. Sakotic, G. Kitic, V. Crnojevic-Bengin, N. Jankovic, "Novel Dual-band Band-Pass Filters Based on Surface Plasmon Polariton-like Propagation Induced by Structural Dispersion of Substrate Integrated Waveguide", Scientific Reports, *8, 8332*, 2018.

[24] Z. Sakotic, M. Drljaca, G. Kitic, N. Jankovic, and N. Cselyuszka, "LTCC Dual-band Bandpass Filter Based on SPP-like Propagation in Substrate Integrated Waveguide," The 18th International Conference on Smart Technologies IEEE EUROCON 2019, Novi Sad, Serbia, 2019

[25] N. Cselyuszka, Z. Sakotic, V. Crnojevic-Bengin, V. Radonic and N. Jankovic, "Microwave Surface Plasmon Polariton-Like Sensor Based on Half-Mode Substrate Integrated Waveguide for Highly Sensitive Dielectric Constant Detection", IEEE Sensors Journal, *18, 24, 9984-9992*, 2018.

[26] M. Drljaca, Z. Sakotic, N. Cselyuszka, V. Crnojevic Bengin, N. Jankovic, "Diplexer Based on Surface Plasmon Polariton-like Propagation Induced by Structural Dispersion of Substrate Integrated Waveguide," 13th International Congress on Artificial Materials for Novel Wave Phenomena – Metamaterials 2019, Rome, Italy, Sept, 2019, submitted

[27] C. Mariotti, W. Su, B. Cook, L. Roselli, M. Tentzeris, "Development of Low Cost, Wireless, Inkjet Printed Microfluidic RF Systems and Devices for Sensing or Tunable Electronics," IEEE Sensors Journal, *15, 6, 3156-3163*, 2015.

[28] R. Blakey, A. Morales-Partera, "Microwave dielectric spectroscopy – A versatile methodology for online, non-destructive food analysis, monitoring and process control", Engineering in Agriculture, Environment and Food, *9, 3, 264-273*, 2016.

[29] D. Khaled, N. Novas, J. Gazquez, R. Garcia, F. Manzano-Agugliaro, "Fruit and Vegetable Quality Assessment via Dielectric Sensing", Sensors, *15, 7, 15363-15397*, 2015.

[30] G. Ayissi Eyebe, B. Bideau, N. Boubekeur, É. Loranger, F. Domingue, "Environmentally-friendly cellulose nanofibre sheets for humidity sensing in microwave frequencies", Sensors and Actuators B: Chemical, *245, 484-492*, 2017.

[31] A. Benleulmi, N. Sama, P. Ferrari F. Domingue, "Substrate Integrated Waveguide Phase Shifter for Hydrogen Sensing", IEEE Microwave and Wireless Components Letters, *26, 9, 744-746*, 2016.

[32] W. Chen, K. Stewart, R. Mansour, A. Penlidis, "Novel undercoupled radio-frequency (RF) resonant sensor for gaseous ethanol and interferents detection", Sensors and Actuators A: Physical, *230, 63-73*, 2015.

[33] T. Chretiennot, D. Dubuc, K. Grenier, "Microwave-Based Microfluidic Sensor for Non-Destructive and Quantitative Glucose Monitoring in Aqueous Solution", Sensors, *16, 10, 1733*, 2016.

[34] A. Ebrahimi, W. Withayachumnankul, S. Al-Sarawi, D. Abbott, "Microwave microfluidic sensor for determination of glucose concentration in water," IEEE 15th Mediterranean Microwave Symposium (MMS), 2015.

[35] M. Islam, M. Islam, M. Samsuzzaman, M. Faruque, N. Misran, "A Negative Index Metamaterial-Inspired UWB Antenna with an Integration of Complementary SRR and CLS Unit Cells for Microwave Imaging Sensor Applications," Sensors, *15, 5, 11601-11627*, 2015.

[36] G. Gennarelli, S. Romeo, M. Scarfi, F. Soldovieri, "A Microwave Resonant Sensor for Concentration Measurements of Liquid Solutions," IEEE Sensors Journal, *13, 5, 1857-1864*, 2013.

[37] T. Kim, M. Y. Choi, H. W. Park, and J. H. Park, "Measurent and Analysis of Deterioration in the Automotive Engine Oil," Spring Proceeding of KSNT (Korean Society of Non-Destructive Testing, *220-224*, 2013.

[38] K. M. Park, "Oil deterioration sensor," US Patent *55400086*, 1996.

[39] E. Zanelato, F. Machado, A. Rangel, A. Guimarães, H. Vargas, E. da Silva, A. Mansanares, "Investigation of Biodiesel Through Photopyroelectric and Dielectric-Constant Measurements as a Function of Temperature: Freezing/Melting Interval," International Journal of Thermophysics, *36, 5-6, 924-931*, 2014.

[40] J. De Souza, M. Scherer, J. Cáceres, A. Caires, J. M'Peko, "A close dielectric spectroscopic analysis of diesel/biodiesel blends and potential dielectric approaches for biodiesel content assessment," Fuel *105, 705-710*, 2013.

[41] A. Sonkamble, R. Sonsale, M. Kanshette, K. Kabara, K. Wananje, A. Kumbharkhane, A. Sarode, "Relaxation dynamics and thermophysical

properties of vegetable oils using time-domain reflectometry", European Biophysics Journal, *46, 3, 283-291,* 2016.

[42] A. Khaled, S. Aziz, F. Rokhani, "Capacitive sensor probe to assess frying oil degradation," Information Processing in Agriculture, *2, 2, 142-148,* 2015.

[43] S. Junior, L. Paiter, J. Galvão, D. Roque, E. Chaves, "Sensor and Methodology for Dielectric Analysis of Vegetal Oils Submitted to Thermal Stress," Sensors, *15, 10, 26457-26477,* 2015.

Comparison of Various Geometries of Nonuniform Helical Antennas

Jelena Dinkić, Dragan Olćan, *Member*, *IEEE*, and Antonije Đorđević

***Abstract*—We present comparison of nonuniform helical antennas with linear, exponential, and piecewise-linear variations of the turn radius and the pitch angle. Parameters that define the geometry of these antennas are optimized at a single frequency and in a frequency range in order to maximize the gain. The obtained optimal designs are compared and some practical aspects are commented.**

***Index Terms*—Nonuniform helical antennas, numerical modeling, optimization.**

I. Introduction

HELICAL antennas have been known for more than 70 years [1]. During that period, different types of geometry were investigated and published in the open literature. Uniform helical antennas [1], [2] have the simplest geometry. Further, nonuniform helical antennas were investigated in order to improve the antenna performances [3]–[8].

In this paper we consider three geometry types of nonuniform helical antennas, with linear, exponential, and piecewise-linear variations of the turn radius and the pitch angle. For each geometry and different axial antenna lengths, we perform sets of optimization at a single frequency and in a frequency range. From the optimization results, the optimal antennas for all considered geometries are identified. These optimal antennas are further compared by considering various practical aspects.

The rest of the paper is organized as follows. Section II defines the considered geometries, simulation models, and optimization setups. Section III presents results of the single-frequency optimization. Section IV compares antennas obtained by the optimization in a frequency range. Finally, section V concludes the paper.

II. Geometries, Simulation Models, and Optimization Setup

The typical geometry of nonuniform helical antennas is shown in Fig. 1. The geometry of the helical antenna is defined by the turn radius, *r*, the pitch angle, φ (or pitch, *p*), and the overall axial antenna length, *L*. The conductor is considered to be a wire with a circular cross-section of the radius r_{w}, which is uniform along the conductor.

In this paper we consider three classes of geometries of nonuniform helical antennas.

The first class are helical antennas with linearly varying geometrical parameters (LG). The turn radius and the pitch angle are linear functions of the axial coordinate ($0 \le z_k \le L_k$):

$$r = (r_{k+1} - r_k)\frac{z_k}{L_k} + r_k\,,\ \varphi = (\varphi_{k+1} - \varphi_k)\frac{z_k}{L_k} + \varphi_k\,, \qquad (1)$$

where $k = 1$ and $L_1 = L$. Parameters r_1 and r_2 are the radii of the first and the last turn, and φ_1 and φ_2 are the pitch angles of the first and the last turn, respectively. During the optimization, the parameters r_1, r_2, φ_1, and φ_2 serve as the optimization variables.

The second class are helical antennas whose geometrical parameters have piecewise-linear variations (PWLG). We primarily focus on antennas that consist of three linear segments. Each segment is defined in a similar way as in (1). Hence, the radius and the pitch angle of each turn can be calculated from (1), where *k* is the index of the segment, $k \in \{1,2,3\}$. The optimization variables are r_j and φ_j, $j \in \{1,2,3,4\}$. In addition, we optimize the axial length of the two lower linear segments, whereas the length of the third segment is $L_3 = L - L_1 - L_2$, where L_1, L_2, and L_3 are the axial lengths of segments (Fig. 1).

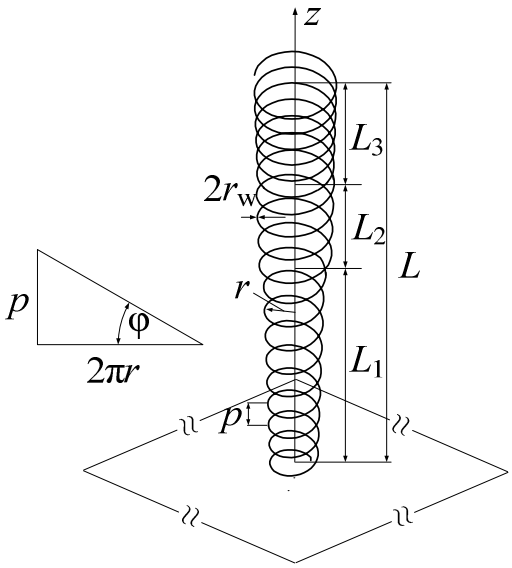


Fig. 1. Sketch of nonuniform helical antenna.

The third considered class of helical antennas are those with exponential variation of geometrical parameters (EG). The turn radius and the pitch angle are exponential functions of the axial coordinate (*z*):

$$r = A_{\mathrm{r}} + B_{\mathrm{r}}\mathrm{e}^{C_{\mathrm{r}}z}\,,\ \varphi = A_{\varphi} + B_{\varphi}\mathrm{e}^{C_{\varphi}z}\,, \qquad (2)$$

where $A_{\mathrm{r}} = r_1 - B_r$, $B_{\mathrm{r}} = (r_2 - r_1)/\left(\mathrm{e}^{C_{\mathrm{r}}L} - 1\right)$,

Jelena Dinkić, Dragan Olćan, and Antonije Đorđević are with the School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, P.O. Box 35-54, 11120 Belgrade, Serbia (e-mail: jdinkic@etf.bg.ac.rs, olcan@etf.bg.ac.rs, and edjordja@etf.bg.ac.rs).

$A_\varphi = \varphi_1 - B_\varphi$, and $B_\varphi = (\varphi_2 - \varphi_1)/\left(\mathrm{e}^{C_\varphi L} - 1\right)$. In this case, the optimization variables are r_1, r_2, φ_1, φ_2 (the same variables as defined for LG antennas), with additional variables C_r and C_φ.

All simulations are performed in software WIPL-D [9]. Since WIPL-D can analyze only straight-line wire segments, in order to reduce the computation time, we approximate perfectly circular turns with square turns as described in [10]. Since the focus of this paper is on the design of the helical conductor, and not on the design of the ground plane, all considered antennas are located above an infinite perfectly conducting ground plane. All antennas are fed by a generator located on a short vertical wire segment between the ground plane and the helical conductor.

Parameters for the three antenna geometries are optimized in order to maximize the partial gain [11] for the circular polarization at a single frequency or in a frequency range.

A two-step optimization is performed. The reliable choice for this optimization is PSO [12]–[13] followed by Nelder-Mead simplex [14], as indicated in [15]. For the linearly and the exponentially varying geometrical parameters, we implement PSO with 2000 iterations, followed by Nelder-Mead simplex with 200 iterations launched form the best found PSO solution. For the antenna with piecewise-linear variations of the geometrical parameters, 5000 PSO iterations are followed by 200 Nelder-Mead simplex iterations.

For the optimization in the frequency range, the same combinations of optimization algorithms are utilized, but they are repeated five times. The best solution of those five independent optimizations is taken to be the optimal solution.

Simulations and optimizations are performed on a PC (Intel Core i5-4690 processor). Typically, the single-frequency simulation lasts around 0.3 s, whereas the simulation in the considered frequency range lasts around 1 s. The results presented in this paper are obtained from the optimizations that contain around 500000 electromagnetic (EM) solver calls.

III. Single-Frequency Optimization

Firstly, the optimization for each geometry is performed at a single frequency, 300 MHz. The overall axial antenna length is $L = 3\lambda$, where λ is the wavelength at the operating frequency ($\lambda \approx 1\,\mathrm{m}$), and the conductivity of the wire conductor is $\sigma = 58\,\mathrm{MS/m}$ (copper). The models of the optimal antennas are shown in Fig. 2. Although the optimization is performed at a single frequency, in Fig. 3 the gain of the optimal antennas is presented in a frequency range in order to illustrate the frequency response. The gain of the LG and EG antennas at 300 MHz is 0.5 dB lower than the gain of the PWLG antenna. However, the PWGL antenna has a sharp maximum of the gain at 300 MHz, while the gain of the LG and EG antennas is a smooth function of frequency. In practice, this can be critical due to the tolerances in manufacturing because the frequency response of PWGL antenna may shift.

From the practical point of view, an important characteristic is the overall length of the wire conductor, which was not taken into the account during the optimization because it is of a secondary interest. With respect to this criterion, the PWLG antenna has the shortest conductor length (27.63 λ), whereas the LG and the EG antennas require longer conductors (39.02 λ and 33.50 λ, respectively). Note from Fig. 2 that the optimal LG antenna has significantly more turns at the lower end of the antenna compared to the PWLG and EG antennas.

However, the bandwidth of the PWLG antenna is significantly narrower than for the LG and EG antennas, which makes this comparison somewhat unfair.

For those reasons, we perform optimization in a frequency range with the aim to obtain as high as possible gain in a reasonable frequency range. Since the discrepancies of the gain are significant at higher frequencies, instead of the frequency range symmetrically positioned around the operating frequency, the frequency range considered during the optimization is slightly shifted towards higher frequencies.

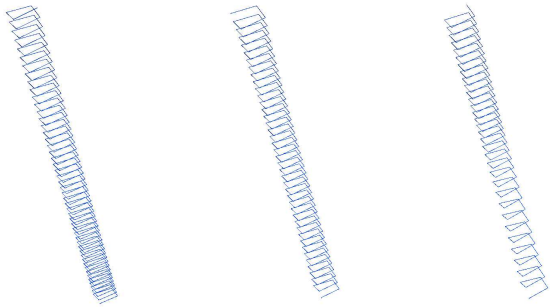


Fig. 2. Models of the optimal antennas (single-frequency optimization) whose geometrical parameters are (a) linear, (b) exponential, and (c) piecewise-linear functions of axial coordinate.

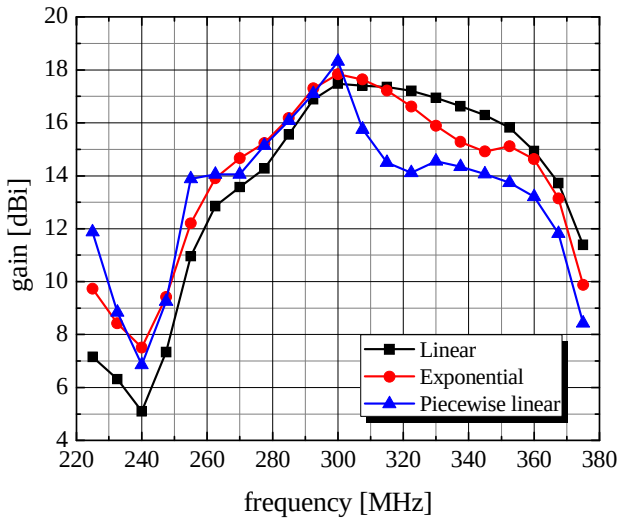


Fig. 3. Gain of the optimal antennas (single-frequency optimization).

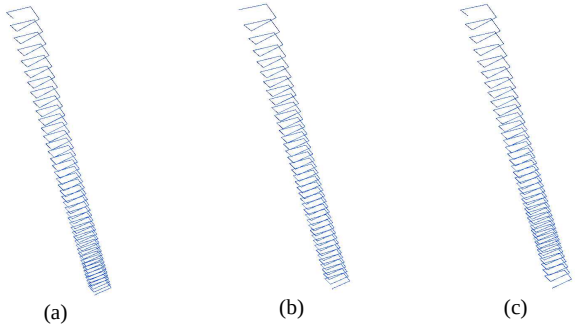
IV. Optimization in Frequency Range

For the optimization in a frequency range, we consider seven equidistantly spaced frequencies within the range from 280 MHz to 340 MHz. For each frequency, we calculate the partial gain for the circular polarization. The cost function is the arithmetic mean of the cost functions

calculated at each considered frequency in the same way as for the optimization at a single frequency.

We perform optimization in the frequency range for the same antenna axial length as in the previous section ($L = 3\lambda$, where λ is the wavelength at 300 MHz), and the wire conductivity is $\sigma = 58\,\mathrm{MS/m}$.

The optimal antennas for all considered geometries are shown in Fig. 4. The gain and the axial ratio of those antennas are presented in Fig. 5. Small differences between the antennas shown in Fig. 2 and Fig. 4 can be noticed. The conductor lengths of the antennas shown in Fig. 4a, b, and c are 34.88 λ, 31.79 λ, and 32.97 λ, respectively. These total wire lengths differ for less than 10 % among each other. It can be seen from Fig. 4 that all considered geometries have more densely spaced turns and smaller turn radii near the feeding point, compared to the end of antenna.



(a) (b) (c)

Fig. 4. Models of the optimal antennas (optimization in frequency range) whose geometrical parameters are (a) linear, (b) exponential, and (c) piecewise-linear functions of axial coordinate.

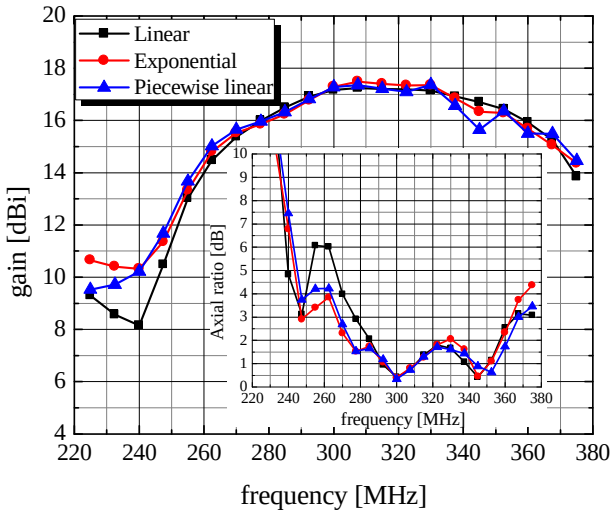


Fig. 5. Gain and axial ratio of the optimal antennas (optimization in frequency range).

Obviously, the optimization performed in a frequency range resulted in a wider bandwidth. However, this improvement is accompanied in some cases with degradation of other characteristics. The optimal antennas achieve practically the same gain from 260 MHz to 330 MHz, regardless of the geometry (the gain differs for less than 0.5 dB in the whole considered frequency range and for all considered antennas). Further, from Fig. 6 it can be noticed that the PWLG antenna optimized in a frequency range achieves around 1 dB lower gain at 300 MHz then the PWLG antenna optimized at a single frequency, and its wire conductor is around 20% longer. For the EG antenna the decrease of the gain at 300 MHz is around 0.5 dB, but the wire conductor is slightly shorter and the bandwidth is wider. In case of the LG antenna, the optimization in a frequency range results in a slightly wider bandwidth and shorter conductor, accompanied with a small decrease of the gain at 300 MHz (0.3 dB).

The high gain at 300 MHz and short conductor are the main advantages of the PWLG antennas optimized at a single frequency. However, those advantages are not present in case of the PWLG antennas optimized in a frequency range.

Hence, the following investigation for various antenna axial lengths will be performed only for LG and EG antennas because their geometry is simpler. Hence, they require fewer optimization parameters and allow easier fabrication.

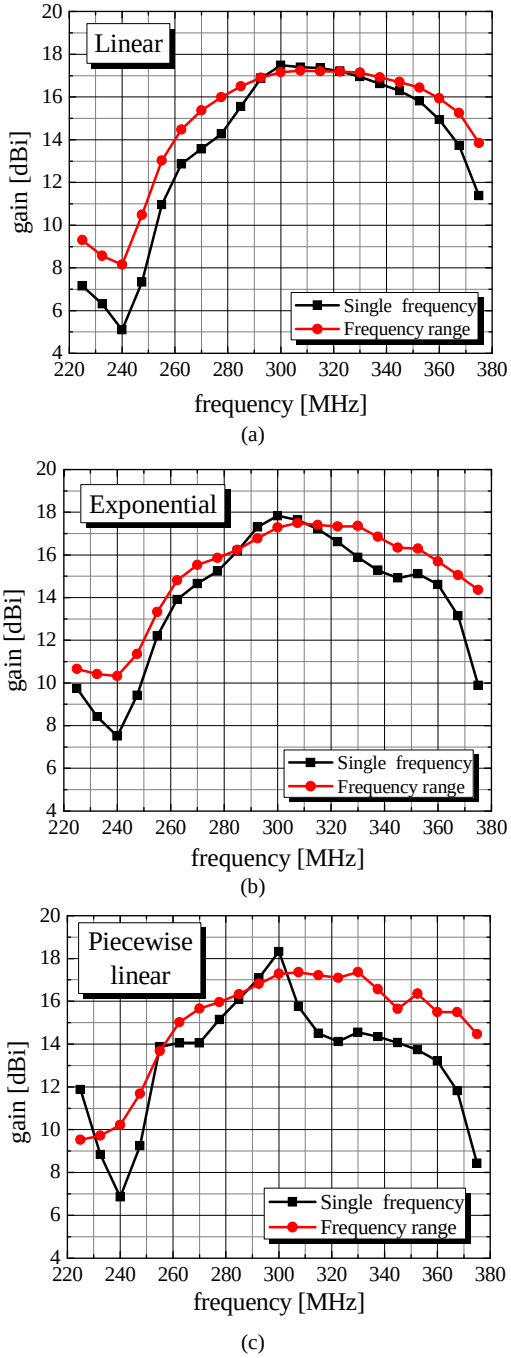


Fig. 6. Gain of the optimal (a) LG, (b) EG, and (c) PWLG antenna optimized at single frequency and in frequency range.

Optimization in a frequency range for the LG and the EG antennas is performed for different axial antenna lengths (2 λ, 2.5 λ, 4 λ, and 5 λ). Fig. 7 compares the gain of those antennas for each axial length, whereas Table I lists the lengths of the wire conductors for the considered antennas. The optimal EG antennas achieve a slightly higher gain at the operating frequency (300 MHz), especially for longer antennas (around 0.15 dB). For shorter antennas ($L \leq 3\lambda$) the optimal EG antennas also require shorter wire conductors than the optimal LG antennas. However, for longer antennas the situation is reversed.

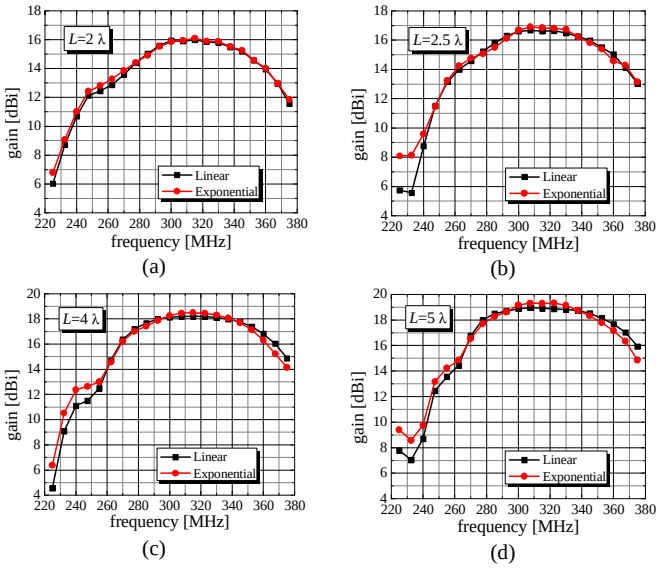


Fig. 7. Gain of the optimal antennas optimized in frequency range: (a) L=2 λ, (b) L=2.5 λ, (c) L=4 λ, and (d) L=5 λ.

TABLE I
CONDUCTOR LENGTHS AND MAXIMAL GAIN OF OPTIMAL ANTENNAS

L	LG antennas		EG antennas	
	conductor length	max gain [dBi]	conductor length	max gain [dBi]
2 λ	26.7 λ	16.0	26.7 λ	16.1
2.5 λ	31.8 λ	16.7	31.2 λ	16.9
3 λ	34.9 λ	17.2	31.8 λ	17.5
4 λ	44.0 λ	18.2	45.8 λ	18.5
5 λ	53.2 λ	19.0	54.9 λ	19.3

V. CONCLUSION

In this paper various geometries of nonuniform helical antennas are investigated and compared. It is shown that the optimization at a single frequency yields antennas with linearly-varying geometrical (LG) parameters that are reasonably broadband. The bandwidth of antennas whose geometrical parameters have exponential (EG) and piecewise-linear (PWLG) variations is narrower, although the gain is slightly higher. The optimization in a frequency range broadens the bandwidth, but the gain of all antennas is almost equal. Further, electrically shorter EG antennas require shorter conductors than LG antennas. However, for electrically longer antennas the situation is reversed. The results presented in this paper impose the conclusion that the optimization in a frequency range aimed at maximizing the partial gain should also include the conductor length into the cost function, with an appropriate weight, in order to increase the gain, but also to keep the required conductor length as short as possible, which is of practical interest. That optimization will be the scope of our future work.

ACKNOWLEDGMENT

This work was supported by the Serbian Ministry of Science and Technological Development under grant TR 32005 and by the Serbian Academy of Sciences and Arts under grant F133.

REFERENCES

[1] J.D. Kraus, "Helical beam antennas," *Electronics*, 20, pp. 109-111, April 1947.

[2] A.R. Djordjevic, A.G. Zajic, M.M. Ilic, and G.L. Stuber, "Optimization of helical antennas," *IEEE Antennas Propag. Mag.*, vol. 48, no. 6, pp. 107–115, Dec. 2006.

[3] J.L. Wong and H.E. King, "Broadband quasi-taper helical antennas," *IEEE Trans. Antennas Propag.*, vol. 27, no. 1, pp. 72–78, 1979.

[4] H.M. Elkamchouchi and A.I. Salem, "Helical antennas with nonuniform helix diameter," *Proceedings of the Eighteenth National Radio Science Conference. NRSC'2001 (IEEE Cat. No.01EX462)*, 2001, pp. 143–152.

[5] H.M. Elkamchouchi and A.I. Salem, "Effects of geometrical parameters, loading, and feeding on nonuniform helical antennas," *Proceedings of the Nineteenth National Radio Science Conference*, 2002, pp. 90–100.

[6] R. Golubovic, A. Djordjevic, D. Olcan, and J. Mosig, "Nonuniformly-wound helical antennas," *Proc. EuCAP 2009*, Berlin, Germany, 2009, pp. 3077-3080.

[7] K. Jimisha and S. Kumar, "Optimum design of exponentially varying helical antenna with non uniform pitch profile," *Procedia Technol.*, vol. 6, pp. 792–798, 2012.

[8] C.H. Chen, E.K. N. Yung, B.J. Hu, and S.L. Xie, "Axial mode helix antenna with exponential spacing," *Microw. Opt. Technol. Lett.*, vol. 49, no. 7, pp. 1525–1530, Jul. 2007.

[9] WIPL-D, Belgrade, Serbia. (2017). WIPL-D Pro v11.0—3D EM Solver. [Online]. Available: www.wipl-d.com.

[10] J.Lj. Dinkić, D.I. Olćan, A.R. Djordjević, and A.G. Zajić, "High-Gain Quad Array of Nonuniform Helical Antennas," *International Journal of Antennas and Propagation*, vol. 2019, Article ID 8421809, 12 pages, 2019.

[11] *IEEE Standard Definitions of Terms for Antennas* (IEEE Std 145-1983).

[12] J. Kennedy and R. Eberhart, "The particle swarm," in *Swarm Intelligence*, 1st ed., Morgan Kaufmann, 2001, ch. 7, pp. 287–326.

[13] J. Robinson and Y. Rahmat-Samii, "Particle swarm optimization in electromagnetics," *IEEE Trans. on Antennas and Propagation*, vol. 52, pp. 397–407, 2004.

[14] J.A. Nelder and R. Mead, "A simplex method for function minimization," *The Computer Journal*, 7, pp. 308-313, 1965.

[15] J.Lj. Dinkić, D.I. Olćan, A.G. Zajić, and A.R. Djordjević, "Comparison of optimization approaches for designing nonuniform helical antennas," *Proceedings of 2018 IEEE AP-S Symposium on Antennas and Propagation and URSI CNC/USNC*, Boston, MA, USA, 2018, pp. 1581–1582.

Two-dimensional Green's function for the Truncated Wedge in Terms of an Improper Integral

Dragan Filipović, Tatijana Dlabač

***Abstract*—In this paper two-dimensional Green's function for the truncated wedge is derived by separation of variables in Laplace's equation in cylindrical coordinates. The sign of the separation constant is chosen so as to exclude periodical particular solutions (harmonics). Hence, Green's function is sought in the form of an improper integral (which is essentially a summation of continual harmonics). An unknown function in the integral is found from the boundary conditions, and the integral itself, although rather complex, is found in a closed form. This method is in contrast with the conventional one, where Green's function involves an infinite summation of discrete harmonics.**

***Index Terms*—Green's function, truncated wedge, improper integral.**

I. Introduction

SEPARATION of variables in Laplace's equation is a standard way for determining two-dimensional Green's function for various cylindrical domains [1] – [3]. Usually, the sign of the separation constant is chosen so as to make some of the particular solutions (harmonics) periodic and Green's function is obtained in the form of an infinite summation of discrete harmonics.

However, it is quite legitimate to choose the separation constant with the opposite sign which gives rise to nonperiodical harmonics. In this case it is not possible to find Green's function in the form of an infinite summation and a representation in the form of an improper integral (essentially an infinite summation of continual harmonics) is used instead.

As an example of the latter, less conventional approach, two-dimensional Green's function for the truncated wedge is derived in the present paper. The involved integral is found in a closed form yielding Green's function in the same closed form, as obtained in [4] by the conventional way.

II. Two possible ways of separating variables in two-dimensional Laplace's equation in cylindrical coordinates

Separation of variables in two-dimensional Laplace's equation in cylindrical coordinates leads to two equations

$$\frac{r}{R}\frac{\partial}{\partial r}\left(r\frac{\partial R}{\partial r}\right)=\lambda \tag{1}$$

$$\frac{1}{\Theta}\frac{\partial^2\Theta}{\partial\theta^2}=-\lambda\,, \tag{2}$$

where λ is the separation constant, and $R(r)$ and $\Theta(\theta)$ constitute the potential, i.e.

$$V(r,\theta)=R(r)\cdot\Theta(\theta)\,.$$

Let $\lambda=k^2$ be positive. Then, the particular solutions of (1)–(2) are $r^{\pm k}$ (radial harmonics) and cos$k\theta$, sin$k\theta$ (angular harmonics), and the potential is sought in the form of an infinite summation of these (discrete) harmonics

$$V(r,\theta)=\sum_k\left(A_k r^k+B_k r^{-k}\right)\left(C_k\cos k\theta+D_k\sin k\theta\right),$$

where the unknown coefficients are determined from boundary conditions.

It is instructive to see how the potential looks quite different if the separation constant is chosen to be negative $(\lambda=-k^2)$ The particular solutions of (1) - (2) in this case are sin(k·ln$r$), cos(k·ln$r$) (radial harmonics) and sh$k\theta$, ch$k\theta$ (angular harmonics) (see Appendix A for a derivation of the radial harmonics). Neither of these harmonics is a periodical function, hence it is impossible to find the potential in the form of an infinite summation. Instead, we seek for a solution in the form of an improper integral, which is essentially an infinite summation of continual (in $k\in(0,\infty)$) harmonics, defined above. Thus, the potential should have the form

$$V(r,\theta)=\int_0^\infty\left[A(k)\,\mathrm{sh}k\theta+B(k)\,\mathrm{ch}k\theta\right]$$
$$\cdot\left[C(k)\cos(k\ln r)+D(k)\sin(k\ln r)\right]dk\,.$$

Note that the unknown coefficients with the harmonics are now functions to be determined from boundary conditions.

III. Two-dimensional Green's function for the truncated wedge in the form of an improper integral

Let there be given a cylindrical domain whose cross-section is a truncated wedge of radius R and angle α. By definition, Green's function is the potential at the point (r, θ), produced by the unit line charge running parallel and passing through the point (r', θ'), provided the boundary is kept at zero potential. The wedge is partitioned into two

Dragan Filipović is with the Faculty of Electrical Engineering, University of Montenegro, Dzordza Vasingtona bb, 81000 Podgorica, Montenegro (e-mail: draganf@ucg.ac.me).

Tatijana Dlabač is with the Faculty of Maritime Studies, University of Montenegro, Dobrota 36, 85330 Kotor, Montenegro (e-mail: tanjav@ucg.ac.me).

subdomains by the surface $\theta = \theta'$ (Fig. 1).

As stated in the previous section, the potentials in subdomains 1 and 2 can be sought in the forms

$$V_1(r,\theta) = \int_0^\infty A(k)\,\mathrm{sh}k\theta \sin\left(k\ln\frac{r}{R}\right)dk\,,\ \theta \le \theta'\,, \qquad (3)$$

$$V_2(r,\theta) = \int_0^\infty B(k)\,\mathrm{sh}\,k(\alpha-\theta)\sin\left(k\ln\frac{r}{R}\right)dk,\ \theta \ge \theta'\,, \qquad (4)$$

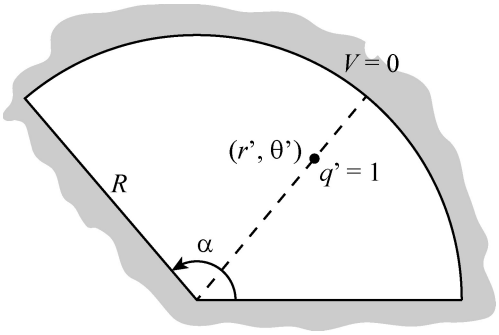


Fig. 1. Truncated wedge with unit line charge partitioned into two subdomains

which ensure that the boundary conditions are met (V_1=0 for θ=0 and r=R, V_2=0 for θ=α and r=R). In (3) – (4) $A(k)$ and $B(k)$ are unknown functions that should be determined.

Continuity of the potential requires V_1=V_2 for $\theta = \theta'$, whence

$$B(k) = A(k)\frac{\mathrm{sh}k\theta'}{\mathrm{sh}k(\alpha-\theta')},$$

and

$$V_2(r,\theta) = \int_0^\infty A(k)\frac{\mathrm{sh}k\theta'}{\mathrm{sh}k(\alpha-\theta')}\mathrm{sh}k(\alpha-\theta)\sin\left(k\ln\frac{r}{R}\right)dk\,. \quad (5)$$

The unknown function $A(k)$ in (3) and (5) can be determined from the boundary condition for the normal components of the electric field at the interface $\theta = \theta'$

$$\frac{1}{r}\left(\frac{\partial V_1}{\partial\theta} - \frac{\partial V_2}{\partial\theta}\right)_{\theta=\theta'} = \frac{1}{\varepsilon_o}\cdot\delta(r-r')\,, \qquad (6)$$

where the δ - function accounts for the presence of the unit line charge at the point (r', θ'). Some properties of the δ – function used below are listed in Appendix B. By using (3) and (5), after some elementary manipulations (6) becomes

$$\frac{1}{r}\int_0^\infty kA(k)\frac{\mathrm{sh}k\alpha}{\mathrm{sh}k(\alpha-\theta')}\sin\left(k\cdot\ln\frac{r}{R}\right)dk = \frac{1}{\varepsilon_o}\cdot\delta(r-r')\,. \quad (7)$$

Next, we multiply (7) by $\sin(m\ln\frac{r}{R})$, m>0 and integrate with respect to r from 0 to R. Then, after interchanging the order of integration (7) is transformed into

$$\int_0^\infty kA(k)\frac{\mathrm{sh}\,k\alpha}{\mathrm{sh}\,k(\alpha-\theta')}\left(\int_0^R \frac{\sin(k\ln\frac{r}{R})\sin(m\ln\frac{r}{R})}{r}dr\right)dk = \quad (8)$$
$$= \frac{1}{\varepsilon_0}\sin(m\ln\frac{r'}{R})\,.$$

The inner integral in (8) can be expressed in terms of the δ function by the change of variables $\ln\frac{r}{R} = u$

$$\int_0^R \frac{\sin\left(k\ln\frac{r}{R}\right)\sin\left(m\ln\frac{r}{R}\right)}{r}dr = \int_{-\infty}^0 \sin ku\sin mu\ du =$$
$$= \frac{1}{2}\int_{-\infty}^\infty \sin ku\sin mu\ du = \quad (9)$$
$$= \frac{1}{4}\int_{-\infty}^\infty\left[\cos(k-m)u - \cos(k+m)u\right]du =$$
$$= \frac{1}{2}\pi\left[\delta(k-m) - \delta(k+m)\right].$$

When (9) is substituted into (8), we obtain

$$m\cdot A(m)\frac{\mathrm{sh}\,m\alpha}{\mathrm{sh}\,m(\alpha-\theta')}\frac{\pi}{2} = \frac{1}{\varepsilon_o}\cdot\sin\left(m\cdot\ln\frac{r'}{R}\right),$$

whence

$$A(k) = \frac{2}{\pi\varepsilon_0}\cdot\frac{\mathrm{sh}\,k(\alpha-\theta')}{k\cdot\mathrm{sh}k\alpha}\cdot\sin\left(k\cdot\ln\frac{r'}{R}\right), \qquad (10)$$

and the potentials (3) and (5) become

$$V_1(r,\theta) = V_1(r,r',\theta,\theta') =$$
$$= \frac{2}{\pi\varepsilon_0}\int_0^\infty \frac{\mathrm{sh}\,k(\alpha-\theta')}{k\ \mathrm{sh}\,k\alpha}\sin(k\ln\frac{r'}{R})\mathrm{sh}\,k\theta\,\sin(k\ln\frac{r}{R})dk\,, \quad (11)$$
$$V_2(r,\theta) = V_2(r,r',\theta,\theta') =$$
$$= \frac{2}{\pi\varepsilon_0}\int_0^\infty \frac{\mathrm{sh}\,k\theta'}{k\ \mathrm{sh}\,k\alpha}\sin(k\ln\frac{r'}{R})\mathrm{sh}\,k(\alpha-\theta)\sin(k\ln\frac{r}{R})dk\,. \quad (12)$$

Equations (11) – (12) constitute Green's function for the considered domain

$$G(r,r';\theta,\theta') = \begin{cases} V_1(r,r',\theta,\theta'), & \theta \le \theta' \\ V_2(r,r',\theta,\theta'), & \theta \ge \theta' \end{cases}$$

It can be observed from (11) – (12) that V_2 is obtained from V_1 by interchanging θ and θ', so that V_1 suffices to describe Green's function

$$G(r,r';\theta,\theta') = \begin{cases} V_1(r,r',\theta,\theta'), & \theta \le \theta' \\ V_1(r,r',\theta',\theta), & \theta \ge \theta' \end{cases} \quad (13)$$

IV. Green's function in an explicit closed form

The integral in (11) can be found in a closed form. This is done in Appendix C; here we state the result

$$\int_0^\infty \frac{\operatorname{sh} k(\alpha-\theta')}{k \operatorname{sh} k\alpha} \sin\left(k \ln\frac{r'}{R}\right) \operatorname{sh} k\theta \sin\left(k \ln\frac{r}{R}\right) dk = \\ = \frac{1}{8} \ln\left[\frac{\operatorname{ch}\left(\frac{\pi}{\alpha}\ln\frac{r}{r'}\right) - \cos\left(\frac{\pi}{\alpha}(\theta+\theta')\right)}{\operatorname{ch}\left(\frac{\pi}{\alpha}\ln\frac{r}{r'}\right) - \cos\left(\frac{\pi}{\alpha}(\theta-\theta')\right)} \cdot \right. \\ \left. \cdot \frac{\operatorname{ch}\left(\frac{\pi}{\alpha}\ln\frac{r\cdot r'}{R^2}\right) - \cos\frac{\pi}{\alpha}(\theta-\theta')}{\operatorname{ch}\left(\frac{\pi}{\alpha}\ln\frac{r\cdot r'}{R^2}\right) - \cos\frac{\pi}{\alpha}(\theta+\theta')} \right]. \quad (14)$$

Interchanging θ and θ' does not affect (14), hence $V_2(r,r',\theta,\theta') = V_1(r,r',\theta',\theta) = V_1(r,r',\theta,\theta')$ and from (11), (13) – (14), Green's function is

$$G(r,r',\theta,\theta') = \frac{1}{4\pi\varepsilon_0} \ln\left[\frac{\operatorname{ch}\left(\frac{\pi}{\alpha}\ln\frac{r}{r'}\right) - \cos\left(\frac{\pi}{\alpha}(\theta+\theta')\right)}{\operatorname{ch}\left(\frac{\pi}{\alpha}\ln\frac{r}{r'}\right) - \cos\left(\frac{\pi}{\alpha}(\theta-\theta')\right)} \cdot \right. \\ \left. \cdot \frac{\operatorname{ch}\left(\frac{\pi}{\alpha}\ln\frac{r\cdot r'}{R^2}\right) - \cos\frac{\pi}{\alpha}(\theta-\theta')}{\operatorname{ch}\left(\frac{\pi}{\alpha}\ln\frac{r\cdot r'}{R^2}\right) - \cos\frac{\pi}{\alpha}(\theta+\theta')} \right].$$

This result is identical to the one obtained in [4] by summing an infinite series in a closed form.

V. Conclusion

This paper presents a derivation of two-dimensional Green's function for the truncated wedge. Two-dimensional Laplace's equation is solved by separating variables in cylindrical coordinates. The sign of the separation constant is chosen so that no periodical harmonics exist which makes an improper integral a suitable form for Green's function, contrary to the conventional summation form. The integral which determines Green's function is found in a closed form

Appendix A

For $\lambda = k^2$ (1) becomes

$$r^2 \cdot R_r'' + r \cdot R_r' + k^2 R = 0 . \quad (A1)$$

The change of variables $r = e^t$ transforms (A1) into

$$R_t'' + k^2 R = 0 ,$$

with particular solutions cos(*kt*) and sin(*kt*). Therefore, the radial harmonics, i.e. the solutions of (A1) are cos(*k*·ln*r*) and sin(*k*·ln*r*).

Appendix B

In deriving (8) – (10) we used the following properties of the δ - function [5]

$$\int_a^b f(x)\,\delta(x-x')dx = \begin{cases} f(x'), x' \in (a,b) \\ 0, x' \in (-\infty,a) \cup (b,+\infty) \end{cases}$$

and

$$\delta(x) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \cos xt\, dt = \frac{1}{\pi}\int_{-\infty}^{0} \cos xt\, dt = \frac{1}{\pi}\int_0^{+\infty} \cos xt\, dt .$$

Appendix C

To evaluate the integral in (11), we start from [6]

$$F(A,B,C) = \int_0^\infty \frac{\operatorname{sh} Ax \cdot \operatorname{sh} Bx}{x \operatorname{sh} Cx} dx = \frac{1}{2} \ln \frac{\cos\frac{\pi}{2C}(A-B)}{\cos\frac{\pi}{2C}(A+B)}, \quad (C1)$$

where

$$A + B < C .$$

The next step is evaluation of

$$G(A,B,C,D) = \int_0^\infty \frac{\operatorname{sh} Ax \cdot \operatorname{sh} Bx}{x \cdot \operatorname{sh} Cx} \cos Dx\, dx = \\ = \operatorname{Re} \int_0^\infty \frac{\operatorname{sh} Ax \cdot \operatorname{sh}(B + jD)}{x \cdot \operatorname{sh} Cx} dx = \operatorname{Re} F(A, B + jD, C) .$$

By using (C1) we readily find

$$G(A,B,C,D) = \frac{1}{4} \ln \frac{\operatorname{ch}\frac{\pi D}{C} + \cos\frac{\pi}{C}(A-B)}{\operatorname{ch}\frac{\pi D}{C} + \cos\frac{\pi}{C}(A+B)} . \quad (C2)$$

Finally, the integral in (11)

$$I(A,B,C,D,F) = \int_0^\infty \frac{\operatorname{sh} Ax \cdot \operatorname{sh} Bx}{x \cdot \operatorname{sh} Cx} \sin Dx \cdot \sin Fx \cdot dx ,$$

with $A = \alpha - \theta'$, $B = \theta$, $C = \alpha$, $D = \ln\frac{r}{R}$ and $F = \ln\frac{r'}{R}$ is evaluated by using the identity $\sin Dx \cdot \sin Fx = \\ = \frac{1}{2}\left[\cos(D-F)x - \cos(D+F)x\right]$ and (C2)

$$I(A,B,C,D,F)=\frac{1}{2}\left[G(A,B,C,D-F)-G(A,B,C,D+F)\right]=$$

$$=\frac{1}{8}\ln\left[\frac{\mathrm{ch}\left(\frac{\pi}{\alpha}\ln\frac{r}{r'}\right)+\cos\left(\frac{\pi}{\alpha}(\theta+\theta')\right)}{\mathrm{ch}\left(\frac{\pi}{\alpha}\ln\frac{r}{r'}\right)+\cos\left(\frac{\pi}{\alpha}(\theta-\theta')\right)}\cdot\right.$$

$$\left.\cdot\frac{\mathrm{ch}\left(\frac{\pi}{\alpha}\ln\frac{r\cdot r'}{R^2}\right)+\cos\frac{\pi}{\alpha}(\theta-\theta')}{\mathrm{ch}\left(\frac{\pi}{\alpha}\ln\frac{r\cdot r'}{R^2}\right)+\cos\frac{\pi}{\alpha}(\theta+\theta')}\right].$$

REFERENCES

[1] W. R. Smythe, *Static and Dynamic Electricity*, 3rd ed., McGraw-Hill, New York, 1968.

[2] W. K. Panofsky, M. Phillips, *Classical Electricity and Magnetism*, 2nd ed., Addison – Wesley, Reading, Massachusetts, 1962.

[3] J. D. Jackson, *Classical Electrodynamics*, John Wiley, New York 1962.

[4] D. Filipović, T. Dlabač, V. Durković, Two-dimensional Green's function for a truncated wedge, 4th IcETRAN, Kladovo, Serbia, June 5 – 8, 2017

[5] V. Vladimirov, *Distributions en physique mathematique*, Editions Mir, Moscou, 1979.

[6] P. Prudnikov, Yu. A. Brychkov, O. I. Marichev, *Integrals and Series*, Science Publishers, Moscow, 1981 (in Russian)

On Efficient Evaluation of Pole-Free Sommerfeld Integrals

Nikola Basta and Branko Kolundžija, *Fellow, IEEE*

***Abstract*—An approach to computation of pole-free Sommerfeld integrals is proposed using the example of free-space Green function in lossless media. Besides the cancellation of the branch-point singularity based on a change of variables, the proposed approach includes optimal choice of the lower endpoint of the tail subdomain and definition of thresholds that improve accuracy and efficiency of tail-integral evaluation. Additionally, formulas for estimation of the required number of integration points for given accuracy are given. The presented techniques are verified through numerical examples with large range of source-observer distances.**

***Index Terms*—Sommerfeld integral; free space; Green's function;**

I. Introduction

THE increasing demand for accurate solving of large-scale electromagnetic scenarios, e.g. in [1]–[2] calls for further investigation of strategies that address computation of the Sommerfeld integrals (SIs) [3]–[5] for a wide range of source-observer distances. Sommerfeld integrals are important tool for computation of the electromagnetic field in the presence two or more layered materials.

The associated integrands are singular, oscillatory and slow decaying, which makes the evaluation of the SIs demanding. In this work, we focus on numerical computation of SIs [6]–[13] for a large range of source-observer distances, from 10^{-3} to 10^{3} free-space wavelengths. Typically, the singularities, i.e. poles and branch points, are circumvented by integrating over a deformed path [6]–[10], [14], [15]. In this study, integration over the real-axis is applied [6], [16] due to its simplicity, robustness and resistance to the problem of high-valued Bessel function of complex argument [13]. Since the numerical treatment of the poles close to the integration path is well covered in the literature [6], [7], [13], our intention here is to consider thoroughly the branch-point singularities [16], [17]–[19], the effect of which has shown to be significant in numerous cases of interest, e.g. low-loss materials in the form of microwave dielectric laminates or weakly conductive earth. The proposed technique for cancellation of the branch-point singularity relies on a square-root change of variables.

Once the integrand is freed of singularities, the question of calculation of the semi-infinite integral tail remains. The position of the tail beginning can be important for the efficiency of the evaluation and related to the treatment of branch points, yet we were unable to find precise guidelines for its choice [9]–[12]. Therefore, expressions for optimal choice of the lower endpoint of the tail subdomain and of the threshold for tail truncation are proposed.

Not much work has been dedicated to construction of formulas that predict minimum number of integration points for requested accuracy, which is important for efficient evaluation of SIs, particularly for large source-observer distances. We propose simple formulas for estimation of required number of integration points for given accuracy.

Since the key numerical properties of the SIs remain the same regardless of the field component they represent, the techniques established here for the free-space SI can be transferred to other pole-free SIs.

Nikola Basta is with the School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail: nbasta@etf.rs).

Branko Kolundzija is with the School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail: kol@etf.rs).

II. Properties of the Proposed Approach

The particularities of pole-free SIs can be observed fairly through the integral associated to the scalar spherical wave in homogenous free space. Its solution is the well-known free-space Green's function [1], thus the relative error of the numerical evaluation can be calculated precisely. Therefore, we will illustrate our approach by computing integrals of general form

$$I(k_{\rho 1}, k_{\rho 2}; \rho, z) = \int_{k_{\rho 1}}^{k_{\rho 2}} \frac{k_\rho}{\sqrt{k_\rho^{\,2} - k^2}} \mathrm{e}^{-\sqrt{k_\rho^{\,2}-k^2}\,|z|} J_0(k_\rho \rho) \mathrm{d}k_\rho \qquad (1)$$

so that the SI associated to free-space GF can be expressed as

$$I(0, \infty) = \frac{\mathrm{e}^{-\mathrm{j}kr}}{r} = g(r), \qquad (2)$$

where J_0 is the Bessel function of zeroth order, $k = \omega\sqrt{\varepsilon_0 \mu_0} = 2\pi/\lambda$ is the free-space wave number, $r = \sqrt{\rho^2 + z^2}$ and ρ and z are the horizontal and vertical coordinate of the observation point, respectively, as shown in inset of Fig. 1. For purpose of brevity, we have omitted dependence of the integral on ρ and z in (2) and in all further references of (1).

In this work, we apply the Gauss-Legendre (GL) quadrature formula for numerical evaluation of the integral. There are several challenges in numerical evaluation of (2). Besides the singularity at the branch point, $k_\rho = k$, there is also the oscillatory nature of the integrand, due to the Bessel function factor for any k_ρ, and the exponential factor for $k_\rho < k$. The oscillatory nature of the integrand and the fact

that the domain of integration is semi-infinite raise the question of the computation efficiency. To overcome these difficulties, the integral (2) is divided into three parts, i.e., $I(0,\infty) = I_1 + I_2 + T$, where $I_1 = I(0,k)$, $I_2 = I(k,b)$, $T = I(b,\infty)$, and b designates the lower endpoint of the tail. An example of the integrand plot is illustrated for $\rho = 3\lambda$ and $z = 0.05\lambda$ in Fig. 1.

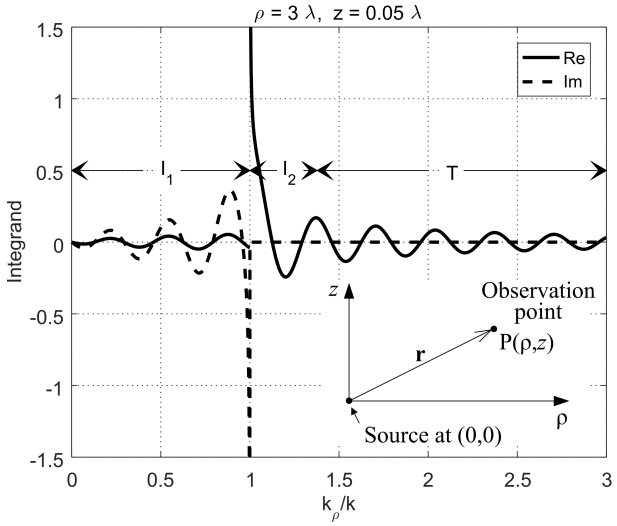


Fig. 1. Example of the integrand function in (1) with decomposition of the problem to three integrals, I_1, I_2 and T, over respective subdomains.

A. Cancellation of branch-point singularity

The integrals over finite subdomains, I_1 and I_2, have a common singular endpoint at $k_\rho = k$, where the integrand is unbounded. For overcoming this branch-point singularity, several techniques can be found in published literature, e.g. [16], [17]–[19]. However, most of these techniques are either too cumbersome or are not robust enough for the large range of source-observer distances that is of interest here.

For an efficient evaluation of I_1 and I_2 we propose a singularity cancellation technique by square-root change of variables $s = \sqrt{k_\rho^2 - k^2}$. If applied to (1), we obtain

$$I(k_{\rho 1}, k_{\rho 2}) = \int_{s_1}^{s_2} \mathrm{e}^{-s|z|} J_0(\rho\sqrt{s^2 + k^2})\mathrm{d}s, \tag{3}$$

where $s_{1,2} = \sqrt{k_{\rho 1,2}^2 - k^2}$. Note that the integrand in (3) is now nonsingular. This enables increase of accuracy and savings of integration points, and therefore of computation time. We notice that in the first subdomain, s is purely imaginary, whereas in the second subdomain, it is purely real. The integration paths before and after the change of variables are presented in the corresponding complex planes in Fig. 2. The singularity $k_\rho = k$ in the k_ρ-plane is mapped to the nonsingular point at the origin, $s = 0$, in the s-plane. It is worth noting that both integrand in (3) and its first derivative are bounded, despite the fact that the argument of the Bessel function, $\rho\sqrt{s^2 + k^2}$, suffers from the first-derivative discontinuity at $s = \mathrm{j}k$.

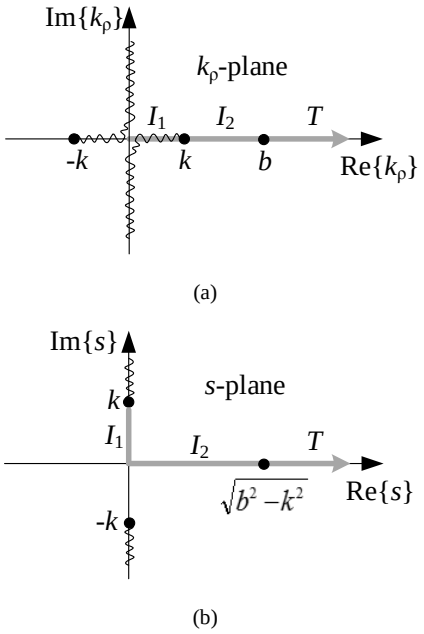


Fig. 2. Integration paths (grey line) and branch cuts (wavy line) for corresponding subdomain integrals in lossless media: (a) In k_ρ-plane, before the change of variables. (b) In s-plane, after the change of variables.

B. Lower endpoint of the tail subdomain

Among the methods for numerical integration of the tail, one can basically distinguish two approaches. The first one is the *integrate-and-sum* approach [11], [12], which proposes dividing the tail domain into equal segments. The partial integrals over these segments represent a series of numbers, the sum of which is to be evaluated. The second approach would be to evaluate the tail integral over the entire semi-infinite subdomain by means of a dedicated quadrature formula, which usually incorporates a variable transformation [6], [16]. In this paper, we adopt the former approach without using any change of variables for the tail. Integral T is approximated as a sequence of partial sums

$$\begin{aligned} T &\approx T_n = \sum_{i=1}^{n} t_i, \\ t_i &= I(b + i\Delta k_\rho, b + (i-1)\Delta k_\rho), \end{aligned} \tag{4}$$

where b is the tail beginning, Δk_ρ is the size of the segments and n is the number of segments. Partial integrals t_i in (4) can be evaluated using the GL quadrature. Depending on the segments' breakpoints, series t_i can be monotonous or alternating, and sequence T_n can converge slower or faster. In an extensive analysis we have considered several series-acceleration techniques: Shanks' transformation, Aitkin's Δ^2 process, Cesàro means [19]–[21] and weighted averages (WA) [11],[23]. The analysis has confirmed, as indicated in [11], [22], [23] that the combination of (i) *half-period* long segments, (ii) placement of breakpoints at the extrema [22], and (iii) application of WA acceleration algorithm is the optimal choice regarding complexity, convergence rate, computational cost and

robustness for a wide range of parameters ρ and z. The WA algorithm is based on a repeated (recursive) linear transformation of the sequence

$$T_n^{(l)} = \frac{T_n^{(l-1)} + \eta_n^{(l-1)} T_{n+1}^{(l-1)}}{1+\eta_n^{(l-1)}}, \qquad \begin{matrix} 1 \le n \le N \\ l = 1,2,...,n-1 \end{matrix} \tag{5}$$

where l is the level of the transformation, coefficient η_n is derived from the asymptotic expansion of the sequence [11] as

$$\eta_n^{(l)} \approx e^{z\pi/\rho}\left(1+\frac{2l}{b\rho/\pi+n}\right) \tag{6}$$

and N is the maximal reached number of segments, n, resulting from a certain accuracy criterion. Thus, the best estimate of the tail integral is $\hat{T} = T_1^{(N-1)}$.

Now, we propose setting b to the *first extremum after the second zero*, exceeding $k_\rho = k$. Note that, with such framework, b depends on ρ, which produces a *floating* endpoint of the tail subdomain. One important advantage of this framework is that the complexity of the integrand over the second subdomain $0 < k_\rho < b$ is kept low, with maximally one full oscillation occurring within. This implies that for given accuracy, small and rather constant number of integration points with respect to ρ will be required for the second subdomain. Additionally, in our experiments such choice of b has proven to be far enough from the branch-point singularity, which leaves the tail convergence unaffected. The extrema of Bessel function can be found approximately in different ways, e.g. through Newton-Raphson's rule or as mean of two neighboring zeros. In our study, we have used the extrema of the large-argument approximation

$$J_0(u) \approx \sqrt{\frac{2}{\pi u}}\cos\left(u-\frac{\pi}{4}\right), \tag{7}$$

thus, obtaining their positions at

$$k_{\rho,ex}(n) \approx b + n\frac{\pi}{\rho}. \tag{8}$$

Finally, the tail beginning is defined as

$$b = \left(\left(\left\lfloor \frac{k\rho}{2\pi} \right\rfloor + 2\right)\pi + \frac{\pi}{4}\right)\frac{1}{\rho}, \tag{9}$$

where $\lfloor \cdot \rfloor$ denotes the *floor* function.

C. Thresholds for truncation of tail integral

With purpose of control and savings in computing resources, we propose placing certain thresholds in the part of the algorithm, which is responsible for the calculation of the tail.

For large ratios z/ρ the coefficient η_n of the WA approach in (6) can reach arbitrarily high values. We, therefore, limit the maximum value of η_n to 10^{10}, thus avoiding working with extremely large numerical dynamic range.

If the ratio z/ρ is moderate or small, a cut-off number of required tail segments needs to be introduced. We use relative error for determination of the cut-off number. This criterion is introduced in a *while loop*, which halts further integration when the desired error level is reached.

However, for large *z*, the exponential factor of the integrand in (1) can be rather small, which allows us to discard the tail segments, the contribution of which is below certain threshold. We have set the threshold for the value of the exponential factor to 10^{-30} and obtained the limit for the tail subdomain

$$k_{\rho,\mathrm{th}} = \sqrt{k^2 + \left(\frac{-30}{|z|\log_{10}\mathrm{e}}\right)^2}. \tag{10}$$

We stress that value $k_{\rho,\mathrm{th}}$ can be smaller or larger than b, which, in former case, allows us to completely discard the tail, thus affecting only the integral of the second part, I_2. That way we save computational resources, which would be otherwise used for summing series of rather small numbers. The integral of the second part changes its limits and is then redefined as $I_2 = I(k, k_{\rho,\mathrm{th}})$.

D. Relative error

If an evaluation of our SI is denoted by A, then the relative error of that evaluation is defined as

$$RF = \frac{|A - A_{\mathrm{ref}}|}{|A_{\mathrm{ref}}|}, \tag{11}$$

where A_{ref} is some reference value. Additionally, we define the number of significant digits with which the evaluation is performed as $\chi = -\log_{10}(RF)$.

We point out that for the error analysis we choose the reference value always to be the total value of the GF, $A_{\mathrm{ref}} = g(r)$, regardless of the subdomain that is being analyzed. This is very important when analyzing the convergence of the tail integral, where very small cut-off number of required segments can be achieved, if the tail turns out to be numerically much less significant than the finite subdomains with respect to $g(r)$.

However, in the computation algorithm itself, the tail truncation is achieved using the relative error with respect to the preceding iteration as criterion function. In that case, the relative error is computed using $A = T_1^{(n)}$ and $A_{\mathrm{ref}} = T_1^{(n-1)}$, where index n designates the number of half-period segments included in the evaluation of the SI, according to (4) and (5).

III. Analysis of Subdomain Partial Integrals

In this section we provide a short overview of qualitative and quantitative properties of the SI in terms of relative error, using GL quadrature formula. For each of the three parts we propose formulas that estimate the required number of integration points for a given accuracy. The space of interest is a 2-D set of distance coordinates, given by $(\rho, z) \in [10^{-3}\lambda, 10^{3}\lambda] \times [10^{-3}\lambda, 10^{3}\lambda] = D$.

A. First subdomain integral

In the first subdomain, where $k_\rho \in [0, k]$, the oscillations occur due to the imaginary nature of the exponent and due to the Bessel function factor in (3). According to (11), the relative error for the first subdomain can be determined for the observable $A(n_{g1}) = \hat{I}_1(n_{g1}) + \tilde{I}_2 + \tilde{T}$, where $\hat{I}_1$ is the evaluation of the integral over the first subdomain, n_{g1} is the number of integration points used and $\tilde{I}_2$ and $\tilde{T}$ are the second subdomain and tail integral evaluations, respectively. Integrals $\tilde{I}_2$ and $\tilde{T}$ are computed with the maximum possible accuracy using enough GL integration points and tail half-periods. In Fig. 3, we show the relative error for the first subdomain with respect to the number of integration points in a subset of D, defined by the line $\rho = z$. In general, we can distinguish three characteristic parts of the relative-error curve: (i) the initial part, where the relative error is larger than one, (ii) the steep (decreasing) part and (iii) the saturation part where some minimal error level ($10^{-16} \div 10^{-10}$) is reached. The slope of the steep part is changing moderately with distance, while the threshold, after which the error starts decreasing abruptly, depends strongly on ρ and z.

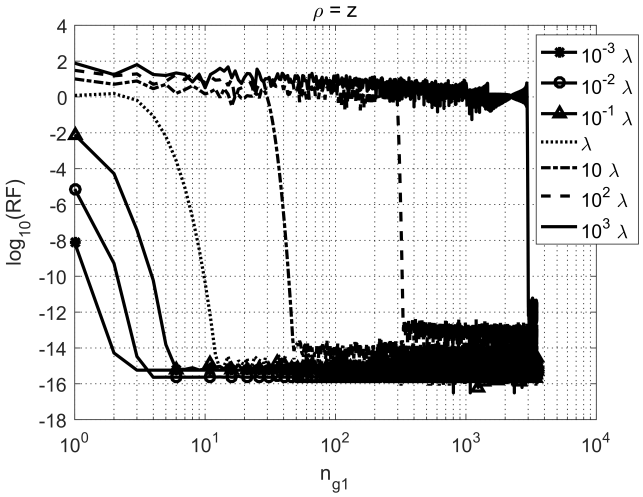


Fig. 3. Relative error of the first subdomain integration vs. number of Gauss-Legendre integration points for different values of ρ (ρ = z).

After a series of simulations and a thorough analysis, we have assembled an empirical expression for estimated number of integration points that is required to reach given accuracy in the first subdomain

$$\hat{n}_{g1} \approx 6 + R_e + \frac{\chi_{\text{desired}}}{3} R_e^{0.3}, \qquad (12)$$

where $R_e = \sqrt{(2.4\rho/\lambda)^2 + (1.8z/\lambda)^2 + 0.4\rho z/\lambda^2}$ and χ_{desired} is the desired number of significant digits.

B. Second subdomain integral

The second subdomain is defined by $k_\rho \in [k, b]$. When the proposed change of variables is applied, we obtain real integrand and real argument. Unlike in the case of the first subdomain, the integrand in the second subdomain changes moderately with respect to k_ρ, but also to s. Parameter ρ dictates the position of breakpoint b, which can impact the size of the subdomain, yet the number of oscillations within is limited, thanks to the floating definition of the breakpoint. On the other hand, parameter z dictates only the damping of the integrand function. For the second subdomain we perform the same error analysis as in previous subsection. We consider the observable $A(n_{g2}) = \tilde{I}_1 + \hat{I}_2(n_{g2}) + \tilde{T}$, and the reference is again $A_{\text{ref}} = g(r)$. Magnitude $\hat{I}_2$ is the evaluated integral over the second subdomain and n_{g2} is the corresponding number of integration points. Integrals $\tilde{I}_1$ and $\tilde{T}$ are computed with the maximum possible accuracy using enough GL integration points and tail half-periods. The plot of the relative error for the second subdomain is given in Fig. 4 for $\rho = z$. We notice that the error characteristic is rather canonical and confined, which implies moderate dependence on distance coordinates within the given set. The error curves can be approximated by a family of slopes which are limited by two extreme cases, roughly determined by $\rho > z$ and $\rho < z$, where the latter case corresponds to the slower convergence.

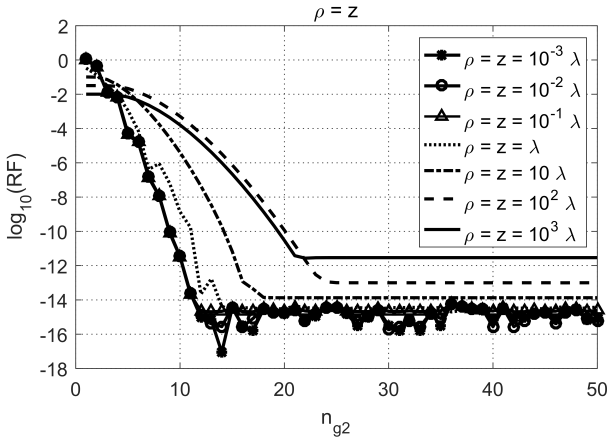


Fig. 4. Relative error of the second subdomain integration vs. number of Gauss-Legendre integration points for different values of ρ (ρ = z).

The numerical nature of the integrand discussed above and performed simulations over D allow us to form an approximate expression for prediction of number of integration points, required for achieving given accuracy in the second subdomain

$$\hat{n}_{g2} \approx \begin{cases} N_1, & \log_{10}\left(\frac{z}{\rho}\right) < B \\ N_1 + (N_2 - N_1)\left(\log_{10}\left(\frac{z}{\rho}\right) - B\right), & \text{elsewhere} \\ N_2, & \log_{10}\left(\frac{z}{\rho}\right) > C \end{cases} \quad (13)$$

where $N_1 = \chi_{\text{desired}} + 3$, $N_2 = N_1 + 10$, $B = -1.2$, and $C = -0.2$. Note in (13) that the two levels, N_1 and N_2, correspond to the slowest and fastest slopes in Fig. 3.

C. Tail subdomain integral

In the tail subdomain, the integrand and, more importantly, the exponent in (1) are real (Fig. 1), making a non-monotonous decreasing function whose damping is dictated by both ρ and z, whereas its oscillation depends exclusively on ρ. For computation of the tail, we do not use any change of variables, but rely on its convergence properties. The tail evaluation is based on partial integrals over half-period segments, yet the shape of the integrand within them is not qualitatively affected by the distance coordinates ρ and z, thus a very weak dependence on the coordinates is expected. However, it should be noted that there is a strong relation to ρ and z in the summation phase of the tail computation, due to their direct impact on the convergence of sequence in (4). Therefore, we propose an expression for approximate number of integration points per half-period that is needed for given accuracy in terms of number of significant digits, χ_{desired}, as

$$\hat{n}_{g,hp} \approx \frac{\chi_{\text{desired}}}{2} + 2. \quad (14)$$

IV. Efficiency of Complete-Domain Integration

In this section we assemble together all three parts of the integral defined in (2) and then look into their total impact on the error level. Based on the analysis in section III, we can fairly predict the required number of integration points for each part, where the number of half-period segments for the tail is automatically determined, for any given error level, RF_{desired}. We will show how well the proposed prediction matches the computation.

In order to verify the algorithm as a whole, we compute the total relative error for given accuracy, RF_{desired}, based on the predicted number of integration points as

$$RF_{\text{tot}}(\rho, z) = \frac{\hat{I}_1(\hat{n}_{g1}) + \hat{I}_2(\hat{n}_{g2}) + \hat{T}(\hat{n}_{g,hp}) - g(r)}{g(r)}. \quad (15)$$

where $\hat{I}_1$, $\hat{I}_2$ and $\hat{T}$ are the respective integral parts obtained by numerical integration, $\hat{n}_{g1}$, $\hat{n}_{g2}$ and $\hat{n}_{g,hp}$ are calculated with empirical equations (12), (13), and (14), respectively. In Fig. 5, the obtained total relative error is shown in the entire set D for $RF_{\text{desired}} = 10^{-10}$. We can see that the error is slightly bellow the desired limit. In Fig. 6 the total estimated number of integration points, $\hat{n}_{g,tot} = \hat{n}_{g1} + \hat{n}_{g2} + N\,\hat{n}_{g,hp}$, is presented with respect to distances, where N is the reached number of tail half-periods. In addition, Fig. 7 shows percentage of the relative difference between the estimated and the exact minimal required number of integration points for the given accuracy,

$$\delta(n_{g,tot})[\%] = 100 \cdot \frac{\hat{n}_{g,tot} - n_{g,tot}}{n_{g,tot}}. \quad (16)$$

It is evident that the estimated number of integration points is only about 10 % higher than the exact number of required points in almost entire space D. The largest error of the estimation occurs in a narrow region where ρ and z are small and of similar values. However, this is not critical, since the required number of integration points in that region is rather low.

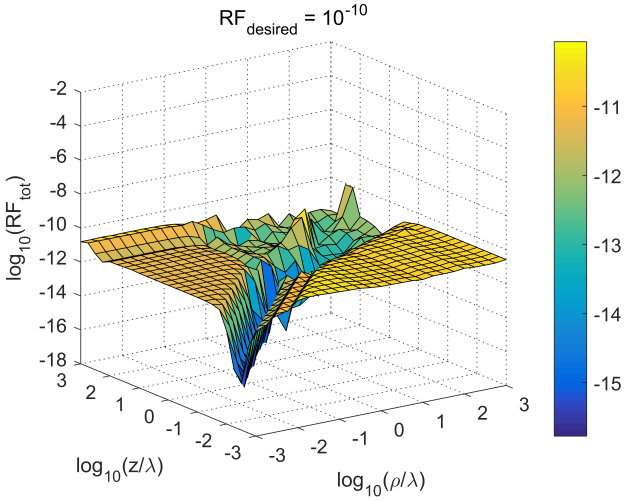


Fig. 5. Total relative error for $RF_{\text{desired}} = 10^{-10}$ vs. normalized ρ and z.

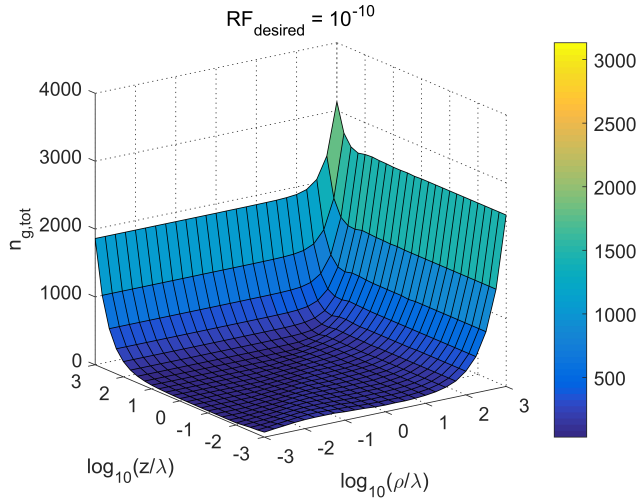


Fig. 6. Total estimated number of integration points for $RF_{\text{desired}} = 10^{-10}$ vs. normalized ρ and z.

We can conclude that for the given change of variables, the proposed equations yield a robust and efficient prediction of

the required number of integration points for free-space SIs.

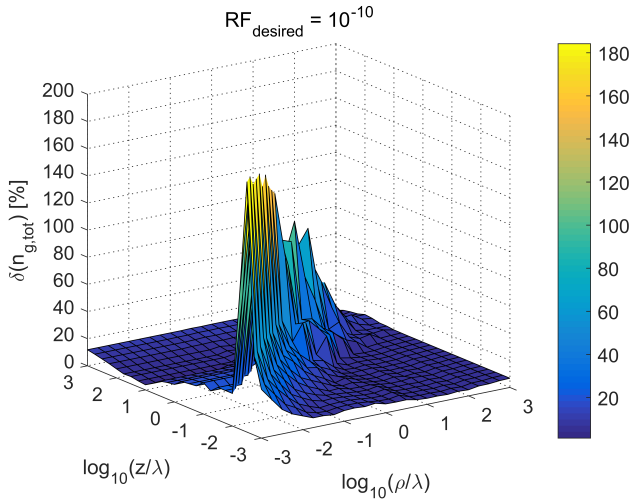


Fig. 7. Relative difference of the estimated and exact required number of integration points vs. normalized ρ and z.

V. Conclusion

A method for computation of pole-free Sommerfeld integrals for a wide range of source-observer distances is presented using the example of free-space Green's function in lossless media. The integral is divided in a standard manner by the branch-point singularity and the lower end point of the tail.

The branch-point singularity is cancelled using a novel variable transform, thus minimizing the number of Gauss-Legendre integration points required for desired accuracy.

The tail beginning is defined as the first extremum after the second zero exceeding the branch point singularity. In this way the tail integral is unaffected by the singularity and the complexity of the second-subdomain integral is kept low. The tail computation relies on the established weighted-averages algorithm that is proven to be very efficient within the given framework. The elements of the number series, to which the weighted-averages method is applied, are partial integrals calculated over intervals defined by the successive extrema, which are simply determined using the sinusoidal large-argument approximation of the Bessel function of the first kind. The definition of thresholds for the weights of the weighted-averages algorithm and for the truncation of the integral provides additional robustness, accuracy and savings of integration points.

Overall, an efficient evaluation of the integral is achieved for a wide range of source-observer distances, up to 1000 free-space wavelengths. Since the Sommerfeld integrals that belong to the same representation have similar numerical properties, the techniques proposed for the free-space case can be transferred to other scenarios, e.g. the half-space problems.

References

[1] R Trembinski and D. A. McNamara, "The engineering modelling of electromagnetic wave scattering from sea ice by surface-based radar", *2018 IEEE Int. Symp. Ant. Prop. (APS/URSI)*, Boston, MA, Jul. 2018.

[2] B. M. Kolundzija, M. S. Tasic, D. I. Olcan, D. P. Zoric, and S. M. Stevanetic, "Advanced techniques for efficient modeling of electrically large structures on desktop PCs," *Applied Computational Electromagnetics Society Journal, Special Issue on Computational Electromagnetics Workshop, CEM 11*, Vol. 27, No. 2, pp. 123-131, Feb. 2012.

[3] A. N. Sommerfeld, "Über die Ausbreitung der Wellen in der Drahtlosen Telegraphie," in *Ann. Physik*, vol. 28, pp. 665-736, 1909.

[4] H. Weyl, "Ausbreitung elektromagnetischer Wellen über einem ebenen Leiter," *Ann. Phys.* vol. 60, pp. 481-500, 1919.

[5] Van der Pol and K. F. Niessen, "Über die Ausbreitung von Electromagnetischer Wellen ueber eine Ehene Erde," *Ann. Phys.,6*, pp. 273-294, 1930.

[6] J R. Mosig and F E. Gardiol, A Dynamical Radiation Model for Microstrip Structures, In: Peter W. Hawkes, Editor(s), *Advances in Electronics and Electron Physics*, Academic Press, 1982, Volume 59, pp. 139-237.

[7] P. Gay-Balmaz, and J. R. Mosig, Three-dimensional planar radiating structures in stratified media. *Int. J. Microw. Mill.-Wave Comput.-Aided Eng.*, 7, pp. 330–343, 1997.

[8] E. Simsek, Qing Huo Liu and Baojun Wei, "Singularity subtraction for evaluation of Green's functions for multilayer media," in *IEEE Transactions on Microwave Theory and Techniques*, vol. 54, no. 1, pp. 216-225, Jan. 2006.

[9] R. Golubovic-Niciforovic, A. G. Polimeridis and J. R. Mosig, "Fast computation of Sommerfeld integral tails via direct integration based on double exponential-type quadrature formulas," in *IEEE Transactions on Antennas and Propagation*, vol. 59, no. 2, pp. 694-699, Feb. 2011.

[10] R. Golubovic, A. G. Polimeridis and J. R. Mosig, "Efficient algorithms for computing Sommerfeld integral tails," in *IEEE Transactions on Antennas and Propagation*, vol. 60, no. 5, pp. 2409-2417, May 2012.

[11] K. A. Michalski, "Extrapolation methods for Sommerfeld integral - tails," in *IEEE Transactions on Antennas and Propagation*, vol. 46, no. 10, pp. 1405-1418, Oct 1998.

[12] J. Mosig, "The weighted averages algorithm revisited," in *IEEE Transactions on Antennas and Propagation*, vol. 60, no. 4, pp. 2011-2018, Apr. 2012.

[13] A. Álvarez Melcón, "Applications of the integral equation technique to the analysis and synthesis of multilayered printed shielded microwave circuits and cavity backed antennas," Ph.D. dissertation EPFL, Lausanne, Switzerland, 1998.

[14] N. Alexopoulos and D. Jackson, "Fundamental superstrate (cover) effects on printed circuit antennas," in *IEEE Transactions on Antennas and Propagation*, vol. 32, no. 8, pp. 807-816, Aug. 1984.

[15] K. A. Michalski and J. R. Mosig, "The Sommerfeld half-space problem revisited: from radio frequencies and Zenneck waves to visible light and Fano modes," *Journal of Electromagnetic Waves and Applications* Vol. 30, Iss. 1, 2016.

[16] I. D. Koufogiannis, A. G. Polimeridis, M. Mattes and J. R. Mosig, "Real axis integration of Sommerfeld integrals with error estimation," *2012 6th European Conference on Antennas and Propagation (EUCAP)*, pp. 719-723, Prague, 2012.

[17] V. Volskiy, G. A. E. Vandenbosch, R. G. Nićiforović, A. G. Polimeridis and J. R. Mosig, "Numerical integration of Sommerfeld integrals based on singularity extraction techniques and double exponential-type quadrature formulas," *2012 6th European Conference on Antennas and Propagation (EUCAP)*, Prague, 2012, pp. 3215-3218.

[18] W. A: Johnson and D. G. Dudley, "Real axis integration of Sommereld integrals: Source and observation points in air", *Radio Sci.*, vol. 18, no. 2, pp. 175-186, 1983.

[19] V. V. Petrovic, A. J. Krneta and B. M. Kolundzija, "Singularity extraction for reflected Sommerfeld integrals over a multilayered media," *2013 21st Telecommunications Forum Telfor (TELFOR)*, Belgrade, 2013, pp. 648-651.

[20] G. H. Hardy, *Divergent Series*, Oxford University Press, London, 1973.

[21] E. J. Weniger, "Nonlinear sequence transformations for the acceleration of convergence and summation of divergent series," in *Comput. Phys. Rep.*, vol. 10, pp. 189–371, Dec. 1989.

[22] S. K. Lucas and H. A. Stone, "Evaluating infinite integrals involving Bessel functions of arbitrary order," in *Journal of Computational and Applied Mathematics*, vol. 64, no. 3, pp. 217-231, Dec. 1995.

[23] K. A. Michalski and J R. Mosig, "Efficient computation of Sommerfeld integral tails - methods and algorithms", *JEMWA*, vol. 30, no. 3, pp. 281-317, 2016.

The Influence of Corona on the Lightning Surge Propagation Along Transmission Lines

Milan Ignjatović, Jovan Cvetić and Dragan Pavlović

***Abstract*— Corona is the partial discharge that occurs around the wires and edges in inhomogeneous electric field. Minimum intensity of the electric field for the impact ionization of the gas is around 2.6 MV/m in dry air. In power systems, the corona is the unwanted effect caused by overvoltages. In this study the propagation of the overvoltage wave due to negative lightning along the transmission line is numerically simulated. The effect of the corona is modeled by the drift-diffusion-reaction equations for the electrons, the positive and the negative ions.**

***Index Terms*— Corona, lightning overvoltage, transmission lines, QV curve, streamers**

I. Introduction

THE lightning strike is one of the most common causes of power interruption on transmission lines and substations. Since it is a natural phenomenon, lightning occurrence can not be foreseen, but its performance on power systems can be estimated. The insulation strength is selected to minimize the risk of the failure, taking into account the statistical data of the lightning occurrence in the wider area and the cost of the insulation construction [1]. There are methods to estimate the magnitude and steepness of the average incoming surge that arrives at substations [2].

When the lightning strikes the transmission lines, two initiating events can happen - the shielding failure and the backflash. The shielding failure occurs when the lightning hits the phase conductor directly avoiding the shield wire. If the lightning hits the shield wire, the injected overvoltage can be greater than the critical flashover voltage of the insulation and the backflash occurs.

For the direct lightning strikes the main effect that influence the shape of the overvoltage waveform is the corona. It is a partial discharge that occurs around the conducting wire when the voltage is above the corona inception threshold described by the Peek's formula [3]. The corona effects the delay of the part of the impulse that is above the corona threshold and the attenuation of the overvoltage amplitude due to losses in the ionized gas in the area around the wire.

The main goal of the paper is to numerically simulate the propagation of the overvoltage wave along the transmission line including the corona. The effect of corona is taken into consideration by the drift-diffusion-reaction model which represents the set of the continuity equations for each type of particles (electrons, positive and negative ions, neutrals, etc.) that participate in the discharge. The model enables the inclusion and the analysis of detailed physical and chemical reactions that are taking place in the area around the wire of the transmission line.

Milan Ignjatović is with the School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail: ignjatovic@etf.rs).

Jovan Cvetić is with the School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail: cvetic_j@etf.rs).

Dragan Pavlović is with the School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail: dragan.lab3@etf.rs).

II. Theory

The equations describing voltage and current along the overhead wire taking into account the effect of corona [4] are

$$\frac{\partial V(t,x)}{\partial x} + L\frac{\partial I(t,x)}{\partial t} = 0, \qquad (1)$$

$$\frac{\partial I(t,x)}{\partial x} + \frac{\partial Q(t,x)}{\partial t} = 0, \qquad (2)$$

These equations represent telegrapher's equations for lossless transmission line where $Q(t,x)$ is the total line charge density. In vacuum, the charge-voltage dependence is linear $Q(t,x) = CV(t,x)$. The inductance and the capacitance per line length, L and C respectively, can be calculated by formula

$$L = \frac{\mu_0}{2\pi}\operatorname{acosh}\left(\frac{h}{a}\right),\ C = 2\pi\varepsilon_0 \Big/ \operatorname{acosh}\left(\frac{h}{a}\right), \qquad (3)$$

where a is the wire radius and h is the height of the wire above the ground.

When the voltage intensity is greater than the critical value for the corona inception, the charge voltage dependence becomes nonlinear with the hysteresis. This function is represented by charge-voltage (QV) curve. In order to study the pulse propagation, it is necessary to know the accumulated charge on the transmission line for the given voltage when corona is formed and to determine the QV curve for every position along the transmission line. For the given time evolution of voltage at some position along the line $V(t,x)$, one can obtain the total line charge density Q by solving the continuity equations for charged species in air in cylindrical geometry [5]

$$\frac{\partial n_e}{\partial t}+\frac{1}{r}\frac{\partial\left(r\Gamma_e\right)}{\partial r}=S_{ph}+n_e(\alpha-\eta_2-\eta_3)\left|W_e\right|-n_e n_p\beta+k_{\text{det}}n_{O_2^-} \quad (4)$$

$$\frac{\partial n_p}{\partial t}+\frac{1}{r}\frac{\partial\left(r\Gamma_p\right)}{\partial r}=S_{ph}+n_e\alpha\left|W_e\right|-(n_e+n_{O^-}+n_{O_2^-})n_p\beta\,, \quad (5)$$

$$\frac{\partial n_{O^-}}{\partial t}+\frac{1}{r}\frac{\partial\left(r\Gamma_{O^-}\right)}{\partial r}=n_e\eta_2\left|W_e\right|-n_{O^-}n_p\beta\,, \quad (6)$$

$$\frac{\partial n_{O_2^-}}{\partial t}+\frac{1}{r}\frac{\partial\left(r\Gamma_{O_2^-}\right)}{\partial r}=n_e\eta_3\left|W_e\right|-n_{O_2^-}n_p\beta-k_{\text{det}}n_{O_2^-} \quad (7)$$

In this study, we take into consideration the concentration of electrons n_e, positive ions n_p, negative O^- and O_2^- ions, n_{O^-} and $n_{O_2^-}$, respectively. On the right hand side of the equations (4)-(7) the terms representing the gain and the loss of the particles due to electron impact ionization, two-body attachment, three-body attachment, recombination and detachment described by the coefficients α, η_2, η_3, β and k_{det} respectively are given. The term S_{ph} denotes the generation of electrons and positive ions through photoionization. Further, W_e, W_p and W_n are the velocities of electrons, positive and negative ion drifts, respectively. The fluxes of charged particles are given by

$$\Gamma_e=W_e n_e-D\frac{\partial n_e}{\partial r},\ \Gamma_p=W_p n_p,\ \Gamma_{O^-}=W_n n_{O^-},\ \Gamma_{O_2^-}=W_n n_{O_2^-}, \quad (8)$$

where Γ_e is the flux of electrons, Γ_p is a flux of positive ions, Γ_{O^-} and $\Gamma_{O_2^-}$ are the fluxes of O^- and O_2^- ions, respectively, D is the diffusion coefficient for electrons. We neglected the diffusion of heavy ions, so the ionic current contains only a drift component. The values of the transport and the reaction coefficients depend on the ratio of the value of the electric field intensity and the concentration of air molecules [6], so the continuity equations (4)-(7) are coupled with the Poisson equation using the potential Φ

$$\nabla^2\Phi=\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial\Phi}{\partial r}\right)=-\frac{e\left(n_p-n_e-n_{O^-}-n_{O_2^-}\right)}{\varepsilon_0}\,. \quad (9)$$

Electric field is calculated assuming the electrostatic approximation

$$E=-grad\ \Phi=-\frac{\partial\Phi}{\partial r}\,. \quad (10)$$

When the concentrations of the charge particles and the electric field intensity are calculated, the total line charge density is

$$Q(t)=\int_0^t I_p(t')dt', \quad (11)$$

where I_p is the current per cylinder length, obtained from Sato-Morrow [7] formula for cylindrical geometry

$$I_p=\frac{2\pi e}{\log\frac{R_2}{R_1}}\int_{R_1}^{R_2}\Gamma dr+\frac{2\pi\varepsilon_0}{\log\frac{R_2}{R_1}}\frac{\partial V}{\partial t}. \quad (12)$$

Final goal is to investigate whether the drift-diffusion model in the cylindrical geometry can be used for the simulation of QV curves for the overvoltage calculations in different geometry for transmission lines. Noda [8] measured the QV curves for the overhead wire 1.83 m above the ground. In order to try to estimate the QV curve measured by Noda in wire to plane geometry we ran the simulation in our wire to cylinder program. The radius of the wire is same as in the experiment, $R_1=5$ mm, and the radius of the outer cylinder is adopted to be the height of the wire above the ground, which is $R_2=1.83$ m. Calculated values for the total line charge density were higher than the measured values, which is expected since the capacitance of the cylindrical capacitor is greater than the capacitance of the wire above the perfectly conducting plane. Therefore, the result for the line charge density should be multiplied by the coefficient k, which is the ratio of the capacitance of the wire above the plane C_p and the capacitance of the cylindrical capacitor C_0

$$C_p=\frac{2\pi\varepsilon_0}{\cosh^{-1}\left(R_2/R_1\right)},\ k=\frac{C_p}{C_0}. \quad (13)$$

In this way, we obtained a very good agreement with the experimental results. The QV curve obtained by the correction of the calculated results and the data measured by Noda for 570 kV impulse are depicted in Fig. 1.

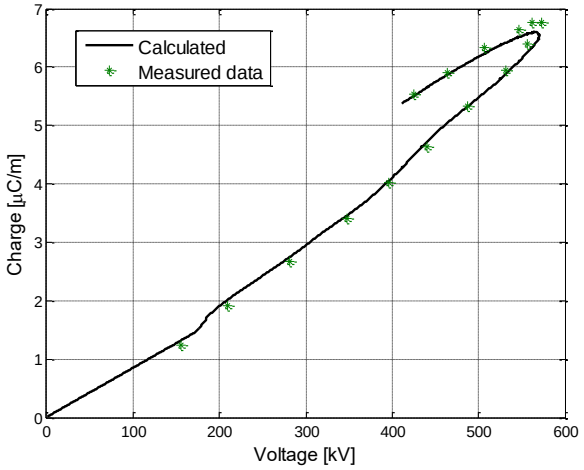


Fig. 1. Calculated QV curve and the measured data by Noda [8]

The final result of the simulation shows excellent agreement with the experimental measurement of the QV curve. It represents affirmation that the drift diffusion model can be used for the calculation of QV curves, which are necessary to complete the equations (1) and (2) in order to study the pulse propagation along the transmission line.

In order to solve the equations (4)-(7) and (9) numerically, great number of time steps is necessary. For the simulation of the 5 μs corona discharge, a million time steps is needed.

Therefore, it would be very time consuming to simulate the corona at every position along the transmission line. So, the corona discharge is simulated only on some nodes along the line and the values between those nodes are obtained by linear interpolation.

III. Results and Discussion

The calculations have been performed for the single overhead wire of the radius $a = 21\,\text{mm}$ placed at the height $h = 15\,\text{m}$ above the ground. The goal is to try to obtain the same results as in the experiments preformed by Wagner et al [9]. The surge generator was located at the one end of the line and the voltage pulse is injected which defines the boundary condition $V(x=0,t)$. Since the pulse can not propagate with the speed greater then the speed of light, two more boundary conditions can be defined $I(x > ct) = 0,\ V(x > ct) = 0$.

The surge generator voltage impulse and the shape of the voltage impulse after the wave travelled the distance of 622 m are given in the Fig. 2. Unfortunately, first result of the simulation did not give the solution that has satisfactory agreement with the measured impulse. On a positive note, simulation confirmed all effects of the corona- the retardation and the attenuation of the impulse and correct value of the corona threshold.

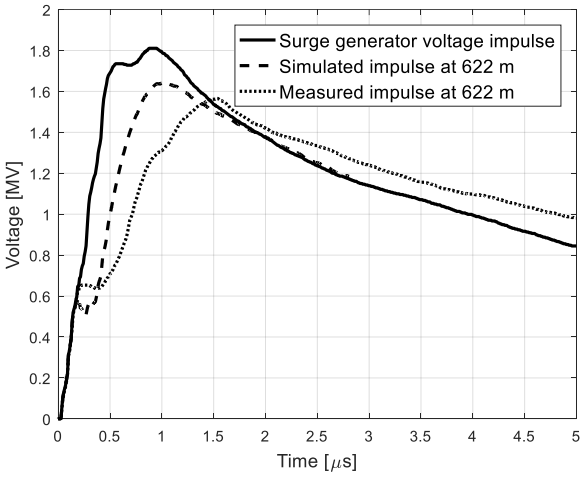


Fig. 2. Simulation of the propagation of the surge pulse at 622 m

In order to obtain better results, we conducted the investigation why results are different. By comparing results of the QV curves with the engineering models it was concluded that drift-diffusion model gives lower values of the charge for wires with the radius greater than 10 mm. For thinner wires drift-diffusion model gives satisfactory solution, as shown in Fig. 1. Therefore, in our example the generated charge is less than it should be and corona effect is not strong enough to obtain agreement with the measured values.

This difference in the results for the thin and wide wires can be explained by the effect of streamer. The structure of the space charge around the wire is not the radially symmetrical fluid, as it is assumed by drift-diffusion model. In reality, the space charge is composed of thin filaments of the discharge – the streamers. These filaments have the head where photoionization occurs and they propagate through space leaving behind the weakly ionized tail [10]. Their significant characteristic is that they can propagate even when the electric field intensity is smaller than the critical electric field for the impact ionization (around 2.6 MV/m for dry air). The minimal value of the electric field for the propagation of the positive streamers is 0.5 MV/m, and for the negative streamers it is around 1 MV/m.

Due to streamers additional charge is generated in the area farther form the wire where the electric field has lower intensity values, which can not be taken into account with the original drift-diffusion model. In order to include this effect, we modified the values of the Townsend coefficient α which represents the number of positive ion-electron pairs generated per unit length. For the atmospheric pressure, the Townsend coefficient depends only on the electric field intensity. The modification is done so that the Townsend coefficient has values high enough for the ionization when the electric field intensity is lowered to 1 MV/m. This procedure gave the result that is in a good overall agreement with the experimental results, as can be seen in Fig. 3.

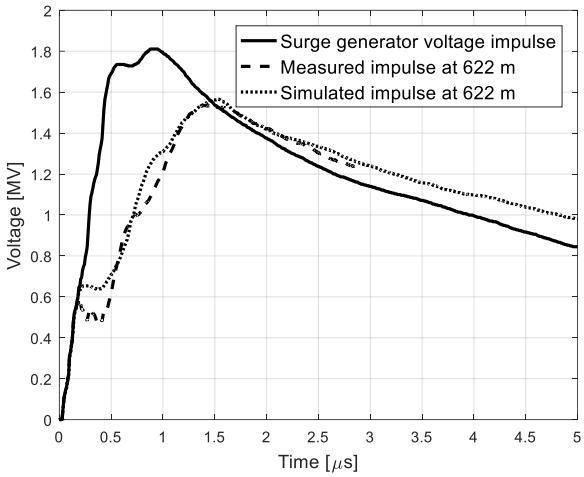


Fig. 3. Simulation of the propagation of the surge pulse at 622 m with the modified Townsend coefficient value

IV. Conclusion

The drift-diffusion model has been used to simulate corona discharge due to lightning surge on transmission lines. In order to obtain a good agreement with the experimental results, reaction coefficient for the impact ionization needs to be modified. It is necessary to include the effect of the streamers, the thin filaments of the discharge that can propagate in the area of lower electric field intensity. The streamers increase the generated charge around the wire and enhance the effects of the corona.

Acknowledgment

Ministry of Science and Technological Development of the Republic of Serbia supported this work under contracts No. 171007 and 37019.

REFERENCES

[1] A. R. Hileman, *Insulation Coordination for Power Systems,* Boca Raton, USA, Taylor & Francis, 1999

[2] K. H. Weck and A. J. Eriksson, @simplified Procedures for Determining Representative Substation Impinging Lightning Overvoltages", CIGRE Paper 33-16, 1988

[3] F. W. Peek, *Dielectric phenomena in HV engineering*, McGraw-Hill, 1929

[4] V. Cooray an N. Theethayi, "Pulse Propagation Along Transmission Lines in the Presence of Corona and Their Implication to Lightning Return Strokes", IEEE Transactions on Antennas and Propagation, Vol. 56, No. 7, pp. 1948-1959, July 2008.

[5] A. Luque, U. Ebert, "Density models for streamer discharges: Beyond cylindrical symmetry and homogenous media", Journal of Computational Physics, Vol. 231, 2012, pp. 904-918.

[6] R. Morrow, J. J. Lowke, "Streamer propagation in air", J. Phys. D: Appl. Phys., Vol. 30, pp. 614-627, 1997

[7] R. Morrow, N. Sato, "The discharge current induced by the motion of charged particles in time-dependent electric fields; Sato's equation extended", J. Phys. D: Appl. Phys, Vol.32, pp. L20-L22, 1999.

[8] T. Noda, T. Ono, H. Matsubara, H. Motoyama, S. Sekioka and A. Ametani, "Charge-Voltage Curves of Surge Corona on Transmission Lines: Two Measurement Methods", IEEE Transactions on Power Delivery, Vol. 18, No. 1, pp. 307-314, Jan. 2003.

[9] C.F.Wagner, I. W. Gross, B. L. Lloyd, "High-Voltage Impulse Tests on Transmission lines", AIEE Trans., 1954., pp. 196-210.

[10] Y. Raizer: "*Gas discharge physics*", Springer-Verlag, Berlin, Germany, 1991

UWB Printed Monopole Antenna With and Without the Reflector

Dragan Nikolić and Miodrag Tasić, *Member IEEE*

***Abstract* - Measured reflection coefficient of a typical UWB monopole antenna is heavily dependent on the measurement setup. The antenna is connected to the network analyzer (NA) using coaxial cables. During the measurement, currents along the inner and the outer conductor of the coaxial cables are unbalanced, causing parasitic radiation. The NA chassis also radiates, so the entire measurement setup behaves like a single antenna under test. As a consequence, a simulated results can be very different from the measured ones, and the single antenna can behave very differently when mounted in different environments. Also, minor errors in the antenna fabrication can cause significant differences in the measured reflection coefficients of similar prototypes.**

In this paper we examine if a reflector connected to the ground of the UWB monopole antenna can provide stabilization in the antenna characteristics, simulated and measured, and the satisfactory matching between reflection coefficients of different prototypes.

***Index term* - UWB antenna; antenna measurements; reflection coefficient; electromagnetic modeling.**

I. INTRODUCTION

The rapid development of UWB (Ultra-Wideband) technology [1] begins as the appropriate standards were adopted in the United States [2] and Europe [3]. UWB devices work in the frequency range from 3.1 to 10.6 GHz (FCC band).

A printed monopole antenna with a circular radiating element [4] can achieve a reflection coefficient less than -10dB, across the entire UWB band. With smaller changes in the shape of the radiating element and the ground plane, it is possible to improve the characteristics of the antenna, but the basic structure of the antenna remains unchanged: one side of the antenna is radiating element with approximately circular shape and a feeder, the other is a substrate with a small ground plane. Minor errors in the antenna fabrication can lead to significant differences between reflection coefficients (return loss) of different prototypes (for the same antenna model). Size and construction of such antenna are the main reason for the difference in results.

Dragan Nikolić – School of Electrical Engineering, University of Belgrade, Kralja Aleksandra Blvd. 73, 11120 Belgrade, Serbia (e-mail: nikolicdragansiki@gmail.com).

Miodrag Tasić – School of Electrical Engineering, University of Belgrade, Kralja Aleksandra Blvd. 73, 11120 Belgrade, Serbia (e-mail: tasic@ etf.bg.ac.rs).

Diameter of the radiating element is about 22 mm, which is about 4 times less than the wavelength at the highest frequency in UWB band. Small ground of antenna causes the current to flow back from the radiator to the outer surface of the coaxial cable. This results in secondary radiation which can influence measurement accuracy [5].

In this paper we examine if a reflector connected to the ground plane of the UWB printed monopole antenna can increase stability of the simulated and measured results, and also lead to the satisfactory matching between different prototypes of the same antenna model. We start with the UWB printed monopole antenna without the reflector. The antenna was modeled using software for polygonal modeling AW Modeler [6], whereas the electromagnetic analysis of the model was performed in the software package for electromagnetic modeling and analysis WIPL-D Pro [7]. The radiating element was modeled using NURBS curves, as explained in [8]. Two prototypes were fabricated and measured, showing considerable differences in the reflection coefficients. Then, we added reflectors to both antennas, and the matching between measured reflection coefficients was excellent.

Different simulation models and measured results of these antennas are shown in the following section.

II. UWB ANTENNA DESIGN

Initial UWB printed monopole antenna (without a reflector) was modeled in [8]. The antenna was made using the substrate with relative permittivity 3,5 and thickness 0,762 mm. The radiating element was modeled using NURBS curve, with three different parameters for the shape control. The antenna is fed trough a microstrip line attached to a SMA connector Optimal design is chosen in a way to minimize reflection coefficient in the frequency range of interest (from 3.1 to 10.6 GHz). The layout for the optimal design is shown in Fig. 1.

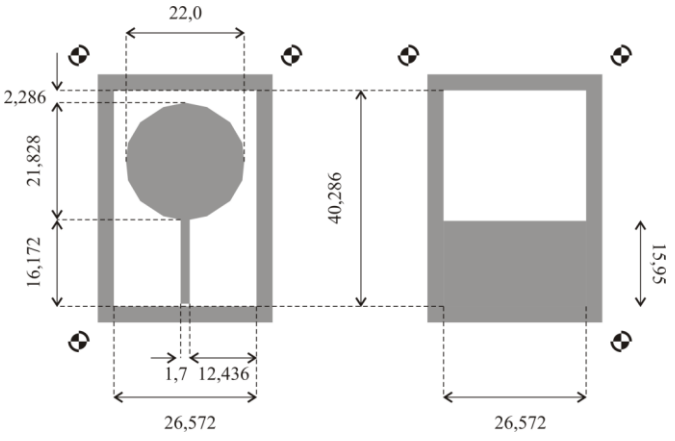


Fig. 1. Layout of UWB printed monopole antenna (top and bottom view)

The model of the antenna, with precise modeling of a SMA connector, is shown in Fig. 2.

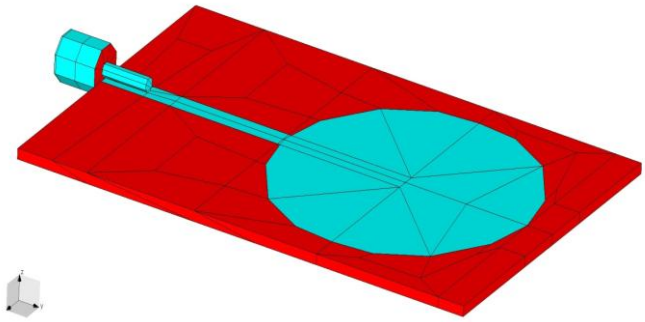


Fig. 2. Model of UWB printed monopole antenna without reflector

The simulated return loss of the UWB printed monopole antenna without reflector is shown in Figure 3. Return loss of the modeled antenna is practically less than -10dB across the entire UWB band.

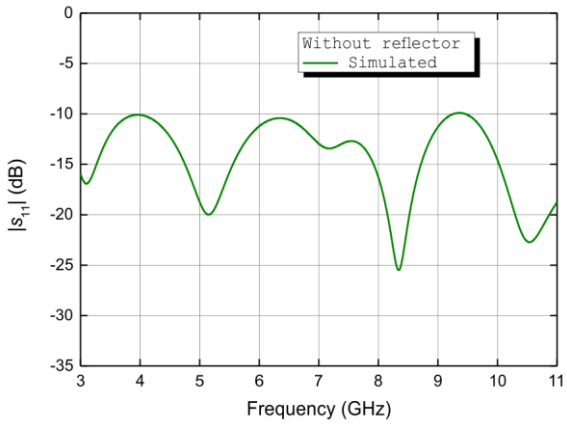


Fig. 3. Simulated return loss of UWB printed monopole antenna without the reflector

Two prototypes were fabricated, and the photo of the antennas, from both sides, with SMA connectors attached, is shown in Fig. 4.

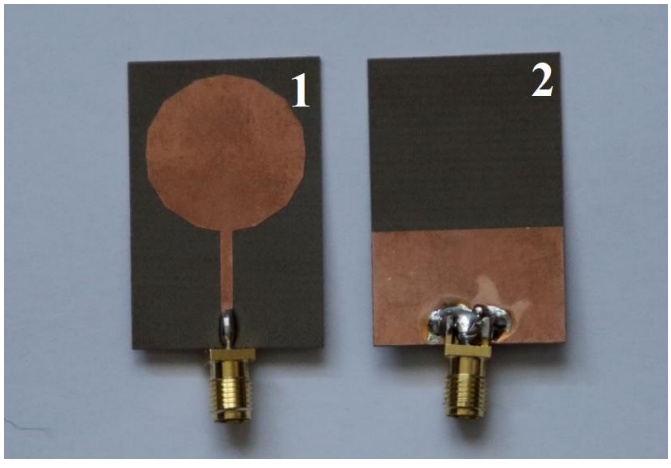


Fig. 4. Photograph of prototypes without the reflector (top view of the first prototype and bottom view of the second prototype)

The measured return loss of the prototypes are shown in Fig. 5. The measured return losses are significantly different between 4 GHz and 8 GHz. Also, both measured results are significantly different from the simulated one.

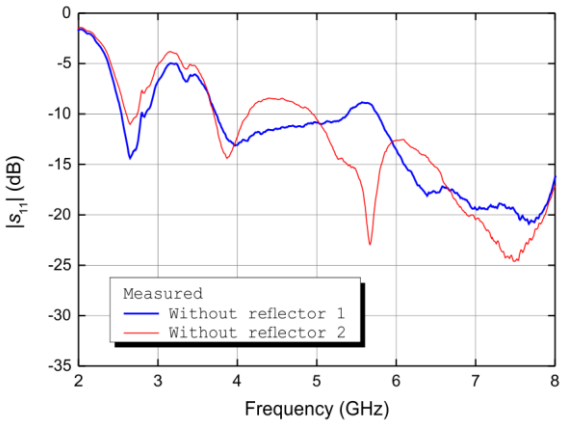


Fig. 5. Measured return loss of two prototypes of UWB printed monopole antenna without reflector

The model of the same UWB printed monopole antenna, but with a reflector, is shown in Fig. 6. The reflector is made from the dielectric substrate with relative permittivity 4,4 and thickness 1,75 mm. The metallization exists on both sides of the substrate, and the metallic vias are made around the connector, electrically connecting the both sides of the substrate. Our analysis show that such reflector is electrically equivalent to the pure metallic reflector, but it is better for mechanical processing and soldering. The dimensions of the reflector are optimized to achieve minimal reflection coefficient.

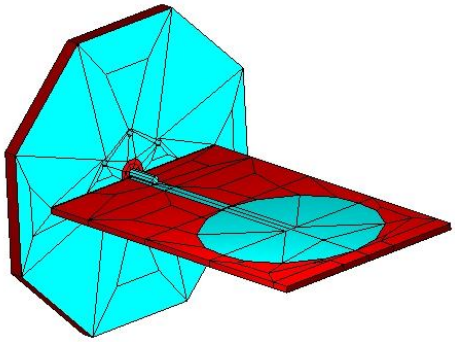


Fig. 6. Model of the UWB printed monopole antenna with the reflector

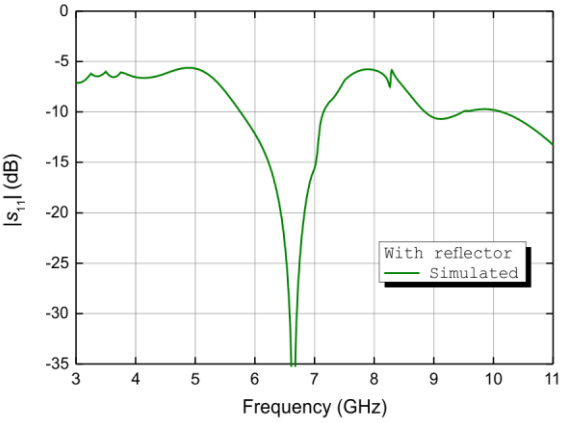


Fig. 7. Simulated return loss of monopole UWB antenna with reflector

The simulated return loss of the UWB printed monopole antenna with the reflector is shown in Fig. 7. There is some degradation in the simulated return loss characteristics compared to the antenna without the reflector.

Two prototypes of the UWB printed monopole antenna with the reflector are fabricated, ant the photos of the antennas are shown in Fig. 8 (top side), and Fig. 9 (bottom side).

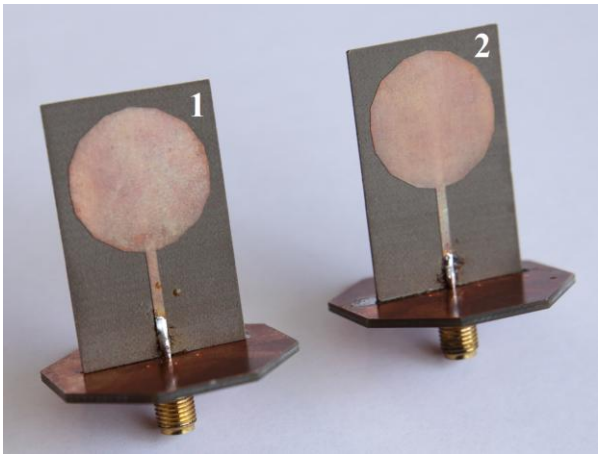


Fig. 8. Photograph of UWB printed monopole antennas with the reflector (top view)

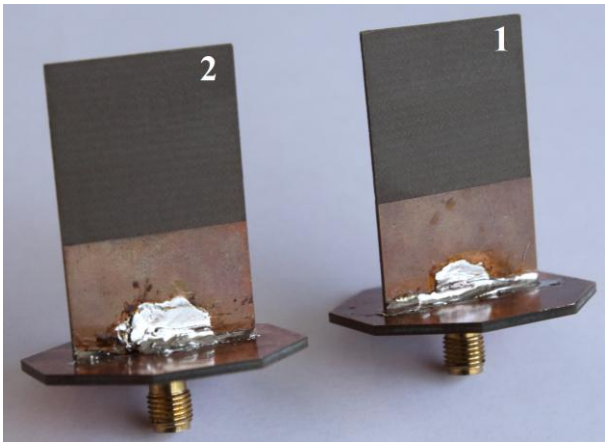


Fig. 9. Photograph of UWB printed monopole antennas with the reflector (bottom view)

Return losses of prototypes with the reflector are shown in Fig. 10. We see that matching between two antennas is almost perfect across the entire UWB band, unlike the prototypes without the reflector.

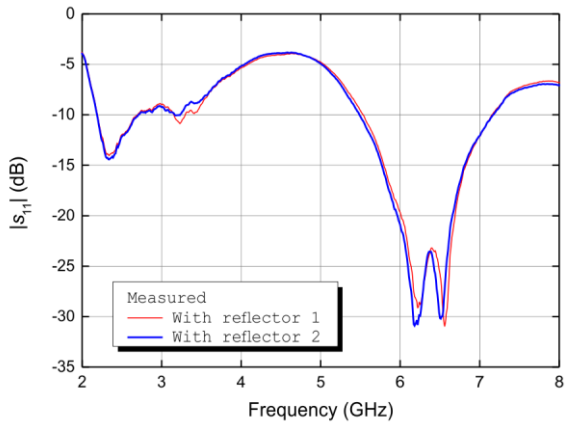


Fig. 10. Measured return loss of prototypes with the reflector

In order to investigate more closely such difference in behavior between the models with and without the reflector, we created the measurement models of the antennas. These models include coaxial cable and a network analyzer (NA). The measurement model for the antenna with the reflector is shown in Fig. 11. The reflection coefficient of the antenna only (without the coaxial cable and NA) is obtained by de-embedding of the parameters, as described in [7]. The de-embedding plane is shown in Fig. 12. The similar model is created for the antenna without the reflector. Finally, the NA boxes are removed from both models, so we can isolate the influence of the NA on the measured results.

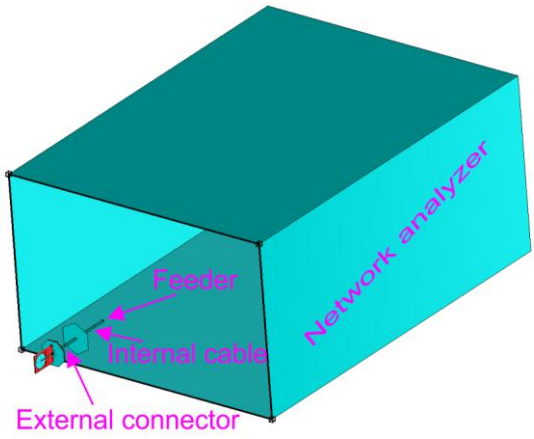


Fig. 11. Measurement model of the antenna with the reflector

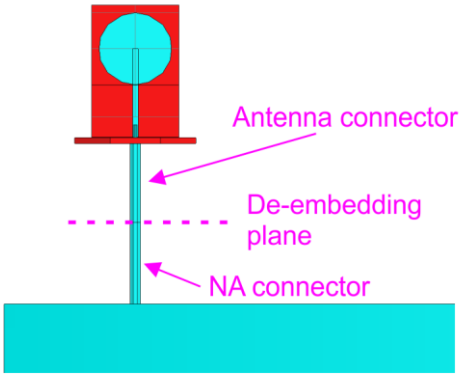


Fig. 12. De-embedding plane in the measurement model

The simulated reflection coefficients for the measurement models of the antenna without the reflector are shown in Fig. 13, whereas corresponding results for the antenna with the reflector are shown in Fig. 14.

We can see that simulated results for the antenna with the reflector are very stable: by removing the NA, only slight change in reflection coefficient is observed. The result with the NA has clear resemblance of the measured results: all resonances are replicated, albeit with somewhat different levels.

On the contrary, the results for the antenna without the reflector are not stable at all, and since measured results for two prototypes are very different, it is hard to establish clear reference for "exact" result.

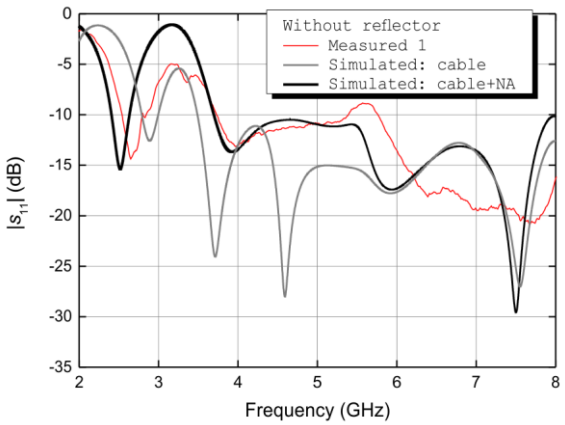


Fig. 13. Simulated reflection coefficients for the measurement model of the antenna without the reflector

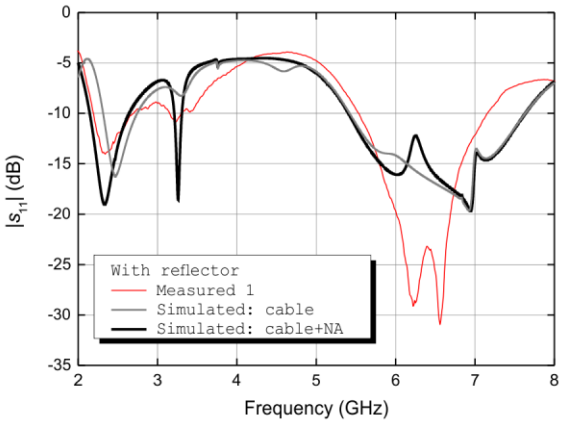


Fig. 14. Simulated reflection coefficients for the measurement model of the antenna with the reflector

III. CONCLUSIONS

Additional reflector changes return loss of the UWB printed monopole antenna, but the stability of the simulated and measured results for the antenna with the reflector is much better. So, it can be attached to various devices without significant changes in the characteristics. The question remains if such antenna with the reflector can be design to fulfill -10dB reflection coefficient in the whole UWB frequency range.

REFERENCES

[1] R. J. Fontana, "Recent system applications of short-pulse ultra-wideband (UWB) technology," in *IEEE Transactions on Microwave Theory and Techniques*, vol. 52, no. 9, pp. 2087-2104, Sept. 2004.

[2] "Revision of part 15 of the commission's rules regarding ultra-wideband transmission systems," FCC report and order, adopted February 14, 2002, released July 15, 2002.

[3] "The harmonised conditions for devices using Ultra-Wideband (UWB) technology in bands below 10.6 GHz," ECC decision, approved 24 March 2006.

[4] Jianxin Liang, C. C. Chiau, Xiaodong Chen and C. G. Parini, "Study of a printed circular disc monopole antenna for UWB systems," in *IEEE Transactions on Antennas and Propagation*, vol. 53, no. 11, pp. 3500-3504, Nov. 2005.

[5] L. Liu, Y. F. Weng, S. W. Cheung, T. I. Yuk and L. J. Foged, "Modeling of cable for measurements of small monopole antennas," 2011 Loughborough Antennas & Propagation Conference, Loughborough, 2011, pp. 1-4.

[6] https://www.wipl-d.com/products.php?cont=add-on-tools/aw-modeler

[7] https://www.wipl-d.com/index.php

[8] V. Mojic, *Electromagnetic modeling of printed monopole antennas for frequency range from 3.1 to 10.6 GHz*, in Serbian, University of Belgrade-School of Electrical Engineering, Belgrade, 2017.

Аналитичко решење Волтерине интегралне једначине прве врсте за генералисани модел повратног удара са путујућим струјним извором

Драган Павловић, Јован Цветић, Градимир Миловановић и Милан Игњатовић

***Апстракт*—Разматрали смо аналитичку методу за решавање специјалне Волтерине интегралне једначине конволуционог типа која се примењује у прорачуну функције пражњења канала код генералисаног модела повратног удара са путујућим струјним извором (GTCS модел). У оквиру аналитичког метода, примењена је Лапласова трансформација уз употребу Meijer-G функције. Разматрани су и методи за инверзну нумеричку Лапласову трансформацију.**

***Кључне речи*—повратни удар, GTCS модел, Волтерина интегрална једначина, Лапласова трансформација.**

I. Увод

Проучавање динамике пражњења канала муње је од великог значаја за физику плазме, као и за електроинжењерску праксу. После детаљног прегледа свих метода који се користе у литератури, одлучено је да се за проучавање динамике канала користи инжењерски GTCS модел [1]. Како би започели овај процес, неопходно је прво израчунати функцију пражњења канала из Волтерине интегралне једначине [2].

Решавање ове једначине представља студију о нелинеарној, нехомогеној Волтериној интегралној једначини прве врсте са импулсним функцијама. Волтерине интегралне једначине прве врсте су саме по себи слабо условљени проблеми.

Математички термин добро условљен проблем, произилази из дефиниције коју је дао француски математичар Жак Адамар [2]. Према Адамару, проблем је добро условљен ако: проблем има решење (егзистенција), ово решење је јединствено одређено (јединственост) и решење непрекидно зависи од улазних података, тј. мале промене улазних података не доводе до велике промене решења (стабилност). Ако један од ових услова није испуњен, проблем је слабо условљен.

Ова слаба условљеност чини нумеричко решење веома тешким, а то значи да мала грешка у улазним подацима може да доведе до дивергентног решења. Упркос извесним ограничењима, једначина је нумерички решена методом модификоване композитне трапезне формуле [3]. Метод је формиран тако да обезбеди висок степен тачности, уз минималне апроксимације и малу сложеност алгоритма, што је и постигнуто. Међутим, оно што је проблематично је пролазак нумеричког решења кроз нулу пре краја интервала, као и извесни пикови на самом почетку интервала, што не проистиче из физике електричних пражњења.

Због поменутих проблема, неопходно је решити Волтерину једначину у аналитичком облику и одредити природу решења.

II. Поставка проблема

Циљ рада је израчунавање функције пражњења канала $f(t-z/v^*)$. Полазимо од Волтерине интегралне једначине прве врсте [4]:

$$i_0(t)=\int_0^{v^*t} q_0'(z)\cdot\frac{\partial}{\partial t}f(t-z/v^*)dz. \qquad (1)$$

по непознатој функцији $f(t-z/v^*)$, где су познате функције: функција струје у тачки удара $i_0(t)$, $q_0'(z)$ укупно подужно наелектрисање у корони пре повратног удара, t је апсолутно време и представља почетак повратног удара, z је висина канала, v је брзина повратног удара, а v^* је редукована брзина повратног удара. Функција $f(t-z/v^*)$ се може записати као $f(u)$. Аргумент функције $f(u)$ је генералисано време и оно се

Драган Павловић – Електротехнички факултет, Универзитет у Београду, Булевар краља Александра 73, 11020 Београд, Србија (електронска пошта: dragan.lab3@etf.bg.ac.rs).

Јован Цветић – Електротехнички факултет, Универзитет у Београду, Булевар краља Александра 73, 11020 Београд, Србија (електронска пошта: cvetic_j@etf.rs).

Градимир Миловановић – Математички институт САНУ, Кнез Михаилова 36, 11020 Београд, Србија (електронска пошта: gvm@mi.sanu.ac.rs).

Милан Игњатовић – Електротехнички факултет, Универзитет у Београду, Булевар краља Александра 73, 11020 Београд, Србија (електронска пошта: ignjatovic@etf.rs).

може изразити као $u = t - z / v^*$.

За моделовање струје у тачки удара користили смо Хајдлерову функцију (једначина (3) из рада [3]), а за моделовање функције подужног наелектрисања смо користили функцију (5) из рада [3]).

Аналитичко решење мора да задовољи следећа четири услова која следе из саме поставке модела [4]:

$$f(0) = 0 \tag{2}$$

$$f(u) \geq 0 : u \geq 0 \tag{3}$$

$$\lim_{u\to\infty} f(u) = 0 \tag{4}$$

$$\frac{df(u)}{du} \leq 0 : u \geq 0. \tag{5}$$

Полазимо од једначине (1) и уводимо следећу смену $f_1(t - z / v^*) = \partial / \partial t \; f(t - z / v^*)$. После примењене смене, једначина (1) постаје линеарна, нехомогена Волтерина интегрална једначина прве врсте:

$$i_0(t) = \int_0^{v^* t} q_0'(z) \cdot f_1(t - z / v^*) dz. \tag{6}$$

Примењујући математичке трансформације, долазимо до интегралне једначине конволуционог типа коју ћемо решавати методом Лапласове трансформације:

$$i(t) = \int_0^t q(t) f_1(t-\tau) \cdot d\tau = \int_0^t q(t-\tau) f_1(t) \cdot d\tau. \tag{7}$$

где су $i(t) = i_0(t) / v^*$ и $q(\tau) = q_0'(v^* \cdot \tau)$.

III. Метод Лапласове трансформације

Теоријски приступ у решавању Волтерине интегралне једначине подразумева примену Лапласове трансформације. Ово је стандардни метод за проналажење решења Волтерине интегралне једначине конволуционог типа. Лапласова трансформација функције $f_1(t)$ је функција $F_1(\mathrm{s})$:

$$F_1(\mathrm{s}) = L\left[f_1(t) \right](s) = \int_0^\infty e^{-st} f_1(t) dt. \tag{8}$$

Применом Лапласове трансформације на конволуциону једначину добијамо:

$$I(s) = Q(s) \cdot F_1(s), \tag{9}$$

где су $I(s)$ и $\mathrm{Q}(s)$ Лапласове трансформације функција $i(t)$ и $q(t)$, респективно.

За дате функције, директна Лапласова трансформација није могућа. Из овог разлога за трансформације користимо Meijer-G функцију [5], [6]. Да би одредили трансформацију, потребно је знати следећи интеграл:

$$G_k(\alpha) = \int_0^\infty e^{-\alpha x} \frac{x^k}{1+x^k} dx \quad \left(\mathrm{Re}\,\alpha > 0,\, k \in \mathbb{N}\right). \tag{10}$$

Ако изразимо интеграл (10) преко Meijer-G функције, добијамо:

$$G_k(\alpha) = \int_0^\infty e^{-\alpha x} \frac{x^k}{1+x^k} dx = \\ = \frac{1}{\sqrt{\kappa (2\pi)^{\kappa-1}}} G_{1,k+1}^{k+1,1}\left(\frac{\alpha^k}{x^k} \middle| \begin{matrix} -\frac{1}{k} \\ -\frac{1}{k}, 0, \frac{1}{k}, \frac{2}{k}, \ldots, \frac{k-1}{k} \end{matrix} \right). \tag{11}$$

У суштини, овај резултат представља Лапласову трансформацију функције $L\left[\psi_k(t)\right](s) = G_k(s)$. Према особинама Лапласове трансформације можемо писати:

$$L\left[\psi_k(t) \cdot e^{-\gamma t} \right](s) = G_k(s + \gamma). \tag{12}$$

Сада смо у могућности да пронађемо Лапласове трансформације $I(s) = L\left[i(t)\right](s)$ и $Q(s) = L\left[q(t)\right](s)$, за функције $i(t)$ и $q(t)$, респективно:

$$I(s) = A\tau_1 G_n\left(\tau_1\left(s + \tau_2^{-1}\right)\right), \tag{13}$$

$$Q(s) = Q_0' \left\{ \left(1 + \frac{\lambda_{d1}}{v^*} s + \frac{\lambda_{d2}}{v^{*2}} s^2\right) \hat{\tau}_1 G_m\left(\hat{\tau}_1\left(s + \hat{\tau}_2^{-1}\right)\right) - \frac{\lambda_{d2}}{v^{*2}} \delta_{m,1} \right\}, \tag{14}$$

где су параметри I_0, v^*, η, τ_1, τ_2, λ_1, λ_2, λ_{d1} и λ_{d2} дефинисани у раду [3], док је параметар $A = I_0 / (\eta \cdot v^*)$, а $\delta_{m,1}$ је Кронекеров делта симбол.

На основу једначине (9), као и израза за Лапласове трансформације (13) и (14) добијамо:

$$F_1(s) = \frac{K G_n(\tau_1(s + \tau_2^{-1}))}{(1 + \tau_{d1} s + \tau_{d2}^2 s^2) G_m(\hat{\tau}_1(s + \hat{\tau}_2^{-1})) - \left(\frac{\tau_{d2}}{\hat{\tau}_1}\right)^2 \delta_{m,1}}, \tag{15}$$

где је $K = A(\tau_1 / \hat{\tau}_1) / Q = I_0(\tau_1 / \hat{\tau}_1) / (\eta \cdot v^* \mathrm{Q}_0')$.

Коначно решење можемо добити као инверзну Лапласову трансформацију од израза (15):

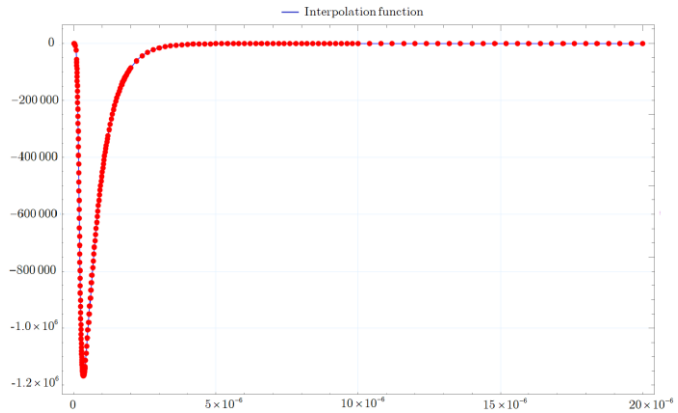
$$f_1(t) = L^{-1}\left[F_1(s)\right](t) = \frac{1}{2\pi i}\int_{\Gamma} e^{st}F_1(s)ds. \qquad (16)$$

У општем случају, овај комплексни интеграл се не може израчунати у аналитичком облику. У таквим случајевима морамо користити неке од метода [7] за приближно израчунавање интеграла (16). Ми смо се у овом раду одлучили да користимо методу Гавера [8] за израчунавање нумеричке инверзне Лапласове трансформације.

IV. Примери

Волтерина интегрална једначина је аналитички решена за велики број примера. Посебно су спроведени прорачуни за функције пражњења језгра, а посебно короне атмосферског пражњења. Ми ћемо у овом раду навести два примера која су урађена.

У примеру 1 смо користили следеће улазне параметре за функцију струје: $I_0 = 13000\,\mathrm{A}$, $\eta = 0{,}87$, $n = 5$, $\tau_1 = 0{,}24\cdot 10^{-6}\,\mathrm{s}$ и $\tau_2 = 3{,}4\cdot 10^{-6}\,\mathrm{s}$. За функцију подужног наелектрисања смо користили параметре: $Q_0' = -1.922\cdot 10^{-4}\,\mathrm{C/m}$, $\lambda_1 = 9\,\mathrm{m}$, $\lambda_2 = 255\,\mathrm{m}$, $\lambda_{d1} = 45\,\mathrm{m}$, $\lambda_{d2} = 0\,\mathrm{m}$ и $m = 4$. Резултат прорачуна за овај пример је приказан на слици 1.



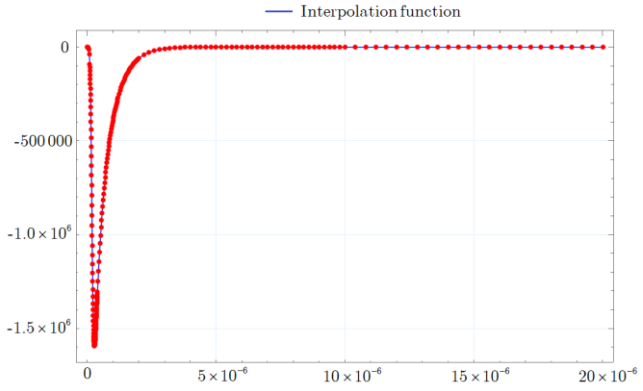
Сл. 1. Први извод функције пражњења добијен Гаверовим методом за пример 1.

Примењујући Гаверов метод добили смо нумеричке вредности решења на целом интервалу. При израчунавању поделили смо цео интервал на следећих шест подинтервала:

$$[0,\ t_6] = [0,\ t_1] \cup [t_1,\ t_2] \cup [t_2,\ t_3] \cup [t_3,\ t_4] \cup [t_4,\ t_5] \cup [t_5,\ t_6], \qquad (17)$$

где вредности $t_1, t_2, t_3, t_4, t_5, t_6$ представљају границе подинтервала. Прорачун је изведен на интервалу $[0,\ t_5] = [0,\ T_{\max}] = \left[0,\ 20\cdot 10^{-6}\right]$. Први подинтервал има границе $[0,\ t_1] = [0,\ T_{\max}/200]$, а корак на овом интервалу је $h_1 = T_{\max}/1000$. Други подинтервал има границе $[t_1,\ t_2] = [26T_{\max}/5000,\ T_{\max}/50]$, а корак на овом интервалу је $h_2 = T_{\max}/5000$. Трећи подинтервал има границе $[t_2,\ t_3] = [21T_{\max}/1000,\ 6T_{\max}/100]$, а корак на овом интервалу је $h_3 = T_{\max}/1000$. Четврти подинтервал има границе $[t_3,\ t_4] = [31T_{\max}/1000,\ T_{\max}/10]$, а корак на овом интервалу је $h_4 = T_{\max}/500$. Пети подинтервал има границе $[t_4,\ t_5] = [11T_{\max}/1000,\ T_{\max}/2]$, а корак на овом интервалу је $h_5 = T_{\max}/100$, док последњи, шести, подинтервал има границе $[t_5,\ t_6] = [26T_{\max}/50,\ T_{\max}]$, а корак на овом интервалу је $h_6 = T_{\max}/50$. Израчунавања су вршена у аритметици двоструке прецизности. Поступак за нумеричку инверзију траје 31 минут. Израчунавање инверзне Лапласове трансформације за овај пример је извршено у програму Mathematica, Ver. 11.3, на рачунару који има следећу конфигурацију: Intel Core i3-3220 CPU, 12GB RAM, 500 GB Sdd, Windows 10 Pro.

Приказаћемо још један пример. За примеру 1 смо користили следеће улазне параметре за функцију струје: $I_0 = 13000\,\mathrm{A}$, $\eta = 0{,}87$, $n = 5$, $\tau_1 = 0{,}24\cdot 10^{-6}\,\mathrm{s}$ и $\tau_2 = 3{,}4\cdot 10^{-6}\,\mathrm{s}$. За функцију подужног наелектрисања смо користили параметре: $Q_0' = -1.979\cdot 10^{-4}\,\mathrm{C/m}$, $\lambda_1 = 4.5\,\mathrm{m}$, $\lambda_2 = 255\,\mathrm{m}$, $\lambda_{d1} = 45\,\mathrm{m}$, $\lambda_{d2} = 0\,\mathrm{m}$ и $m = 1$. Резултат прорачуна за овај пример је приказан на слици 2.



Сл. 2. Први извод функције пражњења добијен Гаверовим методом за пример 2.

С обзиром да је облик улазних функција сличан као у претходном примеру, користили смо исту расподелу тачака да би дошли до крајњег решења. Резултат је добијен у аритметици двоструке прецизности, а поступак за нумеричку инверзију траје 35 минута. Израчунавање инверзне Лапласове трансформације за овај пример је извршено у програму Mathematica, Ver. 11.3, на рачунару који има следећу конфигурацију: Intel Core i3-3220 CPU,

12GB RAM, 500 GB Sdd, Windows 10 Pro.

V. Закључак

Прикладнији приступ за решавање интегралних једначина је примена релевантних нумеричких метода, међутим због слабе условљености датог проблема, једначина се мора решити аналитички. Аналитичко решење за функцију пражњења канала захтева употребу Meijer-G функције, због немогућности директног преласка у Лапласов домен, а затим и коришћење Гаверовог метода за инверзну Лапласову трансформацију. Добијена решења сматрамо егзактним и користићемо их као потврду да ли је одређени нумерички поступак довољно прецизан.

Захвалница

Овој рад је настао као део планираних активности на десеминацији резултата на пројекту ТР 37019 Електродинамика атмосфере у урбаним срединама Србије. Министарства просвете, науке и технолошког развоја Републике Србије за пројектни циклус 2011-2019.

Литература

[1] J. M. Cvetic, B. V. Stanic, “An improved return stroke model with specified channel-base current and charge distribution along lightning channel”, ICEAA, Torino, Italy, 1995.

[2] J. Hadamard, “Sur les problèmes aux dérivées partielles et leur signification physique“, Princ. Univ. Bulletin, pp. 49–52., 1902.

[3] Д. Павловић, Г. Миловановић, Ј. Цветић, Н. Мијајловић, М. Игњатовић, “Нумеричко решавање Волтерине интегралне једначине прве врсте за генералисани модел повратног удара са путујућим струјним извором”, 61 конф. ЕТРАН, Кладово, 5 до 8. јуна 2017., пп. АП1.3. 1-5.

[4] Jovan M. Cvetić, “Model povratnog udara atmosferskog pražnjenja sa specificiranom strujom u tački udara i raspodjelom naelektrisanja duž kanala,” doktorska disertacija, Elektrotehnički fakultet Univerziteta u Beogradu, Srbija, jul 1996.

[5] H. Bateman, A. Erd´elyi, Higher Transcendental Functions, Vol. I, Krieger, New York, 1981.

[6] http://mathworld.wolfram.com/MeijerG-Function.html.

[7] G.V. Milovanović, A.S. Cvetković, “Numerical inversion of the Laplace transform”, Facta Univ. Ser. Elec. Energ. 18, pp. 515 – 530, 2005.

[8] P. P. Valko and J. Abate, “Comparison of sequence accelerators for the Gaver method of numerical Laplace transform inversion,” Comput. Math. Appl., vol. 48, pp. 629–636, 2004.

Abstract

We analyzed the analytical method for solving the special kind of Volterra integral equation of the convolution type, which is applied in the calculation of the channel discharge function in the generalized traveling current source return stroke model (GTCS model). Within the analytical method, Laplace transformation was applied using the Meijer-G function. The methods for the inverse numerical Laplace transform were also considered.

Numerical Solution of Volterra Integral Equation of First Kind for GTCS Return Stroke Model

Dragan Pavlovic, Jovan Cvetic, Gradimir Milovanovic and Milan Ignjatovic

Lokalizacija tačkastih izvora elektromagnetskog zračenja tehnikom retkih signala

Marija Stevanović, *Member, IEEE*, Jelena Dinkić i Antonije Đorđević

***Apstrakt*—Tehnika retkih signala primenjena je za lokalizaciju tačkastih izvora elektromagnetskog polja. Iako posmatrani problem lokalizacije spada u klasu "loše-definisanih" problema, korišćenjem l_1 regularizacije, tj. predznanja da je broj izvora mali, moguće je odrediti nepoznate lokacije i momente električnih i magnetskih dipola. Predloženi algoritam za lokalizaciju testiran je na podacima dobijenim numeričkim simulacijama bez šuma, kao i sa dodatim šumom.**

***Ključne reči*—Električni dipol, lokalizacija, magnetski dipol, tehnika retkih signala.**

I. Uvod

Električki mali izvor elektromagnetskog polja može se u opštem slučaju modelovati pomoću tri ortogonalna električna i tri ortogonalna magnetska dipola čiji se centri nalaze u istoj tački [1]. Ovakav model se može primeniti za lociranje aktivnih logičkih kola u FPGA (engl. *Field-Programmable Gate Array*) integrisanim sistemima kao u [2], [3], gde je razmatrana lokalizacija magnetskih dipola u kvazistacionarnom magnetskom polju korišćenjem standardnih metoda optimizacije.

U cilju pripreme metode za obradu rezultata eksperimenata, u ovom radu primenjujemo obradu retkih signala za lokalizaciju električki malih izvora brzopromenljivog prostoperiodičnog elektromagnetskog polja na osnovu poznavanja polja u malom broju tačaka. Ekvivalentne dipole tražimo na diskretnim pozicijama, u okviru uniformne pravougaone mreže pretraživanja. Smatrajući da se u svakom čvoru mreže nalaze tri električna i tri magnetska dipola, razvijen je linearan elektromagnetski model koji povezuje momente dipola na mreži i elektromagnetsko polje u posmatranim tačkama. Model koristi analitičke izraze za zračenje dipola u prisustvu savršeno provodne ravni. S obzirom na to da je broj nepoznatih momenata mnogo veći od broja merenja, problem spada u klasu "loše-definisanih" problema (engl. *ill-posed problems*). Međutim, ukoliko iskoristimo predznanje da je broj dipola sa značajnim momentima mali, odnosno, mnogo manji od dimenzija mreže za pretraživanje, problem se može rešiti primenom l_1 regularizacije [4]. Predloženi algoritam je dodatno poboljšan uvođenjem normalizacije kojom se ujednačavaju slabljenja prenosa između različitih dipola u mreži i tačaka u kojima se posmatra polje.

Marija Stevanović, Jelena Dinkić i Antonije Đorđević – Univerzitet u Beogradu – Elektrotehnički fakultet, Bulevar Kralja Aleksandra 73, 11120 Beograd, Srbija (e-mail: mnikolic@etf.bg.ac.rs, jdinkic@etf.bg.ac.rs i edjordja@etf.bg.ac.rs).

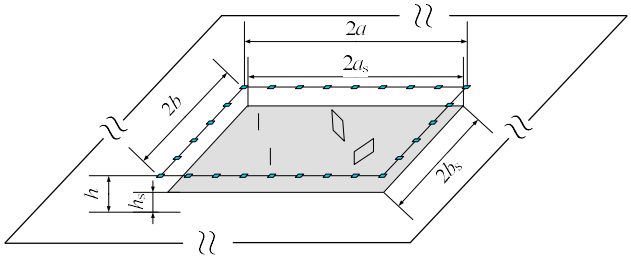
II. Model sistema

U cilju rešavanja inverznog elektromagnetskog problema lokalizacije izvora brzopromenljivog prostoperiodičnog elektromagnetskog polja u vakuumu, posmatra se sistem prikazan na sl. 1. Izvori elektromagnetskog polja smešteni su iznad beskonačne, savršeno provodne ravni, na visini $h_\mathrm{s} = 5\,\mathrm{mm}$ u odnosu na provodnu ravan, u pravougaoniku dimenzija $2a_\mathrm{s} \times 2b_\mathrm{s} = 97{,}5\,\mathrm{mm} \times 78\,\mathrm{mm}$. Izvori elektromagnetskog polja su električni i magnetski dipoli postavljeni na proizvoljnim lokacijama i proizvoljno orijentisani. Radna učestanost je $f = 1\,\mathrm{GHz}$.

Cilj lokalizacije je određivanje pozicije izvora, kao i električnih momenata električnih dipola i magnetskih momenata magnetskih dipola.

Elektromagnetsko polje određuje se u tačkama oko prostora u kome se nalaze izvori. Smatramo da se u svakoj tački nalazi virtuelni senzor za određivanje elektromagnetskog polja (označen kvadratićem na sl. 1). Senzori su postavljeni duž stranica pravougaonika dimenzija $2a \times 2b = 100\,\mathrm{mm} \times 80\,\mathrm{mm}$, na visini $h = 30\,\mathrm{mm}$ iznad provodne ravni. Duž svake duže stranice pravougaonika ekvidistantno je raspoređeno po 10 senzora, a duž svake kraće stranice po 8 senzora, što ukupno čini $N_\mathrm{s} = 36$ senzora.

Svaki senzor meri po tri kompleksne Dekartove komponente električnog i magnetskog polja. Dakle, na mestu svakog senzora poznato je 6 kompleksnih skalarnih veličina.



Sl. 1. Skica sistema.

III. Analitički model

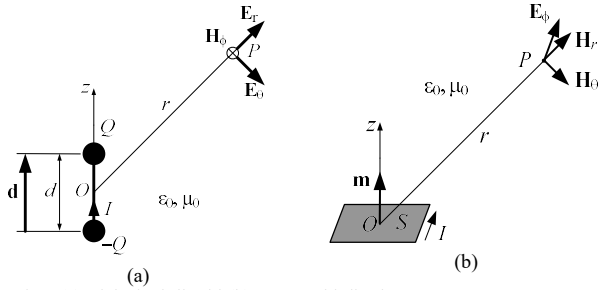
Izrazi za brzopromenljivo električno i magnetsko polje električnog i magnetskog dipola izvode se polazeći od zakasnelih potencijala [1].

Na sl. 2a prikazan je električni dipol (Hercov dipol). Pretpostavljamo da je kompleksna struja $\underline{I}$ uniformna duž ose dipola. Električno i magnetsko polje u tački P posmatranog dipola dati su izrazima:

$$\underline{\mathbf{E}} = \underline{E}_r \mathbf{u}_r + \underline{E}_\theta \mathbf{u}_\theta$$
$$= \frac{(\mathrm{j}\beta)^2}{4\pi\varepsilon_0 r}\left(\left(1+\frac{3}{\mathrm{j}\beta r}+\frac{3}{(\mathrm{j}\beta r)^2}\right)\left(\underline{\mathbf{p}}\cdot\mathbf{u}_r\right)\mathbf{u}_r\right)\mathrm{e}^{-\mathrm{j}\beta r}$$
$$-\frac{(\mathrm{j}\beta)^2}{4\pi\varepsilon_0 r}\left(1+\frac{1}{\mathrm{j}\beta r}+\frac{1}{(\mathrm{j}\beta r)^2}\right)\underline{\mathbf{p}}\,\mathrm{e}^{-\mathrm{j}\beta r}, \quad (1)$$

$$\underline{\mathbf{H}} = \underline{H}_\phi \mathbf{u}_\phi = \frac{(\mathrm{j}\beta)^2 \underline{\mathbf{p}}\times\mathbf{u}_r}{4\pi\varepsilon_0 r}\sqrt{\frac{\varepsilon_0}{\mu_0}}\left(1+\frac{1}{\mathrm{j}\beta r}\right)\mathrm{e}^{-\mathrm{j}\beta r}, \quad (2)$$

gde je $\beta = 2\pi f / c_0$, r rastojanje između dipola i tačke u kojoj se računa polje, $\underline{\mathbf{p}} = \dfrac{\underline{I}\mathbf{d}}{\mathrm{j}\omega}$ kompleksni električni moment dipola, $\omega = 2\pi f$ kružna učestanost, ε_0 i μ_0 permitivnost i permeabilnost vakuuma, respektivno, a $\mathbf{u}_r$, $\mathbf{u}_\theta$ i $\mathbf{u}_\phi$ ortovi sfernog koordinatnog sistema.



Sl. 2. (a) Električni dipol i (b) magnetski dipol.

Na sl. 2b prikazan je magnetski dipol (strujna petlja). Uvedena je slična pretpostavka kao u slučaju električnog dipola: kompleksna struja $\underline{I}$ je uniformna duž petlje. Električno i magnetsko polje strujne petlje u tački P dati su izrazima:

$$\underline{\mathbf{E}} = -\frac{(\mathrm{j}\beta)^2 \underline{\mathbf{m}}\times\mathbf{u}_r}{4\pi r}\sqrt{\frac{\mu_0}{\varepsilon_0}}\left(1+\frac{1}{\mathrm{j}\beta r}\right)\mathrm{e}^{-\mathrm{j}\beta r}, \quad (3)$$

$$\underline{\mathbf{H}} = \underline{H}_r \mathbf{u}_r + \underline{H}_\theta \mathbf{u}_\theta$$
$$= \frac{(\mathrm{j}\beta)^2}{4\pi r}\left(\left(1+\frac{3}{\mathrm{j}\beta r}+\frac{3}{(\mathrm{j}\beta r)^2}\right)\left(\underline{\mathbf{m}}\cdot\mathbf{u}_r\right)\mathbf{u}_r\right)\mathrm{e}^{-\mathrm{j}\beta r}$$
$$-\frac{(\mathrm{j}\beta)^2}{4\pi r}\left(1+\frac{1}{\mathrm{j}\beta r}+\frac{1}{(\mathrm{j}\beta r)^2}\right)\underline{\mathbf{m}}\,\mathrm{e}^{-\mathrm{j}\beta r}, \quad (4)$$

gde je $\underline{\mathbf{m}} = \underline{I}\,\mathbf{S}$ kompleksni magnetski moment magnetskog dipola, a $\mathbf{S}$ površina površi ograničene petljom, čija je orijentacija pravilom desne zavojnice vezana za referentni smer struje.

U sistemu opisanom u prethodnom odeljku, uticaj beskonačne, savršeno provodne ravni, na osnovu teoreme likova, može se modelovati likom odgovarajućeg izvora.

IV. Merni model

Za potrebe formiranja mernog modela, neophodno je definisati mrežu tačaka u kojima će se vršiti pretraživanje. Mreža se prostire po površi pravougaonika $(2a_\mathrm{s} \times 2b_\mathrm{s})$, tj. u oblasti u kojoj se nalaze izvori. Mreža tačaka za pretraživanje, korišćena u ovom mernom modelu, sastoji se od 980 tačaka (35×28 tačaka).

Tokom pretraživanja se pretpostavlja da se u svakom čvoru mreže za pretraživanje nalaze ortogonalni električni i magnetski dipoli. Bez umanjenja opštosti, pretpostavićemo da se provodna ravan nalazi u xOy ravni Dekartovog koordinatnog sistema. U tom slučaju, možemo smatrati da dominantan doprinos elektromagnetskom polju stvaraju samo električni dipoli postavljeni duž z ose Dekartovog koordinatnog sistema, kao i magnetski dipoli čiji su momenti paralelni x i y osi Dekartovog koordinatnog sistema. Polja originala i lika električnih dipola postavljenih duž x i y ose, kao i magnetskog dipola čiji je magnetski moment postavljen duž z ose, praktično se poništavaju. Stoga definišemo da se u svakom čvoru mreže za pretraživanje nalazi samo po jedan električni dipol postavljen duž z ose i samo po dva magnetska dipola čiji su magnetski momenti postavljeni duž x i y.

U tehnici obrade retkih signala [4], [5] pretpostavlja se da su svi dipoli pobuđeni. Međutim, značajne električne, odnosno magnetske momente ima samo mali broj dipola.

Merni model je definisan jednačinom:

$$[\mathbf{b}] = [\mathbf{A}][\mathbf{d}], \quad (5)$$

gde je $[\mathbf{b}]$ vektor merenih podataka, $[\mathbf{A}]$ matrica sistema, a $[\mathbf{d}]$ nepoznati vektor.

Na mestu svakog senzora određuju se tri komponente magnetskog i tri komponente električnog polja:

$$[\mathbf{b}] = \begin{bmatrix} \mathbf{h}_x \\ \mathbf{h}_y \\ \mathbf{h}_z \\ \mathbf{e}_x \\ \mathbf{e}_y \\ \mathbf{e}_z \end{bmatrix}, \quad (6)$$

$$\mathbf{h}_\mathrm{p} = \left[H_\mathrm{p}(\mathbf{r}_1) \quad \cdots \quad H_\mathrm{p}(\mathbf{r}_{36})\right]^\mathrm{T}, \quad (7)$$

$$\mathbf{e}_\mathrm{p} = \left[E_\mathrm{p}(\mathbf{r}_1) \quad \cdots \quad E_\mathrm{p}(\mathbf{r}_{36})\right]^\mathrm{T}, \quad (8)$$

gde p označava x, y ili z komponentu, a $\mathbf{r}_n$ poziciju n-tog senzora $(n = 1,2,3,\ldots,36)$.

Nepoznati vektor $[\mathbf{d}]$ sadrži magnetske i električne momente dipola koji se nalaze u čvorovima mreže po kojoj se vrši pretraživanje:

$$[\mathbf{d}] = \begin{bmatrix} \mathbf{m}_x \\ \mathbf{m}_y \\ \mathbf{p}_z \end{bmatrix}, \ \mathbf{m}_\mathrm{p} = \left[m_{\mathrm{p},1} \quad \cdots \quad m_{\mathrm{p},980}\right]^\mathrm{T}, \quad (9)$$

$$\mathbf{p}_z = \left[p_{z,1} \quad \cdots \quad p_{z,980}\right]^\mathrm{T}, \quad (10)$$

gde p označava x ili y komponentu magnetskog momenta.

Matrica sistema $[\mathbf{A}]$ je definisana na sledeći način:

$$[\mathbf{A}] = \left[\mathbf{A}_{m_x} \quad \mathbf{A}_{m_y} \quad \mathbf{A}_{p_z}\right], \quad (11)$$

$$\mathbf{A}_{m_x}=\begin{bmatrix}\mathbf{H}_{x,m_x}\\ \mathbf{H}_{y,m_x}\\ \mathbf{H}_{z,m_x}\\ \mathbf{E}_{x,m_x}\\ \mathbf{E}_{y,m_x}\\ \mathbf{E}_{z,m_x}\end{bmatrix},\ \mathbf{A}_{m_y}=\begin{bmatrix}\mathbf{H}_{x,m_y}\\ \mathbf{H}_{y,m_y}\\ \mathbf{H}_{z,m_y}\\ \mathbf{E}_{x,m_y}\\ \mathbf{E}_{y,m_y}\\ \mathbf{E}_{z,m_y}\end{bmatrix},\ \mathbf{A}_{p_z}=\begin{bmatrix}\mathbf{H}_{x,p_z}\\ \mathbf{H}_{y,p_z}\\ \mathbf{H}_{z,p_z}\\ \mathbf{E}_{x,p_z}\\ \mathbf{E}_{y,p_z}\\ \mathbf{E}_{z,p_z}\end{bmatrix},\quad (12)$$

gde svaka submatrica (12) predstavlja doprinos odgovarajućeg magnetskog, odnosno električnog momenta.

Submatrice (12) se mogu detaljnije predstaviti u obliku:

$$\mathbf{H}_{\mathrm{p,s}}=\begin{bmatrix}H_{\mathrm{p}}(\mathbf{r}_1,\mathbf{t}_1,\mathrm{s}) & \dots & H_{\mathrm{p}}(\mathbf{r}_1,\mathbf{t}_{980},\mathrm{s})\\ \vdots & \ddots & \vdots\\ H_{\mathrm{p}}(\mathbf{r}_{36},\mathbf{t}_1,\mathrm{s}) & \dots & H_{\mathrm{p}}(\mathbf{r}_1,\mathbf{t}_{980},\mathrm{s})\end{bmatrix},\quad (13)$$

$$\mathbf{E}_{\mathrm{p,s}}=\begin{bmatrix}E_{\mathrm{p}}(\mathbf{r}_1,\mathbf{t}_1,\mathrm{s}) & \dots & E_{\mathrm{p}}(\mathbf{r}_1,\mathbf{t}_{980},\mathrm{s})\\ \vdots & \ddots & \vdots\\ E_{\mathrm{p}}(\mathbf{r}_{36},\mathbf{t}_1,\mathrm{s}) & \dots & E_{\mathrm{p}}(\mathbf{r}_1,\mathbf{t}_{980},\mathrm{s})\end{bmatrix},\quad (14)$$

gde se p, $\mathrm{p}\in\{x,y,z\}$, odnosi na komponente magnetskog i električnog polja, dok s, $\mathrm{s}\in\{m_x,m_y,p_z\}$, označava odgovarajući moment magnetskog, odnosno električnog dipola. Dalje, $\mathbf{t}_l$ predstavlja poziciju l-te tačke u mreži, $l=1,2,3,\dots,980$. Odatle su dimenzije submatrica $\mathbf{H}_{\mathrm{p,s}}$ i $\mathbf{E}_{\mathrm{p,s}}$ jednake 36×980. Izrazi za električno i magnetsko polje dobijaju se iz (1)–(4) kada se uvrsti $r=\|\mathbf{t}_l-\mathbf{r}_n\|$, $\mathbf{u}_r=\dfrac{\mathbf{t}_l-\mathbf{r}_n}{r}$ i primeni teorema likova kako bi se u obzir uzeo i uticaj provodne ravni.

Kako je broj čvorova u mreži za lokalizaciju mnogo veći od broja senzora, lokalizacija električnih i magnetskih izvora, kao i procena njihovih intenziteta, primer je "loše-definisanog" problema. Za rešavanje ovog problema može se iskoristiti l_1 regularizacija [4], [5]. Funkcija koja se minimizira je:

$$\left[\hat{\mathbf{d}}\right]=\min_{\mathbf{d}}\left\{\|[\mathbf{b}]-[\mathbf{A}][\mathbf{d}]\|_2^2+\gamma\|\mathbf{d}_{\mathrm{s}}\|_1\right\},\quad (15)$$

$$\mathbf{d}_{\mathrm{s}}=\begin{bmatrix}\sqrt{m_{x,1}^2+m_{y,1}^2+m_{z,1}^2}\\ \vdots\\ \sqrt{m_{x,980}^2+m_{y,980}^2+m_{z,980}^2}\end{bmatrix},\quad (16)$$

gde $\hat{\mathbf{d}}$ označava estimaciju nepoznatog vektora. Prvi član u (15) predstavlja kvadratnu grešku između merenih rezultata i modela, drugi član favorizuje rešenje koje je retko, a regularizacioni parametar γ balansira između ovih zahteva.

U cilju sprečavanja isticanja čvorova mreže za pretraživanje koji se nalaze bliže nizu senzora u odnosu na čvorove koji su dalje, uvedena je normalizacija. Submatrice (11) pišemo u sledećem obliku:

$$\mathbf{A}_{\mathrm{s}}=\begin{bmatrix}\mathbf{a}_{\mathrm{s},1} & \dots & \mathbf{a}_{\mathrm{s},980}\end{bmatrix},\quad (17)$$

gde se svaka kolona odnosi na jednu tačku mreže. Kao i malopre, $\mathrm{s}\in\{m_x,m_y,p_z\}$. Definišemo vektore čiji su elementi kvadratne norme svake od kolona (17):

$$\mathbf{w}_{m_x}=\begin{bmatrix}w_{m_x,1} & \dots & w_{m_x,980}\end{bmatrix}^{\mathrm{T}},\ w_{m_x,l}=\|\mathbf{a}_{m_x,l}\|_2,\quad (18)$$

$$\mathbf{w}_{m_y}=\begin{bmatrix}w_{m_y,1} & \dots & w_{m_y,980}\end{bmatrix}^{\mathrm{T}},\ w_{m_y,l}=\|\mathbf{a}_{m_y,l}\|_2,\quad (19)$$

$$\mathbf{w}_{p_z}=\begin{bmatrix}w_{p_z,1} & \dots & w_{p_z,980}\end{bmatrix}^{\mathrm{T}},\ w_{p_z,l}=\|\mathbf{a}_{p_z,l}\|_2.\quad (20)$$

S obzirom na to da u svakoj tački mreže imamo dva magnetska dipola, uvodimo vektor:

$$\mathbf{w}_m=\begin{bmatrix}w_{m,1} & \dots & w_{m,980}\end{bmatrix}^{\mathrm{T}},\ w_{m,l}=\sqrt{{w_{m_x,l}}^2+{w_{m_y,l}}^2}.\quad (21)$$

Normalizovana matrica sistema glasi:

$$[\mathbf{A}_{\mathrm{n}}]=\begin{bmatrix}\mathbf{A}_{m_x,\mathrm{n}} & \mathbf{A}_{m_y,\mathrm{n}} & \mathbf{A}_{p_z,\mathrm{n}}\end{bmatrix},\quad (22)$$

$$\mathbf{A}_{m_x,\mathrm{n}}=\begin{bmatrix}\dfrac{\mathbf{a}_{m_x,1}}{w_{m,1}} & \dots & \dfrac{\mathbf{a}_{m_x,980}}{w_{m,980}}\end{bmatrix},\quad (23)$$

$$\mathbf{A}_{m_y,\mathrm{n}}=\begin{bmatrix}\dfrac{\mathbf{a}_{m_y,1}}{w_{m,1}} & \dots & \dfrac{\mathbf{a}_{m_y,980}}{w_{m,980}}\end{bmatrix},\quad (24)$$

$$\mathbf{A}_{p_z,\mathrm{n}}=\begin{bmatrix}\dfrac{\mathbf{a}_{p_z,1}}{w_{p_z,1}} & \dots & \dfrac{\mathbf{a}_{p_z,980}}{w_{p_z,980}}\end{bmatrix}.\quad (25)$$

Na ovaj način ujednačava se slabljenje prenosa između senzora i različitih tačaka mreže. Finalna optimizaciona funkcija je:

$$\left[\hat{\mathbf{d}}\right]=\min_{\mathbf{d}}\left\{\|[\mathbf{b}]-[\mathbf{A}_{\mathrm{n}}][\mathbf{d}]\|_2^2+\gamma\|\mathbf{d}_{\mathrm{s}}\|_1\right\}.\quad (26)$$

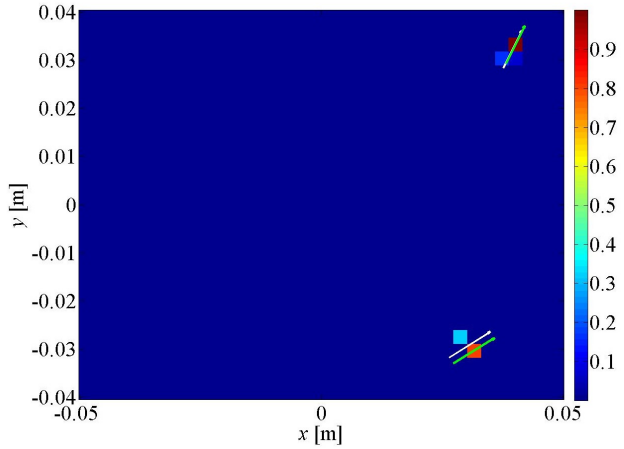
Za rešavanje (26) korišćen je CVX, softverski paket za konveksno programiranje [6], [7].

V. Rezultati lokalizacije

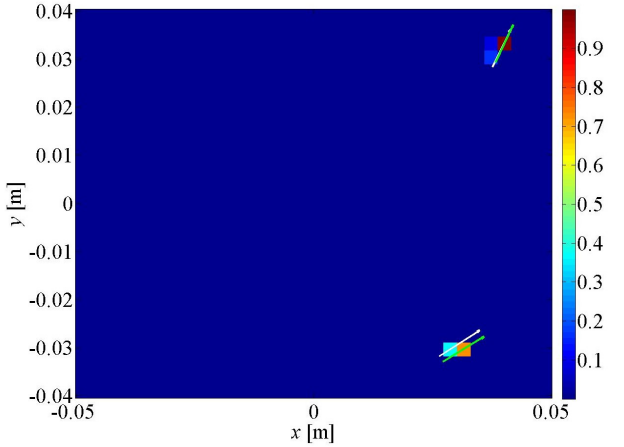
Opisani algoritam za lokalizaciju električnih i magnetskih izvora testiran je na primerima u kojima su bila prisutna dva i tri izvora. Lokacije izvora, kao i magnetski i električni momenti izvora, birani su proizvoljno. Električno i magnetsko polje izvora, na pozicijama senzora, računato je korišćenjem (1)–(4) pre primene algoritma za lokalizaciju. Pritom, pozicije izvora se ne poklapaju sa pozicijama tačaka za pretraživanje. Ovi podaci su tokom lokalizacije tretirani kao rezultati merenja. Algoritam za lokalizaciju je testiran i u slučaju kada je računatim podacima dodat Gausov šum, što je ekvivalentno dodavanju šuma na izlazu senzora.

Na sl. 3 i 6 prikazani su rezultati lokalizacije dva i tri izvora bez dodatog šuma, na sl. 4 i 7 su rezultati lokalizacije uz dodat šum, pri čemu je odnos signal-šum *SNR* = 20 dB, dok su na sl. 5 i 8 predstavljeni rezultati lokalizacije pri odnosu signal-šum *SNR* = 10 dB. Bele strelice na sl. 3–8 označavaju vektore magnetskih momenata stvarnih izvora, dok zelene strelice označavaju magnetske momente lokalizovanih izvora. Lokacije strelica odgovaraju lokacijama izvora. U tabeli I dati su odnosi z komponenti električnih momenata lokalizovaih i stvarnih izvora.

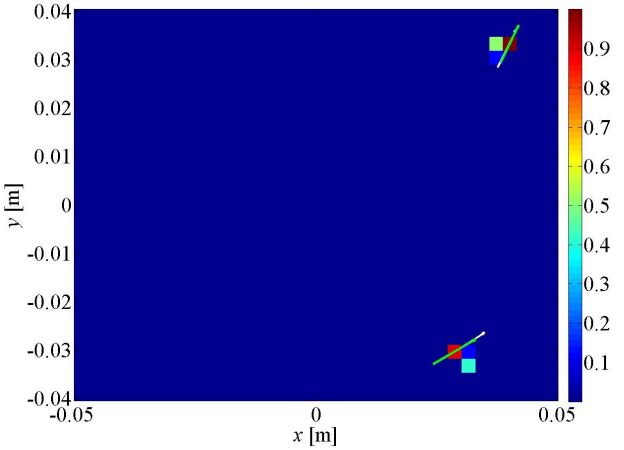
Predstavljeni rezultati potvrđuju da je opisanim algoritmom moguće izvršiti lokalizaciju, kao i procenu jačine elektromagnetskih izvora.



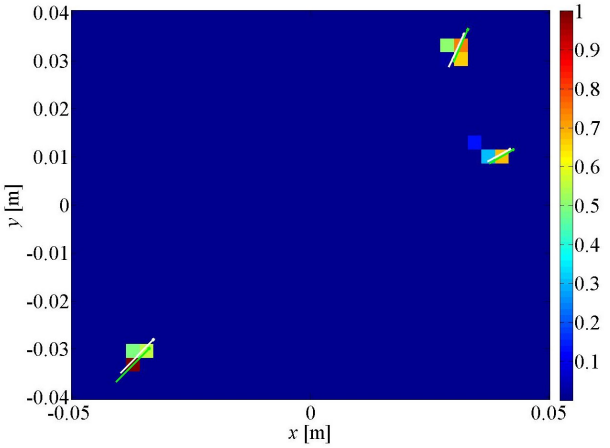
Sl. 3. Rezultati lokalizacije dva izvora bez dodatog šuma.



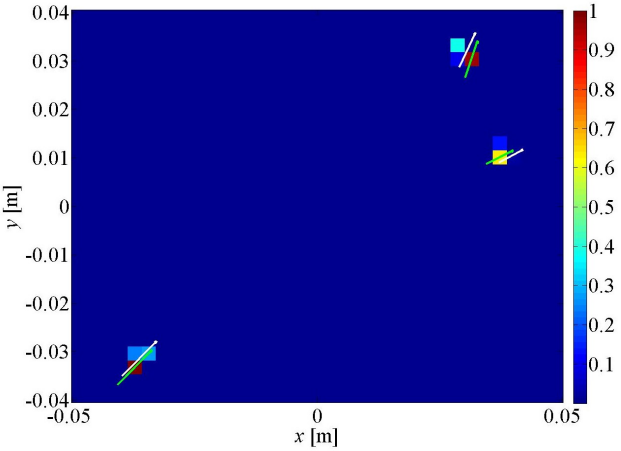
Sl. 4. Rezultati lokalizacije dva izvora sa dodatim šumom, za odnos signal-šum *SNR*=20 dB.



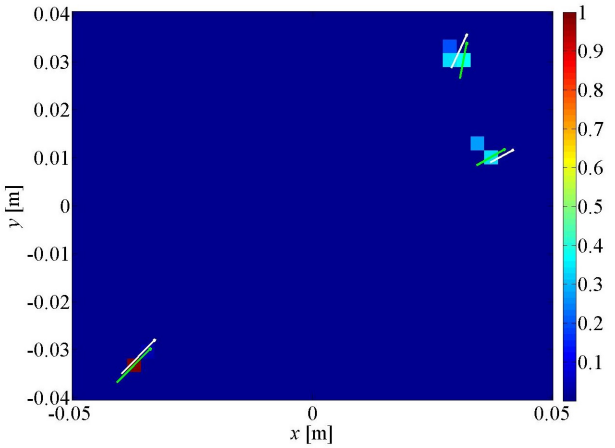
Sl. 5. Rezultati lokalizacije dva izvora sa dodatim šumom, za odnos signal-šum *SNR*=10 dB.



Sl. 6. Rezultati lokalizacije tri izvora bez dodatog šuma.



Sl. 7. Rezultati lokalizacije tri izvora sa dodatim šumom, za odnos signal-šum *SNR*=20 dB.



Sl. 8. Rezultati lokalizacije tri izvora sa dodatim šumom, za odnos signal-šum *SNR*=10 dB.

TABELA I
ODNOS *Z* KOMPONENTI STVARNIH I LOKALIZOVANIH ELEKTRIČNIH MOMENATA

$\frac{p_{z,s}}{p_{z,l}}$	bez šuma	*SNR*=20 dB	*SNR*=10 dB
2 izvora	1,02	1,02	1,00
	0,97	0,97	0,98
3 izvora	0,99	0,92	0,95
	1,03	0,93	0,87
	1,00	0,96	0,92

VI. ZAKLJUČAK

U radu je prikazan algoritam za lokalizaciju električki malih izvora elektromagnetskog polja primenom analitičkih izraza za zračenje električnih i magnetskih dipola. Primenjena je l_1 regularizacija, koja je posebno pogodna kada je broj izvora mali. Numeričke simulacije su pokazale sposobnost algoritma da odredi pozicije izvora, kao i odgovarajuće momente u slučaju dva i tri izvora. Rezultati izloženi u radu predstavljaju pripremu za budući eksperiment u kome će umesto rezultata numeričkih simulacija biti korišćeni rezultati merenja. Kao senzori za merenje polja biće korišćeni električni dipoli i magnetske petlje. Senzori mogu biti postavljeni na fiksnim lokacijama ili mogu biti fiksirani za pokretnu robotsku ruku koja ih pozicionira na željenu lokaciju. Signal sa senzora biće meren mernim prijemnikom, pri čemu je očekivani nivo signala reda veličine $-100\,\mathrm{dBm}$.

ZAHVALNICA

Ovaj rad je delimično finansiran iz projekta TR32005 Ministarstva prosvete, nauke i tehnološkog razvoja Republike Srbije i iz projekta F133 Srpske akademije nauka i umetnosti.

LITERATURA

[1] A. R. Đorđević, "Elektromagnetika," Beograd, Srbija: Akademska misao, 2008.

[2] F. Werner, D. Chu, A. Đorđević, D. Olćan, M. Prvulović, A. Zajić, "A method for efficient localization of magnetic field sources excited by execution of instructions in a processor," *IEEE Trans. on Electromagnetic Compatibility*, vol. 60, no. 3, pp. 1-10, 2017.

[3] F. Werner, A. Đorđević, D. Olćan, M. Prvulović, A. Zajić, "Experimental validation of localization method for finding magnetic sources on IoT devices," *in Proc. of 2018 International Symposium on Electromagnetic Compatibility (EMC EUROPE)*, Amsterdam, Netherlands, 2018.

[4] M. Nikolic Stevanovic, L. Crocco, A. Djordjevic, A. Nehorai, "Higher-order sparse microwave imaging of PEC scatterers," *IEEE Trans. on Antennas and Propagation*, vol. 64, no. 3, pp. 988–997, 2016.

[5] D. Malioutov, M. Cetin, A. Willsky, "Sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Trans. on Signal Processing*, vol. 53, no. 8, pp. 3010–3022, 2005.

[6] Michael Grant and Stephen Boyd. CVX: Matlab software for disciplined convex programming, version 2.0 beta. http://cvxr.com/cvx, September 2013.

[7] Michael Grant and Stephen Boyd. Graph implementations for nonsmooth convex programs, Recent Advances in Learning and Control (a tribute to M. Vidyasagar), V. Blondel, S. Boyd, and H. Kimura, editors, pages 95-110, Lecture Notes in Control and Information Sciences, Springer, 2008. http://stanford.edu/~boyd/graph_dcp.html.

ABSTRACT

Sparse processing is used for the localization of electromagnetic point sources. Although this localization belongs to the class of ill-posed problems, by utilizing the l_1 regularization, i.e., a priori knowledge that the number of sources is small, it is possible to predict the locations and moments of the electric and magnetic dipoles. The proposed algorithm was tested on noiseless and noisy numerically simulated data.

Localization of Electromagnetic Point Sources by Sparse Processing

Marija Stevanović, Jelena Dinkić, and Antonije Đorđević

Recent Results on Modeling and Control Methods Applying the "Fractional" Approach

Guido Maione, *Senior Member, IEEE*

***Abstract*—This paper illustrates some recent results the author obtained in fractional modeling and control of complex engineering systems. Obviously the description is not exhaustive and does not consider many other appreciable results given by the literature. However, by the considered applications, the author aims at proving how the paradigm of fractional-order modeling and control as opposed to the usual integer-order modeling and control can offer benefits that are sometimes unpredicted or underestimated by researchers and practitioners.**

***Index Terms*—Fractional-order models, fractional-order controllers, injection system, common rail, compressed natural gas, electro-injector.**

I. Introduction

NOWADAYS the scientific field of control systems engineering has received interesting applications in many engineering problems. Consequently, it is strategically important to know both classic tools, like PID control, and innovative methods for automatic control. It is not only a matter of industrial automation for products manufacturing or regulation and control of well-known processes involving physical variables (temperature, flow, pressure). Namely, there are many contexts (robotics, mechatronics, cyber-physical systems, networked systems, autonomous vehicles, smart cities, smart grids, energy efficiency in industry, applications with IoT and big data, healthcare systems, home and building automation, intelligent transport systems, etc.) that require effective and innovative control approaches.

This paper examines some applications in the field of automotive systems. Automotive industry is the scene of a continuous improvement of technologies that are developed and finalized to limit fuel consumption and operating costs, reduce harmful emissions (CO_2, CO, NO_x, HC, particulate matter, noise, etc.) and adhere to increasingly strict regulations. It is known that a goal in EU for 2030 is reducing greenhouse gases (GHGs) emission by at least 40% below 1990 levels, thanks to the increase of share of renewable energy, and improving energy efficiency by 27%. Moreover, the goal for 2050 is reducing GHGs emission by 80%-90% and, in particular, GHGs emission from transport by 60%.

To achieve the mentioned goals, several innovative technologies have been introduced. For example, in powertrain control, fuel injection systems, combustion systems, exhaust after-treatment devices. Alternative fuels, electrified vehicles, hybrid powertrains combining an internal combustion engine and an electric traction motor were also made available. An important contribution has also been given by intelligent transport systems, e.g. in optimization of vehicle configuration and transport tasks to improve energy efficiency, hence to reduce emissions.

To reduce fuel consumption and emissions, the automotive engines considered here are based on two peculiar approaches that employ the Common Rail Injection System technology. The first uses the conventional Diesel fuel. The second is based on compressed natural gas (CNG).

In the first case, all the components of the Common Rail Diesel Injection System (CRDIS) have received increased attention, especially in designing advanced electronic control units and electro-injectors. Moreover, efforts were spent to improve combustion strategies and to realize sensors and production process technologies for obtaining low polluting Diesel engines. The performance of CRDIS can be analyzed in terms of the available power, fuel consumption, noise reduction, pollutant emission reduction, etc. In more details, note that, to achieve an efficient combustion, the electronic control unit (ECU) of a CRDIS must accurately meter the amount of fuel and the air-fuel mixture that is injected into the cylinders. In addition, an injection rate shaping (IRS) strategy is frequently employed: the distance between main injections is very short to practically obtain an equivalent single injection according to a desired profile of the flow rate, which gives a more accurate fuel metering and a considerable emissions reduction. As for the electro-injectors, an accurate model is beneficial for optimizing their layout, parameters, and operation and for controlling IRS.

In the second case, CNG is considered because of low cost, availability in many countries, and the obvious capability to reduce pollution from harmful gaseous emissions and fuel consumption. However, for efficient combustion, accurate metering of the air-gas mixture is required by controlling both injection timing and injection pressure at specified levels. In the first case, precise adjustment is possible by the electro-injectors. In the second case, obtaining the desired levels is a challenge and defining the most adequate control technique is an open problem. Namely, the gas compressibility makes the working point of the CNG injection system change a lot, on dependence of the speed and power requested by the driver, so that complex phenomena affect the performance.

II. Why Fractional-Order Models and Controls?

As mentioned in the introduction, modeling, identification,

Guido Maione is with the Department of Electrical and Information Engineering, Polytechnic University of Bari, Via E. Orabona, 4, 70125 Bari, Italy (e-mail: guido.maione@poliba.it).

optimization and control can effectively help to solve the engineering problems of reduction of emissions and fuel consumption in automotive engines.

In particular, fractional-order models and controllers can be used. They derive from fractional calculus [32, 35], which is an ever-growing field that started from a visionary idea of Leibniz generalizing derivatives with integer order to derivatives with non-integer orders and sharing this possibility in a famous exchange of letters with L'Hôpital [39]. Later on, many scientists contributed to the field, both from the theory and from the applications point of view, in particular in automatic control. The reader is referred to [33, 36, 19, 29, 30, 3, 4, 5, 6, 10, 21], to cite just a few contributions.

But why fractional-order modeling and control should be used? In short words, these models and controllers obtain benefits with respect to integer-order ones. Namely, fractional-order differential equations (for linear or nonlinear systems) or fractional-order transfer functions (for linear time-invariant systems) are advantageous because of two main reasons:

- By using fractional calculus, fractional-order models may better describe certain real dynamical processes as heat flow, viscoelastic materials, the electro-magnetic waves propagation in biological tissues, anomalous diffusion, long-term memory effects, etc., that are efficiently represented by non-integer order models.
- The tentative to fit all the processes and systems to integer-order models can be unsuccessful or require too much effort. On the contrary, fractional calculus gives an opportunity for new ideas and better results. As Heaviside said, *"There is a universe of mathematics lying in between the complete differentiations and integrations"* [13].

As an example, in this paper, a fractional-order model of the high-pressure Diesel flow in an electro-injector of a CRDIS will be described.

On the other hand, fractional-order controllers (FOC) are exploited because they offer several advantages. Namely:

- According to Bode, a non-integer order integrator is the ideal open-loop transfer function for ensuring robustness to gain variations in the design of feedback amplifiers [2]. This ideal function was mostly equivalent to $G(s) = A_0/[1+(s/\omega_0)^{\alpha}]$, where Bode chose $\alpha = 5/3$ and achieved a phase margin of 30° for large variations of A_0. The same robustness feature can be obtained to other parametric variations as well. Also Tustin realized this benefit in motion control of massive objects subject to saturation due to power limitation [43]. Namely, Tustin suggested $G(s) = (\omega_{gc}/s)^{\alpha}$, where ω_{gc} is the gain crossover frequency and $\alpha = 1/5$. Then the slope of the asymptotic magnitude diagram was –30 dB/decade and a nearly constant phase margin of 45° was obtained over a wide range around ω_{gc}, i.e. for $0.2\ \omega_{gc} < \omega < 1.4\ \omega_{gc}$.
- Control loops compensated by FOC usually guarantee the ISO-damping feature, i.e. the closed-loop step response maintains the same maximum overshoot with gain variations. This key feature is strictly connected to the flatness of the Bode phase diagram around the gain crossover frequency, i.e. to a typical nearly constant phase margin in a sufficiently wide range around such frequency.
- Adequate realization of FOC is important. Bode designed controllers by a suitable integer number of stages to properly approximate the ideal open-loop gain in a specified frequency range. In the same way, the FOC irrational transfer function must be realized by means of a rational transfer function approximating the FOC [23, 24, 26, 27, 28, 34, 37]. The rational analog realizations must be stable, minimum-phase functions which usually show interlaced zeros and poles along the negative real half-axis in the *s*-domain [25].
- FOC improve the tradeoff between stability robustness and dynamic performance in control loops. Namely, specifications are more easily satisfied by the further design degrees of freedom provided by the non-integer orders of integration and differentiation.
- FOC can be very useful to cope with model uncertainty, disturbances, etc., and provide more flexibility in controller design by the non-integer orders of differentiation and integration.

Many successful applications of fractional-order modeling and control exists. Here, it is shown how to represent the fuel flow in an electro-injector, which is part of a common rail Diesel injection system. More in details, the obtained fractional-order system has a better ability to describe and predict a kind of wave propagation inside the injector, and hence the consequent flow rate injected into the cylinder.

Moreover, the fractional-order PI (FOPI) control of the common rail pressure in the injection system of a CNG engine will be designed and realized. The proposed control approach shows improvements with respect to the usually employed approach, which is based on gain scheduled PI controllers.

III. Fractional-Order Modeling Fuel Flow in an Electro-Injector of a CR Diesel Injection System

This section provides a fractional-order model for the fuel flow in an electro-injector of a common rail Diesel injection system. First of all, the injection system and the fuel dynamics in the electro-injector are described. Then, an integer-order model of the wave pressure propagation in the injector is introduced and a fractional-order model of the same wave pressure propagation is presented. Finally, after some information on the numerical method used to approximate time-fractional partial differential equations, simulation results compare the integer- and fractional-order models.

A. The Injection System and Electro-injector

The CRDIS is very often employed in automotive engines, especially in the light- or heavy-duty Diesel engines. To optimize combustion inside the engine cylinders, the injection system is designed to accurately meter both the fuel amount and air/fuel ratio.

The fuel flows inside the system at different pressures. In a low-pressure circuit, fuel is delivered from a tank to a first low-pressure pump providing fuel to a second high-pressure pump, which is necessary to increase pressure. In a high-pressure circuit, fuel flows from the high-pressure pump to the common rail volume, through a specific valve, and then to the electro-injectors. In a backflow circuit, fuel from the high-pressure pump, the common rail volume and the injectors is sent back to the tank. An Electronic Control Unit (ECU) controls the process in all operating conditions to regulate the pressure in the common rail to constant, high reference values. If a higher vehicle speed and an engine torque are required, the ECU determines the amount of fuel that must be injected and the necessary reference value of the common rail pressure. Moreover, the ECU commands the solenoid valve of the electro-injectors: this valve is responsible to let the fuel enter the injectors, hence to increase or decrease the pressure forces in mechanical elements of the injectors; the vertical displacement of these elements determines the opening or closing of the injectors then fuel injection into the cylinders. Basically, the ECU utilizes a PID control to achieve regulation of common rail pressure with an acceptable transient response (fast transient with limited overshoot and undershoot).

An efficient combustion into the cylinders depends on the air-fuel mixture. The ECU meters the air-fuel ratio and shapes the injected flow rate in two ways: *a*) by adjusting the opening (closing) time intervals of injectors; *b*) by regulating the fuel pressure in the common rail at the specified reference levels. Moreover, note that particular injection rate shaping strategies are commonly used. Namely, the distance between consecutive injections is established so that a single injection is obtained according to a desired *profile* of the flow rate, which gives a more accurate fuel metering and a considerable reduction of emissions. Obviously, a special attention must be paid to the electro-injectors operation and an accurate model is beneficial for both design and optimization of injectors (i.e. their layout, parameters, and operation) and for controlling the injection by shaping the fuel flow rate.

To properly describe the fuel flow in the electro-injector, both an accurate representation of nonlinear fuel dynamics and of the mechanical deformation of relevant parts are necessary. The fuel dynamics can be described by a standard integer-order model or by an innovative fractional-order model, which is based on fractional-order partial differential equations that describe the wave pressure propagation inside a specific part of the injector, as shown below. The model can be further refined by optimizing the injector layout and its mechanical parameters. Namely, internal parts of the injector are subject to pressure forces, and hence to mechanical deformations affecting the fuel flow. These parts constitute the so-called plunger-needle mechanical coupling. In details, several cylindrical parts, with different length and sections and different mechanical parameters (stiffness and mass), are linked in a top-down fashion to form the plunger and the needle. The needle element is responsible to open or close the injector nozzles, then to inject or not the fuel.

Now let us see in which way the fuel flows inside the electro-injector. Basically, two circuits can be distinguished: a *control circuit* and a *feeding circuit*. In the control circuit, fuel arriving from the common rail passes through orifices and enters into a "control chamber", in which the fuel pressure determines if the plunger-needle element is pushed down or up. In the feeding circuit, fuel arriving from the common rail flows through a high-pressure pipe, enters an accumulation volume and then a terminal "SAC" (from French *cul-de-sac*) volume, from which it is injected into the cylinders if the nozzles are open (Fig. 1).

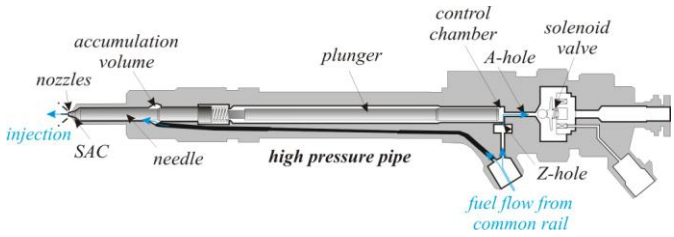


Fig. 1. Electro-injector in a common rail Diesel injection system, with its characteristic pipe to deliver fuel from the common rail volume.

More in details, fuel flow is regulated by a solenoid valve, which is employed to change the pressure acting on the plunger in the control chamber. If the valve is open, the fuel flows from the control chamber to a low-pressure zone. Then pressure on the plunger decreases and the mechanical element, namely the needle part, is pushed up by the pressure in the accumulation volume. In this case, the injector is opened because the accumulation volume is connected to the SAC volume and fuel is injected. On the contrary, if the valve is closed, the high pressure fuel from the common rail increases the pressure in the control chamber that pushes down the plunger (and the needle). In this case, injection is stopped because the needle part closes the nozzles at the end of the SAC volume.

Note that the fuel is compressible, that the feeding pipe is subject to elastic deformation, and that distributed friction losses occur. Moreover, since the needle in the injector is opened and closed very quickly, a kind of water hammer effect generates in the injector pipe. Also cavitation through the holes is considered. Finally note that a pressure wave is generated and damped because of viscous effects. As shown below, this wave propagation is better described by the proposed fractional-order model.

B. Integer-order Model

The complete model is obtained by partitioning the injector in connected volumes in which fuel is accumulated. For some of these volumes, a lumped parameters representation is suitable because the pressure is uniform and time-varying and there is no wave propagation. Then ordinary differential equations are sufficient. For other volumes, namely the elements of the injector pipe, a distributed parameters model is necessary because pressure is nonuniform and time-varying and a wave propagation is experimentally observed. For these

volumes, partial differential equations are used.

Other important assumptions are that the moving elements, i.e. the components of the plunger-needle mechanical coupling, are subject to elastic deformation. Moreover, fuel parameters depend on pressure but not on temperature.

To summarize, the mathematical model is obtained by applying the continuity equation, the momentum equation, the Newton's second law of motion, and the Faraday's law.

The lumped parameters representation is used to represent how pressure p changes with volume v due to the motion of mechanical elements and with intake and outtake flow rates q_i. In general, it can be written:

$$\frac{dp}{dt} = -\frac{K_f}{v}\left(\frac{dv}{dt} - q_l + \sum\nolimits_i q_i\right), \tag{1}$$

where K_f is the Bulk modulus for compressible fuel, q_l is the leakage flow rate, $q_i = \mathrm{sgn}(\Delta p_i)\, c_d\, A_0 \sqrt{2\,\rho_i^{-1}\,|\Delta p_i|}$, in which Δp_i is the pressure gradient across the flow section A_0, c_d is the discharge coefficient, and, finally, ρ_i is the fuel density.

As regards the motion of mechanical elements in the injector, the instantaneous flow sections and volumes depend on the axial displacement of the moving elements. In particular, the displacement of the valve shutter changes the section between the control chamber and the low-pressure zone in which the valve is located. Moreover, the displacement of the plunger-needle coupling changes the section between the accumulation volume and the SAC volume. Moreover, the Newton's second law of motion can be applied to the valve anchor-shutter coupling and to the plunger-needle coupling. Note that, to represent elastic deformation and contact forces, the plunger-needle is modeled as a series of mass-spring-damper systems.

If the high-pressure pipe between the common rail and the accumulation volume is considered, a distributed parameters representation is necessary to describe the wave pressure propagation. Then the classical Navier-Stokes partial differential equations can be used to model fuel dynamics. They originate the standard continuity and momentum equations that are respectively specified as follows:

$$\frac{\partial p}{\partial t} + c_0^2\,\rho_0^2\,\frac{\partial u}{\partial x} = 0\,, \tag{2}$$

$$\frac{\partial u}{\partial t} + \frac{1}{\rho_0}\frac{\partial p}{\partial x} - (\lambda + 2\mu)\frac{\partial^2 u}{\partial x^2} = 0\,, \tag{3}$$

where $p = p(t,x)$ is the fuel pressure depending on time t and on the location x on the (unique) direction of propagation along the pipe, $u = u(t,x)$ is the fuel wave velocity, $c_0 = c_0(p)$ is the speed of sound, $\rho_0 = \rho_0(p)$ is the density, $\lambda = \lambda(p)$ is the dilatation viscosity, $\mu = \mu(p)$ is the dynamic viscosity. If the Stokes' assumption holds, then $\lambda = -2\mu/3$ and the extended wave equation is obtained:

$$\frac{\partial^2 p}{\partial t^2} = c_0^2\,\frac{\partial^2 p}{\partial x^2} + \frac{4\mu}{3\rho_0}\,\frac{\partial}{\partial t}\left(\frac{\partial^2 p}{\partial x^2}\right), \tag{4}$$

which provides a classical integer-order model of wave pressure propagation in the considered pipe.

C. Fractional-order Model

Although there is not a clear evidence of fractional fluid dynamics, the model can be changed to a non-integer order one. Namely, in this way, a better fitting to experimental data will be obtained, as shown below. Moreover, it was shown that since several fluids are characterized by hereditary dependencies of power type [44], then fractional calculus tools based on non-integer-order time derivatives become essential.

The first idea is to extend equations (2) and (3):

$$\frac{\partial^\alpha p}{\partial t^\alpha} + c_0^2\,\rho_0^2\,\frac{\partial u}{\partial x} = 0\,, \tag{5}$$

$$\frac{\partial^\beta u}{\partial t^\beta} + \frac{1}{\rho_0}\frac{\partial p}{\partial x} - \frac{4\mu}{3}\frac{\partial^2 u}{\partial x^2} = 0\,, \tag{6}$$

in which the non-integer orders are $0 < \alpha \le 1$ and $0 < \beta \le 1$. The limit unitary value allows to perform comparisons with the integer-order model.

Remark. One could question about the physical meaning of equations (5)-(6). A physical explanation of (5) and (6), showing how conservation laws are preserved, can be obtained by starting from the Navier-Stokes equations and using fractional viscosity, but it is not the scope of this paper. However, as shown below, this model provides a better fitting to experimental data than a classical integer-order model.

To the best of the author's knowledge, equations (5)-(6) can not be analytically solved. Then an approximating numerical method is employed by using boundary conditions, i.e. the instantaneous pressures and flows at the pipe inlet and outlet sections. In particular, the input pressure is available from measurements of the common rail pressure, while the output pressure in the accumulation volume can be obtained by integration of (1). The method is based on finite differences [12, 11]. Discretization of time-fractional derivatives uses the Grünwald-Letnikov scheme for its excellent stability properties. The approximations are inserted in the momentum and continuity equations in a mixed implicit/explicit way.

Discretization of space derivatives is based on partitioning the pipe in several distinct volume cells. A staggered grid is obtained. Pressure is computed at the center of each cell, and speed is computed on every face between two adjacent cells. For details about the fully discretized fractional-order partial differential equations and their solution, see [12, 11].

D. Comparison by Simulation of Non-Optimized Integer-Order and Fractional-Order Model

Simulation was performed in the MATLAB/Simulink environment. Experimental data for the injected flow rate are available in the following conditions: the exciting time interval of injectors is 700 μs, the common rail reference pressure is 800, 1200, and 1600 bar. Then simulation will compare results with these data. Moreover, the real common rail pressure is used as input to the pipe. A constant uniform pipe section is assumed, the time discretization points are $3000 \leq N \leq 10000$, the space discretization points are $20 \leq M \leq 100$. The non-integer orders α and β are varied to evaluate the effect on model prediction capability.

A first test with $\alpha = \beta$ was made in the most common condition (1600 bar). See Fig. 3 in [12]. The best fitting to experimental data was obtained with $\alpha = \beta = 0.8$. Lower values provide an output in an injection interval that is too small, which is not acceptable for the consumption and emissions. The case $\alpha = \beta = 1$ gives the integer-order model.

A second test with $\alpha \neq \beta$ was made, by setting one order to 1 and letting the other vary as non-integer number. Also in this case, data from the 1600 bar condition were used. See Fig. 4 in [12]. The best results were obtained with $\alpha = 1$, $\beta = 0.95$ or $\alpha = 0.85$, $\beta = 1$. The results somewhat improve the fitting with respect to the case $\alpha = \beta$.

A third test with $\alpha \neq \beta$ was made, in this case allowing any non-unitary, non-integer values for the two orders. All the three cases of experimental data were considered for comparison (1600, 1200 and 800 bar). See Fig. 6 in [12]. The best results were provided by $\alpha = 0.85$, $\beta = 0.98$. The matching to experimental data from a real electro-injector is acceptable if one considers the simplifying assumptions (uniform pipe section, non-optimized values of mechanical parameters of plunger and needle) and the approximation by the numerical method to solve the fractional partial differential equations. Moreover, the model allows prediction of the injected flow rate in several working conditions.

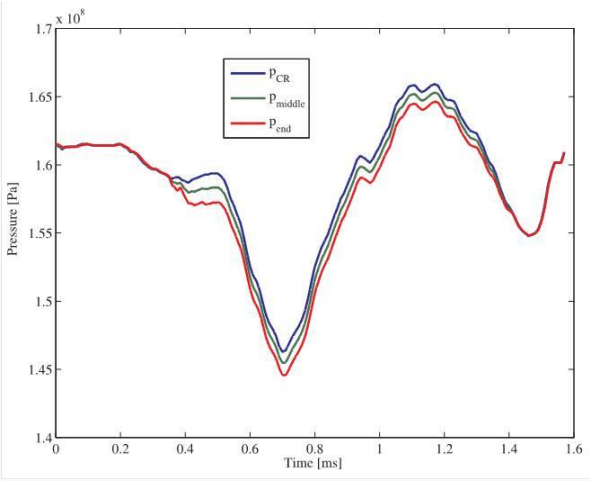


Fig. 2. Pressure trend in three different pipe sections.

To further confirm that results from the fractional-order model are correct, the pressure trend in three different pipe sections is shown in Fig. 2. The sections are at the input, middle and output of the pipe (the input pressure is the common rail pressure). It is clear that a wave propagation is obtained in all sections.

E. Optimization of Fractional-order Model

A further improvement is obtained by optimizing the non-integer orders α and β and the values of mechanical parameters in the representation of the plunger and needle as a series of ideal spring-mass-damper systems. This optimization is important because the deformation of the plunger-needle coupling, which is due to pressure forces, changes the cross-section of flow through the nozzles, namely the cross-section below the needle tip.

In particular, the coupling is divided in seven cylinder elements, each being characterized by a different section and mass m_i ($i = 1, \ldots, 7$). The plunger is formed by five cylinders, the needle by two cylinders. As a consequence, the first and last mass elements in the series model are characterized by one-half of the mass of the first and last cylinders, respectively, whereas every other mass element in the model has half of the masses of two adjacent cylinders (see Table I). Moreover, springs (with elastic constants) and dampers (with viscous damping coefficients) connect the masses as specified in Table I and only one spring links the plunger and needle. The issue is finding the optimal parameters that improve model prediction and correspondence to real measurements of the injected flow rate.

TABLE I
MECHANICAL MODEL PARAMETERS

Mass	Value
M_1	$m_1/2$
M_2	$(m_1+m_2)/2$
M_3	$(m_2+m_3)/2$
M_4	$(m_3+m_4)/2$
M_5	$(m_4+m_5)/2$
M_6	$m_5/2$
M_7	$m_6/2$
M_8	$(m_6+m_7)/2$
M_9	$m_7/2$
Spring, damper coefficients	Element location
k_1, c_1	Between M_1 and M_2
k_2, c_2	Between M_2 and M_3
k_3, c_3	Between M_3 and M_4
k_4, c_4	Between M_4 and M_5
k_5, c_5	Between M_5 and M_6
k	Between M_6 and M_7
k_6, c_6	Between M_7 and M_8
k_7, c_7	Between M_8 and M_9

For each mass M_i in the model, the following general equation is derived from the Newton's second law:

$$M_i\,\ddot{z}_i + c_{i-1}(\dot{z}_i - \dot{z}_{i-1}) + c_i(\dot{z}_i - \dot{z}_{i+1}) + \\ + k_{i-1}(z_i - z_{i-1}) + k_i(z_i - z_{i+1}) = \sum_i F_i\,, \tag{7}$$

where z_i is the deformation of M_i, $k_i = E\,A_i / l_{0i}$, with E being the Young's modulus, A_i the cross-section area of the considered cylinder, l_{0i} the initial length of the cylinder, $c_i = 0.01\sqrt{k_i\,M_i}$ [9] and the right side of (7) specifies pressure forces. Solving the system of equations resulting from (7), provides the displacement z_9, which determines the position of the needle tip, i.e. the outflow section from the SAC volume to the nozzles, then the injector closing/opening, and depends on all the pressures and the valve command.

Given the relation between c_i and k_i, the parameters under optimization are k_i, for i = 1, …, 7. In particular, an evolutionary technique named Differential Evolution is employed to overcome the nonlinearity and complexity of the problem. If $\mathbf{x} = [k_1\ k_2\ k_3\ k_4\ k_5\ k_6\ k_7\ \alpha\ \beta]$, the cost function is

$$J(\mathbf{x}) = \sum_{t_j} \left| y_{\mathrm{mod}}(t_j, x) - y_{\mathrm{exp}}(t_j) \right|, \tag{8}$$

where $y_{\mathrm{mod}}(t_j, \mathbf{x})$ is the output (injected flow rate) predicted by the simulation model at the sampling time instant t_j and depending on the current parameter values in $\mathbf{x}$, while $y_{\mathrm{exp}}(t_j)$ is the experimental output value at t_j. The initial parameter values are given by the theoretical values specified by geometrical dimensions. Minimization of $J(\mathbf{x})$ is subject to the following constraints:

$$\begin{aligned} & k_7 < k_2 < k_3 < k_5 < k_4 < k_1 < k_6\,, \\ & 0 < \alpha < 1, \\ & 0 < \beta < 1, \\ & \left(1 - \frac{pv_1}{100}\right)\mathbf{x}_{\mathrm{nom}} < \mathbf{x} < \left(1 + \frac{pv_2}{100}\right)\mathbf{x}_{\mathrm{nom}}\,, \end{aligned} \tag{9}$$

where the first inequalities express the geometrical differences among the cylinders that compose the plunger and needle hence the different stiffness. The last inequalities define the search space for each component of $\mathbf{x}$, by expressing the maximum allowed variation as a function of the nominal parameter values in $\mathbf{x}_{\mathrm{nom}}$ given by the previously specified theoretical values. The percentage variations $0 < pv_1 < 100$ and $0 < pv_2 < 100$ are set at the beginning of the optimization.

Differential Evolution (DE) [42, 45, 38, 8, 7] efficiency allows to save time and to perform optimization with time and computational constraints, moreover DE has superior performance and robustness with respect to other evolutionary techniques [45]. DE works as follows. It improves a population of N_{pop} possible solutions $\{\mathbf{x}_{ig},\ i = 1, \ldots, N_{\mathrm{pop}}\}$ through evolution of successive generations (g = 0, 1,…, $G_{\max}$) aiming at a solution as close as possible to the optimum [38]. The search space is defined by $\mathbf{x}_{\min} \le \mathbf{x}_{ig} \le \mathbf{x}_{\max}$, where the minimum and maximum are specified by (9).

The initial population (g = 0) is randomly taken from a uniform distribution

$$\mathbf{x}_{i0}(j) = \mathbf{x}_{\min}(j) + r\,[\mathbf{x}_{\max}(j) - \mathbf{x}_{\min}(j)]\,, \tag{10}$$

for j = 1, …, 9, i.e. for each component of $\mathbf{x}$, where r is a random number between 0 and 1. Then it is evaluated as every generation, by using $J(\mathbf{x}_{i0})$ for all i. Evolution goes until the optimum or the final maximum generation is reached. N_{pop} and $G_{\max}$ are usually determined by trial and error.

In each generation, DE performs mutation, crossover and selection. Mutation is to better explore the search space. For each candidate $\mathbf{x}_{ig}$, it creates a mutant $\mathbf{v}_{ig}$ by using other three candidate solutions:

$$\mathbf{v}_{ig} = \mathbf{x}_{i1g} + F\,[\mathbf{x}_{i2g} - \mathbf{x}_{i3g}]\,, \tag{11}$$

where $0 < F < 2$ is a scaling factor, and $i1$, $i2$, and $i3$ are mutually exclusive integer numbers, distinct from the integer number i, and each between 1 and N_{pop}.

Crossover increases diversity and includes best solutions from the previous generation. It creates a new solution $\mathbf{u}_{ig}$ that inherits at least one component from $\mathbf{v}_{ig}$:

$$\mathbf{u}_{ig}(j) = \begin{cases} \mathbf{v}_{ig}(j) & \text{if } r \le CR \text{ or } \quad j = j_r \\ \mathbf{x}_{ig}(j) & \text{if } r > CR \text{ and } \quad j \ne j_r \end{cases} \tag{12}$$

where $0 < CR < 1$ is a crossover ratio, r is a random number from the uniform distribution between 0 and 1, and $j_r \in \{1,\ldots,9\}$ is taken from the uniform distribution.

Selection determines the best solutions that must be considered in the new generation $h = g+1$:

$$\mathbf{x}_{ih}(j) = \begin{cases} \mathbf{x}_{ig} & \text{if } J(\mathbf{x}_{ig}) < J(\mathbf{u}_{ig}) \\ \mathbf{u}_{ig} & \text{if } J(\mathbf{u}_{ig}) \le J(\mathbf{x}_{ig}) \end{cases} \tag{13}$$

The optimized values of parameters are shown in Table II.

TABLE II
OPTIMIZED PARAMETERS

Parameter	Optimized Value
α	0.975
β	0.900
k_1	1.44×10^8
k_2	6.21×10^7
k_3	8.62×10^7
k_4	1.23×10^8
k_5	1.01×10^8
k_6	3.50×10^8
k_7	7.68×10^7

F. Comparison by Simulation of Optimized Integer-Order and Fractional-Order Model

Two phases are distinguished. A first phase in which the model is tested in the condition of 1600 bar. A second "validation" phase in which the model is tested in the other conditions of 1200 and 800 bar. In all cases, the excitation time interval of injectors is kept the same (700 μs). The injected flow rate is measured for about 1.2 ms.

The injection flow rate predicted by the model is compared with the available experimental flow rate in the mentioned conditions. Moreover, comparison is made with prediction by the nominal model related to non-optimized parameter values for coeffiecients k_i and c_i.

Finally, as term of comparison, the flow rate given by a simplified model, in which the plunger-needle coupling is considered as a unique rigid body that is not subject to deformation by pressure forces.

In all cases, results confirm that the optimized model gives the best fitting. See Table III that provides the values of the optimization index in all considered cases. As it is shown, the rigid body model gives much higher values because the model prediction is not effective. The nominal model improves the prediction, but not as much as the optimized one that gives a significant optimization function reduction.

TABLE III
OPTIMIZATION INDEX BY VARIOUS MODELS

Model	Optimization Index J
Rigid-body 1600 bar	1486
Nominal 1600 bar	681
Optimized 1600 bar	164
Rigid-body 1200 bar	1529
Nominal 1200 bar	794
Optimized 1200 bar	171
Rigid-body 800 bar	1706
Nominal 800 bar	738
Optimized 800 bar	145

IV. FRACTIONAL-ORDER CONTROL OF COMMON RAIL PRESSURE IN CNG ENGINES

Compressed natural gas (CNG) prototype engines were proposed to reduce costs and consumption, to take advantage of gas reservoirs and wide-spread gas distribution, and to drastically reduce polluting emissions [16].

To achieve combustion efficiency, an accurate metering of the air-fuel mixture is required. To this aim, the metering system is based on the common rail technology and electro-injectors (usually four in number). Two modalities are contemporaneously used.

The first involves the control of injection timing and guarantees a precise adjustment by commanding the opening and closing of electro-injectors. The second controls injection pressure in the common rail volume at required reference levels. This last control problem is a challenge. Namely, difficulty arises mainly from gas compressibility and the working points of the injection system, depending on power and speed requirements, vary a lot. Therefore complex phenomena and disturbances occur.

In the CNG engine injection system (Fig. 3), gas is delivered from a high-pressure tank (40-200 bar) to a common rail volume (4-20 bar). Since the pressure in the tank slowly varies, it can be assumed constant in a large time interval.

Gas flows through different pipes and passes through a mechanical pressure reducer, in which a piston separates a control chamber from a main chamber, in its turn linked to the common rail volume.

A solenoid valve regulates the flow into the control chamber. It may let more gas enter the control chamber, then push the piston down, open an inflow section from the tank to the main chamber, hence to let more gas arrive to the common rail. Otherwise, if the valve closes, less gas in the control chamber is obtained, the gas in the main chamber pushes the piston up and the outflow section to the main chamber is closed under the action of a spring: in this case, less gas arrives to the common rail. The valve and all the process is under control of an electronic control unit (ECU).

The flow injected into the cylinders depends on the common rail pressure and the opening time intervals of injectors, both controlled by the ECU. Namely, the ECU determines the common rail pressure and controls gas flow by setting the injection timing (which depends on the required speed and applied load) and the current driving the valve.

To derive an analytical model, the system is divided in connected control volumes in which the fuel flows and the pressure is uniform. The pipes are uncompressible and not directly represented to describe pressure propagation. Also the transient intervals in which injectors open or close are neglected with respect to the main gas dynamics, as well as the dynamics of mechanical parts of the valve (i.e. the anchor and shutter) because inertia is small and the valve vertical displacement is limited. Finally, the inertia of the piston and the shutter closing the main chamber are neglected.

By applying physical laws (continuity law, conservation of momentum, Newton's second law), a nonlinear mathematical model is obtained in the state-space form [16]. The chosen state variables are the pressures in the control chamber (x_1) and in the common rail (x_2). The main chamber is not considered because its pressure is almost equal to the rail pressure. The available inputs are the command to the valve (u_1) and to the injectors (u_2), the last being considered as a disturbance. The obvious output is the rail pressure: $y = x_2$.

The injection system works in several conditions that depend on the driver power request, the engine speed, and the applied load. On their basis, the output must properly reach a reference level.

In practice, the ECU sets the proper value of the command to injectors and the reference for the output, and controls the process by the valve regulation until the actual output reached the reference one.

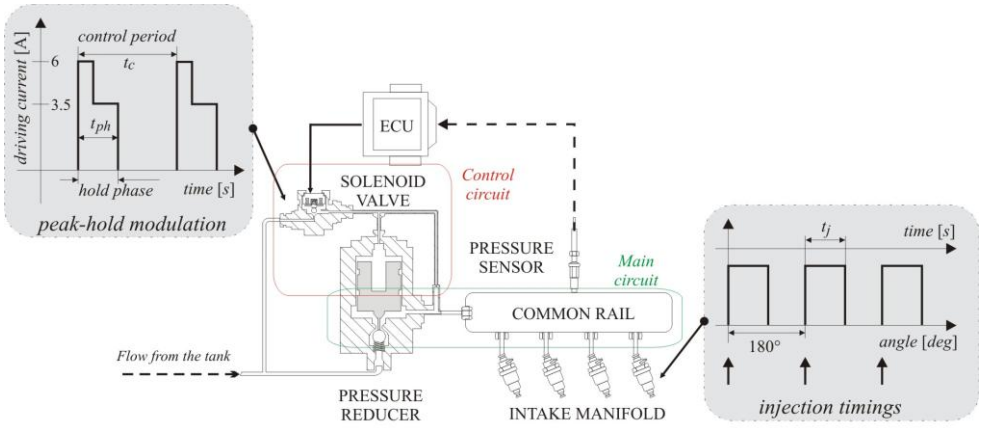


Fig. 3. Operation of the CNG injection system.

Since the system dynamics is through transition among several working points, the equilibrium points of the mathematical model are considered. Then, the variations with respect to the equilibrium values of u_1 and y are considered as input and output. It follows that linearization around working points and Laplace transform of the linearized system provide a transfer function, which can be expressed as a first-order system by considering the dominant pole only:

$$G_p(s) = \frac{K_e}{1+T_e} e^{-L_e s}, \tag{14}$$

where K_e is the equivalent static gain, T_e is the equivalent time constant, L_e is the equivalent dead-time. The triple (K_e, T_e, L_e) depends on the working point, in particular on the pressure in the tank. Hence a family of models is obtained. However L_e can be assumed constant to represent the pressure propagation delay from the main chamber to the common rail.

A. Design of the Fractional-Order Controller

The control methodology employs fractional-order controllers and gain scheduling. Fractional-order PI (FOPI) controllers improve robustness with respect to integer-order PI controllers, for each considered working point. Gain scheduling is used to switch between FOPI controllers that are designed for different working points: namely, the gain scheduling approach is very common in industry and allows switching from one controller to another in order to adapt to various working points.

In more details, the FOPI controllers are designed for specific reference working points by a loop-shaping technique [17], which is reinforced by the *D*-decomposition methodology [18]. The loop-shaping technique allows to achieve a good tradeoff between frequency-domain performance specifications for an optimal feedback system and robustness specifications for a nearly constant phase margin in a sufficiently wide frequency range. The *D*-decomposition approach enhances the robust stability of the closed-loop system.

Remark. In the problem of pressure control of injection systems, PI controllers are mostly applied, as well as in more than 90% of the industrial control loops [1]. This fact shows one more time, namely for automotive industry and for an important purpose, that FOC (in this case FOPI controllers) can give benefits with respect to integer-order controllers. Moreover, the tuning of FOPI controllers [17] takes inspiration and properly modifies popular rules used for PI controllers, namely the symmetrical optimum method [15].

The controller design takes advantage of the Bode's idea on the optimal open-loop frequency response [2]: it consists in shaping the asymptotic gain diagram, mainly in choosing the slope of the segment crossing the frequency axis, and maintaining this slope in a wide frequency interval around the crossover frequency. Hence the phase will have a nearly flat trend and the phase margin will be constant in the same interval. This characteristic is a clear indication of stability robustness even for high gain variations.

Moreover, to obtain an optimal feedback system in the Kalman's sense [14], in a unitary feedback loop with the closed-loop transfer function $F(s) = 1/[1 + G^{-1}(s)]$, a high open-loop gain $|G(j\omega)|$ would be required for each ω, such that $|1 + G^{-1}(j\omega)| \approx 1$ and $|F(j\omega)| \approx 1$. Namely, this condition would imply an almost perfect input-output tracking. To avoid stability problems, $|G(j\omega)|$ is shaped to get high gains at low frequencies and a roll off at high frequencies [40].

Then the practical procedure is in two steps: *a*) to choose the bandwidth in which optimality is desired and determine the crossover frequency where to guarantee a specified phase margin; *b*) to determine the fractional integrator so that the phase plot of the open-loop gain is nearly flat (constant phase margin) in a sufficiently large range around the crossover frequency.

Consider the controller

$$G_C(s) = K_P + \frac{K_I}{s^\nu} = \frac{K_I\,(1+T_I\,s^\nu)}{s^\nu}, \tag{15}$$

where K_P and K_I are the proportional and integral gains, $T_I = K_P/K_I$ and $1 < \nu < 2$, which adds an integer-order integrator (not in the plant) to obtain zero steady-state error to reference step input and to reject disturbance. Then the open-loop frequency response $G(j\omega) = G_c(j\omega)\, G_p(j\omega)$ is given by:

$$G(j\omega) = \frac{K_I K_e [1 + T_I\, \omega^{\nu}\, (C + jS)]\, e^{-j\omega L_e}}{\omega^{\nu}\, (C + jS)\, (1 + j\omega T_e)}, \qquad (16)$$

where $C = \cos(\nu\pi/2)$ and $S = \sin(\nu\pi/2)$. Then the magnitude and the phase are:

$$|G(j\omega)| = \frac{K_I K_e}{\omega^{\nu}} \sqrt{\frac{1 + 2T_I \omega^{\nu} C + T_I^2 \omega^{2\nu}}{1 + \omega^{\nu} T_e^2}}, \qquad (17)$$

$$\langle G(j\omega) = \operatorname{atan}\left(\frac{T_I \omega^{\nu} S}{1 + T_I \omega^{\nu} C}\right) - \operatorname{atan}(\omega T_e) - \omega L_e - \frac{\nu\pi}{2}. \qquad (18)$$

Now the first specification is enforced on the bandwidth ω_B in which to ensure input-output tracking. The gain crossover frequency is estimated by $\omega_C \in [\omega_B/1.7,\ \omega_B/1.3]$, for example $\omega_C = \omega_B/1.5$, according to a rule of thumb [20, 22]. The phase is set to be constant around ω_C. Then the phase margin computed at ω_C, i.e. $PM = \langle G(j\omega_C) + \pi$, should be in its turn constant in a range around ω_C. This requirement implies that

$$PM = \operatorname{atan}\left(\frac{T_I \omega_C{}^{\nu} S}{1 + T_I \omega_C{}^{\nu} C}\right) - \operatorname{atan}(\omega_C T_e) - \omega_C L_e - \frac{\nu\pi}{2} + \pi \qquad (19)$$

is set equal to $PM = \pi - \pi\nu/2 = (2-\nu)\,\pi/2$. As it can be verified after applying simple algebra, the only way to get this result is to choose:

$$T_I = \frac{\omega_C T_e + \tan(\omega_C L_e)}{\omega_C{}^{\nu}[(S - \omega_C T_e C) - (C + \omega_C T_e S)\tan(\omega_C L_e)]}. \qquad (20)$$

Given the phase margin specification PM_S, it holds:

$$PM_S = (2-\nu)\frac{\pi}{2} \Leftrightarrow \nu = 2 - \frac{PM_S}{\pi/2}. \qquad (21)$$

The last two relations show the strict link that is obtained between the fractional order and the phase margin. A specification on the phase margin imposes the value of ν. Moreover, the left side of (20) shows that a selected value of determines the phase margin that can be obtained.

By putting $|G(j\omega_C)|=1$, the integral gain is obtained:

$$K_I = \frac{\omega_C^{\nu}}{K_e} \sqrt{\frac{1 + \omega_C^{\nu} T_e^2}{1 + 2T_I \omega_C^{\nu} C + T_I^2 \omega_C^{2\nu}}}. \qquad (22)$$

B. Realization of the Fractional-Order Controller

Since the fractional integral operator and the associated FOPI controller are infinite-dimensional and irrational, the controller can be realized only by an approximation specified by a rational transfer function. To this aim, many different approximating methods exist, among which the most popular is the Oustaloup's recursive approximation [34].

For a control purpose, it is important that the rational transfer function is characterized by stable poles and minimum-phase zeros. Moreover, zeros and poles are usually chosen alternating on the negative real half-axis: this property is referred to as interlacing between zeros and poles [25]. Also, the number n of zero-pole pairs should be limited to simplify the realization.

An efficient method to obtain stable poles interlaced with non-minimum phase zeros is based on a continued fraction expansion (CFE) [24]. By this method, the truncated CFE that specifies the so-called *convergent* can be converted to a rational transfer function of order n by simple formulas. It holds, for $0 < \nu < 1$:

$$s^{\nu} \approx \frac{p_{n,0}\, s^n + p_{n,1}\, s^{n-1} + \ldots + p_{n,n-1}\, s + p_{n,n}}{q_{n,0}\, s^n + q_{n,1}\, s^{n-1} + \ldots + q_{n,n-1}\, s + q_{n,n}}, \qquad (23)$$

where the numerator and denominator coefficients are provided by the following closed-form expressions [24]

$$p_{n,j} = q_{n,n-j} = (-1)^j \binom{n}{j} (\nu + j + 1)_{(n-j)}\, (\nu - n)_{(j)}, \qquad (24)$$

for $j = 0, \ldots, n$. In such expression, $\binom{n}{j}$ is the binomial coefficient and the other two factors are Pochhammer functions:

$$\begin{cases} (\nu + j + 1)_{(n-j)} = (\nu + j + 1)\,(\nu + j + 2)\cdots(\nu + n), \\ (\nu - n)_{(j)} = (\nu - n)\,(\nu - n + 1)\cdots(\nu - n + j - 1), \end{cases} \qquad (25)$$

in which the first function includes n–j factors and the second function includes j factors. Note that it is set $(\nu+n+1)_{(0)} = 1$, if $j = n$ in the first function, and $(\nu-n)_{(0)} = 1$, if $j = 0$ in the second function. Obviously, if the fractional integral operator $1/s^{\nu}$ must be approximated, the reciprocal of (23) is considered. Usually an order $n = 5$ or 6 is sufficient to obtain a good approximation. If an approximation for $1 < \nu < 2$ is required, then the operator is split as $s^{\nu} = s\, s^{\nu-1}$, such that the factor $s^{\nu-1}$ is approximated according to (23).

Interlacing between simple, real negative poles and zeros is often achieved in an empirical way. Instead, it was proven that the described CFE-based method guarantees interlacing in a systematic way by closed-form formulas [25]. This result was achieved thanks to the special structure of the employed CFE and to the Stieltjes' form of continued fraction [41].

C. Robust Stability Analysis by D-decomposition

The employed loop-shaping technique to design the FOPI controllers guarantees a robust control system. This can be verified by the *D*-decomposition methodology, which is a classical approach for robust stability analysis [31].

Namely, this approach allows to determine the entire domain/set *D* of controller gains leading to a stable closed-loop system. If PI/PID controllers are used, this set is defined in a two- or three-dimensional space and, once the controller gains are fixed, a point in the set is determined. The approach is beneficial for two main reasons:

- It avoids time-consuming stability checks of new controller settings required for new working conditions. Checking should be frequent due to the fact that the reference rail pressure changes, as previously explained.
- If the point associated to the designed controller gains is far from the set boundary, then stability is still ensured for bounded variations that can be graphically observed.

The approach is based on the analysis of the roots of the fractional-order characteristic pseudo-polynomial equation, which can be derived by $F(s) = 1+G(s) = 0$ where $G(s)$ is specified in (16), in the frequency domain:

$$E(s) = (1+T_e\, s)\, s^{\nu} + K_e\, K_I\, (1+T_I\, s^{\nu})\, e^{-L_e s} = 0\,. \quad (26)$$

If all roots of (24) lie in the open left-half of the *s*-plane (LHP roots), then the closed-loop system is BIBO stable.

The set *D* of stabilizing controllers can be defined as follows: if $(K_P, K_I, \nu) \in D$ then all roots lie in the LHP.

D is usually defined by determining its boundaries: the real root boundary (RRB), the infinite root boundary (IRB), and the complex root boundary (CRB). RRB comes from the solutions of $E(s = 0) = 0$: in the present case of a FOPI controller, $K_I = 0$. IRB is defined for $s \to \infty$ and does not exist. CRB comes from the solutions of $E(s = j\omega) = 0$:

$$\begin{cases} K_P(\omega) = \dfrac{(\omega T_e S - C)\sin(\omega L_e) - (S + \omega T_e C)\cos(\omega L_e)}{K_e S}, \\ K_I(\omega) = \dfrac{\omega^{\nu}[\sin(\omega L_e) + \omega T_e \cos(\omega L_e)]}{K_e S}, \end{cases} \quad (27)$$

where $C = \cos(\nu\pi/2)$, $S = \sin(\nu\pi/2)$, and $\omega \in (0,\infty)$. If the fractional order ν is varying, the complete three-dimensional stability domain is obtained. Otherwise, by fixing ν and letting ω to change, a closed curve defines the stability region in the two-dimensional space of coordinates K_P and K_I.

Now assume to consider a specific working condition, which is typically specified by the desired common rail pressure, the injection time interval, and the engine speed. This condition determines the plant parameters (K_e, T_e, L_e), then assume to design a FOPI controller with reference to the condition and the associated reference value of the common rail pressure. FOPI parameters are represented by a point in the stability region (see Fig. 4). This point corresponds to the specified crossover frequency ω_C (then to the bandwidth ω_B) and phase margin PM_S. More in details, it belongs to a relative stability line that can be determined by using the specified phase margin.

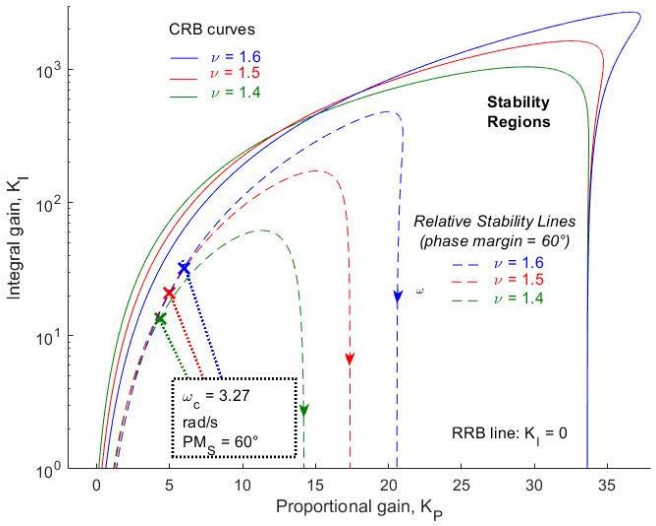


Fig. 4. Stability regions (with CRB curves and IRB line) and relative stability lines for different fractional orders (1.4, 1.5, 1.6) to obtain performance specifications (crossover frequency and phase margin) for a representative working point.

If ν changes, the specified and obtained phase margin changes by (21), then both the relative stability line and the CRB curve of the stability region change (see Fig. 4). However, for each value of ν, the point lies in the stability region and the system is robustly stable. The distance between the CRB curve and the point (its relative stability line) gives a measure of robustness.

Moreover, if the bandwidth specification is modified while the desired phase margin is not, then the point moves along the relative stability line and stays inside the stability region.

One could also consider the relative stability line (not shown in Fig. 4) associated to a gain margin specification GM_S: the point would correspond to the phase crossover frequency.

D. Controllers Gain Scheduling

In industry, to adapt to different working conditions in which the common rail pressure varies to a large extent, it is common to use gain scheduling. Integer-order PI controllers are tuned by heuristic rules and their parameters are adapted by simple gain scheduling. Then, if the injection system operation must change from one condition to another, there is a switching from one PI controller to another, the last being determined by the condition to reach. It is easy to understand that each condition (set of plant parameters) requires a different controller to keep the rail pressure close to the reference value for that condition.

Here the new strategy is to use a FOPI controller for each condition and a new gain scheduling technique. Firstly, a FOPI increases the robustness-performance balance for each

condition. Secondly, the gain scheduling to switch between FOPI controllers is based on a sensitivity analysis of model coefficients.

In details, for each tank pressure of the CNG injection system, the working condition is determined by the reference value of the common rail pressure (p_{CR}) and the average duration of injection (t_{inj}). Assume step variations (worst case) of such variables from an initial to a final working condition. The gain scheduling technique distinguishes two cases:

- *I) Small variations*, in which p_{CR} is bounded by 2 bar and t_{inj} by 6 ms: in this case only one controller is designed with reference to the parameters (K_e, T_e, L_e) of the final working point to reach.
- *II) Large variations*, in which intermediate reference values are considered for p_{CR} and t_{inj}, then intermediate values of (K_e, T_e, L_e) and associated intermediate controllers.

By this technique, a smooth transition between sufficiently close working points is achieved so that each controller guarantees stability in the neighbors of its associated working point and problems are prevented.

E. Comparison by Simulation of Gain Scheduled PI and FOPI Controllers

Simulation tests were performed by using a nonlinear accurate model of the CNG injection system, which was implemented by the AMESim virtual prototyping tool. This tool properly describes complex fluid dynamic phenomena of injection at different working points, and to obtain a simulated system very close to the real one. Moreover, typical value for the reference pressure p_{CR} and injection timing t_{inj} were considered, each yielding a different triple (K_e, T_e, L_e), then a different controller. The aim was to compare PI and FOPI controllers, both gain scheduled in the same way. The most important result should be limiting the overshoot in the actual common rail pressure, i.e. the output of (14). Namely, overshoot would imply too much injected fuel, which alters the air-fuel ratio and increases consumption and emissions.

A first test considered a small step variation in the reference rail pressure (case I above) from 4 to 5 bar, with t_{inj} = 5 ms. Fig. 5 shows the actual rail pressure when controlled by a PI controller or FOPI controllers with different fractional orders of integration. Overshoot increases with the fractional order but it is much less than with PI, which then provides an inaccurate metering of the injected fuel. Moreover, the response by the PI controller shows large, high-frequency oscillations that are determined by large variations or saturation of the control signal (valve command). These results can be explained by the lower sensitivity of FOPI controllers to nonlinearities in the injection process.

A second test applied a large variation in reference pressure from 4 to 10 bar (case II above). Then intermediate reference pressures were considered and the gain scheduling determined the switch between three FOPI or PI controllers. Again, it can be verified that FOPI yielded better and smoother responses with reduced overshoots, and that nonlinearities considerably affect performance of PI controllers (see Fig. 6).

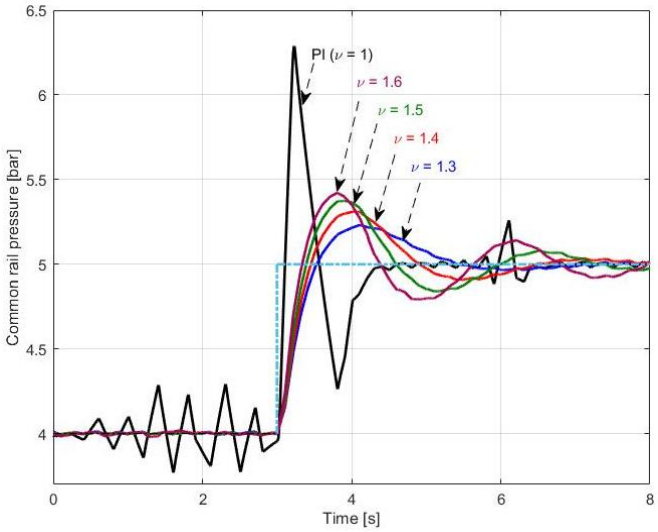


Fig. 5. Rail pressure in CNG injection system in response to a small step.

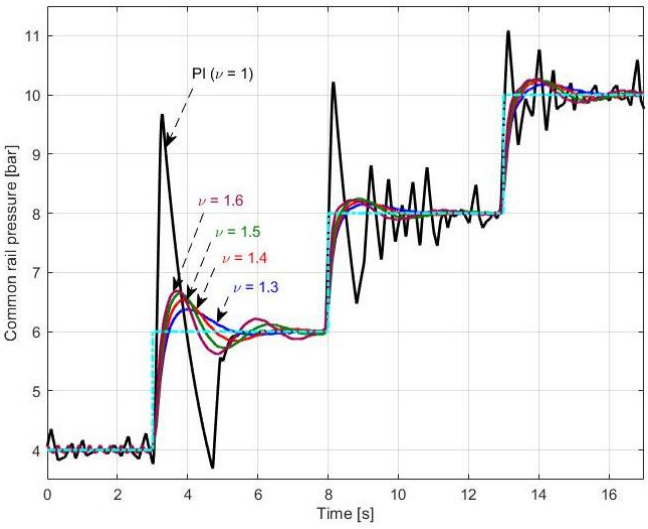


Fig. 6. Rail pressure in CNG injection system in response to a large step.

V. Conclusion

In some engineering applications, like those shown in this paper, the fractional-order models can be more effective than integer-order models in behavior prediction and data fitting.

As for an electro-injector in a CR Diesel injection system, the proposed model is able to match experimental data from a real electro-injector, to predict injected flow rate in several working conditions, and to optimize the parameters for a model-based control of fuel injection. Research is ongoing to improve the model accuracy, to improve the numerical methods that solve fractional partial differential equations, to design model-based control strategies for fuel injection.

Moreover, fractional-order PI controllers can replace PI controllers at a reduced realization cost to improve robustness, performance, sensitivity to nonlinearities, etc. The FOPI controllers can be designed by a loop-shaping method that guarantees robustness (flat Bode phase plot in a sufficient

range around the crossover frequency) to gain changes and performance (bandwidth to pursue an optimal feedback system). Design and realization formulas are given by closed-form expressions that enforce stability, minimum-phase and interlacing properties of the controller. Robust stability of the control system is shown by *D*-decomposition. Finally, gain scheduling is used to cope with various working points of the system and to switch between FOPI controllers.

ACKNOWLEDGMENT

The author thanks dr. Paolo Lino from Polytechnic University of Bari for cooperation in this research.

REFERENCES

[1] K. J. Åström, T. Hägglund, *PID Controllers: Theory, Design, and Tuning*, 2nd ed. Research Triangle Park, NC: Instr. Soc. of America, 1995.

[2] H.-W. Bode, *Network Analysis and Feedback Amplifier Design*. New York: Van Nostrand, 1945.

[3] R. Caponetto, G. Dongola, "A numerical approach for computing stability region of FO-PID controller," *J. Franklin Inst.*, vol. 350, no. 4, pp. 871-889, 2013.

[4] R. Caponetto, G. Dongola, L. Fortuna, I. Petráš, *Fractional Order Systems: Modeling and Control Applications*. Singapore: World Scientific, 2010.

[5] R. Caponetto, G. Dongola, F. Pappalardo, V. Tomasello, "Auto-tuning and fractional order controller implementation on hardware in the loop system," *J. Optimiz. Theory App.*, vol. 156, no. 1, pp. 141-152, 2013.

[6] R. Caponetto, G. Maione, A. Pisano, M. R. Rapaić, E. Usai, "Analysis and shaping of the self-sustained oscillations in relay controlled fractional-order systems," *Fract. Calc. Appl. Anal.*, vol. 16, no. 1, pp. 93-108, 2013.

[7] U. Chakraborty, *Advances in Differential Evolution*. Berlin: Springer-Verlag, 2008.

[8] S. Das, A. Konar, U. Chakraborty, "Two Improved Differential Evolution Schemes for Faster Global Search," 7th Annual Conference on Genetic and Evolutionary Computation, H. Beyer, ed., Washington, DC, ACM-SIGEVO, pp. 991-998, June 25-29, 2005.

[9] C. Dongiovanni, M. Coppo, "Accurate Modeling of an Injector for Common Rail Systems," in *Fuel Injection*, D. Siano, ed., Rejeka, HRV: Sciyo, 2010, pp. 95-119.

[10] M. Ö. Efe, "Fractional order systems in industrial automation – A survey," *IEEE Trans. Ind. Informat.*, vol. 7, no. 4, pp. 582-591, Nov. 2011.

[11] R. Garrappa, P. Lino, G. Maione, F. Saponaro, "Model Optimization and Flow Rate Prediction in Electro-injectors of Diesel Injection Systems", 8th IFAC Int. Symp. Adv. Automotive Control, Norrköping, Sweden, June 19-23, 2016, IFAC-PapersOnLine 49-11 (2016) 484-489.

[12] R. Garrappa, P. Lino, G. Maione, F. Saponaro, "Modeling and Numerical Analysis of Fractional-Order Dynamics in Electro-injectors Pipes", 2015 54th IEEE Conf. on Decision and Control (CDC), Osaka, Japan, pp. 5984-5989, Dec. 15-18, 2015.

[13] O. Heaviside, *Electromagnetic Theory, vol. II, 2nd reprint*. London: Ernest Benn Ltd., 1925.

[14] R. E. Kalman, "When is a linear control system optimal?", *J. Basic Eng.-T. ASME*, vol. 86, series D, pp. 84-90, 1964.

[15] C. Kessler, "Das symmetrische optimum," *Regelungstechnik*, 6, pp. 395–400, 432–436, 1958.

[16] P. Lino, B. Maione, C. Amorese, Modelling and predictive control of a new injection system for compressed natural gas engines, *Control Eng. Pract.*, 16, 1216-1230, 2008.

[17] P. Lino, G. Maione, "Loop-shaping and easy tuning of fractional-order proportional integral controllers for position servo systems", *Asian J. Control*, vol. 15, no. 3, pp. 796-805, May 2013.

[18] P. Lino, G. Maione, "Switching Fractional-Order Controllers of Common Rail Pressure in Compressed Natural Gas Engines", In: E. Boje, X. Xia (Eds.), Proc. 19th IFAC World Congress 2014, Cape Town, South Africa, Aug. 24-29, 2014, IFAC Proceedings on line: Vol. 19, Part 1, pp. 2915-2920.

[19] B. J. Lurie, "Three-parameter tunable tilt-integral-derivative (TID) controller," US patent US5371670, Dec., 06, 1994.

[20] B. J. Lurie, P. J. Enright, *Classical Feedback Control with Matlab*, Control Engineering Series. New York: Marcel Dekker, Inc., 2000.

[21] C. Ma, Y. Hori, "Fractional-order control: Theory and applications in motion control," *IEEE Ind. Electron. Mag.*, vol. 1, no. 4, pp. 6–16, Winter 2007.

[22] J. M. Maciejowski, *Multivariable Feedback Design*. Wokingham, UK: Addison-Wesley, 1989.

[23] G. Maione, "Concerning continued fractions representation of noninteger order digital differentiators," *IEEE Signal Process. Lett.*, vol. 13, no. 12, pp. 725-728, Dec. 2006.

[24] G. Maione, "Continued fractions approximation of the impulse response of fractional order dynamic systems," *IET Control Theory Appl.*, vol. 2, no. 7, pp. 564–572, July 2008.

[25] G. Maione, "Conditions for a class of rational approximants of fractional differentiators/integrators to enjoy the interlacing property", In: S. Bittanti, A. Cenedese, S. Zampieri (Eds.), Proceedings of the 18th IFAC World Congress (IFAC WC 2011), Milan, Italy, Aug. 28 - Sept. 2, 2011, IFAC-PapersOnLine: Vol. 18, Part 1, pp. 13984-13989.

[26] G. Maione, "Closed-form rational approximations of fractional, analog and digital differentiators/integrators," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 3, no. 3, pp. 322-329, Sep. 2013.

[27] G. Maione, Correction to "Closed-form rational approximations of fractional, analog and digital differentiators/integrators," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 3, no. 4, p. 654, Dec. 2013.

[28] K. Matsuda and H. Fujii, "H_∞ optimized wave-absorbing control: Analytical and experimental results," *J. Guid. Control Dyn.*, vol. 16, no. 6, pp. 1146-1153, 1993.

[29] C. A. Monje, B. M. Vinagre, V. Feliu, Y. Q. Chen, "Tuning and autotuning of fractional order controllers for industry applications," *Control Eng. Pract.*, vol. 16, pp. 798-212, 2008.

[30] C. A. Monje, Y. Q. Chen, B. M. Vinagre, D. Xue, V. Feliu, *Fractional-order Systems and Controls: Fundamentals and Applications*. London, UK: Springer-Verlag, 2010.

[31] Yu. I. Neimark. *Ustoichivost' linearizovannykh system upravleniya (Stability of Linearized Control Systems)*. Leningrad: LKVVIA, 1949.

[32] K. B. Oldham, J. Spanier, *The fractional calculus: Integrations and Differentiations of Arbitrary Order*. New York: Academic Press, 1974.

[33] A. Oustaloup, *La Commande CRONE. Command Robuste d'Ordre Non Entièr*. Paris: Hermès, 1991.

[34] A. Oustaloup, F. Levron, B. Mathieu, F. M. Nanot, "Frequency-band complex noninteger differentiator: Characterization and synthesis," *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. 47, no. 1, pp. 25-39, Jan. 2000.

[35] I. Podlubny, *Fractional Differential Equations*. San Diego, CA: Academic Press, 1999.

[36] I. Podlubny, "Fractional order systems and $PI^{\lambda}D^{\mu}$-controllers," *IEEE Trans. Autom. Control*, vol. 44, no. 1, pp. 208-214, Jan. 1999.

[37] I. Podlubny, I. Petráš, B. M. Vinagre, P. O'Leary, L. Dorčák, "Analogue realization of fractional-order controllers," *Nonlinear Dyn.*, vol. 29 (1-4), pp. 281-296, 2002.

[38] K. Price, R. Storn, J. Lampinen, *Differential Evolution—A Practical Approach to Global Optimization*. Berlin: Springer, 2005.

[39] B. Ross, "The development of fractional calculus 1695–1900," *Historia Mathematica*, 4, pp. 75-89, 1977.

[40] S. Skogestad, I. Po1stlethwaite, *Multivariable Feedback Control: Analysis and Design*, 2nd ed. Chichester: Wiley & Sons, Ltd., 2005.

[41] T. J. Stieltjes, *Œuvres Completes de Thomas Jan Stieltjes, Tome II*. Groningen: P. Noordhoff, Pub. par le soins de la Société Mathématique D'Amsterdam, 1918.

[42] R. Storn, K. Price, "Differential Evolution: A Simple and Efficient Heuristic for Global Optimization Over Continuous Spaces," *J. Global Optim.*, 11 (4), pp. 341-359, 1997.

[43] A. Tustin, "The design of systems for automatic control of the position of massive objects," *Proc. Inst. Electr. Eng.*, 105, part C, suppl. no. 1, 1-57, 1958.

[44] V. V. Uchaikin, *Fractional Derivatives for Physicists and Engineers: Volume I Background and Theory*. Berlin Heidelberg: Springer, 2013.

[45] J. Vesterstrøm, R. Thomsen, "A Comparative Study of Differential Evolution, Particle Swarm Optimization, and Evolutionary Algorithms on Numerical Benchmark Problems," Congress on Evolutionary Comp. (CEC2004), Portland, OR, vol. 2, pp. 1980-1987, June 19-23, 2004.

Distributed Consensus-Based Multi-Target Tracking without Measurement Assignment

Srdjan S. Stanković, Nemanja Ilić and Miloš S. Stanković

***Abstract*—In this paper the problem of distributed multi-target tracking in sensor networks is discussed. A new algorithm, based on a combination of probabilistic data association methodology requiring no explicit measurement assignment and consensus communication scheme, is proposed. Unlike standard joint data association methodologies, which enumerate all possible measurement-to-track assignments resulting in a numerical complexity that is combinatorial in the number of measurements and tracks, we propose an approach which is linear in this respect. This is accomplished by constructing a data association scheme which appropriately modifies the clutter spatial density at the location of the measurements and subsequently uses single target tracking filters. The proposed data association algorithm is combined with a multi-step consensus scheme, which provides an adaptive distributed tracking solution when sensors have limited sensing and communication ranges. Numerical experiments illustrate the underlying principles and performance of the proposed algorithm.**

***Index Terms*—Distributed multi-target tracking, Sensor networks, Data Association, Consensus.**

I. Introduction

In recent years sensor networks have attracted significant attention in the research community. One of the main applications has been target tracking, where sensors such as radars, sonars or cameras have been spatially dispersed in order to contribute to the task of wide area surveillance, environment monitoring, disaster response, etc [1]. This task can be solved using a centralized approach, where necessary information for global target state estimation is collected at specific points in the network, called fusion centers. The centralized solution fails in situations when fusion centers become inoperable or when communications with them cease. On the other side, distributed algorithms, where all nodes in the network can be used as global target state estimators, and where global decision is obtained only through local communications between nodes, offer more robust, fault tolerant and scalable solutions.

Consensus scheme has been widely used as one of the most popular algorithms in the domain of distributed state estimation [2], [3], [4], and, more specifically, in the domain of target tracking [5]. One of the main challenges for the successful consensus-based distributed target tracking has been addressing the problem of limited sensing range sensors [6], [7], [8], where at each time instant there can be a vast majority of sensors not observing the target. In addition, there can be sensor measurements coming from some random sources other than target, usually termed clutter; this issue is commonly solved by combining Probabilistic Data Association (PDA) methodologies [9] with consensus-based estimation. Further, the existence of multiple targets complicates the problem even more, and this situation has been solved using the Joint Probabilistic Data Association (JPDA) methodologies [10], [11], [12], [13]. PDA and JPDA approaches can also be extended, by incorporating the concept of target perceivability and target existence, toward Integrated Probabilistic Data Association (IPDA) solutions [14], [15] for track initiation, confirmation and termination, as well as track maintenance [16].

While tracking multiple targets, it is possible that, for each target track, one applies the PDA methodology by assuming that all measurements not associated with that track are false (*i.e.*, they come from clutter) [12]. In practical applications this often leads to track coalescence in cases of spatially close tracks. The JPDA approach assumes that, for each target, measurements not associated with it may come from both other targets and clutter. It creates a list of all feasible joint measurement-to-track assignments which is used for computing measurement association probabilities. Feasible events do not include situations where a single measurement has been assigned to multiple targets, which turns to be beneficial in view of suppressing track coalescence. However, the process of enlisting all feasible measurement-to-track assignments is combinatorial in the number of measurements and the number of tracks, representing a significant computational burden.

In [17] the computational requirements of JPDA have been bypassed by introducing a novel methodology, called linear multi-target (LM) method. Namely, similarly as the PDA approach, LM assumes that, for each target track, all measurements not associated with it come from clutter. However, unlike the PDA approach, which assumes one global level of clutter spatial density (average number of measurements coming from clutter per unit volume), LM modifies this clutter spatial density for each target track with the measurement intensity of other targets. This effectively reduces the degree of association of a measurement to a target when that measurement is strongly associated with another target. As a result, coupling between individual tracks without explicit measurement-to-track combinatorial enumeration process is achieved, linear in the number of measurements and tracks. LM also belongs to the class of IPDA algorithms, incorporating the concepts of target perceivability and target existence.

Regarding the choice of the data association methodology,

S. S. Stanković is with the School of Electrical Engineering, University of Belgrade, Serbia and Vlatacom Institute, Belgrade, Serbia; E-mail: stankovic@etf.rs

N. Ilić, who is the corresponding author, is with the College of Applied Technical Sciences, Kruševac, Serbia and Vlatacom Institute, Belgrade, Serbia; E-mail: nemili@etf.rs

M. S. Stanković is with the Innovation Center, School of Electrical Engineering, University of Belgrade, Serbia; Vlatacom Institute, Belgrade, Serbia; COPELABS, Universidade Lusófona, Lisboa, Portugal and Singidunum University, Belgrade, Serbia; E-mail: milsta@kth.se

consensus-based distributed multi-target tracking algorithms have followed several directions in the literature so far. In [11] the so-called Kalman Consensus Filter (KCF) has been combined with the JPDA methodology. It assumed a simplified scenario of all sensors observing the whole surveillance region. The Information Consensus Filter (ICF) has been combined with the PDA algorithm in [12]. KCF and ICF rely on the exchange between the nodes of the so-called information vectors and matrices which are subsequently used to calculate the target state estimates. In [13], the Adaptive Consensus Filter (ACF) from [18] has been combined with the JPDA methodology. It assumes the exchange of state estimates between the nodes and adaptively tunes the consensus parameters in order to take into account the adopted multi-target scenario with cluttered measurements and limited sensing range sensors.

The goal of this paper is to propose a novel algorithm which combines the LM approach from [17] with the ACF algorithm from [13], [18], in order to obtain a distributed multi-target tracking solution whose computational requirements are significantly smaller than those of solutions incorporating JPDA. The scheme will also allow for extensions involving track-to-track association, as well as integration of track initiation, confirmation and termination with the track maintenance process. It will be shown that the proposed tracking algorithm provides accurate network-wide estimations of the positions of all targets, and that it exhibits the best overall performance, when compared to its version which uses JPDA, as well as to other state-of-the-art algorithms. Also, the illustration of some properties of the proposed algorithm, aimed at clarification of the main ideas upon which it is built, will be given.

The outline of the paper is as follows. Problem setting and the proposed linear multi-target method are discussed in Section II. The adaptive multi-step consensus scheme is presented in Section III. Section IV gives illustration of some of the features of the proposed algorithm, together with the results of characteristic multi-target tracking simulations.

II. Distributed Consensus-Based Linear Multi-Target Tracking Algorithm

In the first part of this Section, we shall present the main concepts of the adopted multi-target tracking strategy, applied to local estimators taken separately. In the second part, the complete distributed algorithm, including the consensus based scheme applicable at the network level, will be given.

A. Problem Formulation. Basic Concepts

Consider the problem of tracking N_T targets by N intelligent sensors forming a network represented by a directed graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, where $\mathcal{N}$ is the set of nodes and $\mathcal{E}$ the set of arcs. Since the scope of this paper is track maintenance only, it is assumed that N_T is known to each node [11], [12], [13]. Each sensor (*e.g.*, radar, sonar, camera) is supposed to have, in general, limited sensing and communication ranges, determining the sets of nodes that can observe the targets (one set for each target) and the set of neighboring nodes exchanging messages. Let $\mathcal{N}_i$ be the in-neighborhood of node i, and $\mathcal{J}_i = \mathcal{N}_i \cup \{i\}$. We shall assume cluttered environment, *i.e.*, a set of measurements is acquired by each sensor without information of their origin: at time t, node i, $i = 1, \ldots, N$, gets m_i measurements, denoted as $z_{i,j}(t)$, $j = 1, \ldots, m_i$. The measurement of a potential target followed by track k, $k = 1, \ldots, N_T$, is present at node i with a probability of detection $P^k_{D,i}$; it is also corrupted by a zero mean white measurement noise, with covariance R_i. Target detection is assumed to be in the selection gate or window with probability $P^k_{W,i}$. Let $z^k_{i,j}(t)$ denote j-th measurement in the selection gate of track k, and let its a priori probability density function (pdf) $p^k_{i,j}(t)$ be Gaussian; let Z^t_i be the set of measurements at node i up to and including the instant t. The clutter spatial density at $z_{i,j}(t)$ is denoted by $\rho_{i,j}(t)$ (see [17] for more details).

A track may follow a target, when it is called true track; otherwise, it is a false track. We introduce the following important random events at time t: $\chi^k_i(t)$ - the event that track k at node i is following a target, *i.e.*, the target k exists; $\chi^k_{i,0}(t)$ - the event that none of the gated measurements are from target k; $\bar{\chi}^k_{i,0}(t)$ - complement of $\chi^k_{i,0}(t)$; $\chi^k_{i,j}(t)$ - the event that measurement $z^k_{i,j}(t)$ is the target detection.

The classical multi-target tracking methodology (*e.g.*, [10]) considers the origin of each acquired measurement. The number of unique joint assignments of all measurements to all tracks is combinatorial in the number of measurements and tracks [17]. The existing approaches to distributed multi-target tracking [11], [13] assume application of the Joint Probabilistic Data Association (JPDA) methodology and suffer from the corresponding computational burden, especially pronounced in networked systems. In order to obtain a more efficient and effective solution, we shall follow in this paper the Linear Multi-target (LM) methodology, proposed in [17]. Essentially, it is based on the single target tracking methodology, allowing for the possibility that a measurement originates either from a target or some other scatterer, which may produce clutter returns or be other targets. In effect, the LM approach is to run a bank of single target tracking filters, achieved though modifying the clutter density for each tracking filter.

Let $P^k_{i,j}(t) = P(\chi^k_{i,j}(t)|Z^{t-1}_i)$ denote a priori probability that measurement j of node i at time t is from target k. According to [17], we shall use the following approximation

$$P^k_{i,j}(t) \approx P^k_{D,i} P^k_{W,i} P(\chi^k_i(t)|Z^{t-1}_i) \frac{p^k_{i,j}(t)}{\rho^k_{i,j}(t)} \Big(\sum_{\mu=1}^{m^k_i(t)} \frac{p^k_{i,\mu}(t)}{\rho^k_{i,\mu}(t)} \Big)^{-1} \tag{1}$$

where

$$\rho^k_{i,j}(t) = \rho_{i,j}(t) + \sum_{\sigma=1, \sigma \neq k}^{N_T} P^k_{D,i} P^k_{W,i} P(\chi^k_i(t)|Z^{t-1}_i) p^\sigma_{i,j}(t);$$

$P(\chi^k_i(t)|Z^{t-1}_i)$ is the a priori probability of target k existence.

We shall assume the Markov chain one model for the propagation of the probability of target existence, see [14], [15] for details. Accordingly, computation of the a posteriori probability of track existence is done by

$$P(\chi^k_i(t)|Z^t_i) = \frac{(1 - \delta^k_i(t)) P(\chi^k_i(t)|Z^{t-1}_i)}{1 - \delta^k_i(t) P(\chi^k_i(t)|Z^{t-1}_i)}, \tag{2}$$

where

$$\delta_i^k(t) = P_{D,i}^k P_{W,i}^k (1 - \sum_{\mu=1}^{m_i^k(t)} \frac{p_{i,\mu}^k(t)}{\tilde{\rho}_{i,\mu}^k(t)}). \quad (3)$$

Following [17], $\tilde{\rho}_{i,j}^k(t)$ is the a priori scatterer measurement density of $z_{i,j}(t)$ given by

$$\tilde{\rho}_{i,j}^k(t) = \rho_{i,j}(t) + \sum_{\sigma=1,\sigma\neq k}^{N_T} p_{i,j}^\sigma(t) \frac{P_{i,j}^\sigma(t)}{1 - P_{i,j}^\sigma(t)}. \quad (4)$$

Introduction of $\tilde{\rho}_{i,j}^k(t)$, instead of $\rho_{i,j}(t)$ utilized in conventional trackers of PDA or IPDA type [9], [10], [14], represents the core of the LM tracking methodology [17]. The aim is to take into account the presence of targets nearby by calculating the probability of target existence using (2). The effect of $\tilde{\rho}_{i,j}^k(t)$ is to decrease the influence of a measurement if it is likely to be the detection of another target, without requiring explicit joint measurement-to-track assignment.

The a posteriori probabilities that the target exists and that measurement $i \geq 0$ is its detection ($i = 0$ denotes the null measurement) are given by

$$P(\chi_i^k(t), \chi_{i,0}^k(t)|Z_i^t) = \frac{1 - P_{D,j}^k P_{W,j}^k}{1 - \delta_i^k(t) P(\chi_i^k(t)|Z_i^{t-1})} P(\chi_i^k(t)|Z_i^{t-1}), \quad (5)$$

$$P(\chi_i^k(t), \chi_{i,j}^k(t)|Z_i^t) = \frac{P_{D,j}^k P_{W,j}^k}{1 - \delta_i^k(t) P(\chi_i^k(t)|Z_i^{t-1})} \frac{p_{i,j}^k(t)}{\tilde{\rho}_{i,j}^k(t)} P(\chi_i^k(t)|Z_i^{t-1}). \quad (6)$$

B. Tracking Algorithm

The proposed tracking algorithm consists of two main parts: a) correction part and b) prediction part.

a) The correction part of the algorithm is defined according to the algorithm from [8], [19], combined with [9], [10], modified using the above exposed LM concept [17]:

$$\xi_i^k(t|t) = \xi_i^k(t|t-1) + L_i^k(t)\tilde{z}_i^k(t), \quad (7)$$

$i = 1, \ldots, N$, where $\xi_i^k(\cdot)$ is an estimate of the true target state $x^k(\cdot)$, generated by *i*-th node

$$\tilde{z}_i^k(t) = \sum_{j=1}^{m_i^k(t)} \beta_{i,j}^k(t)\tilde{z}_{i,j}^k(t), \quad (8)$$

$\tilde{z}_{i,j}^k(t) = z_{i,j}^k(t) - H_i\xi_i^k(t|t-1)$, where matrix H_i is a part of the observation model

$$z_{i,j}^k(t) = H_i x^k(t) + v_i(t), \quad (9)$$

in which $z_{i,j}^k(t)$ is a measurement of node i originated from k-th target and $v_i(t)$ a zero mean white noise term with covariance $R_i > 0$,

$$L_i^k(t) = P_i^k(t|t-1)(H_i^k)^T S_i^k(t)^{-1}, \quad (10)$$

$$P_i^k(t|t) = P_i^k(t|t-1) + [\beta_{i,0}^k(t) - 1]L_i^k(t)S_i^k(t)L_i^k(t)^T + \tilde{P}_i^k(t,t), \quad (11)$$

$$\tilde{P}_i^k(t,t) = L_i^k(t)[\sum_{j=1}^{m_i(t)} \beta_{i,j}^k(t)\tilde{z}_{i,j}^k(t)\tilde{z}_{i,j}^k(t)^T - \tilde{z}_i^k(t)\tilde{z}_i^k(t)^T]L_i^k(t)^T, \quad (12)$$

$$S_i^k(t) = H_i^k P_i^k(t|t-1)(H_i^k)^T + R_i, \quad (13)$$

$$\beta_{i,0}^k(t) = \frac{P(\chi_{i,0}^k(t), \chi_i^k(t)|Z_i^t)}{P(\chi_i^k(t)|Z_i^t)} = \frac{1 - P_{D,i}^k P_{W,i}^k}{1 - \delta_i^k(t)}, \quad (14)$$

$$\beta_{i,j}^k(t) = \frac{P(\chi_{i,j}^k(t), \chi_i^k(t)|Z_i^t)}{P(\chi_i^k(t)|Z_i^t)} = \frac{P_{D,i}^k P_{W,i}^k}{1 - \delta_i^k(t)} \frac{p_{i,j}^k(t)}{\tilde{\rho}_{i,j}^k(t)}. \quad (15)$$

Formally, at the first glance, the correction part of the algorithm is identical to the ACF algorithm proposed and discussed in [16], [13]. The essential difference stems from the new definition of the β-coefficients: in (14) and (15) a new expression for clutter density is introduced, directly taking care of the influence of the existence of other targets (see (4), (3)).

b) The prediction part of the algorithm assumes the standard target model (equal for all targets for the sake of clarity)

$$x^k(t+1) = Fx^k(t) + Ge^k(t), \quad (16)$$

where F and G are constant matrices of appropriate dimensions and $e^k(t)$ Gaussian white noise term, with covariance $Q^k > 0$. The algorithm contains, as its basic part, the following relation for state prediction

$$\xi_i^k(t+1|t) = F\mathcal{C}_{[l]}^k\{\xi_i^k(t|t)\}, \quad (17)$$

where $\mathcal{C}_{[l]}^k\{\cdot\}$ denotes the l-step consensus operator, obtained from the following recursion for an adopted l:

$$\mathcal{C}_{[l]}^k\{\xi_i^k(t|t)\} = \sum_{j\in\mathcal{J}_i} c_{[l],ij}^k(t)\mathcal{C}_{[l-1]}^k\{\xi_j^k(t|t)\}, \quad (18)$$

with $\mathcal{C}_{[0]}^k\{\xi_i^k(t|t)\} = \xi_i^k(t|t)$, $i = 1, \ldots, N$. Coefficients $c_{[l],ij}^k(t)$, $i, j = 1, \ldots, N$, are, in general, dependent on l; they are elements of the $N \times N$ consensus matrix $C_{[l]}^k(t)$, which is row-stochastic and with nonnegative elements, such that $c_{[l],ij}^k(t) = 0$ for $j \notin \mathcal{J}_i$ [2], [4], [20].

Furthermore,

$$P_i^k(t+1|t) = FP_i^k(t|t)F^T + GQ^kG^T. \quad (19)$$

In the case of single node operation (fixed i and $\mathcal{C}_{[l]}^k\{\xi_i^k(t|t)\} = \xi_i^k(t|t)$), the algorithm approximates the posterior state estimate pdf by a single Gaussian pdf

$$p(x_i^k(t)) \sim \mathcal{N}(\xi_i^k(t|t), P_i^k(t|t)), \quad (20)$$

[17]. In general, the presence of the consensus operator (18) implies that (20) does not hold. The overall covariance of the estimates generated by the proposed algorithm is significantly different from the local covariances defined by (11) and (19); it is, in general, a function of all the local estimates and the consensus parameters. It is to be taken into account that, for a large number of consensus steps l, we have $\xi_i^k(t+1|t) \approx \xi_j^k(t+1|t)$,

$i, j = 1, \ldots, N$, and that, consequently, it can be expected that the resulting prediction covariance of the estimates is lower than the majority of the local covariances (provided some general properties hold for the consensus matrices - see the discussion below). This implies that practical implementation of the algorithm is faced with the problem of adequate real-time estimation of the estimation error covariance, having in mind that this covariance is to be incorporated in $p^k_{i,j}(t)$, which plays an essential role in the whole algorithm based on the LM tracking methodology. The problem is that the estimation of this covariance cannot be done in a distributed way - it is feasible only off-line, including all the nodes and their interconnections. However, based on numerous experiments with different target models and diverse problem settings, we have found that $P^k_i(t|t)$ in (20) can be successfully replaced by $\Pi^k_{[l],i}(t|t)$ defined by

$$\Pi^k_{[l],i}(t|t) = [\alpha_k \mathcal{C}^k_{[l]}\{P^k_i(t|t)^{-1}\}]^{-1}, \tag{21}$$

where $\alpha_k \geq 1$. Obviously, we have, for l large enough, $\Pi^k_{[l],i}(t|t) \approx \Pi^k_{[l],j}(t|t)$, $i, j = 1, \ldots, N$, under the same conditions leading to consensus on the estimates themselves.

It is to be noticed that incorporating (21) in the distributed tracking algorithm brings additional communication burden since the estimation error covariances should now also be exchanged between nodes. This requirement is already implicitly present with other consensus-based distributed target tracking algorithms [11], [12]. However, if one wants to lift this burden off toward off-line estimation of error covariances, the facts that different values in (21) are for l large enough very similar, and that (having in mind the nature of the tracking process) they do not vary significantly in time, may be taken into account. Therefore, while computing (4), it is possible to introduce the approximation of (21) by simply using $\Pi^k_{[l],i}(t|t) = \Pi^k_{const}$, where Π^k_{const} is some empirically obtained constant (see Section IV for details).

III. Adaptive Multi-Step Consensus Scheme

One of the essential features of the proposed tracking algorithm is its possibility to adapt behavior to the availability of measurements; this Section deals with the problem of generating consensus matrices $C^k_{[l]}(t) = [c^k_{[l],ij}(t)]$ in (18) that will allow this feature. The underlying idea is directed toward the desirable asymptotic properties of the consensus scheme. It aims to achieve that all nodes have at each time t equal target state estimates, in the form of a weighted average of the target state estimates, in which the weights are designed in such a way as to reflect the current node probability of observing a target. Generalizing analogous considerations from [16], we use $1 - \beta^k_{i,0}(t)$, where $\beta^k_{i,0}(t)$ is given by (14), as an indicator of the i-th node probability of observing the target. Therefore, the formal goal of the multi-step consensus operator in (18) should be to achieve asymptotically (when $l \to \infty$) that

$$\xi^k_i(t+1|t) = \xi^k(t+1|t) = \sum_{i=1}^{N} w^k_i(t)^T F \xi^k_i(t|t), \tag{22}$$

where

$$w^k_i(t) = \frac{1 - \beta^k_{i,0}(t)}{\sum_{j=1}^{N}(1 - \beta^k_{j,0}(t))}. \tag{23}$$

To this end, at each time instant t, we shall apply l consensus steps. In each consensus step, the nodes exchange the scalar variables based on $\gamma^k_i(t) = 1 - \beta^k_{i,0}(t)$, as well as the target state estimates. The variables are used for weighting the communicated estimates. Let $\gamma^k_{[\kappa],i}(t)$ and $\xi^k_{[\kappa],i}(t|t)$, $\kappa = 1, \ldots, l$, be the weight and the estimate of the i-th node connected to the κ-th consensus step, respectively. The algorithm starts from $\gamma^k_{[1],i}(t) = \gamma^k_i(t)$ and $\xi^k_{[1],i}(t|t) = \xi^k_i(t|t)$. The weights $\gamma^k_{[\kappa],i}(t)$ are being exchanged through A_c (satisfying $\lim_{n\to\infty} A^n_c = \mathbf{1}\mathbf{1}^T/N$, where $\mathbf{1}$ is a column vector of ones, see [18], [13] for details), so that the corresponding consensus matrix is given by

$$C^k_{[\kappa]}(t) = \left(A_c \cdot diag\left(\gamma^k_{[\kappa],1}(t), \ldots, \gamma^k_{[\kappa],N}(t)\right)\right)_{rs}, \tag{24}$$

where $(\cdot)_{rs}$ denotes an operator making the resulting matrix row-stochastic. Using the obtained matrix $C^k_{[\kappa]}(t) = [c^k_{[\kappa],ij}(t)]$, $i, j = 1, \ldots, N$, the target state estimates are given by $\xi^k_{[\kappa+1],i}(t|t) = \sum_{j\in\mathcal{J}_i} c^k_{[\kappa],ij}(t)\xi^k_{[\kappa],j}(t|t)$. For the next consensus step, the design requires that the weight of each node corresponds to the sum of the weights it received previously, so that $\gamma^k_{[\kappa+1]}(t) = \begin{bmatrix}\gamma^k_{[\kappa+1],1}(t) & \cdots & \gamma^k_{[\kappa+1],N}(t)\end{bmatrix}^T = A_c \cdot \gamma^k_{[\kappa]}(t)$. At this point one can repeat the described procedure for the total of l consensus steps. It is shown [18] that, for $l \to \infty$, this scheme achieves consensus equivalent to $\lim_{l\to\infty} \xi^k_{[l],i}(t|t) = \sum_{j\in\mathcal{J}_i} c^k_{[\infty],ij}(t)\xi^k_j(t|t)$, where $c^k_{[\infty],ij}(t) = \gamma^k_j(t)/\sum_{j=1}^{N}\gamma^k_j(t)$, which represents the desired result.

To sum it up, the consensus operator $\mathcal{C}^k_{[l]}\{\cdot\}$ in (18) is obtained by applying l consensus steps defined in (24), which is equivalent to the application of a single consensus step whose corresponding consensus matrix $\tilde{C}^k_{[l]}(t)$ is defined by:

$$\tilde{C}^k_{[l]}(t) = C^k_{[l]}(t) \cdot C^k_{[l-1]}(t) \cdot \ldots \cdot C^k_{[1]}(t). \tag{25}$$

The described consensus algorithm can be optimized in the sense of obtaining fast convergence rate, enabling efficient implementation in practice [18]. The procedure aims to minimize the spectral radius of $A_c - \mathbf{1}\mathbf{1}^T/N$ subject to the inherent network constraints (see [18] for details).

IV. Simulations

The problem of distributed tracking of $N_T = 4$ targets via a network of $N = 15$ sensors, each target moving within a 500×500 space, was considered [12], [18], [13]. The sensors were randomly distributed in the space with random orientations resulting in overlapping field-of-views (FOVs), represented by equilateral triangles with the height of 300 units; the communication range was set to 200 units. The initial prior state estimates were randomly selected around initial target states with noise covariance of $P_0 = diag(100, 100, 10, 10)$. The initial error covariance matrices were also set to P_0 for all nodes. The number of consensus steps was set to $K = 10$ and α_k was set to 1. Regarding the parameters included in calculating $\beta^k_{i,j}(t)$ (see [17] for details), false measurements (clutter) were assumed Poisson distributed with the nominal spatial density of $1/32$, gate probability was set to 0.99,

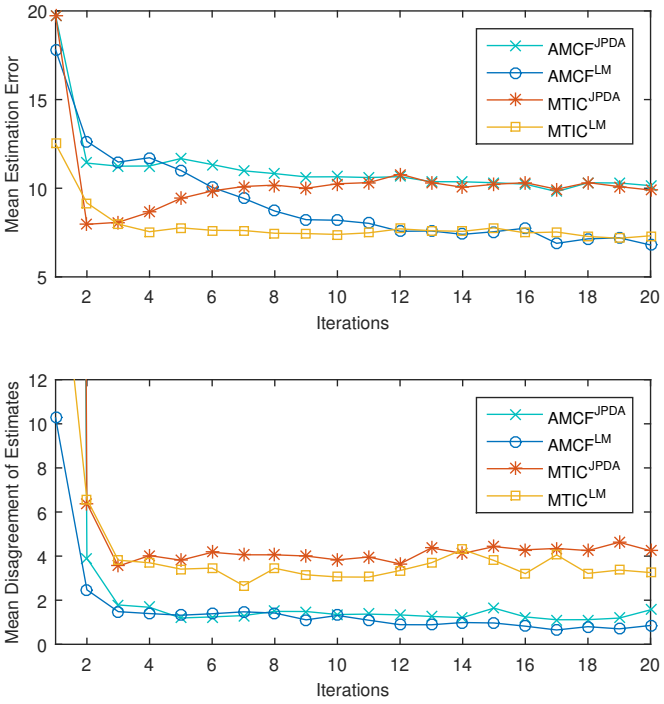


Fig. 1. Average distance per node per target between the true and the estimated positions (top) and average variance of the estimates per target (bottom) versus time.

and probability of detection was calculated by integrating the probability density function of the predicted estimate over the triangular FOV. For other parameters, see, *e.g.*, [12], [18], [13].

An experiment with 100 different randomly selected networks, each with 4 randomly generated tracks, was designed. Estimation error (average distance per node per target between targets positions estimates and the actual positions) and disagreement between the nodes (average variance of the estimates per target) were calculated. The proposed algorithm (named AMCF$^{\text{LM}}$) was also simulated by using JPDA instead of LM (AMCF$^{\text{JPDA}}$). The existing state-of-the-art algorithm from the literature based on the Information Consensus Filter [12] was likewise simulated in two versions (MTIC$^{\text{LM}}$ and MTIC$^{\text{JPDA}}$). It can be seen from Fig. 1 that the proposed algorithm with LM provides the best results both in terms of the average estimation error and the disagreement of estimates. In terms of the estimation error, LM versions of the algorithms provide better results over using JPDA; in terms of the disagreement between estimates, AMCF algorithms exhibit better performance than MTIC algorithms.

In order to illustrate the underlying principles of how LM scheme actually works, an experiment was set with 2 moving targets and with clutter nominal spatial density of $1/4$. The calculated a priori scatterer measurement density $\tilde{\rho}^k_{i,j}(t)$ from (4) is depicted in Fig. 2, for all nodes and all their measurements obtained during the course of the experiment, $i = 1, \ldots, N$, $j = 1, \ldots, m_i$. It can be seen that, for each target, the proposed algorithm indeed amplifies the clutter density of a measurement when that measurement is strongly associated with another target (which subsequently reduces its degree of association to the original target).

As a supplement to the discussion following (21), values of error covariance matrices used in calculating (4) were, for each target, set to two constant values ($\Pi^k_{const} = P_0$ and $\Pi^k_{const} = 10P_0$), in an experiment of tracking 4 moving targets. As can be seen from Fig. 3, this empirical procedure can lead to a successful multi-target tracking scheme, but one should be careful about the quantities chosen, since greater values can lead to track coalescence (right subfigure).

V. Conclusion

In this paper we proposed a novel algorithm for distributed multi-target tracking in sensor networks, based on a combination of: 1) probabilistic data association methodology whose computational requirements are linear in the number of measurements and tracks, and 2) adaptive multi-step consensus scheme, robust in the context of cluttered observations and limited sensing range sensors. The proposed algorithm exhibits better performance than analogous algorithms that use JPDA - this seems to be the consequence of the numerical overflows and underflows inherently present when calculating association probabilities via JPDA [10]. It also exhibits better performance than analogous algorithms from the literature that use Information Consensus Filters (ICF) [12] - this is due to the adopted general consensus strategy that is not restricted to average consensus only.

The future work should address the problem of track-to-track association, which is herein assumed to be perfect (as in [12], [13]). However, this represents a straightforward task, in contrast to the case of using ICF algorithms which do not explicitly exchange state estimates between nodes. Also, the assumed integrated PDA methodology allows for the introduction of track initiation, confirmation and termination phases into the proposed scheme, which would than represent a complete tracking solution.

References

[1] D. E. Maggio and D. A. Cavallaro, *Video Tracking: Theory and Practice*, 1st ed. Hoboken, New Jersey: Wiley Publishing, 2011.

[2] R. Olfati-Saber, A. Fax, and R. Murray, "Consensus and cooperation in networked multi-agent systems," *Proceedings of the IEEE*, vol. 95, pp. 215–233, 2007.

[3] W. Ren and R. Beard, "Consensus seeking in multi-agent systems using dynamically changing interaction topologies," *IEEE Trans. Autom. Control*, vol. 50, pp. 655–661, 2005.

[4] J. N. Tsitsiklis, D. P. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE Trans. Autom. Control*, vol. 31, pp. 803–812, 1986.

[5] R. Olfati-Saber, "Distributed Kalman filtering for sensor networks," in *Proc. IEEE Conf. Decision and Control*, 2007, pp. 5492–5498.

[6] R. Olfati-Saber and N. F. Sandell, "Distributed tracking in sensor networks with limited sensing range," in *Proc. American Control Conference*, 2008, pp. 3157–3162.

[7] A. T. Kamal, J. A. Farrell, and A. K. Roy-Chowdhury, "Information weighted consensus filters and their application in distributed camera networks," *IEEE Transactions on Automatic Control*, vol. 58, no. 12, pp. 3112–3125, Dec 2013.

[8] N. Ilić, M. S. Stanković, and S. S. Stanković, "Adaptive consensus-based distributed target tracking in sensor networks with limited sensing range," *IEEE Transactions on Control Systems Technology*, vol. 22, no. 2, pp. 778–785, March 2014.

[9] Y. Bar-Shalom and E. Tse, "Tracking in a cluttered environment with probabilistic data association," *Automatica*, vol. 11, no. 5, pp. 451 – 460, 1975. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0005109875900217

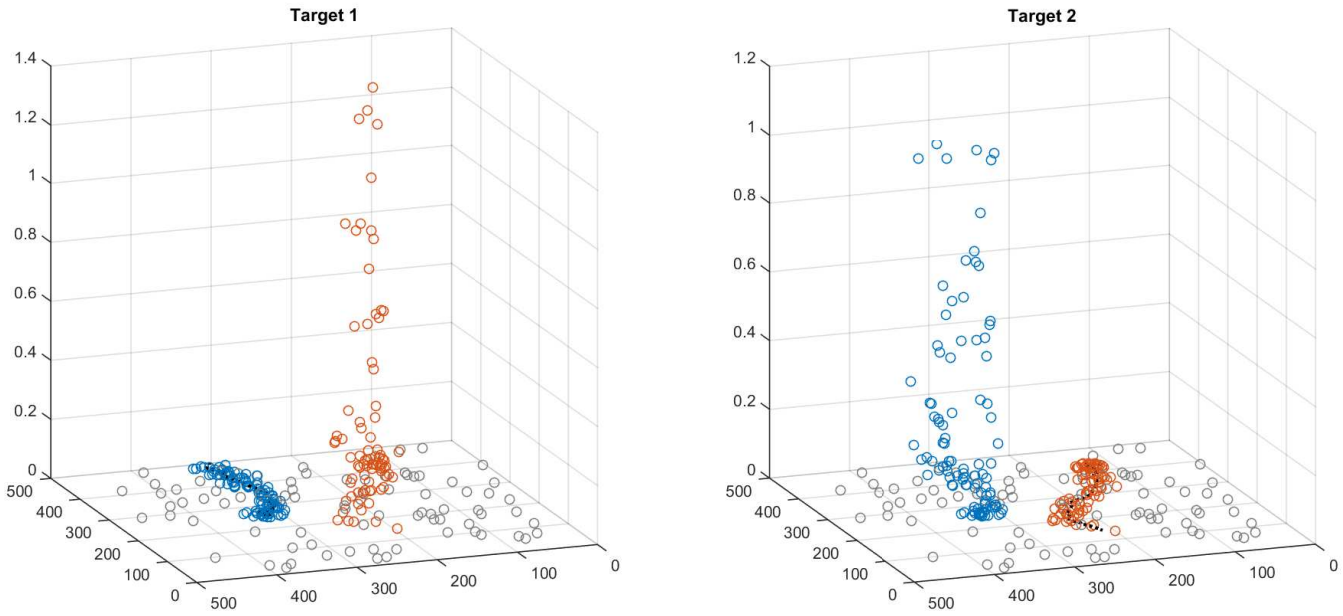


Fig. 2. The a priori scatterer measurement density (shown along z-axis) for two targets (moving in x-y plane) for all measurements (measurements originating from target one, target two and clutter shown in blue, orangered and gray, respectively). Target trajectories are shown by black dotted lines.

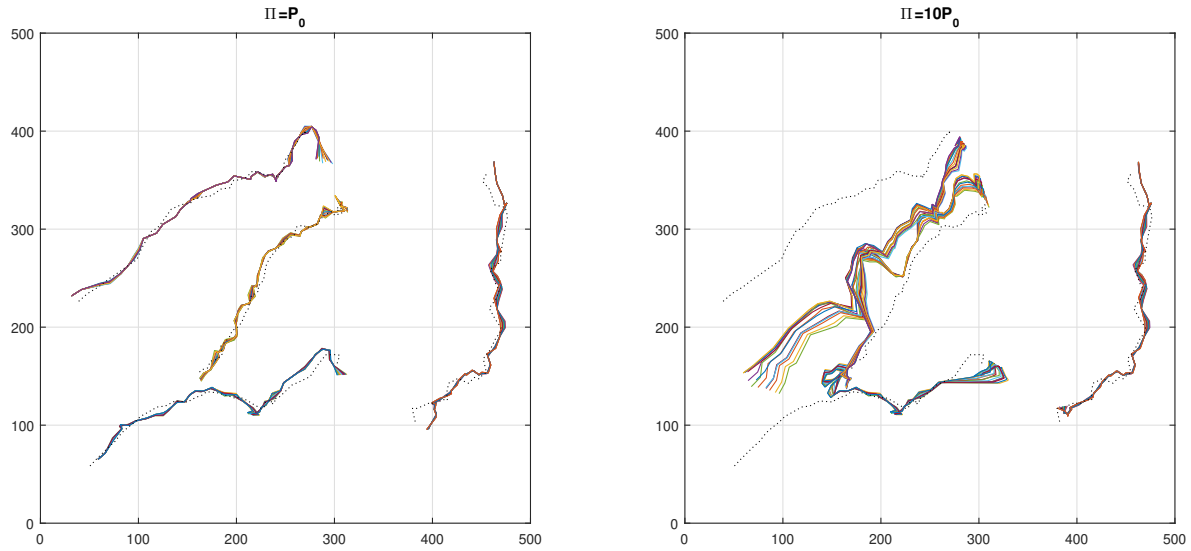


Fig. 3. Targets positions estimates of all nodes (differently colored solid lines), together with targets trajectories (black dotted lines).

[10] T. E. Fortmann, Y. Bar-Shalom, and M. Scheffe, "Sonar tracking of multiple targets using joint probabilistic data association," *IEEE J. of Oceaninc Eng.*, vol. OE-8, pp. 173–184, 1983.

[11] N. F. Sandell and R. Olfati-Saber, "Distributed data association for multi-target tracking in sensor networks," in *2008 47th IEEE Conference on Decision and Control*, Dec 2008, pp. 1085–1090.

[12] A. T. Kamal, J. H. Bappy, J. A. Farrell, and A. K. Roy-Chowdhury, "Distributed multi target tracking and data association in vision networks," *IEEE Trans. Patt. Analysis and Mach. Intel.*, vol. 38, pp. 1397–1410, 2016.

[13] N. Ilić, K. O. Al-Ali, M. Stanković, and S. Stanković, "Distributed multi-target tracking in camera networks using multi-step consensus," in *Proc. 4th IcETRAN Conf.*, 2017.

[14] D. Musicki, R. Evans, and S. Stankovic, "Integrated probabilistic data association," *IEEE Trans. Automat. Contr.*, vol. 39, pp. 1237–1241, 1994.

[15] D. Musicki and R. Evans, "Joint integrated probabilistic data association - JIPDA," in *Proceedings of the Fifth International Conference on Information Fusion. FUSION 2002. (IEEE Cat.No.02EX5997)*, vol. 2, July 2002, pp. 1120–1125 vol.2.

[16] K. O. A. Ali, N. Ilić, M. S. Stanković, and S. S. Stanković, "Consensus-based distributed adaptive target tracking in camera networks using integrated probabilistic data association," *EURASIP Journal on Advances in Signal Processing*, vol. 2018, no. 1, p. 13, Feb 2018. [Online]. Available: https://doi.org/10.1186/s13634-018-0534-z

[17] D. Musicki, B. La Scala, and R. Evans, "Multi-target tracking in clutter without measurement assignment," vol. 1, 01 2005, pp. 716 – 721 Vol.1.

[18] K. O. A. Ali, N. Ilić, M. S. Stanković, and S. S. Stanković, "Distributed target tracking in sensor networks using multi-step consensus," *IET Radar, Sonar Navigation*, vol. 12, no. 9, pp. 998–1004, 2018.

[19] S. S. Stanković, M. S. Stanković, and D. M. Stipanović, "Consensus based overlapping decentralized estimation with missing observations and communication faults," *Automatica*, vol. 45, pp. 1397–1406, 2009.

[20] R. A. Horn and C. A. Johnson, *Matrix Analysis*. Cambridge, England: Cambridge Univ. Press, 1985.

QQ-plot Based Clustering

Željko Nedeljković, Željko Đurović

***Abstract*— Clustering is widely applied, from everyday life to the most diverse fields of scientific activities. Different techniques of clustering are present, with their advantages and disadvantages. This paper examines the possibility of multidimensional QQ-plot based clustering. First, the multidimensional samples are transformed into a scalar form, using the Fibonacci sequence, after that the QQ-plot is formed for the such shaped samples, and then the piecewise linear approximation is constructed. On the basis of the linear segments of the piecewise linear approximation, clusters are segregated. The proposed procedure can be applied multiple times in order to obtain a better result. The initial tests indicate that the proposed method achieves results comparable with conventional clustering methods.**

***Index Terms*—clustering; QQ-plot; Fibonacci sequence.**

I. Introduction

Clustering implies grouping of items, according to a common characteristic, which allows perceiving and drawing conclusions for the entire group, instead of looking at individual cases separately. Grouping is done in such a way that the similarity of data within a group is large, with simultaneous, small similarity of data from different groups. Clustering has a wide application in a variety of fields, and some of the applications in the area of signal processing are pattern recognition, speech processing, image segmentation. Some of the most commonly used clustering methods are K-means [1], Fuzzy C-means [2], Mountain [3], Subtractive [4].

This paper presents a new approach to clustering using QQ (quantile-quantile) plot [5]. The basic application of QQ-plot is a graphical comparison of the distributions of the given sample sets, whereas in this case, it is used for determining points with similar characteristics. In paper [6], the QQ-plot was applied to the problem of estimating the probability density function (pdf). After the formation of the QQ-plot, a piecewise linear approximation is constructed, where the statistics of each linear segment are used to construct a probability density function. This paper relates to the paper [6] in the sense of idea that the points of one linear segment of QQ-plot approximation have the same characteristics. One problem that needed to be solved was the representation of an arbitrary length vector as a scalar, for what the Fibonacci sequence was used [7]. For the obtained scalar values, a QQ-plot was formed, and subsequently, the piecewise linear approximation as well, on the basis of which it was concluded which points belong to the same cluster. Since, during representation of a multidimensional value by one-dimensional, a part of the information was inevitably lost, the procedure was repeated several times for different variations of the formation of scalar representations, and in the end, the aggregation of thus obtained results was performed.

The next chapter summarizes a brief overview of the QQ-plot. The proposed algorithm is presented in chapter three. Test results and conclusions are provided in chapters four and five.

Željko Nedeljković is with the School of Electrical Engineering, University of Belgrade, Kralj Aleksandar's Boulevard 73, 11020 Belgrade, Serbia (e-mail: nz135003p@student.etf.bg.ac.rs).

Željko Đurović – Signals & Systems Division, School of Electrical Engineering, University of Belgrade, Kralj Aleksandar's Boulevard 73, 11020 Belgrade, Serbia (e-mail: zdjurovic@etf.bg.ac.rs).

II. QQ-plot

The QQ-plot was primarily created as a measure of goodness-of-fit. It can be used to determine the degree of probability distributions matching for two given sets of samples, or to determine the matching of the distributions of a given set of samples with a certain theoretical distribution or to compare two theoretical distributions. The QQ-plot is a quantiles graph of one against the other sets of samples. Thereby, the quantile implies a portion of the points (percentage), below the given value. Quantiles of a given set of samples are obtained by sorting, while quantiles of theoretical distribution are obtained using inversion of the cumulative distribution function. For sets of samples that have the same or linearly dependent probability density functions, an approximately linear QQ-plot is obtained. Any deviation from the straight line indicates a mismatch in the distributions of the tested sample sets, and conclusions can be drawn based on the deviation type.

III. Proposed algorithm

The proposed algorithm relies on the idea outlined in the paper [6], where the QQ-plot, formed for a given set of points relative to the theoretical distribution, is interpreted in such a way, that the linear segments of the QQ-plot are observed, and it can be concluded that the points belonging to the same linear segment have the same characteristics, i.e. belong to the same group. In the paper [6], the observed groups are used to determine statistics for the formation of the probability density function, which represents a given set of points. In this paper, the points forming the linear segment are proclaimed representatives of the same cluster.

One limitation of the QQ-plot is that it is applicable to one-dimensional problems only. In practical application, the problems we deal with are usually multidimensional. In compliance with this, this paper has a tendency to cover multidimensional cases. The assumption is that a set of vectors is provided $V_1, V_2, \dots, V_N$, where the dimension of the individual vectors is D. For the purpose of mapping the

vectors of arbitrary length into a scalar value, the Fibonacci sequence was used, similar to the one in the paper [7], $F_n = F_{n-1} + F_{n-2}$, where $F_1 = 5$, $F_2 = 8$ was adopted, thereby forming a Fibonacci numbers' vector of a dimension equal to that of the given vectors $F = [F_1, F_2, \dots, F_D]^T$. A series of indexes is then formed $index = [i_1, i_2, \dots, i_D]$, as one permutation without repetition of integer numbers from 1 to D. The $index$ vector determines the order of multiplication during formation of scalar representation of the input vector $S_n = \hat{V}_n(i_1) \cdot F_1 + \hat{V}_n(i_2) \cdot F_2 + \cdots + \hat{V}_n(i_D) \cdot F_D, \quad n \in [1, N]$, where $\hat{V}_n(i) \in [0,1]$, $i \in [1, D]$ is the normalized value of the i coordinate over all vectors $V_n, n \in [1, N]$. This way, the sequence of the input vectors $V_1, V_2, \dots, V_N$ is represented by the sequence of scalar values $S_1, S_2, \dots, S_N$.

The formation of the QQ-plot begins with the adoption of the theoretical distribution of the samples, and unless there is additional information, it is reasonable to adopt a normal distribution. By forming the inversion of the cumulative distribution function P^{-1}, quantiles of one axis is defined $P^{-1}\left(\frac{n-0.5}{N}\right)$, $n = 1,2, \dots, N$. It is then necessary to sort the scalar representations $S_1, S_2, \dots, S_N$, by which the quantiles of the other axis is determined $y_1, y, \dots, y_N$. By pairing the quantiles, the QQ-plot is determined.

The deviation of the QQ-plot formation from the straight line is expected, but local linear segments that represent groups of points of similar characteristics can be detected. In order to extract linear segments, the piecewise linear approximation of the QQ-plot is constructed. Points belonging to the same linear segment are considered members of the same cluster. In [6], approximation was performed by dividing intervals with the largest error of approximation, which produced a good result in the formation of the probability density function, with the emphasis on the best curve approximation, in order to obtain a better estimation of pdf, and the number of segments was not essentially important. When it comes to the problem that is solving here, the number of segments is important, because it indicates the number of formed groups, so it is necessary to apply a slightly different approach to the formation of linear approximation. In this paper, the set of points is divided into groups of several successive points first, and then, for each group formed in such a way, determination of linear approximation and the error of approximation is performed. Subsequently, iterative conjugation of adjacent linear segments is made, so that the error of approximation remains as small as possible. The stopping criterion is reaching defined number of linear segments. Upon completion of the iterative procedure, it needs to be checked whether there are insufficient points involved in any of the linear segments, in which case, an annexation is applied to one of the adjacent segments. This leads to the final piecewise linear approximation of the QQ-plot.

Based on the formed linear approximation of the QQ-plot, it is possible to draw a conclusion about the belonging of the given samples to clusters. However, if one takes into account that in one of the steps of the algorithm, the vector of arbitrary length has been presented with a scalar, it is clear that a part of the information is lost, with the different components of the training vector being preserved at varying degrees, depending on element of the Fibonacci sequence it is weighted. This leads to the conclusion that, in order to obtain more robust clustering, it makes sense to repeat the previously described steps several times, for the differently formed $index$ vector. In order to form the final result, a square matrix T is formed with dimensionally equal to the number of vectors in the training set N. Each position (n_1, n_2) of the T matrix carries information about the number of occurrences of vectors V_{n_1} and V_{n_2} in the same group (linear segment), for different $index$ vector setups. Maximum value $maxT$, which can appear in the matrix T, is equal to the number of evaluated $index$ vectors. If $T(n_1, n_2) \geq 0.8 \cdot maxT$, it can be concluded that the points V_{n_1} and V_{n_2} are in the same group. Validation is conducted for groups formed in such manner, in terms of the number of contained points. Then, the valid groups are joined to the given number of clusters K, and later, the groups that have not been declared as valid are joined to the valid groups, as well. Conjugation of the groups is performed based on the smallest Euclidean distance of the group centers. This way, the clustering of the given vectors $V_1, V_2, \dots, V_N$ in K clusters was performed.

IV. Results

The proposed algorithm remains an open question when it comes to selecting setups of vector $index$. One possibility is to examine all possible $index$ vector setups, which, for the dimension of the input vectors D, is equal to the number of permutations without repetition for series of integer numbers from 1 to D, which is equal to $D!$. The number of possible setups grows significantly with the increase of D, so it becomes impractical for higher D values. The second option is to randomly select the setups up to the specified number, which meets the requirement of practical applicability. However, the question arises as to whether it is possible to find a measure of goodness of different setups in order to evaluate only those that have the greatest contribution to the final result.

During the following experiments, it will be examined whether it is possible to achieve valid clustering using only one $index$ vector setup. It should also be determined what kind of result is obtained by including all possible $index$ vector setups. Additionally, it makes sense to analyze the change in the quality of the results for a variety of included setups, from one to all.

For simulation purposes, $N = 1000$ three-dimensional samples was generated with three noticeable, balanced groups with normal distributions $G_1 \sim N(\mu_1, \Sigma_1)$, $G_2 \sim N(\mu_2, \Sigma_2)$ and $G_3 \sim N(\mu_3, \Sigma_3)$, where the vectors of mean values are $\mu_1 = [1,1,1]$, $\mu_2 = [6,6,4]$, $\mu_3 = [0,6,6]$ and the diagonal covariance matrices are $\Sigma_1 = diag(2^2, 1, 1)$, $\Sigma_2 = diag(1, 1, 2^2)$, $\Sigma_3 = diag(1, 2^2, 1)$. Fig. 1 illustrates the generated samples, whereby Bhattacharya distances [8] of the

pairs of generated groups of samples are $D_B(G_1, G_2) = 4.83$, $D_B(G_1, G_3) = 4.59$, $D_B(G_2, G_3) = 4.85$.

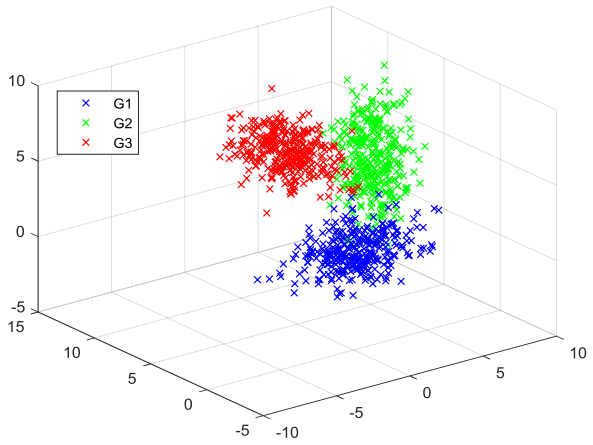


Fig. 1. Generated samples.

Generated vectors of the samples will be transformed into a scalar representation using the Fibonacci sequence using one of the $index$ vector setups: $index_1 = [3, 2, 1]$, $index_2 = [3, 1, 2]$, $index_3 = [2, 3, 1]$, $index_4 = [2, 1, 3]$, $index_5 = [1, 3, 2]$ and $index_6 = [1, 2, 3]$. For the first setup $index_1$, probability density functions of the given groups are presented in Fig. 2. It is clear that the Fibonacci representation has significantly preserved the diversity of the groups, but it is also clear that overlapping of the groups will be reflected on the quality of clustering. Also, by inspecting the Bhattacharya distance of the pairs of the one-dimensional samples obtained this way, it can be observed that the separability is clearly degraded $D_B(G_1, G_2) = 3.41$, $D_B(G_1, G_3) = 0.84$, $D_B(G_2, G_3) = 1.03$. For the formed Fibonacci representation, the QQ-plot has the shape shown in Fig. 3. It is clear that different parts of the curve can be related to the samples that constitute it, but overlapping in certain parts is also evident, which will directly affect the result of the clustering. The result of the clustering based only on first setup is given in Fig. 4, where 16.22% of the samples were clustered differently from their original belonging. Looking at Fig. 4, it is clear that the group 3 was dominantly mixed with groups 1 and 2, which is to be expected on the basis of Figs. 2 and 3, as well as the calculated Bhattacharya distances.

Based on the preceding example, it can be concluded that using only one setup can produce meaningful results. Fig. 5 represents the results for a combination of all possible setups. It is clear that the result is quite good where it comes to just 0.60% of samples in clusters different from the given groups. If the same set of samples is clustered using K-means clustering, 0.70% of wrong clustering is obtained, indicating that the proposed method achieves results in range of the verified methods.

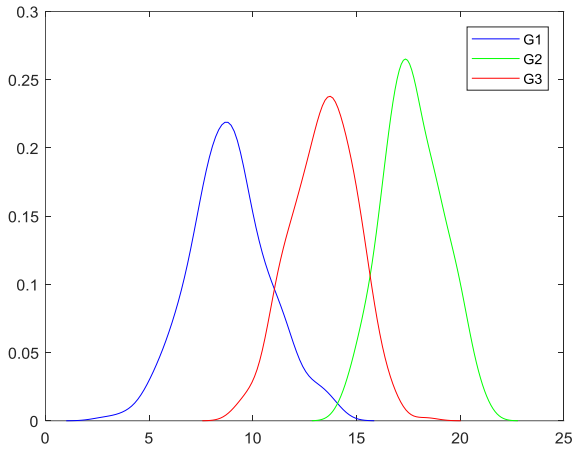


Fig. 2. Pdf of Fibonacci representation for setup 1.

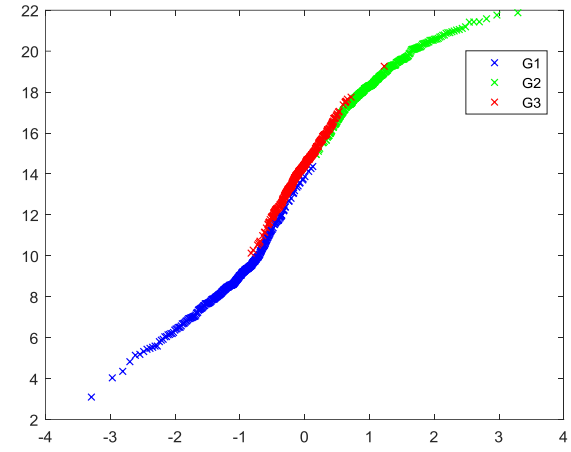


Fig. 3. QQ-plot for setup 1.

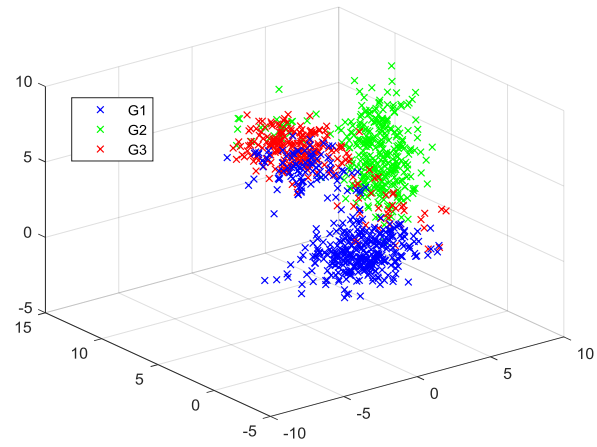


Fig. 4. Clusters based on setup 1 only.

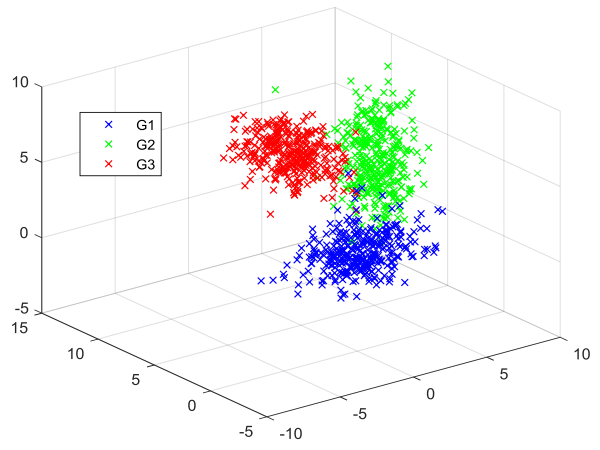


Fig. 5. Clusters based on all setups.

TABLE I
ITERATIVE EXTENSION OF NUMBER OF SETUPS

Number of setups	Included setups	Clustering error
1	3	15.12%
2	3, 4	2.50%
3	3, 4, 1	17.02%
4	3, 4, 1, 6	1.20%
5	3, 4, 1, 6, 2	0.40%
6	3, 4, 1, 6, 2, 5	0.60%

The question as to whether it is necessary to examine all the setups in order to obtain good results arises, also when it comes to vectors of larger dimensions, the examination of all possibilities becomes impractical. Next experiment aimed analyzing the effect of the number of setups on the clustering result, as well as the importance of different setups. The experiment was organized in a way that the number of included setups was iteratively expanded from one to all, adding new setups that maximize the result. The results of the experiment are given in Table 1. Two phenomena are noticed, one is that it is possible to get a result comparable with the result that includes all the setups by using less setups. It can also be noticed that the inclusion of certain setups results in degradation of the end result. It is worth paying attention to researching the impact of different setups on the end result, as well as finding an adequate measure, which will be the subject of future research.

V. CONCLUSION

In wide use, in a wide variety of applications, a variety of clustering methods exists, each with its advantages for a particular type of problem. In this paper, a new method of clustering is proposed, which has yet to be examined regarding the advantages and disadvantages, and possible fields of application. The approach based on the QQ-plot and the Fibonacci numbers proposed in this paper does not look like any of the other well-known clustering methods, leaving room for certain applications, in which the proposed algorithm could be pointed out in relation to the already familiar methods.

The aim of this paper was to show that clustering of multidimensional data can be performed using QQ-plot. Further research for improvement of the proposed steps of the algorithm remains an open question. Also, it is important to solve the issue of engaged setups, which is imposed as a necessary topic of further research.

REFERENCES

[1] J. A. Hartigan and M. A. Wong, "A k-means clustering algorithm," *Applied Statistics*, vol. 28, no. 1, pp. 100-108, 1979.

[2] J. C. Dunn, "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters," *Journal of Cybernetics*, vol. 3, no. 3, pp. 32-57, 1973.

[3] R. R. Yager and D. P. Filev, "Approximate clustering via the mountain method," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 24, no. 8, pp. 1279-1284, Aug., 1994.

[4] S. Chiu, "Fuzzy Model Identification Based on Cluster Estimation," *Journal of the Intelligent and Fuzzy Systems*, vol. 2, pp. 267-278, 1994.

[5] R. Gnanadesikan, *Method for Statistical Dana Analysis of Multivariate Observations*, New York, USA: John Wiley, 1977.

[6] Z. Djurovic, B. Kovacevic and V. Barroso, "QQ-plot based probability density function estimation," Proceedings of the Tenth IEEE Workshop on Statistical Signal and Array Processing, Pocono Manor, USA, pp. 243-247, 16-16 Aug., 2000.

[7] R. Rawat, R. Nayak, Y. Li and S. Alsaleh, "Aggregate Distance Based Clustering Using Fibonacci Series-FIBCLUS," *Web Technologies and Applications. APWeb 2011. Lecture Notes in Computer Science*, vol. 6612, pp. 29-40, Apr., 2011.

[8] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bulletin of the Calcutta Mathematical Society*, vol.35, pp. 99–109, 1943.

Robust Object Tracking based on SURF in Thermal Images

Nataša Vlahović, *Member, IEEE,* Željko Đurović

***Abstract*—This paper describes the problem of single object tracking in thermal image by using SURF (Speeded Up Robust Features) feature descriptor and Kalman filter. Kalman filter is as good as its model, so the measurement and process noise matrices calculation for the specific problem are inevitable and important step in obtaining good tracking results. On the other hand, outliers - deviations from the assumed noise distributions (Gaussian in the case of Standard Kalman filter) can affect and compromise the tracking result of Standard Kalman filter. That is why robust statistics methods are used. In this paper, Robust Kalman filter is modelled using the Huber influence function. The designed Robust Kalman filter is than used along with SURF feature descriptor in the task of pedestrian tracking.**

***Index Terms*— object tracking, SURF descriptor, Kalman filter, Robust Kalman filter, robust statistics, thermal image.**

I. Introduction

Thermal imaging has been associated with military applications only for a long time. Today, thermal sensors are not as expensive as they used to be, their size is decreased, while image quality is significantly increased. With this recent development in the area, thermal imagery is now widely used in civil applications as well. Thermal sensors measure the reflected or emitted radiation of different objects in the scene, and that is why these sensors enable the same visibility in the darkness as in the day light. Thermal sensors are passive, since radiation is emitted by object itself, which is another important advantage of thermal imagery [1]. The infrared wavelength band is divided into: near infrared (NIR, wavelengths 0.7–1 μm), shortwave infrared (SWIR, 1–3 μm), midwave infrared (MWIR, 3–5 μm), and longwave infrared (LWIR, 7.5–12 μm). In this paper, noise characteristics are studied on MWIR thermal camera of the resolution 640x480.

In target tracking applications feature descriptors are often used for feature matching on successive frames. Not all feature descriptors used in images obtained from color cameras can be used in the same way in thermal images, since many image characteristics are different. So, it can be said that there are some misconceptions when talking about object tracking in thermal infrared. The first one is that thermal tracking is hotspot tracking only, but the situation is certainly more complex than that. The other misconception is that thermal infrared tracking is the same as grayscale visual imagery tracking [2]. Thermal images do not have shadows, they have different noise characteristics [3], [4] as well as patterns from variation in material and temperature, not from color such as in color images.

In this paper, the task of single object tracking is addressed. The selected method includes the object description with feature descriptors. SURF (Speeded Up Robust Features) features are found to be very successful in the thermal image domain [5], [6], and that is why these descriptors are chosen. The next step is using the Kalman filter to predict the position of the target in situations where SURF feature descriptor does not give correct results, or when features cannot be found at all. The first step is certainly using SURF detector and Standard Kalman filter for the tracking task. The process and measurement matrix are set according to calculated noise statistics for the observed thermal image.

Since Kalman filter implies a state space model made on some assumptions (measurement and process noise are Gaussian), these assumptions do not have to be true for the observed scenario. And often, this model does not fit the real situation. So minor deviations from the assumptions can affect results. That is why, in this paper, robust procedures are used to signify insensitivity to small deviations from assumptions [7].

II. SURF feature descriptor

Features are some specific structures in an image, which can be corners, or other shapes or blobs. The descriptor chosen for the task of object tracking in thermal image, in this paper, belongs to the family of Spectra Descriptors. SURF (Speeded-up Robust Features) descriptor is developed as a faster version of SIFT descriptor, and it balances the two important requirements in every system: computation speed and performance. SURF detector is scale and rotation invariant [8].

The SURF algorithm is based on the multi-scale space theory and Hessian matrix while it uses its basic approximation. SURF creates a stack, and results in images of the same dimension. Due to the use of integral images, SURF filters the stack using a box filter approximation of the second order Gaussian partial

Nataša Vlahović is PhD student in the School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia. and works at Vlatacom Institute, Milutina Milankovića 5, (e-mail: natasa.vlahovic@ vlatacom.com)

Željko Đurović is with the the School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade (e-mail: zdjurovic@etf.bg.ac.rs).

derivatives. Integral images allow the computation of rectangular box filters in near constant time [5].

Feature descriptor SURF (Speeded-up Robust Features) is chosen from the family of feature descriptors for the task of object tracking in thermal image. Among descriptors from the other families, this detector has shown better performance than descriptors from the other families [5].

SURF features can distinguish light (foreground) and dark (background) patterns due to the gradient orientations used to describe a certain image region. This means that the descriptor of a light region on dark background does not match a dark region on light background and vice versa, which is very useful in the task of pedestrian tracking, since persons are typically lighter than the background objects [8], [6].

III. KALMAN FILTER IN OBJECT TRACKING

A. Standard Kalman filter

Kalman filter represents widely applied algorithm in the field of object tracking, while being one of the most important achievements of linear estimation field [10].

In this paper, Kalman filter models the dynamics of the center of the tracked object, while assuming constant object velocity. The system is represented by the state-space model:

$$x(k+1) = F(k)x(k) + G(k)w(k) \quad (3)$$

$$y(k) = H(k)x(k) + v(k) \quad (4)$$

where *x(k)* represents the state vector and *y(k)* represents observation vector, while *w(k)* is the state noise and *v(k)* is the measurement noise. The state noise *w(k)* and the measurement noise *v(k)* are assumed to be zero mean white noises:

$$E\{w(k)\} = 0; E\{w(k)w(k)^T\} = Q(k)\delta_{kj}; \quad (5)$$

$$E\{v(k)\} = 0; E\{v(k)v(k)^T\} = R(k)\delta_{kj}; \quad (6)$$

where δ_{kj} is the Kronecker's delta symbol and E{.} is the mathematical expectation. *F(k)* is the state transition matrix, and *H(k)* is the observation matrix. If $\hat{x}(k/k-1)$ is the linear least squares estimate of *x(k)* and *P(k/k-1)* denotes the corresponding error covariance matrix, then the standard Kalman filter recursions are:

$$\widehat{x_p}(k+1/k) = F(k)\hat{x}(k/k) \quad (7)$$

$$P_p(k+1/k) = F(k)P_p(k/k)F^T(k) + G(k)Q(k)G^T(k) \quad (8)$$

$$K(k) = P_p(k/k-1)H^T(k)[H(k)P_p(k/k-1)H^T(k) + R(k)]^{-1} \quad (9)$$

$$\hat{x}(k/k) = \widehat{x_p}(k/k-1) + K(k)(y(k) - H(k)\hat{x}(k/k-1)) \quad (10)$$

$$P_p(k/k) = [I - K(k)H(k)]P_p(k/k-1) \quad (11)$$

The initial state *x (0)* is a random value, independent of the future noises *w(k)* and *v(k)* with zero mean and covariance matrix *P (0)*. [11]

B. Robust Kalman filter

Standard Kalman filter, as a statistical method, relies on some assumptions. It gives good results when the initial assumptions are correct: the distribution of noise in observed data is Gaussian. This makes the resulting model more manageable, but, in real situations, these assumptions are usually not correct. The main reason for assuming a normal distribution is that it can be a good and convenient assumption, which enables using optimal statistical methods, such as Kalman filter, among many others. These classical statistics are quite easy for computation. But, the probability density function of noise in real applications deviates from Gaussian, thus making Standard Kalman filter sensitive to outliers in the data, and non-robust. Outliers are data that are atypical for the observed case, and in our case of single object tracking in thermal imagery outliers can be of various nature. Even a single outlier can have a great impact on Standard Kalman filter results.

On the other hand, when using the robust estimation theory, the effect of these outliers can be minimized. If the data contain no outliers the robust method gives approximately the same results as the classical method, while if a small proportion of outliers are present the robust method gives approximately the same results as the classical method applied to the "typical" data. That is why robust methods are reliable method of outlier detection [12].

In this paper, we use the following Robust Kalman filter method:

1) Initialization step:
   - The initial state *x (0)* is a random value, independent of the future noises *w(k)* and *v(k)* with zero mean and covariance matrix *P (0)*.
   - The state noise *w(k)* and the measurement noise *v(k)* are assumed to be zero mean white noises:

$$E\{w(k)\} = 0; E\{w(k)w(k)^T\} = Q(k)\delta_{kj}; \quad (12)$$

$$E\{v(k)\} = 0; E\{v(k)v(k)^T\} = R(k)\delta_{kj}; \quad (13)$$

   - Defining Huber influence function:

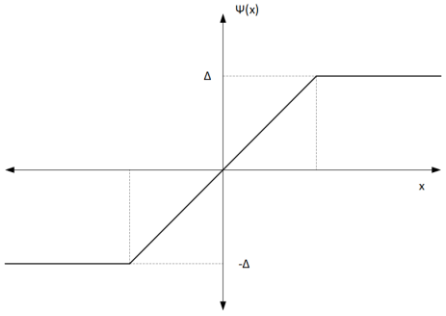


Fig. 1. Huber influence function

$$\psi(x) = \begin{cases} x & , |x| \le \Delta \\ \Delta\, sgn(x) & , |x| > \Delta \end{cases} \quad (14),$$

where Δ is any constant [13].

2) Prediction step:

$$\hat{x}_p(k) = F(k-1)\hat{x}(k-1) \quad (15)$$

$$P_p(k) = F(k-1)P(k-1)F^T(k-1) + G(k-1)Q(k-1)G^T(k-1) \quad (16)$$

3) Robust estimation step:

$$\varepsilon(k) = y(k) - H(k)\hat{x}_p(k) \quad (17)$$

$$S(k) = H(k)P(k)H(k)_k^T + R_k \quad (18)$$

$$d^2(k) = [S_{11}\ S_{22}\ \dots\ S_{nn}] \quad (19)$$

$$v_n(k) = \varepsilon(k) \oslash d(k)\,, \quad (20)$$

Where $\oslash$ is Hadamard division (element by element division)

$$\psi(k) = \psi\big(v_n(k)\big) = min(|v_n(k)|, \Delta) sgn(v_n(k)) \quad (21)$$

$$\omega(k) = \begin{cases} \frac{\psi(v_n(k))}{v_n(k)} & , v_n(k) \neq 0 \\ 1 & , v_n(k) = 0 \end{cases} \quad (22)$$

Based on Robust dynamic stochastic approximation estimator theory [12], an approach based on the definition of a time-varying functional, we can write:

$$K(k) = \omega(k)P_p(k/k-1)H^T(k) * [H(k)P_p(k/k-1)H^T(k) + R(k)]^{-1} \quad (23)$$

$$\hat{x}(k) = \hat{x}_p(k/k-1) + K(k)\varepsilon(k) \quad (24)$$

$$P(k) = [I - K(k)H(k)]P_p(k-1) \quad (25)$$

4) Increment k, the number of iterations, and repeat the procedure from the step 2.

IV. Experimental Results and Discussion

A. Standard Kalman filter results

A starting point in object tracking system design, in this paper, is using SURF features and Standard Kalman filter for the basic system setting. First, SURF features are extracted from the initially set object position region of interest (ROI). The next step is the center position calculation from matched features' locations. This position is Kalman filter measurement vector, with x and y coordinates of the center of the object. Kalman filter predicts the position of the bounding box in the next frame, in accordance with the adopted motion model. In the case where no matched points are found, Kalman filter prediction is used as the center of the object for the next iteration. When the measurement and process noise values are not properly set for the specific problem, Kalman filter cannot give good results [11]. In this case, Kalman filter is used to give predictions when no matched points are found.

Process noise statistics are calculated from the difference image, made from the first two frames of the sequence. The initial measurement noise is calculated from the ground truth data from the given dataset. A difference between ground truth and real measured positions obtained from SURF points matching is calculated. This way, mean values and variances of the measurement and process noise are set.

Standard Kalman filter tracking gives good results when SURF feature matching does not have any obstacles or situations where features are not properly found or matched. This situation is shown in Figure 2 and Figure 3, where the tracking error of x and y position is shown.

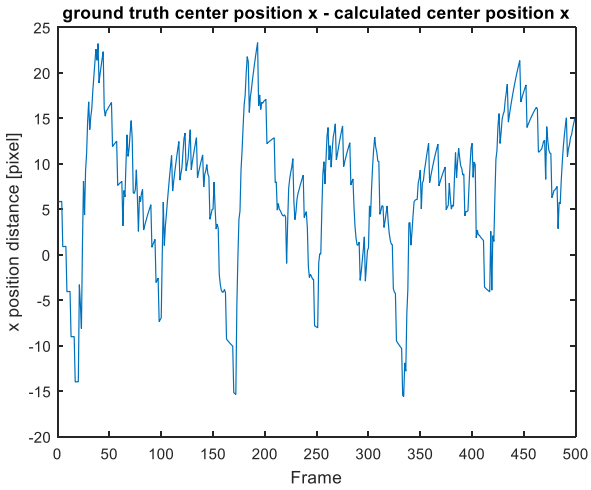


Fig. 2. Standard Kalman filter – dataset 1, error x coordinate

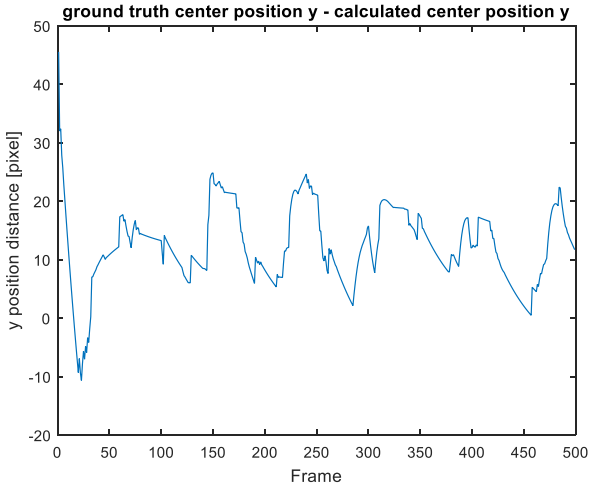


Fig. 3. Standard Kalman filter – dataset 1, error y coordinate

On the other hand, when outliers are present in the data, when SURF feature matching step gives results that step out of calculated noise statistics, Standard Kalman filter does not give good results. These are situations when SURF algorithm finds matching features on some other object in image, or when there

is an obstacle in the image that prevents good matching, as shown in Figure 4.

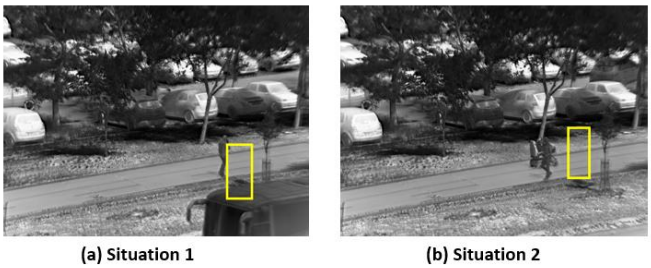


Fig. 4. Standard Kalman filter – estimation error because of an outlier, (a) situation 1 – wrong matching, (b) situation 2 – obstacle in the image

That is why, the next step is designing Robust Kalman filter that ignores outliers in the data.

B. Robust Kalman filter results

The next step, after the Standard Kalman filter design, is Robust Kalman filter design, with the Huber M-function as an influence function.

First, the results for the sequence 1 will be shown. This is a sequence where there are no obstacles, and the comparative results for Standard and Robust Kalman filter are shown. Figure 5 and Figure 6 show ground truth positions as well as Standard and Robust Kalman filter estimations for x and y coordinates, while Figure 7 and Figure 8 show the estimation error in pixels for both coordinates.

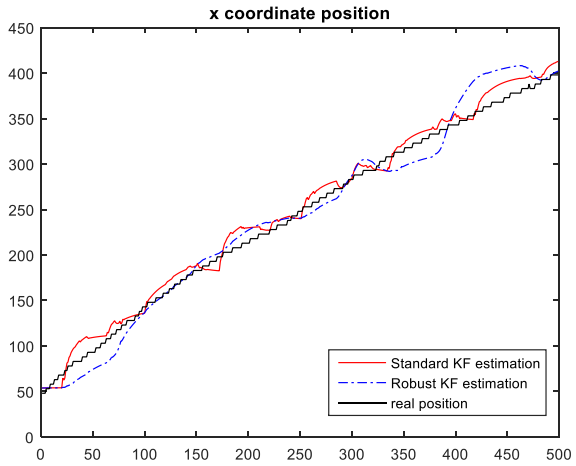


Fig. 5. Comparative representation of estimation results for x coordinate, Standard and Robust KF, sequence 1

SURF features used with Standard Kalman filter for the tracking problem give good performance in the case of expected noise statistics, which is shown in Figures 2 and 3. This is the situation where Robust Kalman filter gives the same or, in some parts, slightly larger error than Standard Kalman filter, but both filters continue tracking the target.

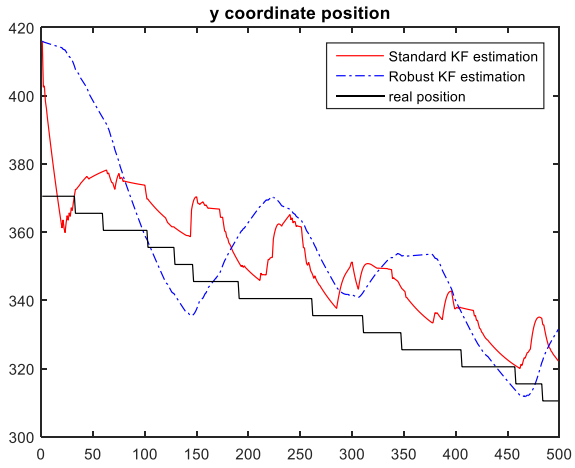


Fig. 6. Comparative representation of estimation results for y coordinate, Standard and Robust KF, sequence 1

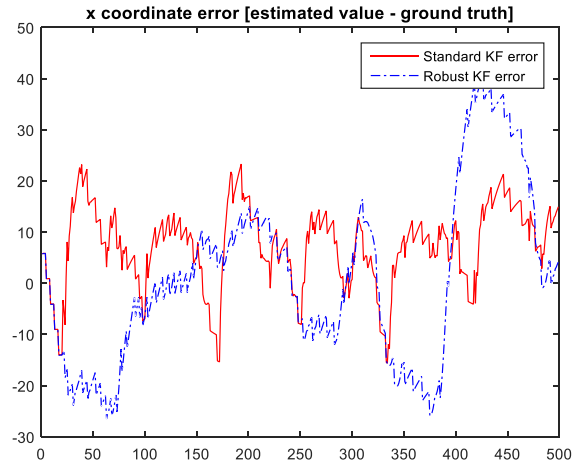


Fig. 7. Estimation error in pixels, results for x coordinate, Standard and Robust KF, sequence 1

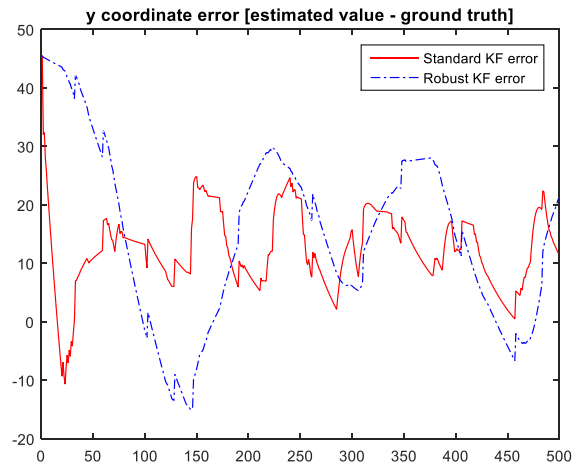


Fig. 8. Estimation error in pixels, results for y coordinate, Standard and Robust KF, sequence 1

As we already mentioned, Standard Kalman filter has problems when dealing with outliers in the data. That is why

for the sequence 2, where SURF algorithm gives imprecise measurements, Robust Kalman filter with Huber influence function gives good results. A few frames from the sequence are shown in Figure 9, Standard (upper three frames) and Robust KF.

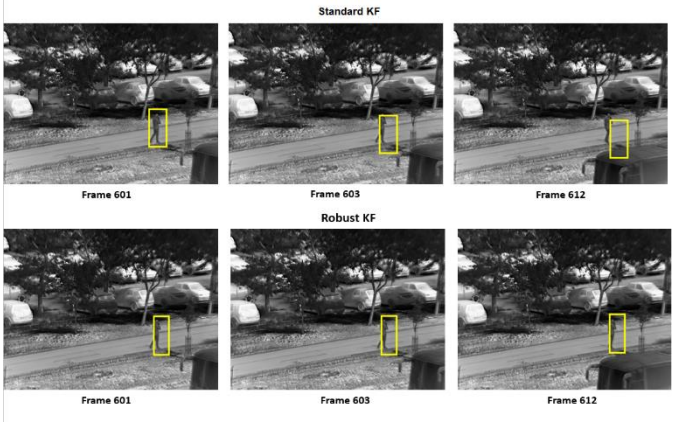


Fig. 9. Standard (upper frames) and Robust KF tracking results, sequence 2

As we can see from the Figure 9, y coordinate has the outlier in measurement data. Figure 10 shows comparative representation of the estimated position for Standard and Robust KF, along with the real coordinates. It points out that Robust Kalman filter does not react on this outlier like Standard Kalman filter does.

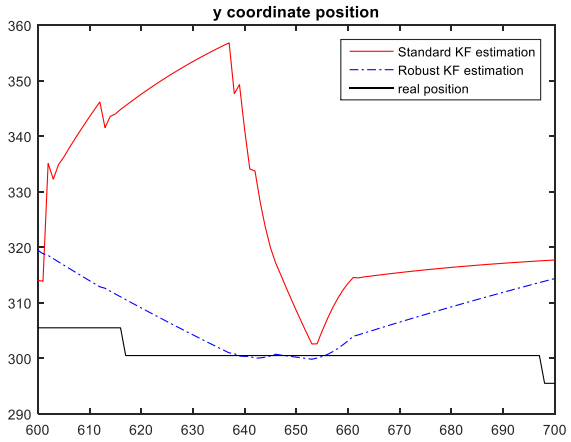


Fig. 10. Comparative representation of Standard and Robust KF, along with ground truth position for y coordinate, sequence 2

The estimation error (deviation of Standard and Robust KF estimation from real positions) for the y coordinate is shown in Figure 11. So, from the results presented in Figure 10 and Figure 11 it is obvious that Standard Kalman filter reacts to outliers in the measurement data poorly, while Robust Kalman filter ignores them while continuing tracking the object of interest.

In the next observed situation 3, there are outliers present in the x coordinate measurement data. This situation also shows that, in the presence of outliers in the data, Standard Kalman filter has lost the target completely, while Robust Kalman filter continues tracking the target.

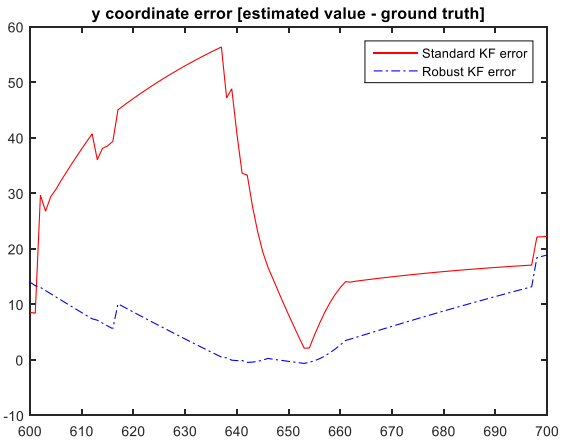


Fig. 11. Estimation error in pixels, results for y coordinate, Standard and Robust KF, sequence 2

For the situation 3, the results for x coordinate only are presented, since it contains outliers. In Figure 12 comparative representation of the estimated position for Standard and Robust KF, and real coordinates is shown (Figure 12), as well as estimation error (Figure 13).

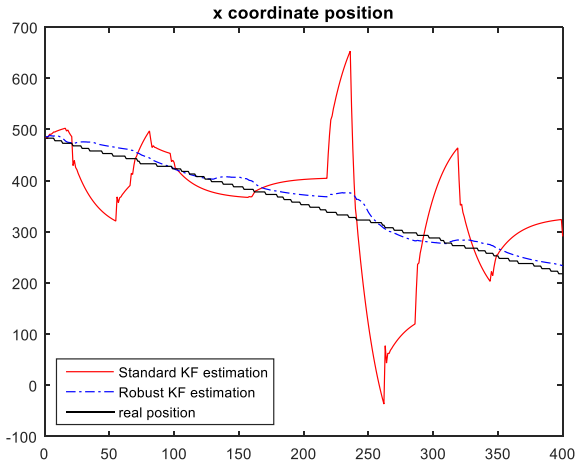


Fig. 12. Comparative representation, x coordinate, sequence 3

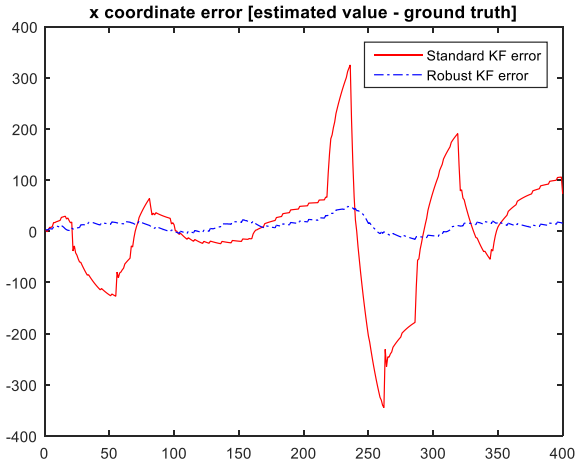


Fig. 13. Estimation error in pixels, results for x coordinate, Standard and Robust KF, sequence 3

In the Figure 14 frames from the real sequence are shown, with estimated position of bounding box in the case of Standard (left image) and Robust Kalman filter (right image).

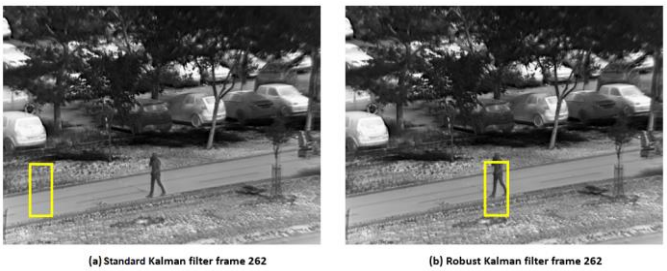


Fig. 14. (a) Standard KF result (left) and (b) Robust KF tracking result, sequence 3

C. Discussion

The situation 1 shows that in the case of the assumed noise distributions without outliers in the data, Standard Kalman filter used with SURF feature matching gives good results, even more precise than Robust Kalman filter.

On the other hand, in the real situations the existence of the outliers in the measurement data is inevitable. That is when the assumptions that Standard Kalman filter is built on are no longer true. Robust Kalman filter gives the solution to the outlier's problems, illustrated in Figure 4.

For the situation 2, the outlier is contained in the y coordinate SURF matching result, which is measurement data for the Kalman filter. SURF feature matching algorithm actually found features not contained in the tracked object. In figures 9, 10 and 11 it is shown that Robust Kalman filter can deal with this type of outliers.

For the situation 3, the outlier is contained in the x coordinate SURF matching result. This outlier is a result of an obstacle that covered the tracked object resulting in the loss of target. In figures 12, 13 and 14 it is shown that Robust Kalman filter can deal with this type of outliers also.

So, the actual tracking results in presence of outliers in the data show that Standard Kalman filter can easily lose the target, while Robust Kalman filter continues tracking the pedestrian as if nothing happened.

V. Conclusion

In this paper, a problem of single object tracking with SURF feature descriptor in thermal image is shown. Even though SURF descriptor can be good for the tracking problem in thermal imaging, it often happens that no feature is found in the image, or that some obstacles affect the matching results. That is why Kalman filter is used, to give object position predictions when the SURF algorithm fails.

In this step of Kalman filter design, it is important to emphasize that Kalman filter noise statistics need to be set for the specific problem, or Kalman filter cannot give good results. For that reason, noise statistics for the observed type of thermal image are calculated (measurement and process noise). SURF features and Standard Kalman filter give good results when there are no outliers in the data.

On the other hand, outliers can affect the results of Kalman filter estimation, therefore the whole tracking performance. This is the reason for using Robust Kalman filter with Huber influence function, in this paper. Robust Kalman filter ignores the influence of outliers in the measurement data in a way that great deviations from the previous position are considered as an outlier, therefore their influence is reduced.

Even though Robust Kalman filters are a good solution when dealing with outliers in the data, there are many situations where Huber influence function is not good enough as an influence function. So, in the future work, some other influence functions will be considered. Also, combination of Adaptive and Robust Kalman filters can be a good solution that can adapt dynamically to any different specific types of noise and outliers in the data, which is also a topic to consider in the future.

References

[1] R. Gade and T. B. Moeslund, "Thermal Cameras and Applications: A Survey," *Machine Vision and Applications* , vol. 25(1), pp. 245-262, 2014.

[2] A. Berg, J. Ahlberg and M. Felsberg, "Channel Coded Distribution Field Tracking for Thermal Infrared Imagery," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2016*, 2016.

[3] R. G. Driggers, M. H. Friedman and J. M. Nichols, Introduction to Infrared and Electro-Optical Systems, Artech House, 2012.

[4] M. Vollmer and K.-P. Möllmann, Infrared thermal imaging _ fundamentals, research and applications,, Second Edition ed., Weinheim, Germany, 2018.

[5] N. Vlahović and S. Graovac, "Sensibility analysis of the object tracking algorithms in thermal image," *Scientific Technical Review*, vol. 67, no. 1, 2017.

[6] K. Jungling and M. Arens, "Feature based person detection beyond the visible spectrum," in *IEEE Conference on Computer Vision and Pattern Recognition*, Miami, 2009.

[7] P. J. Huber and E. M. Ronchetti, Robust Statistics, Second Edition, John Wiley & Sons, Inc., Hoboken, New Jersey, 2009.

[8] H. Bay, T. Tuytelaars and L. V. Gool, "SURF: Speeded Up Robust Features," in *Computer Vision and Image Understanding*, 2008.

[9] Ž. Đurović and B. Kovačević, "Robust Estimation with Unknown Noise Statistics," *IEEE TRANSACTIONS ON AUTOMATIC CONTROL*, vol. 44, no. 6, pp. 1292-1296, 1999.

[10] N. Vlahović and Ž. Đurović, "Object Tracking in Thermal Images based on SURF and KLT features," in *5th International Conference IcETRAN 2018*, Palić, 2018.

[11] R. Maronna, R. Douglas Martin and V. Yohai, Robust Statistics: Theory and Methods, John Wiley & Sons, Ltd., 2006.

[12] B. Kovačević, Ž. Đurović and S. Glavaški, "On robust Kalman filtering," International Journal of Control, vol. 56:3 , pp. 547-562, 2007.

The CFAR Contribution on the Radar Target Tracking

Zvonko Radosavljević, Branko Kovačević and Dejan Ivković

***Abstract*—The different kinds of the constant false alarm rate (CFAR) detectors are used in radar receivers to detect targets in the surveillance zone where all of parameters of the statistical distribution of clutter are not known, or where they are nonstationary. In this time, the standard Integrated Track Splitting (ITS) filter is efficient target tracking algorithm, which is capable to integrates multiscan track with probability of target existence, which becomes the measure of track quality in ITS. We investigate the contribution of two used CFAR algorithms (CA-cell averaging and CATM-cell averaging-trimmed mean) to the quality of target tracking. After the theoretical analysis, this contribution was experimentally tested on the example of single target tracking. Preliminary results of numerical simulations are given in this paper.**

***Index Terms*—CFAR, target tracking, radar detection.**

I. Introduction

When the radar operates, they appear a different sources of noise [1]. There are unwanted signals from other sources of radiation, which can occupy the radar display fully and make targets very hard to see. Detector in radar receivers has to be a detector with the adaptive threshold because radars work always in an environment where there are different sources of noise. It must use the adaptive threshold detector, which has a feature that adjusts automatically its sensitivity according to a variety of interference power. Thus it maintains a constant probability of false alarm [2][3].

In radar receivers, most commonly used detector is the constant false alarm rate (CFAR) detector, which can be classified with algorithms that use the averaging technique, with algorithms that use ordering technique, with algorithms which are the combination of the above mentioned techniques and with algorithms that have some kind of a fusion center in their procedures [4]. They are used very often as a detectors of very close targets per azimuth and per range using the Linear and NonLinear Fusion Constant False Alarm Rate (LF-CFAR and NLF-CFAR) detectors and single CA-CFAR (Cell Averaging CFAR), OS-CFAR (Ordered Statistic CFAR) and TM-CFAR (Trimmed Mean CFAR) which are considered in [5].

When tracking targets in clutter, existence and position of targets in the surveillance are a-priori unknown. Unknown association of measurements with appropriate targets (unknown measurement source) is a common problem in multitarget tracking [6]. Automatic track initiation and termination under such conditions requires some knowledge about track existence. A track exists if it is based on measurements from a target (which follows specified dynamic and detection models), and is not a product of random clutter only. If a track follows a target, we shall call it a true track otherwise we shall call it a false track.

The ITS filter is a single target, multiscan tracking algorithm, where the track quality measure, used for false track discrimination, is the probability of target existence [2]. Each ITS track state estimate probability density function (PDF) is approximated by a Gaussian Mixture. ITS algorithm uses apriori clutter density information to discriminate clutter from target measurements [7]. Because of the uncertainties associated with the origins of the various sensor reports, it is impossible to ascertain precisely which measurement at each scan was the correct one. Further, since sensor detection probability is generally below unity, it is also possible that no measurement on the target was received on a given scan. Pruning [3] involves either removing measurement histories with low probability or removing whole subtrees of measurement histories. How different types of CFAR detectors influence of the quality of targets tracking is the subject of the theoretical and experimental analysis in this paper. However, this topic is not often processed in the literature. Mostly, the targets clutter is treat only in the sense of radar and CFAR thresholding [8], without review of tracking. This paper will be examined the effects of the two different CFAR detectors to the target tracking system.

The paper is organized as follows. After the introducing preamble, problem statements are presented in the Section 2. The cell averaging (CA) and cell averaging-trimmed mean (CATM) CFAR detectors are described at the beginning briefly, in the Section 3. The standard steps of known ITS algorithm is given by the Section IV. At the end of this Section, a probability of detection dependence on target tracking, followed by the results of simulation and final conclusions, from the Section V and Section VI, respectively.

II. Problem Statement

Any target tracking scenario is defined from at least two parameters: probability of detection and clutter density. Again, the clutter density is depending on target dynamics and

Zvonko Radosavljević (corresponding author) is with Military Technical institute, R.Resanovica 1, 11030 Belgrade, Serbia (zvonko.radosavljevic@gmail.com).

Branko Kovacević is with the School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail: kovacevic_b@etf.rs).

Dejan Ivković is with Military Technical institute, R.Resanovica 1, 11030 Belgrade, Serbia (e-mail: ivkovic555@gmail.com).

characteristic of sensor. Generally, clutter is defined by number of selection measurement from size of selection gate. At begin, consider the target state $x_k \in R^{n_z}$ at time interval *k*, which evolves according to:

$$x_k = Fx_{k-1} + v_k \tag{1}$$

where *F* is the state propagation matrix and the process noise, v_k is a zero mean and white Gaussian sequence with covariance matrix Q_k. Originally, size of selection gate is depending on measurements matrix *R* and process noise matrix *Q*. False tracks are consumed at begin, whence true tracks are depending on matrix *Q*. Measurement merging is best modeled in the sensor measurement space, while tracking and data association issues are using more often converted measurements. Converted target measurement $y \in R^{n_y}$ is:

$$y_k = Hx_k + w_k \tag{2}$$

Clutter measurements follow the non-uniform Poisson distribution by clutter measurement density (Poisson intensity) ρ_y. Each track is a set of components. The number of components before processing the measurements at scan *k* will be denoted with $C_k \geq 1$. The sufficient statistic of each track is the union of the sufficient statistics of its components. The article is having a practical contribution while approve which optimal depth of component memory is needed for the specific tracking situation

III. Description of CA and CATM-CFAR

The basic parameters of each *CFAR* are the probability of false alarm rate P_{FA}, size of the window detection $N = 2n$, average signal value in cells *Z*, scaling factor of the detection threshold *T* and detection threshold *S*. Scaling factor of the detection threshold *T* is a constant which achieves a desired value of the probability of false alarm for a given size of the window detection *N*. The detection window consists of two groups with the same number of cells *n* that are located on the opposite sides with respect to the cell whose contents are tested. *CFAR* processes signals by averaging of signals in *2n* neighborhood range bins (X_i) and the resulting mean value compares with the signal in range bin which is under test (*Y*).

The cell averaging (*CA*) *CFAR* algorithm [4] consists of two collector for the leading and lagging windows (Fig. 1). Here, *Z* is simply the sum of Y_1 and Y_2. The cell *Y* in the middle of the reference window is a cell under test. Input samples are sent serially into the reference window. At begin, we calculate the mean clutter power level *Z* using the appropriate *CFAR* algorithm, and than we multiply *Z* by a scaling factor *T* which depends on the *CFAR* algorithm and the designed probability of the false alarm rate P_{fa}. Probability of detection P_{DCA} is given by [5]:

$$P_{DCA} = \left(1 + \frac{T}{1 + SNR}\right)^{-N} \tag{3}$$

where *SNR* is signal-to-noise ratio.

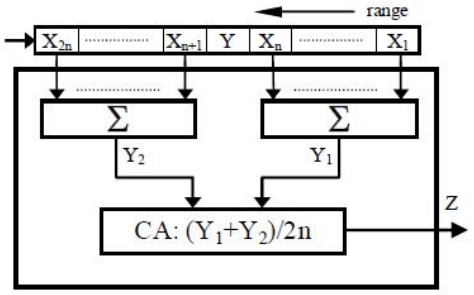


Figure 1 CA-CFAR algorithm

The *CATM-CFAR* [4][5] is cell-averaging-trimmed-mean detector which is optimizes features of some mentioned *CFAR* detectors from different groups depending on the characteristics of clutter and targets, and it is shown in Fig. 2. Main goal of *CATM-CFAR* modelling was increasing the probability of detection at a constant probability of false alarm rate. The cell *Y* in the middle of the reference window is a cell under test. Input samples are sent serially into the reference window.

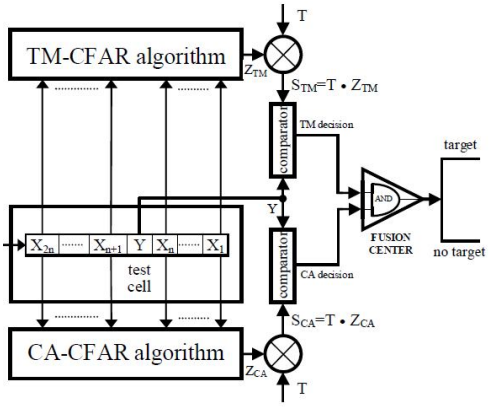


Figure 2 Block diagram of *CATM-CFAR* detector

The first step is to calculate the mean clutter power level *Z* using the appropriate *CFAR* algorithm. The second step is to multiply *Z* by a scaling factor *T* which depends on the *CFAR* algorithm and the designed probability of the false alarm rate P_{FA}. The product *TZ* is the detection threshold *S*. In this time, the *TM-CFAR* detector use first cells in the reference window to be sorted per amplitude. Then it trims T_1 smallest cells and T_2 cells with the highest amplitudes. After that, the summation of the content in the remaining cells is done to obtain *Z*. *CATM CFAR* algorithm is realized by the parallel. The *CA CFAR* detector and the *TM CFAR* detector work simultaneously and independently but with the same scaling factor of the detection threshold *T*. They produce their own mean clutter power level *Z* using the appropriate *CFAR*

algorithm. Next, they calculate their own detection thresholds S_{CA} and S_{TM}. After the comparison with the content in the cell under test *Y*, they decide about target presence independently. The finite decision about target presence is made in the fusion center composed of one"and" logic circuit. If both input single decisions in the fusion center are positive, the finite decision of the fusion center is the presence of the target in the cell under test. In other cases, the finite decision is negative and the target is not at the location which corresponds with the cell under test. Probability of detection P_{DCATM} is given by [5]:

$$P_{DCATM} = \left(1+\frac{T}{1+SNR}\right)^{-N}\prod_{i=1}^{N-T_1-T_2} M_{V_i}\left(\frac{T}{1+SNR}\right) \quad (4)$$

where M_V in (4) is defined as:

$$M_{V_1}(T) = \frac{N!}{T_1!(N-T_1-1)!(N-T_1-T_2)}\cdot\sum_{j=0}^{T_1}\frac{\binom{T_1}{j}(-1)^{T_1-j}}{\frac{N-j}{N-T_1-T_2}+T} \quad (5)$$

and

$$M_{V_i}(T) = \frac{a_i}{a_i+T}, \quad i = 2,3,...,N-T_1-T_2 \quad (6)$$

where a_i is defined as:

$$a_i = \frac{N-T_1-i+1}{N-T_1-T_2-i+1} \quad (7)$$

3.1 Probability of detection of optimal CFAR detector

In radar signals processing, is not possible to use an optimal detector with the fixed optimal threshold S_O to decide target existence, because is it a priori unknown background clutter. A solution is to use the *CFAR* detector which has a constant probability of false alarm. For the optimal detector with the fixed optimal threshold S_O, the probability of detection P_D is given by [4]:

$$P_D = P[Y > S_O \mid H_1 \mid] = \exp[-\frac{S_O}{2\mu(1+SNR)}] \quad (8)$$

where H_1 is the hypothesis the target is present in the radar cell, parameter *SNR* signal to noise ratio and μ background clutter power. The corresponding diagram of dependence of probability of detection versus $\hat{x}_{k|k}$ signal-to-noise ratio *SNR* is given in Fig. 3 according to (3), (4) and (8) for N=16 and P_{fa}=10^{-6}. Finally, we have *CFAR* probability of detection - P_D^{CFAR} as follows [9]:

$$P_D = P\{\chi_{k-1} \mid Z^{k-1}\} = \sum_{c=1}^{C_{k-1}} P\{\chi_{k-1}^c \mid Z^{k-1}\} = P_D^{CFAR} \quad (9)$$

IV. Integrated Track Splitting Algorithm

Assume that each track is the union of components. The estimation state of each component is the output of a filter which is given a single measurement at each scan. The *ITS* filter models each track as a set of components, where each component is defined with a unique measurement history which consists of zero or one measurement received each scan. For each component the state estimate and the a-posteriori probability of component existence are computed recursively. After each scan, a new component is formed from each pair of (existing component; associated measurement) [6]. The probability of the new component existence is the probability that the parent component exists and that the measurement used to create the new component is the target measurement. The probability of target existence, mean and covariance of the state estimate for the track are then calculated and used for track maintenance and track output. Thus each component represents a possible measurement-to-target association history, and components are mutually exclusive. Each component state consists of the probability of the component existence and the component state estimate PDF conditioned on the component existence

A. Target state propagation

Denote the number of components of a track at time *k* with C_k. Let the event that the target exists at time *k*, and therefore the track is a true track, will be denoted by χ_k. A selected set of measurements,z_k, is used to update the track state at scan *k*, with $Z^k = \{z_k; Z^{k-1}\}$ denoting the set of all selected measurements up to and including scan *k* [10]. Thus the a posteriori probability of target existence at time *k - 1* is given by:

$$P\{\chi_{k-1}\left|Z^{k-1}\right.\} = \sum_{c=1}^{C_{k-1}} P\{\chi_{k-1}^c\left|Z^{k-1}\right.\} \quad (10)$$

where $P\{\chi_{k-1}\left|Z^{k-1}\right.\}$ is the probability of existence of the cth component. The a priori probability of component existence is updated using the Markov Chain One for component c and track respectively:

$$P\{\chi_k^c\left|Z^{k-1}\right.\} = \pi_{11}P\{\chi_{k-1}^c \mid Z^{k-1}\} \quad (11)$$

$$P\{\chi_k\left|Z^{k-1}\right.\} = \pi_{11}P\{\chi_{k-1} \mid Z^{k-1}\} \quad (12)$$

where π_{11} denotes the probability that the true track remains a true track, e.g. the target will not disappear.

B. Track State Update

The *PDF* of the target state estimate is the mixture of mutually exclusive component state estimate *PDF*. The component state estimate *PDF* is conditioned on the target existence and component data association history being the correct one and the track state estimate PDF is conditioned on the target existence. Thus the a posteriori track state *PDF* is given by [11]:

$$P\{\chi_k \mid Z^k\} = \frac{\sum_{c=1}^{C_k} P\{\chi_k^c \mid Z^{k-1}\} p^c(x_k \mid Z^k)}{P\{\chi_k \mid Z^k\}} \quad (13)$$

where $p^c(x_k \mid Z^k)$ is the a posteriori component state estimate PDF. At the start of scan k, the ITS filter has Ck components. Associated with each component c is an a priori existence probability $P\{\chi_k^c \mid Z^{k-1}\}$ and an a priori component measurement PDF, $p\{z \mid Z^{k-1}\}$, which is derived from the filtered component state estimate PDF and sensor measurement model. Then the track measurement PDF is given by [12]:

$$p(z_k \mid Z^{k-1}) = \frac{\sum_{c=1}^{C_k} P\{\chi_k^c \mid Z^{k-1}\} p^c(x_k \mid Z^{k-1})}{P\{\chi_k \mid Z^{k-1}\}} \quad (14)$$

Each track selects a set zk of mk candidate measurements, in such a manner that the probability the target measurement is selected, assuming the target exists and is detected, is PW. Denote by zk;i the ith measurement in zk, and by:

$$p_{k,i} = p(z_{k,i} \mid Z^{k-1}) = \frac{\sum_{c=1}^{C_{k-1}} P\{\chi_k^c \mid Z^{k-1}\} p^c(x_k \mid Z^{k-1})}{P\{\chi_k \mid Z^{k-1}\}} \quad (15)$$

its likelihood, where $p^c(z_{k,i} \mid Z^{k-1})$ is likelihood of measurement zk,I with respect to component c. Denote by $\rho_{k,i} \cong \rho\{z_{k,i}\}$ clutter measurement density at zk,i. Define measurement likelihood ratio at time k by [13]:

$$\Lambda_k = 1 - P_D^{CFAR} P_W \sum_{i=1}^{m_k} \frac{p_{k,i}}{\rho_{k,i}} \quad (16)$$

Updated probability of target existence is:

$$P\{\chi_k \mid Z^k\} = \frac{\Lambda_k P\{\chi_k \mid Z^{k-1}]}{1-(1-\Lambda_k)P\{\chi_k \mid Z^{k-1}\}} \quad (17)$$

Each measurement outcome $i \geq 0$ is paired with each old track component *c* to create a new track component. The state estimate of new components is obtained by updating the state prediction pdf of the parent component c with the measurement $z_{k,i}$ (or not updating for *i = 0*). Denoted by k;ithe event that measurement outcome $i \geq 0$ is true. Posterior probability of new component existence is:

$$P\{\chi_k^c, \chi_{k,0}^c \mid Z^k\} = \frac{(1-P_D P_W)P\{\chi_k^c \mid Z^{k-1}\}}{1-(1-\Lambda_k)P\{\chi_k \mid Z^{k-1}\}} \quad (18)$$

$$P\{\chi_k^c, \chi_{k,i}^c \mid Z^k\} = \frac{P_D P_W P\{\chi_k^c \mid Z^{k-1}\}}{1-(1-\Lambda_k)P\{\chi_k \mid Z^{k-1}\}} \frac{p^{c}{}_{k,i}}{\rho_{k,i}} \quad (19)$$

C. Update Components

Each pair *(c, g)* generates a new component for the next scan, where measurement are *g = 1, 2, …G*. The probability of new component $c_n=(c,g)$ is obtained by applying the Bayes formula and is given by:

$$p(x_k \mid \chi_k, Z^k) = \sum_{c_n=1}^{C_k} \xi_k^{c_n} p(x_k \mid \chi_k, c_n, Z^k) \quad (20)$$

State estimate of new component c_n is defined by probability of component $\xi_k^{c_n}$, mean $\hat{x}_{k|k}^{c_n}$ and covariance $P_{k|k}^{c_n}$ of a posteriori state estimate of model *j*, by the following equations (for the *'nul'* mesurements):

$$\xi_k^{c_n} = \beta_{k,0} \xi_{k-1}^{c} \quad (21)$$

$$\hat{x}_{k|k}^{c_n} = \hat{x}_{k|k-1}^{c} \quad (22)$$

$$P_{k|k}^{c_n} = P_{k|k-1}^{c} \quad (23)$$

and apropos for the any other measurements:

$$\xi_k^{c_n} = \beta_{k,1} \xi_{k-1}^{c} p_k^{c_n} p_k \quad (24)$$

$$\hat{x}_{k|k}^{c_n} = \hat{x}_{k|k}^{c,g} \quad (25)$$

$$P_{k|k}^{c_n} = P_{k|k}^{c,g} \quad (26)$$

A new component c_n are obtained by applying Kalman filter state update on measurement with mean $z_k(g)$ and covariance $R_k(g)$ and apriori state estimate of model *j* of old component *c* with mean $\hat{x}_{k|k}^{c,j}$ and covariance $P_{k|k}^{c,j}$. They are standard Kalman filter formulae [1]. The preferred form of track output consists of the mean value and associated error covariance matrix of track trajectory estimate denoted by $\hat{x}_{k|k}$ and $P_{k|k}$, respectively. They are given:

$$\hat{x}_{k|k} = \sum_{c=1}^{C_{k+1}} \xi_k^c \hat{x}_{k|k|}^c \quad (27)$$

$$P_{k|k} = \sum_{c=1}^{C_{k+1}} \xi_k^c [P^c_{k|k} + \hat{x}^c_{k|k} \hat{x}^{c\ T}_{k|k}] - \hat{x}_{k|k} \hat{x}_{k|k}^{\ T} \quad (28)$$

D. Target Tracking Dependence on the CFAR Probability of Detection

In the standard ITS, each target may create one measurements with probability of detection P_D. In each scan the sensor also returns a random number of clutter measurements, parameterized by the clutter measurement density. Tracks are initialized in each scan using measurements from two consecutive scans. Taking into account eaquation (11) and (12) we have expression of the probability of detection for *CA-CFAR* is given as [4]:

$$P_D^{CA} = (1 + \frac{T}{1+SNR})^{-1} \quad (29)$$

According to the reference [5] we have probability of detection:

$$P_D^{CATM} = (1 + \frac{T}{1+SNR})^{-N} \prod_{i=1}^{n-T_1-T_2} M_{V_i}(\frac{T}{1+SNR}) \quad (30)$$

Now, we can define a new measurement likelihood ratio at time *k*, for the *CA* and *CATM CFAR* by the following equations:

$$\Lambda_k^{CA} = 1 - P_D^{CA} P_W \sum_{i=1}^{m_k} \frac{p_{k,i}}{\rho_{k,i}} \quad (31)$$

$$\Lambda_k^{CATM} = 1 - P_D^{CATM} P_W \sum_{i=1}^{m_k} \frac{p_{k,i}}{\rho_{k,i}} \quad (32)$$

respectively

V. Results of Simulations

The application selected for study was a two-dimensional (positions and velocities), four-state aircraft tracking problem in which the sensor observes both position coordinates. The area under surveillance was x = [0; 1000] [*m*] long and y = [0; 400] [*m*] wide. The false measurements satisfied a Poisson distribution with density $\rho = 10^{-4}[scan/m^2]$. Both dimensions were assumed independent, and the sensor measurement errors, target maneuver state excitation errors, and the equations of motion were assumed identical in each dimension.

The ITS parameters are calculated on-line according to the appropriate equations. Period of scanning is *T=1s*. Consider a target motion scenario with non-maneuvering constant velocity (CV) flight mode. Speed is constant 311[*m/s*]. For initialization of tracks we use a 'two point differencing' methodology. The target moves and can appear or disappear in the scene at any time, according to the linear and Gaussian target dynamics. The system input is modeled as follows: vector state $x_k = [x\ \dot{x}\ y\ \dot{y}]$ where x; y are the Cartesian coordinates of the target position, $\dot{x}; \dot{y}$ are the appropriate velocities. Initial target state x0 = [100 12 100 4]. Transition matrix and process noise matrix are given by:

$$\mathbf{F}_1 = \begin{bmatrix} 1 & T & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & T \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (33)$$

$$\mathbf{Q} = q \begin{bmatrix} T^4/4 & T^3/2 & 0 & 0 \\ T^3/2 & T & 0 & 0 \\ 0 & 0 & T^4/4 & T^3/2 \\ 0 & 0 & T^2/2 & T \end{bmatrix} \quad (34)$$

respectively, where *q = 0:252* is a maneuver coefficient. The ITS simulation process is governed by a Markov chain one. Probability of detection, for the CA CFAR and CATM CFAR and given signal to noise ratio (SNR) is 15[dB], is 0,54 and 0,66 [4], respectively [5] .

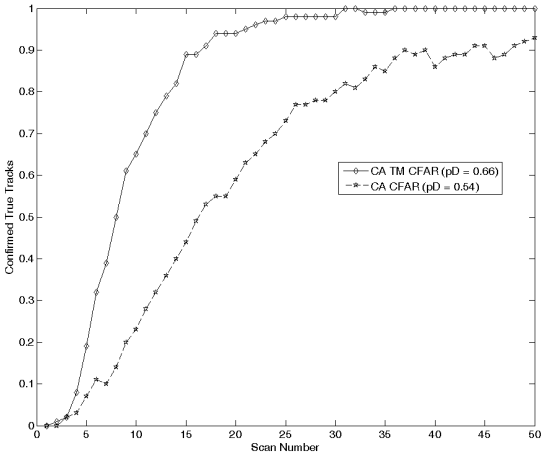


Fig. 3.Comparative confirmed true tracks diagram.

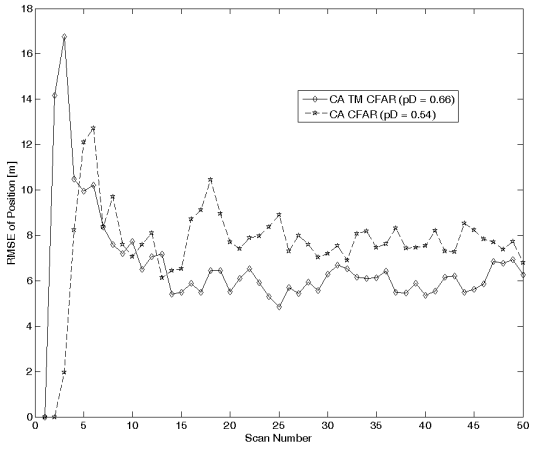


Fig. 4. Comparative RMSE of position diagram.

Respective diagrams of confirmed true tracks and root mean square error are given by the Fig: 3. and Fig: 4. On them is shown influence of two different CFAR algorithms to the target tracking quality. This analysis shows significant improvement of the tracking characteristics by using CATM CFAR algorithms, instead CA CFAR algorithms.

VI. CONCLUSION

The preliminary results of the study of the contribution of different types of CFAR detectors on the characteristics of the single target tracking were given in the paper. The application of the CA CFAR and CATM CFAR algorithms in the radar pre-processing phase and their influence on the distribution of the clatter were examined, through the scattering of the actual confirmed traces in known Integrated Track Splitting algorithm.

The result of single target tracking simulations are showed better tracking performance by the use CATM-CFAR algorithm, related to CA CFAR algorithm. In the future work, the multitarget cases will be considered.

REFERENCES

[1] Reid D.B.: 'An algorithm for tracking multiple targets', IEEE Transaction on Automatic Control, vol. 24, no. 6, 1979, pp. 843-854.

[2] Blackman S.: 'Multiple-target tracking with radar applications', Artech House, 1986.

[3] Challa S., Evans R., Morelande M. and Mušicki D.: 'Fundamentals of Object Tracking', Cambridge University Press, 2011.

[4] D. Ivković, M. Andrić, B. Zrnić, 'Detection of Very Close Targets by Fusion CFAR Detectors', Scientific Technical Review, ISSN 1820-0206, 2016,Vol.66, No.3,pp.50-57.

[5] D. Ivković, M. Andrić, B. Zrnić, P. Okiljević, N. Kozić, CATM-CFAR Detector in the Receiver of the Software Defined Radar, Scientific Technical Review, ISSN 1820-0206, 2014, Vol.64,No.4, pp.27-38.

[6] Mušicki D., Evans R., and Stanković S.: 'Integrated Probabilistic Data Association (IPDA)', IEEE Transaction on Automatic Control, vol. 39, no. 6, 1994, pp. 1237-1241.

[7] Mušicki D. and Evans R.: 'Joint Integrated Probabilistic Data Association – JIPDA', IEEE Transaction on Aerospace and Electronic Systems, vol. 40, 2004, no. 3, pp. 1093-1099.

[8] Rohling H., 'Radar CFAR Thresholding in Clutter and Multiple Target Situations', IEEE Transaction on Aerospace and Electronic Systems Vol. AES-19, 1983, no.4, pp.608-621.

[9] Finn H.M., Johnson R.S., 'Adaptive detection mode with threshold control as a function of spatially sampled clutter level estimate', RCA Rev., 1968, 29, (3), pp.414-464.

[10] Trunk G.V, 'Range resolution of targets using automatic detectors', IEEE Transaction on Aerospace and Electronic Systems, 1978, 14, (5), pp.750-755.

[11] Song T.L., Mušicki D., Kim D.S., and Radosavljević Z.: 'Gaussian mixtures in multi-target tracking: a look at Gaussian Mixture Probability Hypothesis Density and Integrated Track Splitting', IET Proceedings: Radar, Sonar and Navigation, vol. 6, 2012, no. 5, pp. 359-364..

[12] Radosavljević Z., Mušicki D.: Limits of target tracking in heavy clutter, ASIA-Pacific International Conference of Synthetic Aperture Radar APSAR 2011, Seoul, Korea, 2011.

[13] Bujaković D., Andrić M., Mikluc D., Bondžulić B.: 'Parameter Order Selection of Autoregressive Model for Classification of Ground Surveillance Radar Targets', Scientific Technical Review, ISSN 1820-0206, 2016, Vol.66,No.2, pp.3-9.te.

Probability of Detection and False Alarm Density Estimation in Target Tracking Systems With Unknown Measurement Noise Statistics

Asem Elhasaeri, Aleksandra Marjanović, Sanja Vujnović, Goran Kvaščev, and Željko Đurović

***Abstract*—Successful implementation of any moving target detection system depends on precise knowledge of several statistical quantities such as the probability of target detection and false alarms density. These parameters are usually unknown as well as variable and, even though algorithms exist which are able to estimate them, they are further dependent on the knowledge of model parameters. This paper analyzes the effect the unknown measurement noise covariance has on probability of detection and clutter rate estimation in target tracking systems and proposes improvement in a form of noise covariance estimation.**

***Index Terms*—Target Tracking System, Probability of Detection, False Alarms Rate, Covariance Estimation.**

I. Introduction

During the past several decades the scientific community has shown great interest in moving target tracking systems [1]. At first, the motivation for developing these algorithms has been solely for military purposes, but later their applicability has significantly increased to areas such as traffic surveillance systems [2] and biological systems [3]. Currently in the literature many different structures are proposed for moving target tracking systems and among them there are various solutions for target state estimation filters and measurement-to-track association algorithms. Usually successful implementation of these structures heavily depends on precise knowledge of parameters such as the probability of target detection and density of false alarms. This can be quite a drawback due to the fact that these statistical parameters are usually unknown and it is difficult to estimate them because they are non-stationary in time and space. The importance of this problem has been described by [4] where it is stated that the exact knowledge of probability of target detection and density of false alarms is crucial for good behavior of modern target tracking filters such as Probability Hypothesis Density (PHD) and Cardinalized Probability Hypothesis Density (CPHD) filters.

There are several solutions presented in the literature for the problem of probability of target detection and false alarm rate estimation. In [5] Expectation Maximization (EM) algorithm is used to estimate several parameters of Gaussian Multiple Target Tracking Model (MTT), among which are probability of target detection and density of false alarms. Even though this algorithm showed great promise for offline patch procedure, in online setting it was shown to have bias. On the other hand [6] used Joint Probabilistic Data Association (JDPA) filter for estimation of both of these parameters and in [7] further complicated the problem by assuming non-homogeneous false alarm density background. The complexity of aforementioned solutions leaves much to be desired in terms of practical implementation. Recently a new approach for probability of target detection and density of false alarms estimation is proposed [8] which is based on a two-step generalized maximum likelihood (GML) approach and which is numerically simpler due to the fact that each scan generates a maximum number of two hypotheses which need to be tested. Also, this algorithm does not require any specific data association technique which further simplifies things.

Common feature among all of these algorithms is the assumption that the kinematic model parameters are known. This is not always justified, especially when the measurement error covariance is concerned. Within this paper the sensitivity of a two-step GML approach algorithm to the possible errors in measurement noise covariance matrix are analyzed in detail. Furthermore, a Kalman filter is implemented to estimate the correct value of the noise covariance matrix. This simultaneous estimation of probability of target detection and density of false alarms as well as the unknown measurement noise statistics is tested and the results with such corrected parameters are analyzed as well.

This paper is structured as follows. In Section II a detailed description of a moving target tracking system using two-step GML approach is presented. After that, in Section III the sensitivity of this algorithm to changes in measurement noise covariance matrix is analyzed and in Section IV the estimator of the noise covariance is implemented and tested. Finally, the conclusion is presented in Section V.

Asem Elhasaeri is with the School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail: asem_issa@etf.bg.ac.rs).

Aleksandra Marjanović is with the School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail: amarjanovic@etf.bg.ac.rs).

Sanja Vujnović is with the School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail: svujnovic@etf.bg.ac.rs).

Goran Kvaščev is with the School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail: kvascev@etf.bg.ac.rs).

Željko Đurović is with the School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail: zdjurovic@etf.bg.ac.rs).

II. System description

Let us assume that there is a trace whose kinematic characteristics are monitored using Kalman filter model:

$$X[k+1] = \Phi X[k] + \Gamma w[k], \tag{1}$$

$$Y[k] = HX[k] + v[k]. \tag{2}$$

Here $X[k]$ is a trace state vector, $w[k]$ is a stochastic process of a model error, matrices Φ and Γ are state and input matrices, $Y[k]$ is a measurement vector and $v[k]$ is a stochastic process of a measurement noise at k-th scan. By applying Kalman filter state vector predictions can be obtained from one scan to another. The prediction will be denoted as $\hat{X}[k|k-1]$ and covariance error prediction matrix as $S[k|k-1]$.

After the prediction step, an association technique needs to be performed by forming a rectangular gate around the track prediction $\hat{X}[k|k-1]$ with dimensions $K_g\sigma_{ii}, i = 1, \ldots, n$. Here K_g is a gate parameter, σ_{ii} is the i-th diagonal element of covariance matrix $S[k|k-1]$ and n denotes spatial dimension in which the gate is applied. In each scan, in the area encompassed by the gate there can be zero, one or more observations. In an effort to construct a hypothesis whose likelihood maximization will give the estimate of unknown parameters two different cases are analyzed. In the first case the number of observations in k-th scan is $M_k = 0$ and there is only one partial hypothesis:

- H_k^0: target is not detected and the number of false alarms is zero.

In the second case the number of observations in the k-th scan is $M_k \geq 1$ and there are two partial hypotheses:

- H_k^0: target is not detected and the number of false alarms is M_k.
- H_k^1: target is detected and the number of false alarms is $M_k - 1$.

Combination of partial hypotheses in the last N scans produces integral hypotheses and there are $N_h = 2^{v(k-N+1)} \cdot 2^{v(k-N+2)} \cdots 2^{v(k)}$ of them, where $v(l) = \text{sgn}(M_l)$. Note that there is significantly smaller number of hypotheses here than in other multiple hypothesis testing approaches due to the fact that in the case when several observations are detected within a gate, only the one with the smallest statistical distance from the predicted position is analyzed for association purposes. Each integral hypothesis is associated with a likelihood function $L(H^i), i = 1, \ldots, N_h$ and the parameter pair (P_d^i, λ_{fa}^i) is computed which maximizes the likelihood of each integral hypothesis $L^{opt}(H^i) = \max_{P_d, \lambda_{fa}} L(H^i)$. Here P_d is the probability of detection, while λ_{fa} is false alarm density. After that, among all the optimal values of likelihood, the highest one is selected $L^{opt}(H^j) = \max_{i=1,\ldots,N_h} L^{opt}(H^i)$ and optimal estimations (P_d^j, λ_{fa}^j) are obtained from the chosen hypothesis.

Assuming that all the necessary conditions are satisfied [8] (i.e. $\hat{X}[k|k-1]$ is unbiased estimate of $X[k]$ and a Gaussian process, the frequency of false alarms is a Poisson process with mean $\lambda_{fa}V_g$, where V_g is an area of the gate, P_d can be considered constant, etc.), then i-th integral hypothesis can be calculated as

$$L(H^i) = \prod_{j=k-N+1}^{k} L(H_j^{p_j}), \;\; i = 1, \ldots, N_h. \tag{3}$$

Here $L(H_j^{p_j})$ is the likelihood of partial hypothesis in the j-th scan

$$L(H_j^{p_j}) = \begin{cases} (1-P_d)\frac{(\lambda_{fa}V_g)^{M_j} e^{-\lambda_{fa}V_g}}{M_j!(V_g)^{M_j}} & p_j = 0 \\ P_d \frac{(\lambda_{fa}V_g)^{M_j-1} e^{-\lambda_{fa}V_g}}{(M_j-1)!(V_g)^{M_j-1}} f_j & p_j = 1 \end{cases}, \tag{4}$$

with

$$f_j = \frac{1}{(2\pi)^{n/2}\left|S[j|j-1]\right|^{1/2}} e^{-\frac{d(j)^2}{2}}. \tag{5}$$

Here $d^2(j)$ denotes statistical distance between prediction position $X[j|j-1]$ and statistically closest observation Z_j in scan j.

Each integral hypothesis from the last N scans can be represented as a sequance of N binary digits $H^i = [p_1\, p_2 \cdots p_N]$, where p_j takes value 0 if there is no associated observation in scan $k-N+j$ and value of 1 if there is statistically closest observation associated with the trace in that scan. Then, the likelihood of integral hypothesis H^i becomes

$$\begin{aligned} L(H^i) = (1-P_d)^{N-\sum_{j=k-N+1}^{k} p_j} P_d^{\sum_{j=k-N+1}^{k} p_j} \cdot \\ \cdot \frac{\lambda_{fa}^{\sum_{j=k-N+1}^{k} M_j - p_j} e^{N\lambda_{fa}}}{\prod_{j=k-N+1}^{k}(M_j - p_j)!} \prod_{j=k-N+1}^{k} f_j(Z_j)^{p_j}, \end{aligned} \tag{6}$$

where $f_j(Z_j)$ is the probability density function

$$\begin{aligned} f_j(Z_j) = \frac{1}{(2\pi)^{n/2}\left|S[j|j-1]\right|^{1/2}} \cdot \\ \cdot e^{-0.5(Z_j - \hat{X}[j|j-1])^T S^{-1}[j|j-1](Z_j - \hat{X}[j|j-1])}. \end{aligned} \tag{7}$$

Maximization of likelihood as a function of unknown parameters is obtained by solving

$$\frac{\partial L(H^i)}{\partial P_d} = 0, \;\; \frac{\partial L(H^i)}{\partial \lambda_{fa}} = 0, \tag{8}$$

and the results are simple estimates of probability of detection ($\hat{P}_d$) and the density of false alarms ($\hat{\lambda}_{fa}$)

$$\hat{P}_d = \frac{\sum_{j=k-N+1}^{k} p_j}{N}, \hat{\lambda}_{fa} = \frac{\sum_{j=k-N+1}^{k} M_j - p_j}{NV_g}. \tag{9}$$

Further analysis of this estimator [8] shows that there is an estimation bias for both $\hat{P}_d$ and $\hat{\lambda}_{fa}$ which is smaller as the probability of detection increases. For this reason additional steps are implemented in order to reduce this bias.

The first step is estimator bias reduction and it relies on the assumption of nearly deterministic dependency among three parameters: K_g, $\hat{P}_d$ and $\hat{\lambda}_{fa}$. Corrected estimates $\hat{P}_d^c$ and $\hat{\lambda}_{fa}^c$ are obtained as:

$$\begin{aligned} \hat{P}_d^c &= \hat{P}_d + a_P K_g + b_P \hat{P}_d + c_P \hat{\lambda}_{fa} + d_P, \\ \hat{\lambda}_{fa}^c &= \hat{\lambda}_{fa} + a_\lambda K_g + b_\lambda \hat{P}_d + c_\lambda \hat{\lambda}_{fa} + d_\lambda. \end{aligned} \tag{10}$$

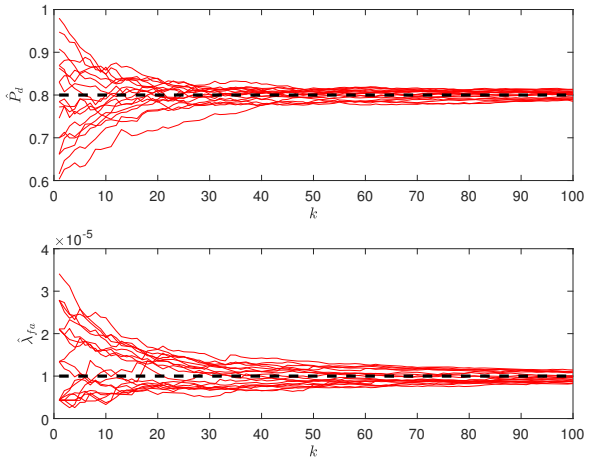


Fig. 1: Estimation results for probability of detection (up) and false alarm density (down) when all the kinematic parameters are known. Red lines represent the output of the algorithm for different initial conditions, while the black line is the correct value.

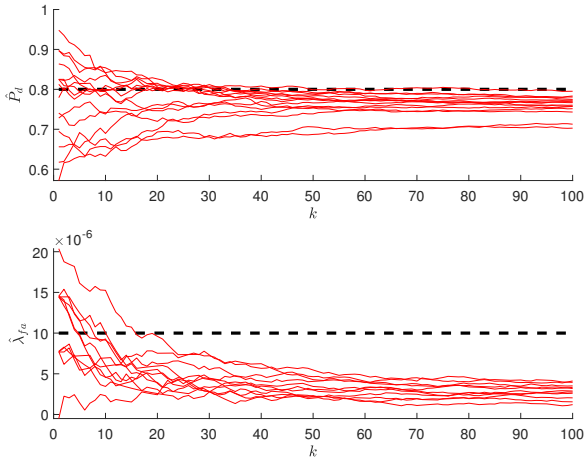


Fig. 2: Estimation results for probability of detection (up) and false alarm density (down) when measurement noise covariance matrix is unknown. Red lines represent the output of the algorithm for different initial conditions, while the black line is the correct value.

The correction parameters a_P, b_P, c_P, d_P and a_λ, b_λ, c_λ, d_λ are obtained by least squares method

$$\phi_i = (XX^T)^{-1}XY_i. \tag{11}$$

Here $\phi_i = [a_i\ b_i\ c_i\ d_i]^T, i \in \{P, \lambda\}$ are unknown correction parameters, $X = [K_g\ \hat{P}_d\ \hat{\lambda}_{fa}\ 1]^T$ is the regression vector and Y_i is the model output. These parameters are estimated for different values of gate width, detection probability and false alarm density.

The second step is estimator variance reduction using a recursive estimator with a variable forgetting factor

$$\begin{aligned}\hat{P}_d^r[k+1] &= \alpha[k]\hat{P}_d^r[k] + (1-\alpha[k])\hat{P}_d^c[k],\\ \hat{\lambda}_{fa}^r[k+1] &= \alpha[k]\hat{\lambda}_{fa}^r[k] + (1-\alpha[k])\hat{\lambda}_{fa}^c[k].\end{aligned} \tag{12}$$

Here $\hat{P}_d^r$ and $\hat{\lambda}_{fa}^r$ are the recursive estimates of probability of detection and false alarm density, and $\alpha[k] = \alpha[\infty] - e^{-k/\tau}(\alpha[\infty] - \alpha[0])$ is the variable forgetting factor with τ as a time constant representing the rate of change, and $\alpha[\infty]$ and $\alpha[0]$ its final and initial values. With these corrections the system for estimating parameters P_d and λ_{fa} is computationally simple with satisfactory performance, under the conditions that all the other parameters are known.

III. Sensitivity analysis

The system for probability of target detection and false alarm density estimation described in the previous section is highly effective and much simpler than other techniques available in the literature; however it does require the complete knowledge of the initial kinematic model from Eq. (1) and (2). Figure 1 shows the behavior of the algorithm for different initial conditions when all the necessary kinematic parameters are known. For fixed probability of detection ($P_d = 0.8$), false alarm density ($\lambda_{fa} = 10^{-5}$) and gate size ($K_g = 2.6$) it is evident that the estimation converges to correct values as the number of scans increases.

It would be interesting to analyze how possible errors in prediction covariance matrix influence the performance of the estimation algorithm. These errors are usually caused by unknown or erroneous information about the measurement noise covariance matrix R. Figure 2 shows how the algorithm behaves when it has erroneous information about the value of R. Fixed parameters are the same as in Fig. 1 and, in this case, the algorithm converges as well, but it has a significant bias. Naturally, the size of this bias depends on the magnitude of misinformation the algorithm has about measurement noise covariance matrix.

IV. Estimator adaptation

Previous chapter demonstrates high sensitivity of a two-step GML approach estimator to unknown measurement noise statistics and, therefore, erroneous covariance error prediction matrix of a model described in Eq. (1) and (2). Since a priori knowledge about the measurement noise is not known, the goal is to estimate it simultaneously with tracking parameters [9]. If we use Eq. (2) to define residual as

$$r[k] = Y[k] - H\hat{X}[k|k-1], \tag{13}$$

then the covariance matrix of this residual vector is

$$E\{r[k]r^T[k]\} = R + HS[k|k-1]H^T. \tag{14}$$

Bearing this in mind, measurement noise covariance matrix can be estimated as

$$\hat{R} = \hat{C}_r - H\hat{S}[k|k-1]H^T, \tag{15}$$

where $\hat{C}_r$ is estimated residual covariance matrix which can be obtained by generating trajectories using kinematic model from Eq. (1) and (2) with the following values:

$$\Phi = \begin{bmatrix}1 & 1 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 1\\ 0 & 0 & 0 & 1\end{bmatrix},\ \Gamma = \begin{bmatrix}0 & 0\\ 1 & 0\\ 0 & 0\\ 0 & 1\end{bmatrix},\ H = \begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 0 & 1 & 0\end{bmatrix}. \tag{16}$$

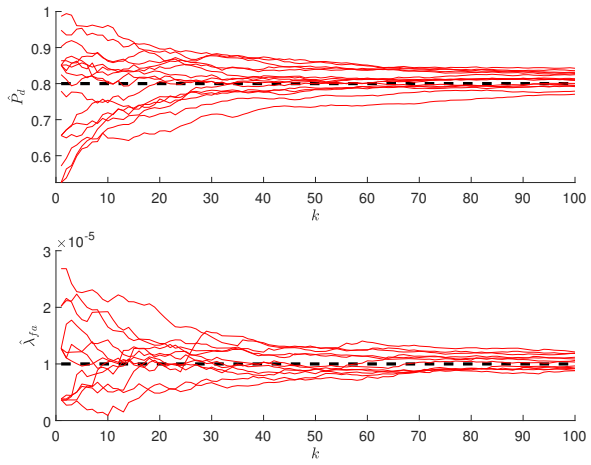


Fig. 3: Estimation results for probability of detection (up) and false alarm density (down) when measurement noise covariance matrix is estimated. Red lines represent the output of the algorithm for different initial conditions, while the black line is the correct value.

Here there are 4 states in vector $X[k]$ which are spatial coordinate x and its velocity, and spatial coordinate y, and its velocity, respectively. In each scan measurement is updated in this way, measurement noise covariance is estimated ($\hat{R}$) and then it is used to estimate parameters $\hat{P}_d$ and $\hat{\lambda}_{fa}$.

By implementing this adaptation to the existing algorithm from Section II, an augmented system is obtained which simultaneously estimates measurement noise covariance matrix ($\hat{R}$) as well as probability of detection ($\hat{P}_d$) and false alarm density estimation ($\hat{\lambda}_{fa}$) in each scan. Results of the algorithm augmented in this way can be seen in Fig. 3.

It is obvious from the figure that performance is significantly better than in the previous case when there was no a priori knowledge about parameter R. For all the initial conditions parameter estimates converge to the exact value of P_d and λ_{fa}. The dynamic of the convergence is similar to the first case in Fig. 1 when there is a priori knowledge about measurement noise, i.e. after as little as 30 scans parameter estimates approach a reasonable vicinity of the exact values. What is also noticeable is that the standard deviation of the estimate error of these parameters is somewhat larger in the case when R is estimated than when the correct values are known.

V. Conclusion

This paper investigates the sensitivity of a two-step GML approach to probability of detection and false alarm density estimation on an unknown measurement noise statistics. The algorithm itself was shown to have satisfactory performance when a priori knowledge about the measurement noise is available, but it has a significant bias when that knowledge is incorrect. For this reason an adaptation to the original approach was proposed which estimates measurement noise covariance matrix in each scan while simultaneously estimating the probability of detection and false alarm density as well. This augmentation of the algorithm was shown to remove the bias and enhance the convergence of the algorithm with respect to the case when estimation is lacking and a priori knowledge about noise statistics is erroneous.

In further research it would be interesting to investigate the case in which knowledge about a model noise is unavailable as well. This case is somewhat more complicated and will probably require more complex noise statistics estimator structure. Also, the existing augmentation of the algorithm can be further improved by introducing adaptive estimation of measurement noise covariance matrix.

Acknowledgments

This research was partially supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia, Contracts No. TR32038 and III42007.

References

[1] S. Blackman and R. Popoli, "Design and analysis of modern tracking systems," *Artech House Radar Library*, 1999.

[2] B. Benfold and I. Reid, "Stable multi-target tracking in real-time surveillance video," in *CVPR 2011*, pp. 3457–3464, IEEE, 2011.

[3] I. Schlangen, V. Bharti, E. Delande, and D. E. Clark, "Joint multi-object and clutter rate estimation with the single-cluster PHD filter," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pp. 1087–1091, IEEE, 2017.

[4] R. P. Mahler, B.-T. Vo, and B.-N. Vo, "CPHD filtering with unknown clutter rate and detection profile," *IEEE Transactions on Signal Processing*, vol. 59, no. 8, pp. 3497–3513, 2011.

[5] S. Yıldırım, L. Jiang, S. S. Singh, and T. A. Dean, "Calibrating the Gaussian multi-target tracking model," *Statistics and Computing*, vol. 25, no. 3, pp. 595–608, 2015.

[6] S. He, H.-S. Shin, and A. Tsourdos, "Joint probabilistic data association filter with unknown detection probability and clutter rate," *Sensors*, vol. 18, no. 1, p. 269, 2018.

[7] X. Chen, R. Tharmarasa, T. Kirubarajan, and M. Pelletier, "Integrated clutter estimation and target tracking using Poisson point process," in *Signal and Data Processing of Small Targets 2009*, vol. 7445, p. 74450X, International Society for Optics and Photonics, 2009.

[8] A. Alhasaeri, A. Marjanović, S. Vujnović, P. Tadić, and Z. Djurović, "On false alarms density and detection profile estimation in target tracking systems," in *SAUM 2018*, 2018.

[9] Z. M. Durovic and B. D. Kovacevic, "Robust estimation with unknown noise statistics," *IEEE Transactions on Automatic Control*, vol. 44, no. 6, pp. 1292–1296, 1999.

On the Performance of the PHD Filter

Predrag Vasilić, Sanja Vujnović, Aleksandra Marjanović, Nikola Popović, and Željko Đurović

***Abstract*—The Gaussian Mixture Probability Hypothesis Density (GMPHD) filter represents a closed form solution to the Probability Hypothesis Density (PHD) filter which solves the problem of multi-target tracking (MTT), namely the tracking of multiple targets using the collection of measurements at each time sample. Each target has a certain probability of detection. Besides the target-caused observations, there are others which represent clutters. The paper examines the performance of the GMPHD filter in a scenario with many false alarms and deviations in the initial assumptions of the filter model parameters. The performance is measured using the Optimal Subpattern Assignment Metric (OSPA), which is broadly used as a standard metric in the estimation of the distance between two sets of vectors.**

***Index Terms*—Multi-target Tracking, Random Sets, Probability Hypothesis Density (PHD) Filter, Gaussian Mixture.**

I. Introduction

Single-target filters such as Kalman filter, assume the presence of only one target and only one noise-contaminated measurement originating from the target at each time sample [1], [2]. Multi-target tracking implies the possibility of more targets in the observation area. Each of these targets has its own time of appearance (birth) and disappearance (death) [3], [4]. Such targets are observed using one or more sensors which provide a collection of observations which does not necessarily include the information about all the targets. Furthermore, it can include some false detections due to various environmental influences and sensor noise. The goal of multi-target filtering methods is to estimate the number of targets and their state at each time sample.

One of the major issues of standard MTT approaches is the association of available measurements to certain targets [4]. Due to the combinatorial nature of the problem, data association implies significant computer resources. Random Finite Set (RFS) theory provides an elegant formulation of MTT problem by avoiding the explicit association of data [5], [6]. The idea is to represent the unknown target state collection as a set of states and the collection of measurements as a set of measurements and to formulate a mathematical tool similar to the Bayesian filtering in a single-target case. RFS theory falls under the field of Point Process theory, which represent the basis for derivation of Mahler's Finite-set statistics (FISST) [6]–[8]. Bayes RFS filter works in the space of sets which limits the application of this algorithm to a small number of targets. Therefore, the approximations of this filter were made in the form of the PHD filter, which operates in the space of vectors of individual targets, and its extension, the Cardinalized Probability Hypothesis Density (CPHD) filter [9], [10]. Even though these filters represent considerably simpler approximations, they contain integrals which do not have a closed-form solution. Different approaches to solving such equations led to many versions of these filters. Special attention was paid to the derivation of the GMPHD filter, whose final equations are analog to those of Kalman's [11].

This paper investigates the robustness of the GMPHD filter in the case of deviation in initial filter parameter estimations. We used the example which is the benchmark for testing of this type of filters, characterized with a high density of false alarms [12] and we provided performance results using the OSPA metric [13]. The robustness was tested in terms of the a prior knowledge about the appearance of targets, the changes in the state transition and measurement covariance matrices, as well as the change in the probability of detection.

The paper is organized as follows. Section II provides a brief overview of the basic concepts of RFS theory, as well as the definitions of the GMPHD filter and the OSPA metric. Section III contains the description of the commonly used test example and the testing of robustness, together with the results. Section IV concludes the paper and points out directions for further research.

II. Filtering formulation

This section gives a formulation of the multi-target filtering problem with the RFS approach. First we introduce the basic concepts of the RFS theory. Then we describe the idea and the properties of the PHD filter. Finally, we provide an overview of the GMPHD filter whose performances are examined in this paper.

A. Multi-target Filtering with Random Finite Sets

Unlike the single-target filtration problems which include only one target, the multi-target case assumes more than one target at each time. Let us assume that there are N_k targets from the space $\mathcal{X} \subseteq \mathbb{R}^{n_x}$ with states $x_{k,1}, x_{k,2}, ..., x_{k,N_k}$ and M_k observations from the space $\mathcal{Z} \subseteq \mathbb{R}^{n_z}$ with measurements $z_{k,1}, z_{k,2}, ..., z_{k,M_k}$ at time k. The ordering of states and observations has no significance, because we do not know

Predrag Vasilić is with the School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail: vasilic@etf.bg.ac.rs).

Sanja Vujnović is with the School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail: svujnovic@etf.bg.ac.rs).

Aleksandra Marjanović is with the School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail: amarjanovic@etf.bg.ac.rs).

Nikola Popović is with the School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail: amarjanovic@etf.bg.ac.rs).

Željko Đurović is with the School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail: zdjurovic@etf.bg.ac.rs).

which observation comes from which target, so it is natural to represent these collections as sets

$$X_k = \{x_{k,1}, x_{k,2}, ..., x_{k,N_k}\} \in \mathcal{F}(\mathcal{X}), \quad (1)$$

$$Z_k = \{z_{k,1}, z_{k,2}, ..., z_{k,M_k}\} \in \mathcal{F}(\mathcal{Z}). \quad (2)$$

Single-target filtering methods deal with the state of only one target x from the space $\mathcal{X} \subseteq \mathbb{R}^{n_x}$ and its observation z from $\mathcal{Z} \subseteq \mathbb{R}^{n_z}$. Therefore, the set X_k is in the collection $\mathcal{F}(\mathcal{X})$ of all finite subsets of $\mathcal{X}$. Similarly, the measurement space is the collection $\mathcal{F}(\mathcal{Z})$ of all finite subsets of $\mathcal{Z}$.

The goal of multi-target filtering methods is to estimate the current target set X_k using the observation sets $Z_{1:k} = (Z_1, Z_2, ..., Z_k)$ [6], [9]. The idea is to create such a mathematical tool which would have the form of the Bayesian filtering, appropriate for sets instead of vectors. This implies the formulation of random statistical set description, which leads to the Random Finite Set theory [6]. Rigorous mathematical definitions of random sets are given by the Point Process theory [5], [14]. Here, we provide only the most important, necessary results [7], [8], [12].

The random set $\mathbf{X}$ is described with a so-called multi-target probability density function

$$\pi(X) = \pi(\{x_1, x_2, ..., x_n\}), \quad (3)$$

which is a function of a set of vectors. The finite set has a random cardinality and each set element represents a random vector variable. The set integral in the region $S \subseteq \mathbb{R}^{n_x}$ is defined to be

$$\int_S \pi(X)\delta X = \pi(\varnothing) + \sum_{n\geq 1}^{\infty} \frac{1}{n!} \times \int_{S^n} \pi(\{x_1, x_2, ..., x_n\}) dx_1 dx_2 ... dx_n \quad (4)$$

and represents the possibility of set elements belonging to S.

Multi-target transition model is defined as follows. Let the multi-target state at time $k-1$ be described with a set X_{k-1}. The random finite set $\mathbf{X}_k$ which is formed using X_{k-1} is equal to

$$\mathbf{X}_k = T_{k|k-1}(X_{k-1}) \bigcup \mathbf{\Gamma}_k, \quad (5)$$

where the random set $T_{k|k-1}(X_{k-1})$ is created from all the targets $x_{k-1} \in X_{k-1}$ with the survival probabilities $p_{S,k|k-1}(x_{k-1})$ and state transition equations

$$f_{k|k-1}(x_k|x_{k-1}). \quad (6)$$

Besides the targets that survived time $k-1$, time k can include some completely new, independent targets. These new targets form the random set $\mathbf{\Gamma}_k$ which can be assigned with a different probability density function depending on the specific problem. The random set $\mathbf{X}_k$ is associated with a multi-target transition model with density function $f_{k|k-1}(X_k|X_{k-1})$.

The observation model is constructed in the similar way. Let us assume that each target $x_k \in X_k$ has the probability of detection $p_{D,k}(x_k)$. The observation set $\mathbf{Z}_k$ represents a union

$$\mathbf{Z}_k = \Theta_k(X_k) \bigcup \mathbf{K}_k, \quad (7)$$

where the random set $\Theta_k(X_k)$ is formed by adding a new observation vector in the event of target detection, which is related to the sensor observation law

$$g_k(z_k|x_k). \quad (8)$$

The random set $\mathbf{K}_k$ represents a set of false alarms and it is described using a certain problem-related distribution [5], [12]. The multi-target density function $g_k(Z_k|X_k)$ is defined for the random set $\mathbf{Z}_k$.

Following the aforementioned transition and observation models, let us define the recursive multi-target Bayes filter

$$p_{k|k-1}(X_k|Z_{1:k-1}) = \int f_{k|k-1}(X_k|X_{k-1}) p_{k-1|k-1}(X_{k-1}|Z_{1:k-1}) \delta X_{k-1}, \quad (9)$$

$$p_{k|k}(X_k|Z_{1:k}) = \frac{g_k(Z_k|X_k) p_{k|k-1}(X_k|Z_{1:k-1})}{\int g_k(Z_k|X) p_{k|k-1}(X|Z_{1:k-1}) \delta X}. \quad (10)$$

The advantage of the Bayesian RFS approach in solving multi-target tracking problems lies in the lack of need for direct data association, which can be observed from the derived prediction and correction equations. The main disadvantage is the complexity and computational intractability of the integrals defined by (9) and (10). Therefore, this filter is applicable only in the case of few targets. Over the years, some of its approximations have been derived, such as the PHD filter.

B. The Probability Hypothesis Density filter

PHD filter relies on the statistical moment of the random finite set [9]. The first statistical moment $v(x)$, also called the Probability Hypothesis Density function, has the following property

$$E\{|\mathbf{X} \cap S|\} = \int_S v(x)dx. \quad (11)$$

In other words, the integral over the area $S \subseteq \mathbb{R}^{n_x}$ gives the expected number of targets in that area. Therefore, the PHD function represents the density of target appearance in the area. Since the PHD function at one point is directly proportional to the expected number of targets at that point, the peak in the PHD function represents an area with the highest expected number of targets. Consequently, it is proportional to the probability of target being at that point. Hence, the idea of the PHD filter is to recursively propagate the corresponding density function $v_{k|k}(x|Z_{1:k})$, instead of the multi-target posterior density function $p_{k|k}(X_k|Z_{1:k})$ [14], [15]. The PHD filter was originally presented in [9], where it was derived using the FISST mathematical tool from Point Process theory [7], [8]. Let us denote the posterior PHD function at time $k-1$ as $v_{k-1|k-1}(x)$ and assume that:

- Each target evolves and generates observations independently of one another.
- The birth RFS and the surviving RFSs are independent of each other.
- The clutter RFS is Poisson and independent of target-originated measurements.
- The prior and predicted multi-target RFS are Poisson.

Consequently, the PHD functions of prediction $v_{k|k-1}(x)$ and correction $v_{k|k}(x)$ are equal to

$$v_{k|k-1}(x) = \int p_{S,k|k-1}(x_{k-1}) v_{k-1|k-1}(x_{k-1}) \times f_{k|k-1}(x|x_{k-1}) dx_{k-1} + \gamma_k(x), \tag{12}$$

$$v_{k|k}(x) = (1 - p_{D,k}(x)) v_{k|k-1}(x) + p_{D,k}(x) v_{k|k-1}(x) \times \sum_{z \in Z_k} \frac{g_k(z|x)}{\kappa_k(z) + \int p_{D,k}(x) v_{k|k-1}(x) g_k(z|x) dx}, \tag{13}$$

where

$p_{S,k|k-1}(x_{k-1})$ =probability of target existence at time k given previous state x_{k-1},
$\gamma_k(x)$ =PHD of the birth RFS $\mathbf{\Gamma}_k$ at time k.
$p_{D,k}(x)$ =probability of detection given a state x at time k,
$\kappa_k(z)$ =PHD of the clutter RFS $\mathbf{K}_k$ at time k.

Functions $f_{k|k-1}(x|x_{k-1})$ and $g_k(z|x)$ are the aforementioned models of transition and observation. The advantage of the PHD filter over the Bayes multi-target filter is the employment of vector space $\mathcal{X}$ instead of the set space $\mathcal{F}(\mathcal{X})$. However, the propagation of the first moment only carries part of the information about the distribution of the random set. The integrals in (12) and (13) generally do not have a closed-form solution, hence some additional assumptions are required. The most popular version of the PHD filter is the Gaussian Mixture Probability Hypothesis Density filter (GMPHD), which introduces following assumptions.

- Each target follows a linear Gaussian dynamical model and the sensor has a linear Gaussian measurement model,

$$f_{k|k-1}(x_k|x_{k-1}) = \mathcal{N}(x_k; F_{k-1}x_{k-1}, Q_{k-1}), \tag{14}$$
$$g_k(z_k|x_k) = \mathcal{N}(z_k; H_k x_k, R_k), \tag{15}$$

- The survival and detection probabilities are state independent

$$p_{S,k|k-1}(x_{k-1}) = p_{S,k|k-1}, \tag{16}$$
$$p_{D,k}(x_k) = p_{D,k}. \tag{17}$$

- The intensity of the birth RFS is Gaussian mixture of the form

$$\gamma_k(x) = \sum_{i=1}^{J_{\gamma,k}} w_{\gamma,k}^{(i)} \mathcal{N}(x; m_{\gamma,k}^{(i)}, P_{\gamma,k}^{(i)}), \tag{18}$$

where $J_{\gamma,k}, w_{\gamma,k}^{(i)}, m_{\gamma,k}^{(i)}, P_{\gamma,k}^{(i)}, i = 1, 2, ..., J_{\gamma,k}$ represent the model parameters of the PHD function of newborn targets. The function $\mathcal{N}(x; m, P)$ denotes a Gaussian density with mean value m and covariance P.

Apart from these assumptions, derivation relies on the two identities of the Gaussian distribution which are used in Bayesian formulation of the Kalman filter [6]. If all previous assumptions stand and if the posterior PDH function at time $k-1$ is equal to

$$v_{k-1|k-1}(x) = \sum_{i=1}^{J_{k-1}} w_{k-1}^{(i)} \mathcal{N}(x; m_{k-1}^{(i)}, P_{k-1}^{(i)}) \tag{19}$$

then the PHD prediction function is given as

$$v_{k|k-1}(x) = v_{S,k|k-1}(x) + \gamma_k(x), \tag{20}$$

where

$$v_{S,k|k-1}(x) = p_{S,k|k-1} \sum_{j=1}^{J_{k-1}} w_{k-1}^{(j)} \mathcal{N}(x; m_{S,k|k-1}^{(j)}, P_{S,k|k-1}^{(j)}), \tag{21}$$

$$m_{S,k|k-1}^{(j)} = F_{k-1} m_{k-1}^{(j)}, \tag{22}$$
$$P_{S,k|k-1}^{(j)} = Q_{k-1} + F_{k-1} P_{k-1}^{(j)} F_{k-1}^T, \tag{23}$$

and $\gamma_k(x)$ is defined by (18). These equations represent the prediction step of the GMPHD filter.

Let PHD prediction function at time k be a polymodal Gaussian function

$$v_{k|k-1}(x) = \sum_{i=1}^{J_{k|k-1}} w_{k|k-1}^{(i)} \mathcal{N}(x; m_{k|k-1}^{(i)}, P_{k|k-1}^{(i)}). \tag{24}$$

Then the posterior PHD function $v_{k|k}(x)$ is also polymodal Gaussian function in the form

$$v_{k|k}(x) = (1 - p_{D,k}) v_{k|k-1}(x) + \sum_{z \in Z_k} v_{D,k}(x; z), \tag{25}$$

where

$$v_{D,k}(x; z) = \sum_{j=1}^{J_{k|k-1}} w_k^{(j)}(z) \mathcal{N}(x; m_{k|k}^{(j)}(z), P_{k|k}^{(j)}), \tag{26}$$

$$w_k^{(j)}(z) = \frac{p_{D,k} w_{k|k-1}^{(j)} q_k^{(j)}(z)}{\kappa_k(z) + p_{D,k} \sum_{l=1}^{J_{k|k-1}} w_{k|k-1}^{(l)} q_k^{(l)}(z)}, \tag{27}$$

$$q_k^{(j)}(z) = \mathcal{N}(z; H_k m_{k|k-1}^{(j)}, R_k + H_k P_{k|k-1}^{(j)} H_k^T), \tag{28}$$

$$m_{k|k}^{(j)}(z) = m_{k|k-1}^{(j)} + K_k^{(j)}(z - H_k m_{k|k-1}^{(j)}), \tag{29}$$

$$P_{k|k}^{(j)}) = (I - K_k^{(j)} H_k) P_{k|k-1}^{(j)}, \tag{30}$$

$$K_k^{(j)} = P_{k|k-1}^{(j)} H_k^T (H_k P_{k|k-1}^{(j)} H_k^T + R_k)^{-1}. \tag{31}$$

Previous equations represent the correction step of the GMPHD filter. The analogy with the Kalman filter equations is direct, due to the Bayesian formulation of the Kalman filter. Let us emphasize that the GMPHD filter has a closed-form solution, which is suitable for implementation. The drawback of the filter is the polynomial increase in the number of Gaussian components through iterations, demanding the heuristics for their pruning which can be found together with set estimation procedure in [11].

Testing of the RFS-based filtering methods requires a metric different then the classic Euclidean, which compares the sets of solutions. The standard metric for these methods is the Optimal Subpattern Assignment Metric (OSPA) [13]. Let $d^{(c)}(x, y) := min(c, |x - y|)$ represent the distance between

vectors x and y bounded by the cut off parameter $c > 0$, then the OSPA metric is defined as

$$\bar{d}_p^{(c)}(X,Y) = \left(\frac{1}{n}\left(\min_{\pi\in\Pi_n}\sum_{i=1}^{m} d^{(c)}(x_i, y_{\pi(i)})^p + c^p(n-m)\right)\right)^{\frac{1}{p}}, \tag{32}$$

if $m \leq n$; $\bar{d}_p^{(c)}(Y,X)$ if $m \geq n$; $\bar{d}_p^{(c)}(X,Y) = 0$ if $m = n = 0$. Set Π_n represents the set of all permutations of n elements. The order parameter p determines the sensitivity of the metric to the outliers, and the cut-off parameter c determines the relative weighting of the penalties assigned to the cardinality and localization errors. Expression (32) includes two factors, which represent the localization and cardinality error. These errors are often in simulations observed separately [12]:

$$\bar{e}_{p,loc}^{(c)}(X,Y) = \left(\frac{1}{n}\left(\min_{\pi\in\Pi_n}\sum_{i=1}^{m} d^{(c)}(x_i, y_{\pi(i)})^p\right)\right)^{1/p}, \tag{33}$$

$$\bar{e}_{p,card}^{(c)}(X,Y) = \left(\frac{c^p(n-m)}{n}\right)^{1/p}, \tag{34}$$

if $m \leq n$,, and $\bar{e}_{p,loc}^{(c)}(X,Y) = \bar{e}_{p,loc}^{(c)}(Y,X), \bar{e}_{p,card}^{(c)}(X,Y) = \bar{e}_{p,card}^{(c)}(Y,X)$ if $m > n$.

III. Perfomance robustness of the GMPHD filter

Previous section defines the prediction and correction parts of the GMPHD filter. Even in the single-target filtering case, there are very few a priori information, namely the parameters that need to be known during the initialization of the algorithm. This includes the parameters of the transition and observation models, target detection probability, process and measurement noise covariance matrices, etc. The quality of the filtration certainly depends on the robustness in terms of parameter changes, because some of the assumptions will not be met in the real-life scenario. Therefore, this is the next thing we should examine.

We used the implementation of the GMPHD algorithm suggested by its author as stated in [11]. Specifically, we used the filter implementation for the example given in [12]. The simulation scenario is the following. The number of targets in a 2D space is unknown. The sensor is collecting the data from the surveillance region $[-1000, 1000] \times [-1000, 1000]m$. The state of each target $x_k = [p_{x,k}, p_{y,k}, \dot{p}_{x,k}, \dot{p}_{y,k}]^T$ consists of position $(p_{x,k}, p_{y,k})$ avd velocity $(\dot{p}_{x,k}, \dot{p}_{y,k})$. The measurement vector $z = [z_{x,k}, z_{y,k}]^T$ received by the algorithm can be a noisy position of the target in the case of detection, or a clutter measurement. The transition and observation models are given by (14) and (15). Let us assume a linear Gaussian model with constant velocity

$$F_k = \begin{bmatrix} I_2 & T_s I_2 \\ 0_2 & I_2 \end{bmatrix}, \qquad Q_k = \sigma_v^2 \begin{bmatrix} \frac{T_s^4}{4} I_2 & \frac{T_s^3}{2} I_2 \\ \frac{T_s^3}{2} I_2 & T_s^2 I_2 \end{bmatrix},$$

where F_k i Q_k are the transition and process covariance matrices, respectively. The sampling period is $T_s = 1s$, and the process noise standard deviation is $\sigma_v = 5\frac{m}{s^2}$. The matrix I_n is an $n \times n$ identity matrix and 0_n is an $n \times n$ zero matrix. The survival probability of each target is $p_{S,k} = 0.99$. The birth of new targets is described using a

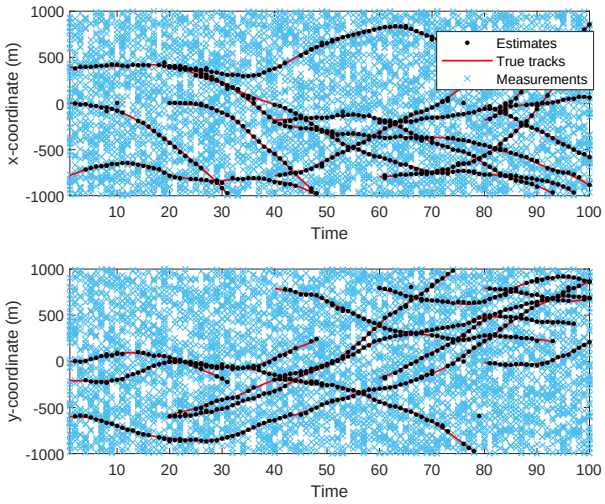


Fig. 1: Simulation results in the case of equivalent filter and model parameters. The scenario includes 12 targets which appear and disappear at different times.

PHD function $\gamma_k(x) = \sum_{i=1}^{4} w_\gamma \mathcal{N}(x; m_\gamma^{(i)}, P_\gamma)$, where $w_\gamma = 0.03, m_\gamma^{(1)} = [0,0,0,0]^T, m_\gamma^{(2)} = [400,-600,0,0]^T, m_\gamma^{(3)} = [-800,-200,0,0]^T, m_\gamma^{(4)} = [-200,800,0,0]^T$ i $P_\gamma = diag([10,10,10,10])^2$. The probability of detection is $p_{D,k} = 0.98$ and the measurement equation satisfies (15). Further on

$$H_k = [I_2 \quad 0_2], \qquad R_k = \sigma_\epsilon^2 I_2,$$

where $\sigma_\epsilon = 10m$ is the standard deviation of the measurement noise. Also, the false alarms are described using a PHD function

$$\kappa_k(z) = \lambda_c V u(z),$$

where $V = 4 \times 10^6 m^2$ is the volume of the surveillance region, and $\lambda_c = 12.5 \times 10^{-6} m^{-2}$ is the average number of false detections in a unit volume, resulting in around 50 false detections in the sensor surveillance area. The function $u(z)$ is a uniform probability density function in the sensor surveillance area.

The other parameters of the implemented GMPHD filter are $T = 10^{-5}, U = 4$ i $J_{max} = 100$, determining the parameters of the pruning and state set estimation algorithms described in [11]. The example of movement and algorithm performance is shown in Fig. 1. In this case the model parameters used for generation of targets, their measurements and clutter measurements are the same as the parameters of the GMPHD filter used for observation. The performance of the algorithm in the case of this many target is obvious. However a more detailed look shows an often error in cardinality. Namely, some of the targets are left out during the estimation process, because the algorithm sets the target states equal to the mean values of the components of the Gaussian PHD function that exceed a preset threshold (typically around 0.5 [11])

The scenario proposed by the authors is interesting because the targets appear at the positions of peaks in the birth intensity function $\gamma_k(x)$, namely these peaks are the mean values of

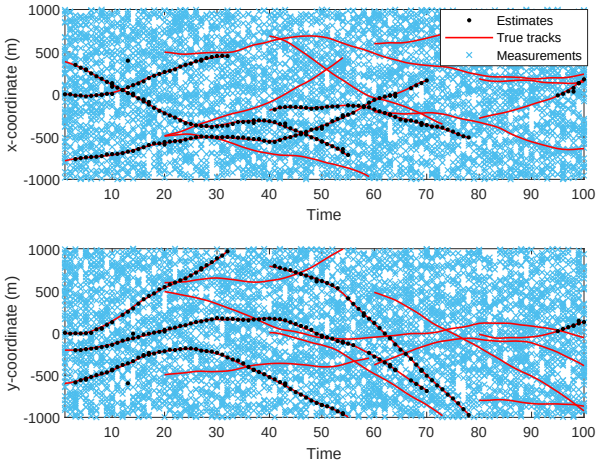


Fig. 2: Simulation results when the target appearance differs from the peak of birth intensity function $\gamma_k(x)$.

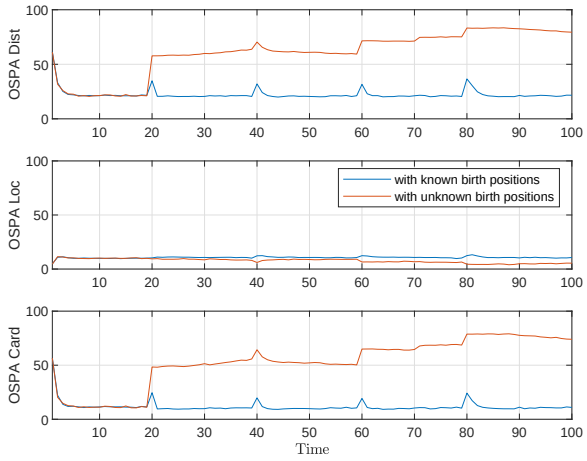


Fig. 3: The average OSPA metric over 1000 MC simulations, when some targets appear at the peaks of birth intensity function, and some of them appear in other locations.

Gaussian components. The constant positions throughout the simulation can represent the airports in a real-life problem in airplane tracking. We tested this algorithm in the case of arbitrary positions of targets' appearance within the sensor surveillance area. The results of the simulation are given in Fig. 2. Four targets are born in the peaks of intensity function and the algorithm tracks them successfully. The other targets are not even recognized. Only in some cases, like the one in the end of the simulation, the algorithm manages to detect and follow the target. The performance of the algorithm, besides the visual inspection, can be perceived using the OSPA metric for comparison of two sets of solutions. We did not consider the target velocity during evaluation, only its position. Due to stochastic noise properties and common errors in cardinality estimation, the OSPA metric during one simulation has many fluctuations and it is not informative. Therefore, we ran 1000 Monte Carlo (MC) simulations and obtained the average OSPA metric. The parameters of the OSPA metric were set to $c = 100$, i $p = 1$. Each MC simulation contained targets born at the same position and the same time as in Fig. 2. The only difference in the trajectories comes from the process noise. We also ran 1000 MC and calculated the average OSPA metric for the scenario shown in Fig. 1, where the targets were born at the same time at the positions of peaks in the intensity function $\gamma_k(x)$. Fig. 3 shows the average OSPA metric together with the localization and cardinality errors for the two mentioned cases. Let us examine the average OSPA metric in the case of known target birth locations. We see the characteristic metric peaks at times $0, 20, 40, ...$, which correspond to the birth of new targets in Fig. 1. The analysis of localization and cardinality errors suggests that those peaks are caused by the cardinality error, because the algorithm needs a few samples before detecting the new target. Let us look at the case of targets being born in arbitrary locations. The cardinality error is the same during the first 20s, because the first three targets appear at the expected positions. Later on the cardinality error increases, since the algorithm does not recognize the unexpected targets. The localization error seems

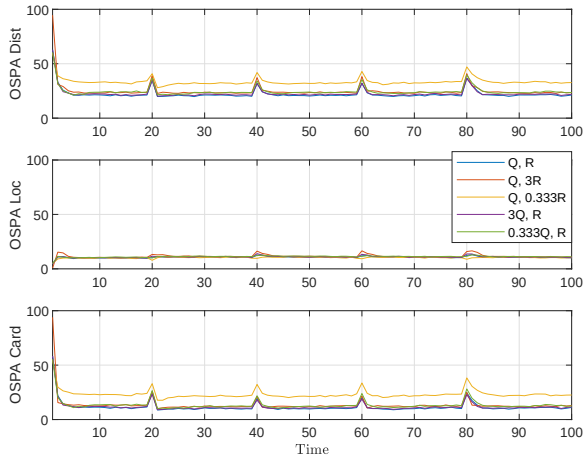


Fig. 4: The average OSPA metric over 1000 MC simulations, when the assumed transition and measurement covariance matrices differ from Q_k i R_k.

to be even smaller when compared to the standard case. The reason for this effect is the small number of detected targets resulting in the small number of elements in the summation (33). The figure shows no significant peaks, suggesting that the algorithm does not make many mistakes. Although Fig. 3 shows the average behavior over 1000 MC simulations, it provides a reliable description of the estimation quality of the simulations from Figs. 1 and 2. The obvious conclusion states that knowing the expected target birth position is of the essential importance to the algorithm performance in the case of a large number of false alarms.

The robustness testing included other parameters as well. For instance, we examined the influence of the underestimated and overestimated process or noise covariance matrix. In other words we ran 1000 MC simulations for four different cases in which the trajectories were generated using the exact matrices Q_k and R_k, while the matrices used by the algorithm

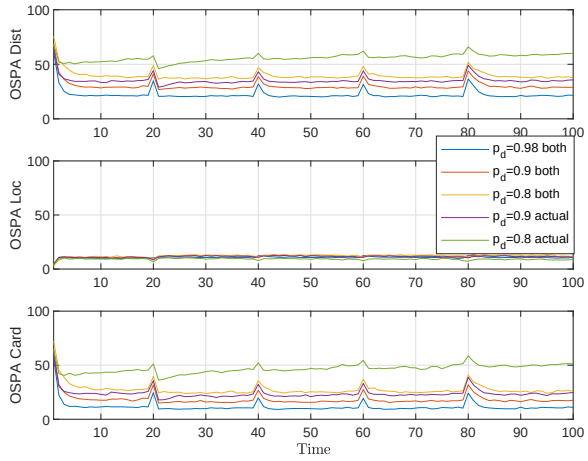


Fig. 5: The average OSPA metric over 1000 MC simulations for different values of probability of detection; *both* means that the assumed and exact probabilities are the same, while *actual* implies that the assumed probability is 0.98, and the exact is smaller.

were changed within some reasonable boundaries. The average OSPA metric with the appropriate legend of covariance matrices is shown in Fig. 4. The scaling of covariance matrix in the range of 0.333 to 3 (the variance is three times smaller or bigger) caused no significant deviations in the OSPA metric. A slight indication of a greater degradation can be observed only for the three times underestimated R_k. This implies that the algorithm is not too sensitive to minor variance changes.

Let us note that the probability of detection $p_{D,k} = 0.98$ is quite large. What if this probability is smaller or the algorithm assumes this large probability in the case when the exact probability is in fact smaller? Again, we ran 1000 MC simulations when the exact probability of detection is equal to the assumed one. We tested two cases: $p_{D,k} = 0.9$ and $p_{D,k} = 0.8$. Next, we analyzed the case when the algorithm assumes high probability of detection $p_{D,k} = 0.98$, and it is actually $p_{D,k} = 0.9$ or $p_{D,k} = 0.8$. All the results are shown in Fig. 5. A degradation in performance quality is expected for the decreasing probability of detection. However, if the algorithm is aware of this fact, the metric increases but the performance is still satisfactory. On the other hand, if a decrease in probability of detection is unknown to the algorithm, there is a significant degradation in performance. This is confirmed by the metric in Fig. 5, when the exact probability of detection is equal to 0.8. Poor filter robustness in terms of cardinality once again shows the need for an improved version of the filter with better estimation of set cardinality, which leads us to the CPHD filter [10].

IV. Conclusion

The paper examines the robustness of the PHD filter in terms of an error in a prior filter parameter estimations. The standard example for RFS-based filter testing is explored. This example is significant because it assumes a large density of false alarms. We obtained satisfactory robustness in the case of changes in process and measurement noise variances. It also allows the change in the probability of detection up to a certain level. The significant drawback of the filter is poor detection in the case when the approximate target birth location is not known in advance. The results of OSPA metrics for error estimation suggest that the biggest performance degradation is caused by a poor set cardinality estimation, because the algorithm is unable to detect the target. Once the targets are detected, the localization error remains small enough, in all of our examples. Future work on performance estimation would include comparative analysis with standard algorithms for target tracking, robustness testing in terms of changes in transition and observation models, etc.

Acknowledgment

This research was partially supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia, Contracts No. TR32038 and III42007.

References

[1] B. Kovacevic and Z. Durovic, *Fundamentals of Stochastic Signals, Systems and Estimation Theory: With Worked Examples*. Springer Publishing Company, Incorporated, 2nd ed. ed., 2008.

[2] R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems," *Transactions of the ASME – Journal of Basic Engineering*, no. 82 (Series D), pp. 35–45, 1960.

[3] S. Blackman, "Multiple target tracking with radar applications," *Dedham, MA, Artech House, Inc., 1986, 463 p.*, vol. -1, 01 1986.

[4] Y. Bar-Shalom and T. E. Fortmann, *Tracking and data association / Yaakov Bar-Shalom, Thomas E. Fortmann*. Academic Press Boston, 1988.

[5] D. Daley and D. Vere-Jones, "An introduction to the theory of point processes. vol. i: Elementary theory and methods. 2nd ed," vol. Vol. 1, 01 2003.

[6] R. P. S. Mahler, *Statistical Multisource-Multitarget Information Fusion*. Norwood, MA, USA: Artech House, Inc., 2007.

[7] R. Mahler, ""statistics 101" for multisensor, multitarget data fusion," *IEEE Aerospace and Electronic Systems Magazine*, vol. 19, pp. 53–64, 2004.

[8] R. P. S. Mahler, ""statistics 102" for multisource-multitarget detection and tracking," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, pp. 376–389, 2013.

[9] R. Mahler, "Multitarget bayes filtering via first-order multitarget moments," *Aerospace and Electronic Systems, IEEE Transactions on*, vol. 39, pp. 1152 – 1178, 11 2003.

[10] R. Mahler, "Phd filters of higher order in target number," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 43, 2007.

[11] B.-N. Vo and W.-K. Ma, "The gaussian mixture probability hypothesis density filter," *Signal Processing, IEEE Transactions on*, vol. 54, pp. 4091 – 4104, 12 2006.

[12] B. Vo, B. VO, and D. Clark, *Bayesian Multiple Target Filtering Using Random Finite Sets*, pp. 75–126. 05 2016.

[13] D. Schuhmacher, B.-T. Vo, and B.-N. Vo, "A consistent metric for performance evaluation of multi-object filters," *Trans. Sig. Proc.*, vol. 56, pp. 3447–3457, Aug. 2008.

[14] J. Kingman, *Poisson Processes*. Oxford Studies in Probability, Clarendon Press, 1992.

[15] A. García-Fernández and B.-N. Vo, "Derivation of the phd and cphd filters based on direct kullback–leibler divergence minimization," *Signal Processing, IEEE Transactions on*, vol. 63, pp. 5812–5820, 11 2015.

Robust Control Design for a 3D Crane System

Anja Buljević, Miloš Miletić, Aleksandra Mitrović, Mirna N. Kapetina, Milan R. Rapaić

***Abstract*—This paper presents an example of controlling complex electromehanical system of 3D crane. Parameters of simplified 3D crane model are estimated using Recursive Least Squares (RLS) algorithm. After that two different control algorithms are applied. The first is fractional-order PI controller which parameters are tuned by Symmetrical Optimum Method, and second one is optimal PI controller which parameters are obtained to satisfy certain measures of the system performance and robustness. Performance of both algorithm are illustrated on the model and real system.**

Keywords: 3D Crane, RLS algorithm, PI control, FOPI control, PSO algorithm, Symmetrical Optimum Method (SOM), Robust control.

I. Introduction

Cranes are used in many industries, factories and warehouses to transfer heavy loads from one place to another. Although, they are generally controlled by human operators, automated systems are able to obtain more precise control. In many papers 3D crane models are considered and various control techniques solving the position control to track desired trajectories and to reduce the load swing have been widely studied in the literature [1], [2], [3].

This paper concerns the time control of 3D crane, based on a simple linear mathematical model. The idea is to realize online parameters estimation for laboratory model of 3D crane and controlling that crane by using different controllers. Recursive least squares algorithm is used for parameters estimation. Depending on what we want to achieve, two types of proportional–integral(PI) controller are used for controlling the 3D crane. Parameters that should be obtained for both controllers are K_p(proportional gain) and K_i(integral gain). Main task is to determine parameters that satisfies some performance and robustness measures of the system. Regulator must follow the reference, eliminate the disturbance and decrease noise impact. Also, regulator must be resistant on the process variations, ie it should be as robust as possible in that sense. Of course, it is almost impossible to satisfy all requirements in the same time, so it is necessary to find some compromise between them.

After Introduction, the paper is organized in the following manner: description of the 3D crane system, process model and parameter estimation procedure are explained in Section II, optimal design method for conventional PI controllers was introduced in Section III, while symmetrical optimum procedure for fractional order PI controllers is introduced in Section IV. The two controllers were compared in Section V, and the concluding remarks are given in the final Section VI.

A. Buljević (anjabuljevic@uns.ac.rs), M. Miletić (m.miletic@uns.ac.rs), A. Mitrović(aleksandra.mitrovic@uns.ac.rs), M. N. Kapetina (mirna.kapetina@uns.ac.rs), M. R. Rapaić(rapaja@uns.ac.rs) University of Novi Sad, Faculty of Technical Sciences, Department of Computing and Control Engineering, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia.

II. Online parameters estimation

Inteco® 3DCrane [4] is laboratory model of 3D crane and it is controlled from a PC. The 3D crane is nonlinear electromechanical system having a complex dynamic behaviour and creating challenging control problems. Its hardware and software can be easily mounted and installed in a laboratory.

3D crane (Fig. 1) consists of a payload hanging on pendulum which length can be changed. The payload can move freely in 3 dimensions (x, y and z-axis). Therefore, 3D crane is driven by 3 DC motors. The payload is lifted and lowered in the z direction. The rail and the cart are capable of horizontal motion along the rail in x direction. The cart is capable of horizontal motion along the rail in the y direction.

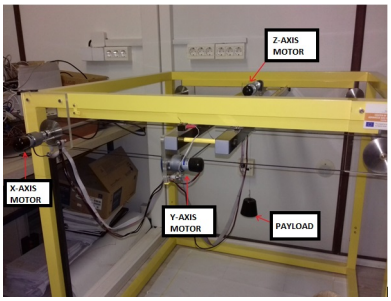


Fig. 1. 3DCrane

Mathematical model of 3D crane has ten differential equations and because of this complexity, it is necessary to linearize the system. For design controllers is enough to use simple linear models which assumed that the oscillations of the payload are so small that the trigonometric relations may be neglected. In this way, we get the transfer function from digital control signal in MATLAB to the measured position. Also, in order to obtain a simpler transfer function, the behavior of the system is observed individually in each axis. The input of transfer function is digital control signal in interval [-1,1](which later transforms by electronics and motors into a pulling force) and output is distance in meters. The simple linear model of the crane dynamics in the X-axis and Y-axis are assumed:

$$G_x(s) = \frac{K_x}{s(T_x s + 1)}$$
$$G_y(s) = \frac{K_y}{s(T_y s + 1)} \quad (1)$$

Generally, linear model can be described by

$$G(s) = \frac{K}{s(Ts+1)} = \frac{K}{Ts(s+\frac{1}{T})} = \frac{K_n}{s(s+p)} \tag{2}$$

where K is K_x or K_y, T is T_x or T_y. Unknown parameters that have been determined are gain K and time constant T. In this paper, we performed online estimation of unknown parameters.

Online estimation algorithm estimates parameters when new data is available during the operation of the physical system. Online parameter estimation is typically performed using a recursive algorithm [5], i.e. model parameters values are determined by recursive least squares algorithm. The results of estimation are compared with parameters obtained from the MATLAB Toolbox.

Recursive least squares algorithm

Recursive least squares algorithm [6], [7], [8] represents mathematical procedure for solving a problems of parameter identification (estimation) in the adopted structure of the model. The appropriate model structure can be written in the following form:

$$y(t) = \varphi_1\theta_1 + \varphi_2\theta_2 + ... + \varphi_n\theta_n \tag{3}$$

$$y(t) = \varphi(t)^T\theta \tag{4}$$

where $y(t)$ is measured output value, φ is vector of known values and θ is vector of unknown values. $y(t)$ is called regressand, elements of vector $\varphi(t)$ are called regressors and variable t represents time.

The basic problem is finding the estimated values of the parameters from the measurements $y(1)$, $y(2)$,..., $y(N)$ and $\varphi(1)$, $\varphi(2)$,...,$\varphi(N)$ where N is number of measurements. A linear equation system can be formed:

$$\begin{aligned} y(1) &= \varphi^T(1)\theta \\ y(2) &= \varphi^T(2)\theta \\ &\vdots \\ y(N) &= \varphi^T(N)\theta \end{aligned} \tag{5}$$

The matrix form is obtained:

$$y = \Phi\theta \tag{6}$$

The linear equation system is solvable by θ if the number of measurement N is equal to the number of variables n. In that case, Φ is square matrix and if it is nonsingular matrix, values of parameter θ can be determined.

However, in practice, the number of measurements is greater than the number of variables ($N > n$) because of the existing of a disturbance and mistakes in model. That means that the linear equation system becomes predetermined and the solution does not exist.

This problem can be solved as a problem of the smallest squares:

$$J = \sum_{k=1}^{N}(y_k - \hat{y}_k)^2 = \sum_{k=1}^{N}(y_k - \sum_{k=1}^{n}\varphi_{k,i}\hat{\theta}_i)^2 \tag{7}$$

In order to minimize criteria function (7) it is necessary to equalize the first derivate with zero:

$$\frac{\partial J}{\partial \hat{\theta}_j} = 0 \quad (\forall j \in 1...n) \tag{8}$$

After that, the following equation was obtained:

$$\Phi^T\Phi\hat{\theta} = \Phi^T y \tag{9}$$

The equation (9) in the literature is called *"normal equation"* [5]. Recursive least squares algorithm is applied to the following mathematical model:

$$\begin{aligned} G(s) = \frac{Y(s)}{U(s)} &= \frac{K}{s(Ts+1)} \\ s^2TY &= -sY + KU \\ \ddot{y}T &= -\dot{y} + Ku \\ \ddot{y} &= -\frac{1}{T}\dot{y} + \frac{K}{T}u \end{aligned} \tag{10}$$

The last equation from (10) represents the general form of the transfer function of all axes of the 3D crane and linear model can be formed:

$$\ddot{y}_f = \begin{bmatrix} -\frac{1}{T} & \frac{K}{T} \end{bmatrix} \begin{bmatrix} \dot{y}_f \\ u_f \end{bmatrix} \tag{11}$$

We can adopt the following relations: $p_1 = -\frac{1}{T}$ and $p_2 = \frac{K}{T}$ and get the mathematical model linear by parameters:

$$\ddot{y}_f = \begin{bmatrix} p_1 & p_2 \end{bmatrix} \begin{bmatrix} \dot{y}_f \\ u_f \end{bmatrix} \tag{12}$$

where $\ddot{y}_f$ is filtered output, $\begin{bmatrix} p_1 & p_2 \end{bmatrix}$ is model linear by parameters (matrix of unknown parameters) and $\begin{bmatrix} \dot{y}_f \\ u_f \end{bmatrix}$ is matrix of known parameters. A transfer function of a filter that was used is $H(s) = \frac{1}{(T_f s+1)^2}$, where T_f is time constant. If noise exists in system, it will have a big impact on parameters, so it is necessary to do filtering. T_f is chosen so that it doesn't cut anything from input signal. In order to obtain unknown values of gain and time constant, it is necessary to perform the following transformation:

$$T = -\frac{1}{p_1} \tag{13}$$

$$K = p_2 T \tag{14}$$

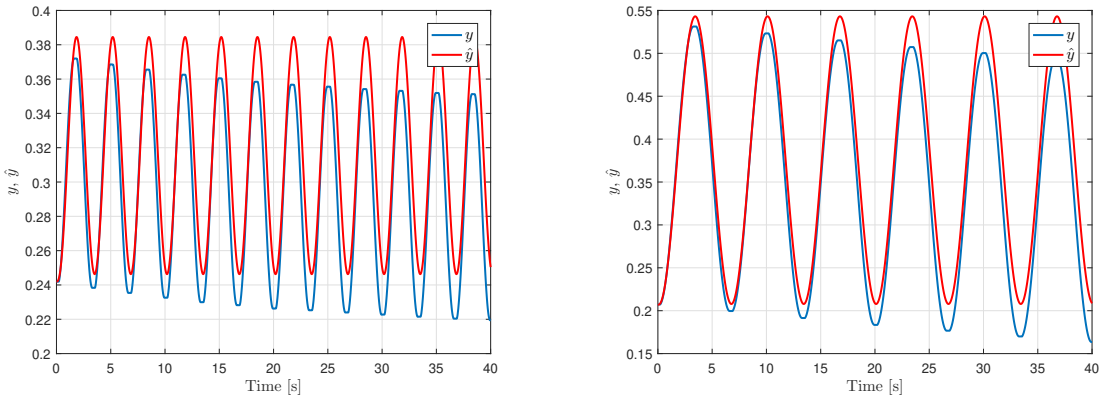
Criteria function that represents deviation of the estimated model from the measurement was minimize in order to obtain unknown values of parameters:

$$J_r = \hat{\ddot{y}}_f - \ddot{y}_f \tag{15}$$

Obtained values of unknown parameters are given in Table I.

The form of input signal was $u(t) = 0.2\sin(\pi t + 0.23) + 0.3\sin(\pi t)$. Comparative view of the crane and model response are shown at Fig. 2a and 2b.

It can be noticed from Fig. 2a and 2b that due to the nonlinearity and imperfection of the system, there are disagreements between the linearised model and the real system. However, it will be shown that linearised model will be sufficient for robust control design.



(a) x-axis (b) y-axis

Fig. 2. Comparison of the actual and estimated output

Table I
Numerical values of unknown estimated parameters

K_x	T_x	K_y	T_y
0.3075	0.1906	0.3176	0.1066

III. Optimal PI Control Design

PI controller [9] is given in form

$$G_c(s) = K_p + \frac{K_i}{s}.$$

The focus of this section is on the system response on the disturbance, which is slow changing variable in time domain, so we can assume that it is constant. In that case, its transfer function is $D(s) = \frac{k}{s}$. Disturbance elimination is the main requirement in most process control loops because it relates to system behaviour in the steady state. Requirements are mathematically formulated by some optimality criterion and accompanying constraints. Before defining measures of system performance and robustness it is necessary to define some characteristic transfer functions. Transfer functions that are used are known as "gang of four" [10], [11]. One of the most common system performance indicator is integral of absolute error(IAE) [10]. Secondly, very similar is integral of error(IE) [10]. Measure of the system performance called Q [10], [11] limits resonant peak of the frequency characteristic. By the limiting Q it is possible to get acceptable values of the IAE. Q is limited by 1.01 [10]. That value was obtained experimentally through a large number of simulations. This measure is defined by

$$Q = \frac{\max_{\omega\geq 0} Q_v(j\omega)}{Q_v(0)} = \max_{\omega\geq 0} \left| \frac{k_i \frac{G_p(s)}{j\omega}}{1 + G_c(j\omega)G_p(j\omega)} \right|_{\omega=\omega_g}$$

Also, it is necessary to define some systems robustness measures. In that purpose most often used is maximal sensitivity M_s [10], defined by

$$M_s = \max_{\omega\geq 0} \left| \frac{1}{1 + G_c(j\omega)G_p(j\omega)} \right|.$$

Measure M_s represents minimal inverse distance of the Nyquist curve from the critical point(-1,0j). Often measure M_s is taken as the main systems robustness criterion. For example, system can have big gain and phase margin, but small maximal sensitivity, so it is easy to destabilize it. Best results are gained if maximal sensitivity is between 1.7 and 2 [10]. Maximal complementary sensitivity M_p [11] is another one systems robustness measure which will be explained. Application of this measure relates to examination of systems robustness in the medium-frequency band. It is used to get bigger damping or bigger phase margin without getting system response slower. The main application of this measure is in unstable processes with integral component. Maximal complementary sensitivity is defined by

$$M_p = \max_{\omega\geq 0} \left| \frac{G_c(j\omega)G_p(j\omega)}{1 + G_c(j\omega)G_p(j\omega)} \right|_{\omega=\omega_p}$$

Next step is to optimize the parameters of PI controller for each axis separately, ie mutual impact between axes is ignored in the optimization. In first section Optimality criterion is the integral of absolute error (IAE) while constraints are Q(limit on resonant peak of frequency characteristic), maximal sensitivity and maximal complementary sensitivity. Optimization was done with the Particle Swarm Optimization (PSO) algorithm [12], [13], [14].

Minimal and maximal allowed value for the M_s and M_p are shown in Table II. Maximal allowed value for Q is 1.01 as it was mentioned earlier. At both cases, the recommended values for the M_s were not used because in that case system response is too fast and it produces intensive swinging of the payload. The regulator has a task to eliminate disturbance. Disturbance is step signal with amplitude 0.3. After optimization, parameters are obtained for each axis. Parameters and corresponding values of the constraint are shown in Table III. This controller will be called *Optimal PI*. Model responses will be shown in Section V where they will be compared with FOPI controller.

Table II
Optimization constraints

	Q^{max}	M_s^{min}	M_s^{max}	M_p^{min}	M_p^{max}
x	1.01	1.3	1.5	1.2	1.4
y	1.01	1.4	1.6	1.2	1.4

Table III
Obtained regulator parameters with corresponding constraints

	k_p	k_i	Q	M_s	M_p
x	25.32	31.1	1	1.5	1.31
y	31.45	28	1	1.6	1.35

IV. Fractional Order PI Control

The PID controller is the most common solution to practical control problems due to its robust performance and simplicity to get the tuning parameters. However, it is sometimes necessary to provide better controller performance, flexibility and more adequate methods for tuning controllers. As a result, fractional-order(FO) controllers were proposed by Podlubny [15] as a generalization of the PID controller with integrator of real order α and differentiator of real order η. FOPI controller [16] is used for controlling 3D crane and it can be described using this formula:

$$G_c(s) = K_p(1 + \frac{1}{(T_i s)^\alpha}), 0 < \alpha < 1. \tag{16}$$

Symmetrical Optimum Method

After parameters identification is done, it makes sense to start controlling the 3D crane. Linear model of the crane dynamics is generally described by (2). There are many methods for tuning parameters for controllers, but we chose to set parameters using Symmetrical Optimum Method(SOM) as it is mentioned in section I. This concept was presented by Kessler [17] in 1958. Later, in 1995. it was modified by Voda and Landau whose modified method ensures obtaining a maximum phase margin in closed loop system [18]. SOM has a few advantages over other methods in aspect of phase and gain margins and sensitivity of system.

In this paper, we use SOM which is explained in [19]. The main principle is based on designing the regulator in order to meet requirements, so that phase margin is as close as it is possible to 37°. That phase margin should be obtained at frequency $\omega_{max} \equiv \omega_{gc}$, where ω_{max} is the frequency at which maximum margin can be reached and ω_{gc} is the gain crossover frequency. Further more, $|W_{sp}| = 1$ should be obtained in the largest possible frequency range, especially in low frequencies. This method also reduces disturbance effects.

The controller design is mainly based on shaping asymptotic gain and choosing the slope of the segment crossing the frequency axis. The gain diagram must maintain this slope in a wide frequency interval around the crossover frequency. The phase margin is constant in that interval and robust stability is guaranteed even for high gain variations.

$$G_{bode}(s) = (\frac{\omega_{gc}}{s})^\alpha.$$

If FO controllers are used, slopes can be $\alpha \cdot 20dB/dec$ (α can be non-integer number, e.g. $\alpha = 0.5 \Rightarrow slope = -10dB/dec$) unlike classic controllers where α has to be integer number. Basic idea for tuning FO controllers parameters using SOM is to choose gain crossover frequency so that phase is maximized. Phase margin is maximum if

$$\frac{d}{d\omega} arg W(j\omega) = 0.$$

If it is not possible to achieve this, then it is possible to analytically solve non-linear minimization function to obtain the tuning formulas for controllers parameters. It is possible to make compromise between dynamic performance and robust stability using α (16). Solution to this problem is given in [19]:

$$C = 1 + \cos(0.5\pi\alpha), S = \sin(0.5\pi\alpha)$$
$$\theta_a = \tan(0.5\pi - 0.5\pi\alpha), \theta_b = \tan(PM_\alpha)$$
$$a = \sqrt{\frac{C(1 + \theta_a\theta_b) - S(\theta_a - \theta_b)}{C(\theta_a - \theta_b) + S(1 + \theta_a\theta_b)}}.$$

Formula for a comes from minimization of a non-linear function that depends on determining the frequency for which the phase is maximum.

Tuning parameters that are obtained are:

$$T_i = a^2 T$$
$$K_p = \frac{1}{KT_i}\sqrt{\frac{1 + a^4}{2a^4 C}}$$
$$PM_\alpha = \arctan\frac{S}{C} - \arctan(a^{-2}) + 0.5\pi(1 - \alpha).$$

PM_α is the maximum achieved phase margin, T_i and K_p are the parameters we were looking for. Obtained parameters using SOM for FOPI controller with $\alpha = 0.7$ by axis are:

1) x-axis: $K_{px} = 6.0389, K_{ix} = 11.9823$
2) y-axis: $K_{py} = 4.9848, K_{iy} = 9.7343$,

and their responses, also by axes, are given in Section V where they are compared with Optimal PI.

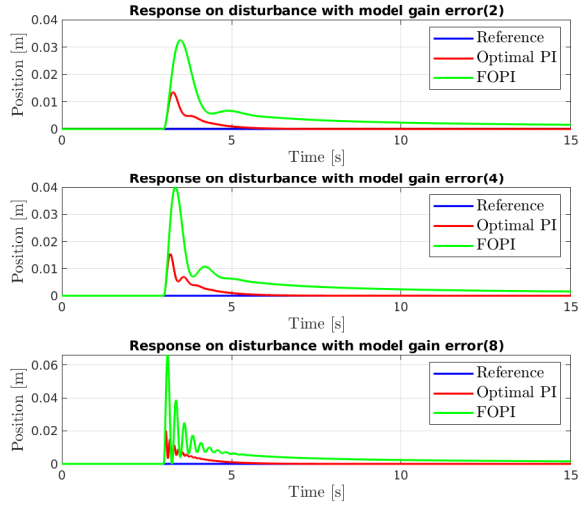
V. Simulation results

In this section, simulations are done on model and real system and their results are shown below.

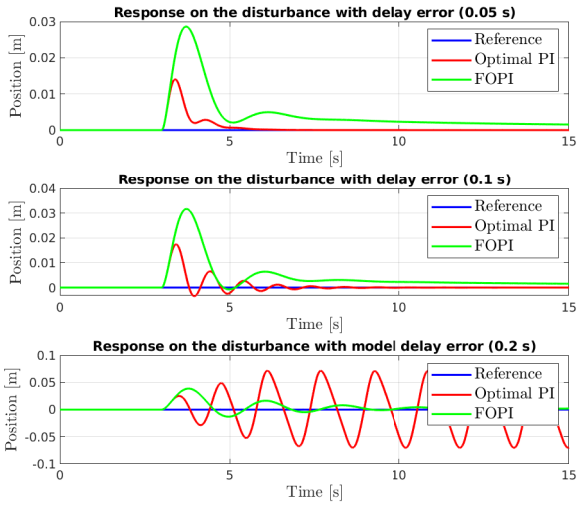
Model results

In this subsection, model responses will be compared when it is controlled by FOPI controller and PI controller described in Section III. Responses will be presented when there is constant disturbance with amplitude 0.3. To test controllers robustness, we will include some errors in modelling (gain error, delay, time constant error).

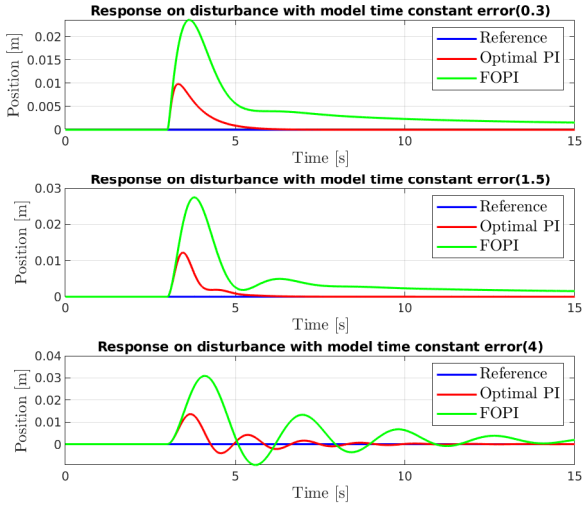
Results will be presented only for x-axis because similar results are obtained for y-axis.



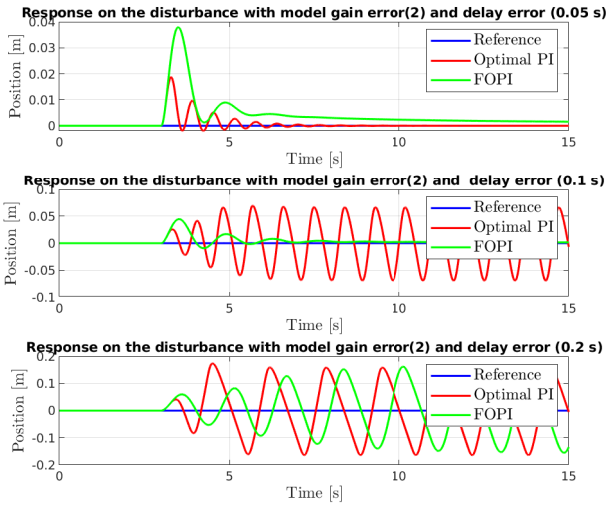
(a) Model responses with gain error



(b) Model responses with delay included in model



(c) Model responses when time constant is not estimated correctly



(d) Model responses with gain error and delay included

Fig. 3. Comparison of model results obtained with Optimal PI and FOPI

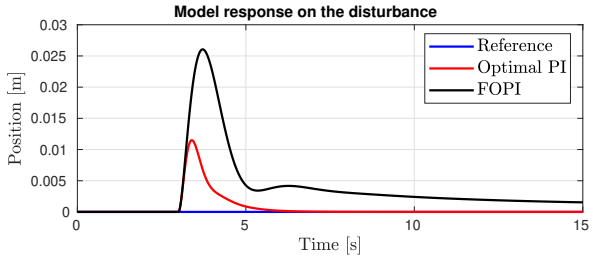


Fig. 4. Model responses on disturbance which starts after 3 (s) with amplitude 0.3

Fig. 4 shows model responses when only disturbance is included. It is obvious that Optimal PI faster eliminates disturbance impact.

Fig. 3a shows responses if there is gain error. Both controllers keep system stable and successfully eliminate disturbance because gain margin is big enough what can be seen at Fig. 5.

Fig. 3b shows responses if there is delay in system model. It is obvious that FOPI controller is more robust in this aspect because it tolerates more delays. Optimal controller can tolerate delay by some time, and after that system becomes unstable. Their delay margins are shown at Fig. 5.

Fig. 3c shows responses when time constant is not estimated correctly. Both controllers are very robust, but Optimal PI has better performance.

Fig. 3d shows responses when there is combined gain error and delay in system model. It is obvious that FOPI has better performance.

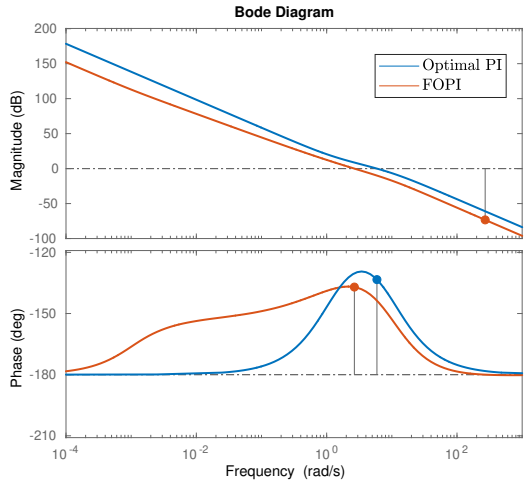
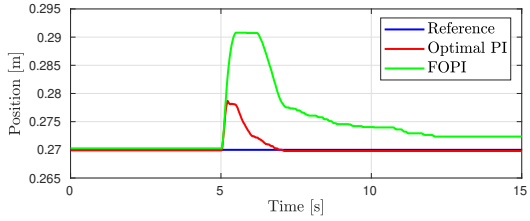
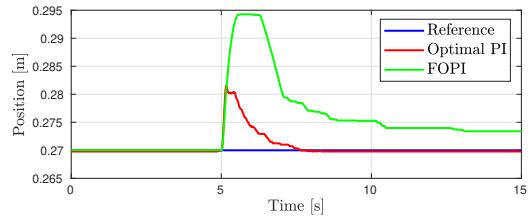


Fig. 5. Bode diagrams for model controlled by FOPI and Optimal PI. FOPI: gain margin is 73.3 (dB), phase margin is 43 (deg), delay margin is 0.282 (s); Optimal PI: phase margin is 46.6 (deg), delay margin is 0.138 (s).



(a) x-axis



(b) y-axis

Fig. 6. Crane responses on disturbance which starts after 5 (s) with amplitude 0.3

Real system results

Obtained parameters for Optimal PI and FOPI regulator are used on real 3D crane. Fig. 6a and 6b show system responses. Disturbance is simulated with step signal with amplitude 0.3, which is added to control signal.

VI. Conclusion

This paper presents one approach of modelling the 3D crane. Original model of 3D crane is very complex, so linearization was done to make model simpler. Obtained model is second-order system. Online parameters estimation was done using recursive least square method. Comparative analysis shows that obtained results for x and y-axis are acceptable. Fractional-order PI regulator was used for controlling the crane and its parameters were tuned by Symmetrical Optimum Method. This method reduces disturbance effects and can tolerate more delays which is proved on the real system. Optimal PI controller was designed to satisfy certain measures of the system performance and robustness. Controller is applied at the real 3D crane and behaviour of system satisfies criteria. Finally, both controllers were compared. It is proved that Optimal PI has better performances, but FOPI is much better if there is bigger delay in system. Although there are reasonable deviations in system model, both controllers fulfill their tasks, so we can conclude that they are robust enough.

VII. Acknowledgment

This work partially supported by Serbian Ministry of Education and Science, grant no. TR32018 (M.N.K., M.R.R.) and grant no. TR33013 (M.R.R.)

References

[1] D. Chwa, "Sliding mode control-based robust finite-time anti-sway tracking control of 3-d overhead cranes," *IEEE Transactionson Industrial Electronics,64,6775–6784*, 2017.

[2] S. Hussein, R. Ghazali, H. Jaafar, and C. Soon, "Analysis of 3d gantry crane system by pid and vsc for positioning trolley and oscillation reduction," *Journal of Telecommunication, Electronic and Computer Engineering,8(7),139–143*, 2016.

[3] X. Wu and X.He, "Partial feedback linearization control for 3-d under actuated overhead crane systems," *ISATransactions,65,361–370.*, 2016.

[4] "3d crane user's manual." http://control.put.poznan.pl/old/sites/default/files/3DCrane.pdf/.

[5] M. Kapetina, *Adaptivna estimacija parametara sistema opisanih iracionalnim funkcijama prenosa*. PhD thesis, Faculty of Technical Sciences, Novi Sad, 2017.

[6] K. J.Astrom, "Adaptive control, pages 437–450," *Springer Berlin Heidelberg, Berlin, Heidelberg*, 1991.

[7] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the ASME–Journal of Basic Engineering, 82(Series D):35–45*, 1960.

[8] P. A. Ioannou and J. Sun, "Robust adaptive control," 2012.

[9] K. Åström and T. Hägglund, "Pid controllers: Theory, design and tuning," *Instrument Society of America, N. Carolina, USA*, 1995.

[10] B. Jakovljević, *Optimalno i suboptimalno podešavanja parametara robusnih linearnih regulatora necelog reda*. PhD thesis, Faculty of Technical Sciences, Novi Sad, 2015.

[11] T. Šekara and M. Mataušek, "Optimization of pid controller based on maximization of the proportional gain under constraints on robustness and sensitivity to measurement noise," *IEEE Transactions on Automatic Control, vol. 54, issue.1, pp.184-189*, 2009.

[12] Željko Kanović, M. R. Rapaić, and Z. D. Jeličić, "Generalized particle swarm optimization algorithm - theoretical and empirical analysis with application in fault detection," *Applied Mathematics and Computation, 217(24):10175–10186*, 2011.

[13] M. Rapaić and Željko Kanović, "Time-varying pso-convergence analysis, convergence related parameterization and new parameter adjustment schemes," *Information Processing Letters, vol. 109, issue 11,pp. 548-552*, 2009.

[14] J. Kennedy and R. Eberhart, "Particle swarm optimization," *Proceedings of IEEE International Conference on Neural Networks, vol. 4, pp. 1942-1948, Perth, Australia*, 1995.

[15] I. Podlubny, "Fractional order systems and $pi^{\alpha}d^{\eta}$ controllers.," *IEEE Transactions on Automatic Control, 44(1):208–214*, 1999.

[16] B. Jakovljević, T. Šekara, M. Rapaić, and Z. Jeličić, "On the distributed order pid controller," *International Journal of Electronics and Communications, vol. 79, pp. 94-101*, 2017.

[17] Kesler, "Regelungstetechnik.," *6, 395–400 and 432–436*, 1958.

[18] I. D. Landau and A. Voda, "A method for the auto-calibration of pid controllers.," *Automatica 31(1), 41-53*, 1995.

[19] G. Maione and P. Lino, "New tuning rules for fractional pi^{α} controllers," *P. Nonlinear Dyn (2007) 49: 251. https://doi.org/10.1007/s11071-006-9125-x*, 2006.

Incident simulator for ADMS performance testing

Nedeljko Stojaković, Marina Stanojević, Darko Čapko, Tatjana Grbić

***Abstract* — The Advanced Distribution Management System (ADMS) is an Industrial Control System specifically designed and developed for the Smart Grid industry. One of the important components of this system is Outage Management Service (OMS) which is responsible for handling unexpected outages or planned maintenance. The most critical time for ADMS, when its response and durability are of exceptional importance, is when a storm occurs. In this paper, we analyze actual storm data, build storm model and propose a simulation of a storm. It is imagined that the utilities could use such simulator in order to test and verify their ADMS according to the real storm conditions.**

***Keywords* — Advanced Distribution Management System; Smart Grid; storm modeling; simulation;**

I. Introduction

THE electric power system is one of the biggest and most important technical system nowadays. Its main goal is to provide customers with electric energy with minimal downtime. Because of this, many electric utilities spend significant amount of money in developing systems such as Advanced Distribution Management System (ADMS). ADMS is used for monitoring, controlling and optimizing electric power distribution system with millions of customers. Outages in electric power system are inevitable, but it is very important to react quickly in order to eliminate their consequences. Component in ADMS that is in charge of outages is called Outage Management System (OMS).

When it comes to closing outage, there is a procedure with a few steps that needs to be executed. Firstly, the dispatcher in control room takes over the incident and makes the Work Order plan for it. In that plan, he needs to specify which field crew is going to be responsible for resolving the incident. The best solution is to find crew which is the nearest to the location of the incident. After this, a crew is notified and they need time to find actual location of the incident and to fix it. Only when crew finishes with all the work, incident can be closed. For ADMS the most critical time is when the storm occurs because then several outages which need to be handled as soon as possible, could be created at the same time. Utility usually needs to send out many crews to restore services during a storm. In this paper, we analyze actual storm data to provide statistical distributions. A storm is modeled as a group of incidents that appear at the same time. As there are several incident types, we analyze them separately. For the analysis, times between arrivals of the same incident type are used and as a result distribution for those times are obtained. After that, the response of the system to the incident, as well as the assessment of the time system to close the incident, is being examined. These distributions can be used in incident generation and later in testing ADMS.

Authors in [1] investigate about impact which natural disasters have on electric power grid and review progress of research field towards methods and tools of forecasting natural disasters, pre-storm operations, and restoration models. One example of model used to predict power interruptions is developed in [2]. It is based on common weather conditions and refined using statistical and deterministic simulations of the model. In [3] automated analysis of distribution systems are presented using cloud-based architecture. Several storm models are evaluated in [4] and authors suggested framework that can be used for predicting outages for emergency-preparedness functions. Literature survey on models and algorithms used to predict hazards and on restoration strategies are presented in [5].

In the next section outage types are described. Section III gives a description of the problem discussed in this paper. Section IV contains information about system architecture. Also, in this section, the statistical distributions of the given input as the results of the research is presented. Section V contains information about data used as an input in storm modeling. And finally, these distributions are used in testing of the system which is described in Section VI.

II. Outages and Incidents in Power Grid

In OMS, outages are modeled as incidents. There are several possible ways to detect outages which lead to several different outage types: Phone Call Created, Supervisory Control And Data Acquisition (SCADA) Created, Work Order and User Created.

A. Phone Call Created incident type

Phone call received from customers is one type of the incident. In this case, it is still unknown what caused outage, where its root is and how many customers are affected. Dispatcher in the control room models incident from the

Nedeljko Stojaković is with the Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21101 Novi Sad, Serbia (phone: 381-66-9382181; e-mail: nedeljko.stojakovic@uns.ac.rs).

Marina Stanojević is with Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21101 Novi Sad, Serbia (phone: 381-63-686827, e-mail: marina.stanojevic@uns.ac.rs).

Darko Čapko is with Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21101 Novi Sad, Serbia (phone: 381-21-4852451; e-mail: dcapko@uns.ac.rs).

Tatjana Grbić is with Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21101 Novi Sad, Serbia (phone: 381-21-4852275; e-mail: tatjana@uns.ac.rs).

information he received from customer and marks it on the network view. If more than one call came from the same area, incidents are grouped and observed as one. The most important in this step is to find outage root.

B. SCADA Created incident type

SCADA incident occurs when the voltage drop is detected on a breaker. ADMS records the event through the SCADA system. With this type, the exact location of the fault is known.

C. Work Order incident type

Work Order incident type is used for modeling planned maintenance. In this case dispatcher knows exact position of the maintenance and models incident to match that position.

D. User Created incident type

Generic type of incident, and it is used when the actual type is not specified.

III. Problem Definition

Conceptual model of the system is presented in Fig. 1.

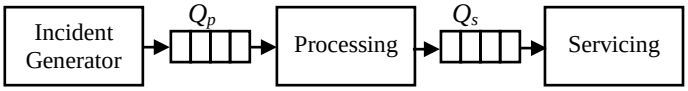


Fig. 1. Conceptual model

The complete process can be treated as a queuing system and multiphase servicing. Waiting is caused by the inability of parallel service in certain service stages.

Incident Generator is used to generate events - incidents at certain times that are determined based on the collected data from the system.

Processing is process from reporting an incident to its acceptance by a dispatcher and assigning to the specific field crew.

Servicing refers to the process from the moment of accepting the incident to its resolution.

The aim of the paper is to simulate the system based on the collected data from the real system. The analysis of the obtained results should determine the number of dispatchers and crews. Criteria for the selection of optimal parameters are minimization of waiting in queues (Q_p and Q_s), utilization of dispatchers and utilization of crews.

IV. System Architecture

In order to obtain the correct results of the simulation, it is necessary to use adequate input data.

Since it analyzes the behavior of the system during incidents in the network, it is first of all necessary to determine the incident events during the simulation. Since data on incidents are stored in the historical database in the ADMS system, the idea is to use this data when estimating the dynamics of generating incidents - input data into the simulation.

In view of this, the process of analyzing the behavior of the system could be divided into 4 phases, as shown in Fig. 2:

1. Extraction incident data from a historical database.
2. Filtering and processing data.
3. Input Modeling - occurrence time of incidents.
4. System simulation and analysis of results.

Fig. 2. System architecture

A. Historical Database

Data related to the incidents that are reported, registered and processed in the ADMS system are stored in a historical database. Data such as moment of occurrence, incident type and location, the time of incident registration, the time of assignment of the incident to the specific team in the field, the time of completion of the incident processing, etc., are kept.

The data recorded in the database can be stored with certain deficiencies and errors, and therefore it is necessary to filter them for further use. For example, SCADA and Phone Call incident could be reported separately while referring to the same outage and thus should be treated as one incident.

B. EFC Process (extraction, filtering, calculations)

The EFC process consists of the following three phases:

1. Data Extraction from the database.
2. Data filtering.
3. Data processing.

During data extraction, raw incident data from the database for the given time period are obtained.

The data filtering process consists of two phases. The first stage identifies information about the same incident (multiple application - idempotence) and eliminates data with possible errors (incident processing time missing or any other significant data). After that, data from the significance of the system behavior analysis (the data are shown in Table I) are extracted from the set of filtered incidents.

TABLE I
DATA OF INTEREST

Symbol	***Data***	***Description***
IT_{ID}	INCIDENT_TYPE_ID	Identification of incident type. e.g. Phone Call, SCADA
t_o	OUTAGE_TIME	Time of occurrence - this represents time when incident happened (e.g. phone call received)
t_c	CREATE_TIME	Time of processing - time when dispatcher confirmed the incident.
t_a	ASSIGNED_TIME	Time of start servicing
t_e	ACTUAL_END_TIME	Close time - when all steps are finished and incident is closed

After data filtering, the data processing phase is next. At this stage, the input data values for each incident are calculated. Input data are:

- The time of occurrence of the incident after the previous incident - the so-called interarrival time – t_d,

- Time of the incident processing (processing time) – t_p,
- Service time of the solving incident – t_s.

These parameters are calculated as follows:

$$t_d(i) = t_o(i) - t_o(i-1), \quad t_o(0) = 0 \tag{1}$$

$$t_p(i) = t_a(i) - t_o(i) \tag{2}$$

$$t_s(i) = t_e(i) - t_a(i), \quad i = 1,2,..., n \tag{3}$$

where i is an ordinal number of incident and n is the total number of incidents.

C. Data modeling

After calculating the parameters given by (1), (2) and (3), it is necessary to model the interarrival times, the processing time and the time of servicing by certain random distributions. The obtained distributions are given as parameters of the simulation model and they are used in the simulation process.

D. Simulation

Simulation model, created in Arena simulation software [6], is presented on Fig. 3. Input blocks represent the generators of incidents divided by type. The next four blocks are Assign modules of Arena Simulation software that are used to define certain attributes and variables in the simulation. Next block presents part of simulation that describes time needed from modeling to incident acceptance.

In the *Servicing* - service time is the time system needs to solve an incident. The last block presents stock which is used for closed incidents that got through the system. Processing and servicing components can have multiple instances.

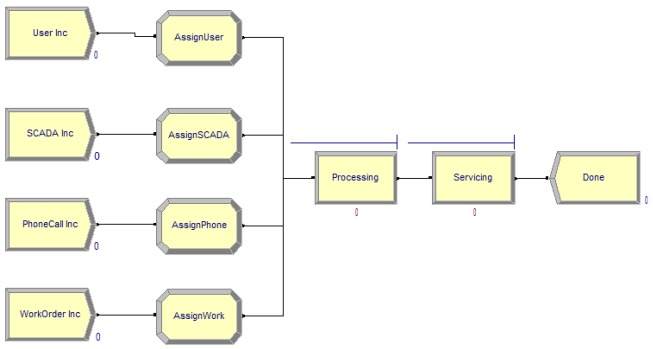


Fig. 3. Simulation model for *Incident Simulator*

V. Simulation Input Data

The incidents on a smaller segment of a typical European network are recorded in period from June 2015 to November 2016. In given period total of 17057 incidents occurred. Table II presents number of incidents per type.

TABLE II
NUMBER OF INCIDENTS PER TYPE

Name	User	SCADA	Phone Call	Work order
Count	1237	4793	8424	2603

After the data filtering, an estimation of the processing time of the servicing time was performed. The distribution and parameters were evaluated using Arena and the following estimates were obtained:

- Processing time - exponential distribution *Exp* (λ=167.214).
- Servicing time - *Gamma* (β=0.397, α=149.546).

In Table III, the months with the most frequent occurrence of incidents by a certain type are given. The idea is to estimate the statistical distribution in the most frequent months for certain incidents and to maximize the system's response time during the simulation in one month. In this way, an assessment could be made of the maximum requirements of the dispatchers and the field crews.

TABLE III
MONTHS IN WHICH CERTAIN INCIDENTS WERE MOST NUMEROUS

Incident Name	***Month/Year***	***Count***	***Distribution***
User	Nov/2016	106	*Exp* (405.32)
SCADA	Dec/2015	429	*Exp* (104.22)
Phone Call	Jan/2016	673	*Exp* (65.30)
Work Order	Dec/2015	127	*Exp* (292.90)

Statistical distribution for processing time, servicing time and incident arrivals (Table III and IV) was estimated using Arena software. Corresponding p-value for Chi-Square test is less than 0.005 and for Kolmogorov-Smirnov test less than 0.01 which means that the distributions are acceptable.

TABLE IV
NUMBER OF INCIDENTS PER TYPE AFTER FILTERING

Incident Name	***Count***	***Distribution***
User	858	*Exp* (898.29)
SCADA	3338	*Exp* (350.44)
Phone Call	6654	*Exp* (306.81)
Work Order	1141	*Exp* (667.44)

VI.

VI. Incident Simulation Results

Two types of simulations are executed:

- Simulated with aggregate statistical parameters - simulation time T_{sim} = 2 years,
- Simulation with maximum occurrences for all types of incidents - T_{sim} =1 month.

The number of service channels is changed:

- Processing - p = 1,2,5,6,7,8,9.
- Servicing - s = 1,2,3,4,5.

The following output parameters are observed:

- Utilization of incident processing (dispatchers) - δ_i,
- Field utilization (field crew) - σ_i,
- Average waiting time in Q_p - W_p and percent of the total number of incidents that waited for processing x%
- Average waiting time in Q_s - W_s, and percent of the total number of incidents that waited for servicing y%.

Results obtained by repeating the simulations 500 times are shown in Tables V and VI.

TABLE V
SIMULATION RESULT FOR MONTH PERIOD

Test no		Utilization				W_s [min] (x%)	W_p [min] (y%)
		Servicing		Processing			
s	p	σ_{min}	σ_{max}	δ_{min}	δ_{max}		
2	5	0.86	0.91	0.97	0.99	361.73 (0.8%)	1170.06 (2.8%)
3		0.45	0.71	0.97	0.99	23.07 (0.05%)	1112.53 (2.7%)
4		0.22	0.65	0.98	0.99	4.03 (0.01%)	1261.60 (3.02%)
5		0.08	0.65	0.93	0.97	**0.92 (0.002%)**	**288.71 (0.67%)**
2	6	0.88	0.91	0.76	0.93	415.56 (0.99%)	115.65 (0.28%)
3		0.48	0.72	0.76	0.93	**28.90 (0.07%)**	**118.43 (0.28%)**
4		0.24	0.66	0.76	0.93	5.50 (0.01%)	119.03 (0.28%)
5		0.10	0.65	0.76	0.93	1.13 (0.003%)	120.69 (0.29%)
2	7	0.88	0.91	0.53	0.88	406.31 (0.96%)	31.25 (0.07%)
3		0.48	0.72	0.54	0.88	**29.02 (0.07%)**	**31.77 (0.07%)**
4		0.25	0.65	0.54	0.88	5.59 (0.01%)	31.44 (0.08%)
5		0.11	0.65	0.53	0.88	1.14 (0.003%)	31.14 (0.07%)
2	8	0.88	0.92	0.36	0.86	405.78 (0.97%)	11.16 (0.03%)
3		0.49	0.72	0.36	0.86	**30.09 (0.07%)**	**10.95 (0.03%)**
4		0.24	0.67	0.36	0.86	5.53 (0.01%)	11.09 (0.03%)
5		0.11	0.65	0.36	0.86	1.15 (0.003%)	11.42 (0.03%)
2	9	0.88	0.92	0.22	0.85	451.40 (1.07%)	4.25 (0.01%)
3		0.49	0.72	0.22	0.85	**29.56 (0.07%)**	**4.18 (0.01%)**
4		0.24	0.66	0.22	0.85	5.67 (0.01%)	4.17 (0.01%)
5		0.11	0.65	0.22	0.85	1.15 (0.003%)	3.94 (0.01%)

The results are shown in Table V (only the minimum and maximum values for service utilization are shown) present the behavior of the system in the case of maximum stress. The total number of incidents in all simulations is ~1350 incidents per simulation. Based on the results shown, it can be concluded that there is a need for minimum $p = 7$ processing services (due to W_p values and Processing utilization) and minimum $s = 3$ servicing services.

TABLE VI
SIMULATION RESULT FOR PERIOD OF TWO YEARS

Utilization						W_s [min]	W_p[min]
Servicing		Processing					
σ_1	σ_2	δ_1	δ_2	δ_3	δ_4		
0.52	*	**0.78**	**0.68**	*	*	**110.7**	**189.5**
0.52	*	0.64	0.49	0.34	*	110.35	24.22
0.52	*	0.60	0.44	0.27	0.14	110.01	4.49
0.38	**0.16**	**0.60**	**0.44**	**0.27**	**0.14**	**6.74**	**4.49**

Based on the obtained results, it can be concluded that, in the case of the first simulation (Table VI), if $s = 1$ servicing service and $p = 2$ processing services are used, the service utilization is correct, but the average waiting time for the incidents to be processed (W_p) and resolved (W_s) is approximately 5 hours in total (average total number of incidents in queues is less than 2 incidents per queue). Increasing the number of processing services to $p = 3$ significantly reduces the waiting time for processing, while increasing the number of servicing services at $s = 2$, reducing the waiting time for servicing, but also reducing their productivity (38% and 16%).

VII. CONCLUSION

During the management of the distribution network, it is necessary to review the need for engaging dispatchers and field crews to timely correct the incidents in the network. In this paper, a simulation and analysis procedure are described that allow the estimation of the necessary resources to effectively solve the problem in the network.

Some of the incidents from different types could have mutual dependence and priority which is not taken into consideration in this paper. For the future work authors plan to give more attention on grouping incidents.

ACKNOWLEDGMENT

The authors are partially supported by the Serbian Ministry of Education, Science and Technological Development, through grants No. 32018, 174009, 32035.

REFERENCES

[1] Y. Wang, C. Chen, J. Wang, R. Baldick, "Research on Resilience of Power Systems Under Natural Disasters-A Review" in *IEEE Transactions on Power Systems*, vol. 31, no. 2, pp. 1604–1613, 2016.

[2] A. I. Sarwat, M. Amini, A. Domijan Jr., A. Damnjanovic, F. Kaleem, "Weather-based interruption prediction in the smart grid utilizing chronological data" in Journal of Modern Power Systems and Clean Energy, vol. 4, no. 2, pp. 308-315, 2016.

[3] J. Lang, S. Pascoe, J. Thompson, J. Woyak, K. Rahimi, R. Broadwater, "Smart Grid Big Data: Automating Analysis of Distribution Systems" in *IEEE Rural Electric Power Conference*, pp. 96-101, 2016.

[4] D. W. Wanik, E. N. Anagnostou, B. M. Hartman, M. E. B. Frediani, M. Astitha "Storm outage modeling for an electric distribution network in Northeastern USA" in *Natural Hazards*, vol. 79, no. 2, pp. 1359-1384, 2015.

[5] A. Castillo, "Risk analysis and management in power outage and restoration: A literature survey" in *Electric Power Systems Research*, vol. 107, pp. 9-15, 2014.

[6] T. Altiok, B. Melamed, "Simulation modeling and analysis with Arena" *Elsevier*, pp. 39-43, 2010.

Comparative analysis of the usage of different image descriptors in object's video tracking

Abdalgalil Alsagair Abdulla and Stevica Graovac

***Abstract*—In this paper we tested a feature-based tracking algorithm for object tracking in video sequences, using different image descriptors (based on color, edge and texture) and particle filtering as a concept. A histogram-based framework is used to describe the object's features, where the object is a window consisting from a tracked vehicle and local background around it. Particle filtering has been proven as very robust one for non-linear and non-Gaussian estimation problems and performs well when clutter and occlusions are present. However, tracker based on single feature may lose the track easily or may start to track the wrong object. One popular remedy for this problem is usage of multiple features. In our approach, we develop the feature based particle filter tracker that relies on the search for the window, whose feature histogram matches a reference feature histogram model as much as possible. This work includes a comparison of tracking performances obtained by usage of different image descriptors separately; showing that some kind of fusion of partial results should be a reasonable solution in the context of analyzed traffic scenarios**

***Key Terms*— object tracking, particle filter, video tracking, image processing, color features, edge features, texture feature, histograms.**

I. Introduction

Tracking of moving objects is required in many applications such as surveillance systems, human-computer interfaces, target tracking, aircraft and car traffic monitoring, security and control. For the most of these tasks, a visual tracking based on processing of video sequences is used [1].

We focused our attention in this work toward visual tracking applications in the area of Intelligent Transportation Systems (ITS) where there are a number of algorithms and systems developed for automatic analysis and/or monitoring of traffic activities. Robust vehicle tracking is essential in traffic monitoring because it is the groundwork to higher level tasks such as traffic control, event detection and tracking [2]. Generally, visual tracking of moving objects is basically realized using object's visual features (e.g. color, texture, and shape) or motion information [3]. For the purpose of reliable vehicle tracking in the sequence of traffic images, typical problems are: variable background and shadows, variable existence of other close objects (moving or stationary), partial or even full occlusion by elements of scene, variable size of a tracked object during the sequence, etc [4], [5], [6]. Moreover, the visual tracking is affected by the problems caused by camera vibration due to wind, and lighting transitions between night/day and day/night.

There are a number of successful results in video tracking based on single or multiple features [4], [5], [7], [8]. Most of the problems encountered in video tracking can be addressed to the modeling of non-linear, non-Gaussian, multi-modal systems or any combination of these. To model accurately the underlying dynamics of a physical system, it is important to include elements of non-linearity and non-Gaussian noise distributions in many application areas. Speaking about of tracking algorithm concept, we have been motivated by the fact that a *Particle Filter* (PF) is very robust for non-linear and non-Gaussian dynamic state estimation problems and performs well when clutter and occlusions are present. PF uses *sequential Monte Carlo methods* based on point mass representations of probability densities that are applied to any state model.

While the color histogram-based particle filtering is the most common method used for object tracking, the existing R&D experiences as well as our analyses, have shown that in these ('single feature') cases tracker may lose the track easily or may start to track the wrong object. That was the reason why we focused our attention toward other features as are orientation of gradients and texture, in order to make the comparison between candidate features.

Tracking problem is specified in Section II, while particular image descriptors are the subject of Section III. In Section IV we have presented the results of usage different descriptors in specified set of test sequences. Concluding remarks are given in Section V.

II. Tracking Problem Formulation

The main task here was to design an algorithm for vehicle tracking using different features in the context of PF, for generated synthetic video sequences as well as for some typical real-world video sequences. Different techniques exist in the literature for video tracking tasks related to traffic monitoring and control. The tracking approach for video data that we apply is based on window-matching techniques consisting in the assessment of the degree of similarity among regions in sequential images. This way, an object may be recognized, and its position inferred in subsequent frames. For the synthetic video sequence, relatively small windows of size equal to 32×32 pixels are used for window matching purposes with the criterion to maximize the proper similarity function, as it was used in [1]. The similarity measure is based on the Euclidian distance between relevant histograms. The proposed approach uses three different types of matching: **color** (RGB)-based matching, **gradient** based matching, and **texture** based matching.

Abdalgalil Alsagair Abdulla is PhD student on the School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade,Serbia(e-mail:abdalgalilsagir@yahoo.com).

Stevica Graovac is retired professor of the School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail:graovac@etf.rs).

This work addresses the problem of tracking single target in video sequences and discusses the obtained results in the presence of typical disturbances.

Particle Filter

Particle Filter is a hypothesis tracker that approximates the filtered posterior distribution by a set of weighted particles. It weights particles based on a likelihood score and then propagates these particles according to a motion model. The basic key idea of particle filters is that, one can find an approximate representation of a complex model (any arbitrary probability distribution function – PDF) rather than an exact representation of a simplified mode (Gaussians). PF is a promising technique because it is concerned with the problem of tracking single and multiple moving objects in video sequences [1], [9], [10] in typical traffic scenes characterized by non-linear motion model and non-Gaussian nature of disturbances. For testing of the tracking of single moving object using PF, we developed a synthetic video sequences and selected some typical real world street video sequences.

Steps in general PF algorithm are:

1. Initialize the state x_1 for the first frame and calculate the reference histogram on proposed features (color, edge and texture) for the target window.
2. Generate a particle set of N particles $\{X^m\}$, m=1: N, around the target point.
3. Predict the new state for each particle using transition model.
4. Compute a histogram distance (Euclidian distances for chosen feature) for each particle.
5. Weight each particle based on histogram distance and normalize the weights.
6. Select the location of target as a particle with minimum histogram distance (maximum weight).
7. Update the target location and particles' positions.
8. Resample the particles for the next iteration
9. Increase the time step and go to step 3.

III. FEATURES BASED TRACKING METHODS

Instead of tracking the entire objects (i.e. pattern of light intensities of pixels inside the tracking window), this approach tracks only the proposed features representing the window around moving objects, from frame to frame. The advantage of this approach is that, even in the presence of partial occlusion, some of the features of the moving object remain visible/recognizable [11]. Several features can be used for this purpose. We have analyzed the application of tracking PF using the features based on:

- Color measurement cue.
- Gradient measurements cue.
- Texture measurements cue.

In all three cases the feature/image descriptor is a particular type of histogram, representing the distribution of color components, gradient orientations, or texture/spatial relationships.

A. Color Feature

Color-based trackers have been proven to be robust. These trackers rely on the deterministic search of a window, whose color content matches a reference histogram. The target object to be tracked is the window initially formed manually around the moving vehicle. Color histogram (RGB) of this window is calculated and used for the purpose of a deterministic search for a matching window [12]. Color histograms have been widely used for tracking problems, because they are relatively robust in the presence of partial occlusion, rotation, and changes in size. They have limits in areas where the background has a similar color as the target object and they have poor performance when the illumination varies.

Tracking the window of interest consists in the context of PF in comparing its histogram with the histograms of the sample positions of particles, using the Euclidian distance as a measure [13]. In this work the histograms are typically calculated in the RGB space using 20 bins for each of color channels (60 bins, in sequence, together).

Fig. 1 shows a typical RGB color histograms for both, the reference window (frame number 4) and the best candidate window (frame number 35).

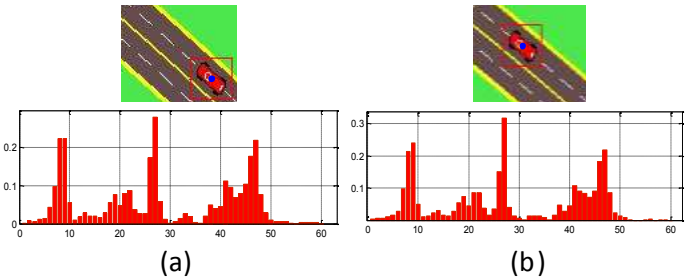


Fig. 1 a) Reference window histogram (k=4), b) Best candidate window histogram (k=35)

The predicted position of the tracked window could be chosen as the position of the particle where the matching of histograms is the best (as in our case) or at the position obtained by weighted averaging of positions of all particles, with the partial influence proportional to the distance between color histogram of the window on particle's position and the reference color window [6]. The smaller the discrepancy between the candidate and the reference models, the higher the likelihood that the object is located inside the candidate region. This discrepancy is measured by calculating the histogram distance. In this work we used Euclidean distance (EC_{dist}) as given in Eq. 1.

$$EC^{i}_{dist}(k) = \sqrt{\sum_{j=1}^{m}(C\_H\_part(j) - C\_H\_ref)^2} \qquad (1)$$

Where: 'i' is the particle index, 'k' is the Frame number, 'j' is bin number, 'm', is total number of bins, 'C_H_ref', is the extracted color histogram of reference window, and 'C_H_part', is the extracted color histogram of a particular particle window to be compared.

The observation likelihood model is used to assign a weight associated to a specific particle (new observation) depending on how similar the reference window histogram and the

particle's histogram are in the case of i^{th} particle [4].To evaluate the similarity between the reference histogram, and the particle's histogram, the similarity criteria based on Euclidian distance is used as in Eq. 2:

$$\mathrm{SIM}^{i}(k) = 1 - \mathrm{EC}^{i}_{\mathrm{dist}}(k) \quad (2)$$

B. Edge Feature

Object edges play a very important role in computer vision/pattern recognition and their orientations describe an important feature specifying the shape of the object of interest. The edge in image is the continuous set of pixels with high intensities of gradient. The edge feature used here is formed as a histogram of gradient orientations for the pixels where the magnitude of gradient is above some specified threshold. Edges are detected using horizontal and vertical Sobel operators K_x and K_y on the gray-scale image. The horizontal gradient G_x and vertical gradient G_y have magnitudes given by:

$$\left.\begin{aligned} G_x(x,y) &= K_x * I(x,y) \\ G_y(x,y) &= K_y * I(x,y) \end{aligned}\right\} \quad (3)$$

Where $K_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}$, $K_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}$

and *I(x,y)* is the candidate window.

The magnitude (G) and phase (θ) of the edges are determined as:

$$\left.\begin{aligned} G(x,y) &= \sqrt{G_x(x,y)^2 + G_y(x,y)^2} \\ \theta(x,y) &= \tan^{-1}\left(\frac{G_y(x,y)}{G_x(x,y)}\right) \end{aligned}\right\} \quad (4)$$

The basic idea in histogram of edge orientation of gradients is that an object shape can be characterized by the specific distribution of gradient orientations along its edges. The histogram bins are equally spaced over the interval -180 ^{o} to +180^{o} for unsigned gradient. There are three steps to find the histogram of edge gradient orientations:

- Find the positions (pixels) where the magnitude (G) is determined to be above some specified threshold (e.g. 80% of maximal magnitude in this window).
- Obtain the phases (θ) of the gradients corresponding to the pixels extracted in previous step;
- Calculate the edge orientation gradients histogram (using 40 bins in this particular case)

C. Texture Cues

Despite there is no unique definition of texture, it is generally agreed that texture provides information about the spatial arrangements of pixel color or grey levels in an image [1], which may be stochastic or periodic, or both. Texture descriptor provides measures of properties such as smoothness, coarseness, and regularity. There are many approaches used for describing the texture such as statistical approach described in [14].

There are various methods for statistical texture description:

- Local Binary Partition
- Histogram and Features
- Co-occurrence Matrices and Features
- Autocorrelation and Power Spectrum

In this work we used a histogram of *Grey Level Co-occurrence Matrix* (GLCM) values at a given offset over an image as texture descriptor.

The GLCM is a tabulation of how often different combinations of pixel brightness values (grey levels) occur in an image. In summary, there are four necessary steps in generating a GLCM matrix:

- Create a framework matrix;
- Decide on the spatial relation between the reference and neighbor pixel (give offset), in our work we use offsets=[0 1].
- Count the occurrences and fill in the framework matrix;
- Normalize the matrix to turn it into probabilities.

The obtained GLCM matrix size is 8x8 (eight classes of gray level), it is converted to a vector of size 64, and probabilities of occurrence being the elements of GLCM, now form the histogram with 64 bins as the texture descriptor.

IV. SIMULATION RESULTS AND DISCUSSION

There are five supposed scenarios for vehicle tracking tests. These are:

Case-I: Tracking the non-maneuvering target car without disturbance (synthetic sequence of images).

Case-II: Tracking a target car with disturbance (car shadow and shadow produced by road environment-(synthetic sequence of images).

Case-III: Tracking a maneuvering target car without disturbance (synthetic sequence of images).

Case-IV: Tracking the car with car ego-shadow and trees' shadow (real video sequence).

Case-V: Tracking the maneuvering car with partial occlusion and full occlusion (real video sequence).

The assumed parameters for synthetic video sequence simulation are:

- Velocity of the target vehicle in all cases is 90 km/h.
- Time interval between frames 50 ms.
- The number of particles used in PF was 150 in Case-I and Case-II, and 200 particles in Case-III.
- The spreading radius of the particles was 10 pixels in Case-I and Case-II, and 15 pixels in Case-III.
- Framework window size is 32×32 pixels
- 1 meter ≈ 7 pixels

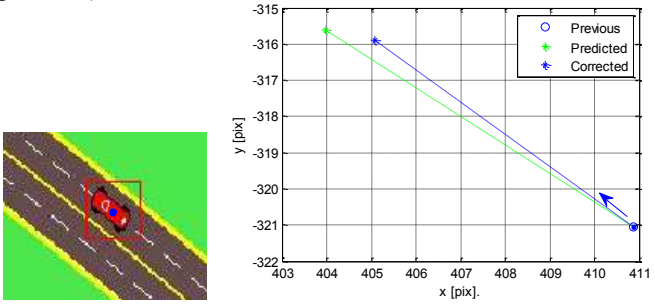
Simulation parameters for the selected real video sequences are:

- Interval between frames - 10.
- The number of particles used in PF was 150 particles for both cases.
- The spreading radius of the particles was 20 pixels in Case-IV and 15 pixels for Case-V.

- Framework window size is 50×50 and 80×70 pixels for Case-IV and Case-V, respectively.
- 1 meter ≈ 63 pixels for Case-IV and 1 meter ≈ 16 pixels for Case-V.

*A. **Case-I:** Tracking the non-maneuvering target car without disturbance.*

This case is verifying the tracking situation where there are no disturbances and no obstacles on the road, as shown in Fig.2a below. Fig.2b shows the tracked trajectory between frame 34 (previous) and frame 35.



(a) (b)

Fig.2 a) target car in the road, b) tracked trajectory between 2-frames

Fig.3 below shows the average similarity (for all particles) and the deviation between the estimated (updated) target positions using the proposed descriptors and the actual vehicle position respectively.

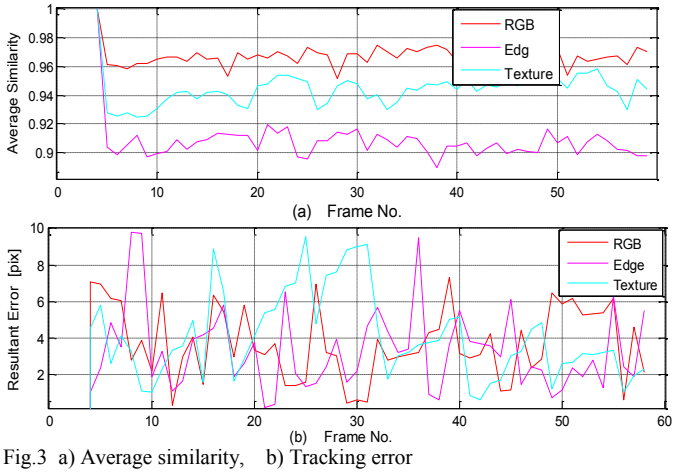
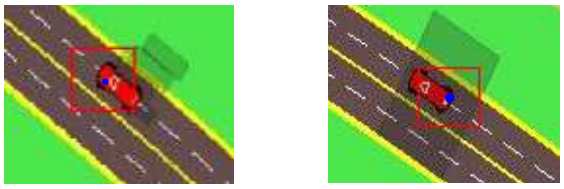


Fig.3 a) Average similarity, b) Tracking error

From Fig.3 we can see that the errors are approximately the same for all three descriptors (root mean square error are 4.18, 3.94 and 4.61 pixels, respectively for color, edge and texture). Average similarity slightly gives the preference to color descriptor.

*B. **Case-II:** Tracking a target car with disturbance (ego-car shadow and shadow produced by road environment)*

The disturbances such as ego-car shadow and tree shadow are inserted in this scenario and they appear abruptly at different intervals. Fig.4a illustrates the target car with shadow produced by itself (k=15), and Fig.4b shows the target car when it is affected by tree shadow (k=45). The obtained average similarity and deviation between the actual and estimated position of the target vehicle respectively can be seen in Fig.5 below.



(a) (b)

Fig.4 a) Target car with its shadow (k=15), b) Target car within Tree shadow (k=45)

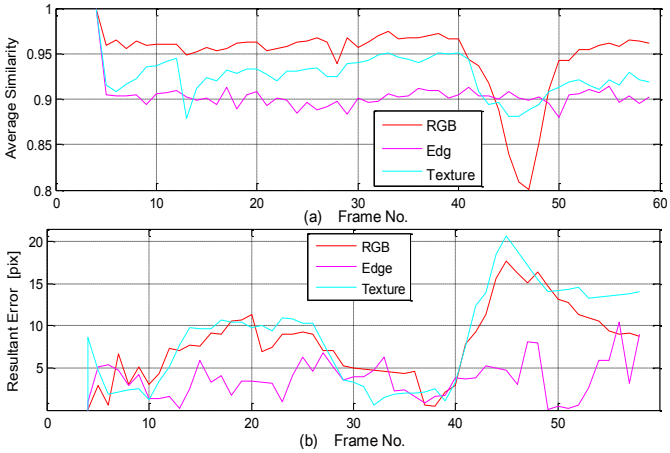


Fig.5 a) Average similarity, b) Tracking error

The error diagram shows that the color and texture descriptors are more sensitive to the shadow in both cases. The Average similarity for color descriptor is dropped to 0.80 at tree shadow interval and the root mean square error is 8.89, 4.32 and 10.10 pixels respectively for color, edge, and texture (giving the preference to edge descriptor).

*C. **Case-III:** Tracking a maneuvering target car without disturbance*

Fig.6 illustrates the tracking situation for maneuvering car bypassing the others: before over-taking (k=25), two cars in parallel (k=39), and after over-taking (k= 55).

(a) (b) (c)

Fig.6 Different status intervals frame (k=25), a) before over taking, b) 2-cars in parallel (k=39), c) after over taking (k=55).

Average similarity and the deviation between the actual and estimated position of the target vehicle respectively, can be seen in Fig.7. The maneuvering stage is occurring between the 4th frame and 70th frame.

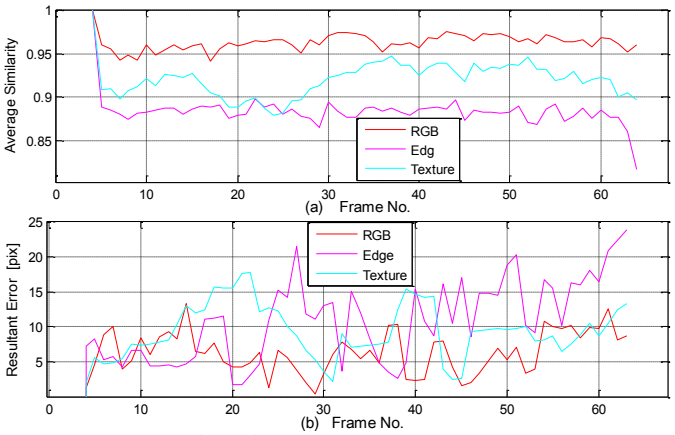


Fig.7 a) Average Similarity, b) Tracking error

We can see from Fig.7 that the color is favorable compared with other descriptors; and the root mean square error is 6.93, 12.14 and 10.05 pixels respectively for color, edge, and texture.

D. ***Case-IV:*** *Tracking the target car in real video sequence with disturbances of car and tree shadow*

This case describes the tracking situation in the existence of car shadow and tree shadow. Fig.8 shows the target window and color histogram respectively for the reference frame (k=235), first frame (k=242), car at tree's shadow frame (k=263) and car after tree shadow frame (k=291).

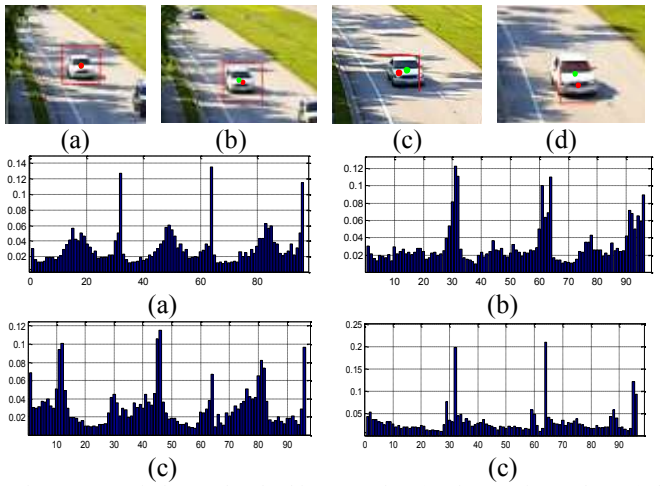
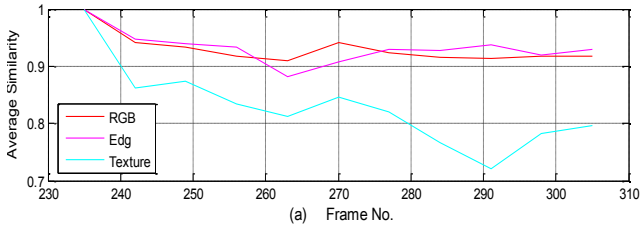


Fig.8 Target window and Color histogram for, a) reference frame (k=235), b) first frame (k=242), c) car at tree shadow frame (k=263), d) car after tree shadow frame (k=291).

The average similarity and the deviation between the actual and estimated position of the target vehicle respectively can be seen in Fig.9.



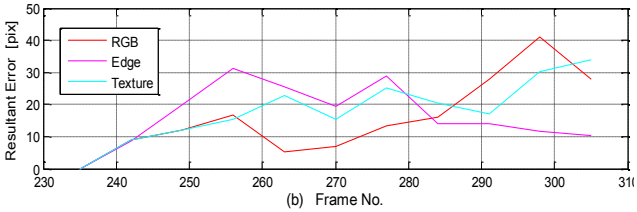


Fig.9 a) Average Similarity, b) Tracking error

We can see from the average similarity in Fig.9 that the color and edge are favorable descriptors; while the root mean square error is 19.65, 18.87 and 20.47 pixels respectively for color, edge, and texture. One should note that there are some differences in comparison to synthetic sequence of this type:

1. Size of the Frame – larger than before
2. Variable shadow all the time
3. Both types of shadows simultaneously
4. Correlation between pixels and meters is different (1 meter = 64 pixels).

E. Case-V: Tracking of maneuvering target car in real video sequence with partial and full occlusion

This case describes the tracking situation of the target vehicle exposed to partial and full occlusions due to the tree beside the road. One may say that this case is the most complex scenario because of:

1. Decreasing the size of car.
2. Partial and even full occlusion.
3. Maneuver during the full occlusion phase.

Fig.10 illustrates some important frames and corresponding color histograms representing these situations.

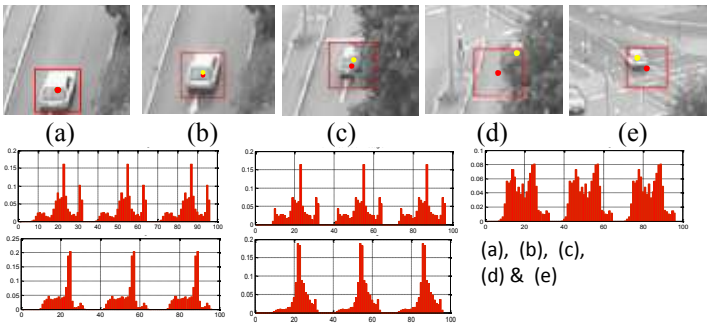
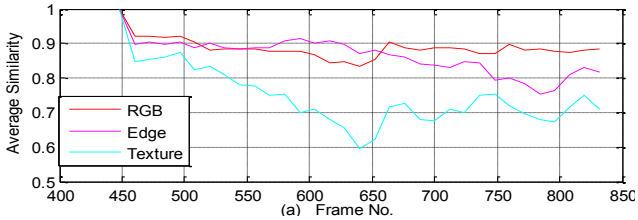


Fig.10 Frame and color histogram for a) Reference (k = 448), b) first frame (k = 460), c) partial occlusion (k = 580), d) full occlusion (k = 628), e) after occlusion (k = 700).

Fig.11 below shows the average similarity and the deviation between the actual and the estimated (updated) target positions using the three descriptors respectively.



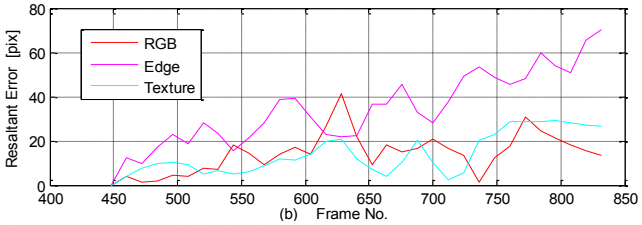


Fig.11 a) Average similarity, b) Tracking error

The root mean square error is 16.91, 38.16 and 16.65 pixels respectively for color, edge and texture; Correlation between pixels and meters is 1 meter ≈16 pixels. We can see in Fig.11 that according to the average similarity, color and edge are favorable descriptors. On the other hand, looking onto error histories, one can see that edge descriptor is practically not useful at all, because tracking based on it lost the tracked car completely after the full occlusion during the interval of k = 544 to k = 664.

Table-I below illustrates the RMS error in meter for all five cases.

TABLE I
RMS OF TRACKING ERRORS IN METERS FOR ALL CASES

	Case I	Case II	Case III	Case IV	Case V
RGB	0.597	1.270	0.990	0.307	1.057
Edge	0.563	0.617	1.734	0.295	2.385
Texture	0.659	1.443	1.436	0.320	1.041

V. CONCLUSION

In this paper, we have presented a comparative analysis of application of object tracking PF algorithm based on three features/image descriptors. From the studded cases, for a single object in synthetic video sequences and in real video, we can conclude the following:

- In the case of non-maneuvering target vehicle without disturbance, the tracking errors are mostly the same for all three descriptors.
- The case of tracking of a target car with its shadow and shadow produced by road environment has shown that the color and texture descriptors are more sensitive in comparison to edge descriptor.
- The maneuvering of undisturbed target case shows the edge descriptor is the most sensitive to rotation as illustrated in Fig.6, and the color cue gives the best accuracy compared to the others.
- In the case of tracking the target car (real video) with its shadow and tree, one can conclude that the color is favorable descriptor.
- The situation of the maneuvering target vehicle with partial and full occlusions has shown in the last case, being the most complex scenario, that the preference should be given to color and texture descriptors in comparison to edge based.

As a final conclusion, it could be said that according to different scenarios tested here, none of analyzed descriptors can be declared as the best one for all cases. To overcome this ambiguity, some fusion of all three tracking results obtained by different descriptors should be made in one integral tracking algorithm based on all three analyzed descriptors.

REFERENCES

[1] L. Mihaylova2, P. Brasnett, N. Canagarajah, D. Bull, "Object Tracking by Particle Filtering Techniques in Video Sequences," in *Advances and Challenges in Multisensor Data and Information*, Amestrdam, IOS Press, Published in cooperation NATO Public Diplomacy Division, 2007, pp. 260-268.

[2] Ms. Susmita A. Meshram, Prof. A. V. Malviya, "Vehicular Traffic Surveillance for Real Time Using Multiple Methodologies.," *International Journal of Scientific & Engineering Research,* vol. 4, no. 5, pp. 1988-1992, 2013.

[3] S. H. Shaikh, K. Saeed, N. Chaki, "Moving Object Detection Using Background Subtraction", Springer, Cham, 2014.

[4] Ng Ka Ki and Edward J. Delp, "New Models For Real-Time Tracking Using Particle Filtering," in *Visual Communications and Image Processing*, 2009.

[5] B. Coifman, D. Beymer, P. McLauchlan, J. Malik, "A real-time computer vision system for vehicle tracking and traffic surveillance," in *Transportation Research Part C: Emerging Technologies6.4 (1998): 271-288*, 1998.

[6] Md. Zahidul Islam, Chi-Min Oh and Chil-Woo Lee, "Video Based Moving Object Tracking by Particle Filter," *International Journal of Signal Processing, Image Processing and Pattern,* vol. Vol. 2, pp. 119-132, 2009.

[7] Y. Dai, B. Liu, "Robust Video Object Tracking Using Particle Filter With Likelihood," in *eprint arXiv:1509.08182, Computer Science - Computer Vision and Pattern Recognition*, 2015.

[8] C. Shen, A. van den Hengel, A. Dick, "Probabilistic Multiple Cue Integration for Particle Filter Based Tracking," Sydney, Proc. of the VIIth Digital Image Comp.: Techniques and Appl., 2003, pp. 399-408.

[9] M. Jaward, L. Mihaylova, N. Canagarajah and D. Bull, "Multiple Object Tracking Using Particle Filters," in *Aerospace Conference,IEEE*, 2006.

[10] A. Almeida, J. Almeida, R. Ara'ujo, "Real-Time Tracking of Moving Objects Using," in *IEEE ISIE*, Dubrovnik, Croatia, 2005.

[11] S. Kamijo, K. Ikeuchi, M. Sakauchi, "Vehicle Tracking in Low-angle and Front-View Images based on Spatio-Temporal Markov Random Field Model," in *18th world Congress on ITS*, Sydney, 2001.

[12] N. Easwar, J. Shah, "Object Tracking using Particle Filter," CIS 601, https://slideplayer.com/slide/7259778/, 2003.

[13] K. Nummiaro, E. Koller-Meier and L. Van Gool, "A Color-based Particle Filte," in *First International Workshop on Generative-Model-Based Vision, in conjunction with ECCV'02, pp*, 2002.

[14] *Texture COMP 9517 Computer Vision,* www.cse.unsw.edu.au/~XCAI7-11/lectures/previous/old/Texture.ppt, 2009.

Gaussian Process Domain Experts for Prediction of Alzheimer's Disease-Related Cognitive Scores

Nikola Popović, Ognjen Rudović, Predrag Vasilić and Predrag Tadić

***Abstract*—We address the problem of predicting Alzheimer's Disease (AD) progression, taking data from past visits as inputs. This is important for timely monitoring of a subjects' cognitive performance, and, consequently, for improving the selection of subjects for new clinical trials. To this end, we introduce a novel prediction model based on the notion of domain adaptive Gaussian Processes (GP). Specifically, in contrast to previous works that employed GPs for this task, here we formulate a mixture of domain-specific GPs, where each GP is trained on a subpopulation with different clinical status (cognitively normal (CN), mild-cognitive impairment (MCI) and AD). Furthermore, by using the probabilistic formulation of GPs, we personalized our model to the target subject (not used to train the model) using his/her previous visits' data, in order to reduce the estimation bias, which is typically pronounced when applying the population-level GP models to the target task. The proposed method was compared to several state-of-the-art GP-based models on a sub-cohort of subjects who participated in the Alzheimer's Disease Neuroimaging Initiative (ADNI) [1]. The models were train to predict the patient's future score on the standard Alzheimer's Disease Assessment Scale–Cognitive Subscale (ADAS-Cog13). We show in the experiments presented here that the proposed domain-specific mixture of adaptive GPs does not lead to large improvements in terms of the point estimate of the ADAS-Cog13 scores compared to existing GP models trained on the whole training population. However, it provides much better uncertainty estimates for its predictions. This is an important advantage of our model as it allows to encode the model's prediction confidence more accurately than the single expert model (i.e. when a GP is trained on all three sub-groups together). We also show that the proposed approach has much lower computational complexity, as the necessary matrix inversions can efficiently be performed on the data-subsets corresponding to the target sub-populations.**

***Index Terms*—Gaussian processes, Alzheimers disease, Personalized models, Gaussian process experts, Machine Learning.**

I. Introduction

RECENTLY, view on Alzheimer's Disease (AD) diagnosis has shifted towards a more dynamic process in which clinical and pathological markers evolve gradually before diagnostic criteria are met [2]. Achieving accurate predictions of AD progression is a significant and difficult challenge. There is wide variability in data available per subject, inherent per-person differences, and the disease has a slowly changing nature. The most widely used general cognitive measure in clinical trials is the Alzheimer's Disease Assessment Scale-cognition sub-scale (ADAS-Cog) [3], which assesses multiple cognitive domains including language, memory, orientation and praxis [4]. The ADAS-Cog has proven important for target clinical assessments, so in this article we focus on a machine learning method for predicting the future values of this score based on previously seen subjects. We specifically use the modified ADAS-Cog 13-item scale [3], which includes all original ADAS-cog items, with the addition of a number of cancellation tasks and a delayed free recall task, for a total of 85 points. Higher scores mean greater severity. These additional items were added to increase the number of cognitive domains and range of symptom severity without a substantial increase in the time required for administration.

One of the main challenges in a clinical assessment of a subject at risk of developing AD is the ability to accurately predict the future cognitive scores of interest for clinical trials, which for instance aim to find the most effective treatments for AD. An algorithm that could predict future cognitive scores such as ADAS-Cog13 would bring great value to the assessment procedure, and would have a potential to improve the efficiency of clinical trials, which are typically expensive and lengthy. For example, out of hundreds of clinical trials, costing billions of dollars, fewer than 1% have proceeded to the regulatory approval stage and non have managed to prove a disease-modifying effect [5], [6]. Improvement in the ability to accurately identify subjects at early stages of the disease, where the treatment has a bigger chance to be effective, promises more success. Thus, having models that can make accurate predictions of the progression of cognitive scores such as ADAS-Cog13 is of great importance.

In this paper, we use data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) [1], which was processed for the TADPOLE Challenge [5], to evaluate models which predict the ADAS-Cog13 score for the next doctor visit. The dataset is very sparse, with different combinations of features missing for different subjects. For about 80% of the subjects, partial records are missing from the dataset. Accurate forecasting of cognitive decline and related measures of disease progression is a very difficult challenge given the wide variability in available data per subject, slowly changing nature of the disease and inherent per-person differences.

We use various Gaussian Process (GP) models for predicting the ADAS-Cog13 score for the next doctor visit based on the current visit. GP are nonparametric, probabilistic models which also offer the predictive uncertainty in the form of the predictive distribution variance [7]. The uncertainty enables better quantification of the prediction performance. We first evaluate the results achieved using regular and personalized

N. Popović, P. Vasilić, P. Tadić – School of Electrical Engineering, University of Belgrade, Bul. Kralja Aleksandra 73, 11120 Belgrade, Serbia (e-mail: {npopovic,vasilic,ptadic}@etf.bg.ac.rs).

O. Rudović – MIT Media Lab, 75 Amherst St, 02139 Cambridge, MA, USA (e-mail: orudovic@mit.edu).

GP which have been used to tackle this challenge [2], [8]. We then propose a novel GP domain-expert approach that is based on the notion of domain-adaptive and distributed GPs [9], [10]. While previous GP-based approaches use a single GP expert to model the whole population under the study, here we use the domain knowledge to derive the GP expert models for each subpopulation (the subjects diagnosed as CN, MCI and AD), as we expect different dynamics in the data of these subjects. Specifically, we build a mixture of three domain-specific GP experts (CN, MCI and AD) as well as the target expert (a GP applied to the target subject's data only), using the variance-based weighting scheme, as in [10]. As in the previous works [2], [8], we also show the effects of the model personalization on the prediction performance. We show that the proposed GP expert approach achieves relatively good predictions with much better uncertainty estimates, while being much more computationally efficient.

The rest of the paper is structured as follows: in Section II, we provide a theoretical overview of GPs, personalized GPs and GP expert models. Section III describes the dataset used, and provides further details about the models, their evaluation, and shows the experimental results. Finally, Section IV concludes the paper.

II. Gaussian Processes

Gaussian processes are very flexible probabilistic, non-parametric models. Unlike classical Machine Learning algorithms, their output is not a point estimate, but a distribution. This is a good property since besides the point estimate, the model also expresses how certain it is in the prediction. For example, if the predictive distribution is tightly concentrated around some value it means that the model is confident in that prediction.

First, we will look at classical Gaussian processes for regression as described in [7]. Then, an overview of Gaussian process adaptation technique [2], [10] which makes personalized predictions will be given. Finally, we will look at Gaussian process domain experts technique [9], [10].

A. Gaussian Processes for Regression

Strictly speaking, a Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution [7]. A Gaussian process $f(\mathbf{x})$ is completely specified by its mean $m(\mathbf{x})$ and covariance function $k(\mathbf{x}, \mathbf{x}')$ which are defined as

$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})], \tag{1}$$

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))], \tag{2}$$

and it is denoted as

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')), \tag{3}$$

where the mean function is usually taken to be zero.

The usual setting is that only noisy measurements of the function value are obtained at some input locations

$$y = f(\mathbf{x}) + \varepsilon. \tag{4}$$

The noise ε is assumed to be independent and identically distributed Gaussian noise with zero mean and variance σ_n^2. We will denote the training set inputs with $\mathbf{X}^{(v)} = \{\mathbf{x}_{n_v}^{(v)}\}_{n_v=1}^{N_v}$ and outputs with $\mathbf{Y}^{(v)} = \{\mathbf{y}_{n_v}^{(v)}\}_{n_v=1}^{N_v}$. The joint distribution of all the training measurements along with one test function value is

$$\begin{bmatrix} \mathbf{Y}^{(v)} \\ f_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K} + \sigma_n^2\mathbf{I} & \mathbf{k}_* \\ \mathbf{k}_*^{\mathrm{T}} & k_{**} \end{bmatrix}\right), \tag{5}$$

where $\mathbf{K} = \mathbf{K}(\mathbf{X}^{(v)}, \mathbf{X}^{(v)})$, $\mathbf{k}_* = \mathbf{k}(\mathbf{X}^{(v)}, \mathbf{x}_*)$ and $k_{**} = k(\mathbf{x}_*, \mathbf{x}_*)$. We can then condition this joint Gaussian distribution on the observations to obtain a posterior Gaussian distribution

$$p(f_*|\mathbf{X}^{(v)}, \mathbf{Y}^{(v)}, \mathbf{x}_*, \boldsymbol{\theta}) \sim \mathcal{N}(\mu_*, \sigma_*^2), \tag{6}$$

with the mean μ_* and variance $V_* = \sigma_*^2$

$$\mu_* = \mathbf{k}_*^{\mathrm{T}} \left(\mathbf{K} + \sigma_n^2\mathbf{I}\right)^{-1} \mathbf{Y}^{(v)}, \tag{7}$$

$$V_* = k_{**} - \mathbf{k}_*^{\mathrm{T}} \left[\mathbf{K} + \sigma_n^2\mathbf{I}\right]^{-1} \mathbf{k}_*, \tag{8}$$

where $\boldsymbol{\theta}$ are the hyper-parameters of the Gaussian process—parameters of the prior covariance function and the variance of the noise σ_n^2. When using Gaussian processes for regression, we first specify a prior distribution with (1) and (2). The prior should be based on our belief about the underlying function. We then use the training data $\left(\mathbf{X}^{(v)}, \mathbf{Y}^{(v)}\right)$ to obtain the posterior predictive distribution (7) and (8), at inputs $\mathbf{x}_*$ that are of interest.

One of the most commonly used covariance functions is the squared-exponential kernel

$$k_{\mathrm{SE}}(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2l^2}\right), \tag{9}$$

where σ_f^2 represents the variance of the process and l represent the characteristic length-scale—how much $\mathbf{x}$ and $\mathbf{x}'$ have to be far apart to become uncorrelated. The parameters σ_f^2 and l are the hyper-parameters of the squared-exponential covariance function.

The hyper-parameters $\boldsymbol{\theta}$ will be learned by maximizing the log-likelihood

$$\begin{aligned} \log p(\mathbf{Y}^{(v)}|\mathbf{X}^{(v)}, \boldsymbol{\theta}) = &-\frac{1}{2}\mathbf{Y}^{\mathrm{T}(v)}\mathbf{K}_y^{-1}\mathbf{Y}^{\mathrm{T}(v)} \\ &-\frac{1}{2}\log|\mathbf{K}_y| - \frac{n}{2}\log(2\pi), \end{aligned} \tag{10}$$

where $\mathbf{K}_y = \mathbf{K} + \sigma_n^2\mathbf{I}$. The process of learning the hyper-parameters is called training. It requires computing the inverse and the determinant of $\mathbf{K}_y$, both of which have $\mathcal{O}(N_v^3)$ complexity.

B. Personalized Gaussian Processes (PGP)

Assume that we have access to a large collection of labeled *source* data and a smaller set of labeled *target* data. For example, when forecasting the development of a patients disease, the target data would be all the data collected from that patient so far and the source data would consist of the collected data of all other patients. The source data is

denoted with $\mathcal{D}^{(s)} = \left(\mathbf{X}^{(s)}, \mathbf{Y}^{(s)}\right)$, where $\mathbf{X}^{(s)} = \{\mathbf{x}_{n_s}^{(s)}\}_{n_s=1}^{N_s}$ and $\mathbf{Y}^{(s)} = \{\mathbf{y}_{n_s}^{(s)}\}_{n_s=1}^{N_s}$. The target data is denoted with $\mathcal{D}^{(t)} = \left(\mathbf{X}^{(t)}, \mathbf{Y}^{(t)}\right)$, where $\mathbf{X}^{(t)} = \{\mathbf{x}_{n_t}^{(t)}\}_{n_t=1}^{N_t}$ and $\mathbf{Y}^{(t)} = \{\mathbf{y}_{n_t}^{(t)}\}_{n_t=1}^{N_t}$.

This PGP algorithm consists of three steps [9]:

1) Train a GP on the source data by maximizing the marginal likelihood $p(\mathbf{Y}^{(s)}|\mathbf{X}^{(s)}, \boldsymbol{\theta})$ to learn the hyper-parameters $\boldsymbol{\theta}$. The posterior distribution is then obtained by applying equations (7) and (8).
2) Use the posterior obtained on the source data as a prior for the GP of the target data $p(\mathbf{Y}^{(t)}|\mathbf{X}^{(t)}, \mathcal{D}^{(s)}, \boldsymbol{\theta})$.
3) Correct the posterior distribution to account for the for the target data $\mathcal{D}^{(t)}$ as well.

That is, our training data will be the target set and our prior will be the predictive distribution of applying a GP to the source data.

Now, the conditional prior of the target data (given the source data) from step 2) is given by applying equations (7) and (8) on $\mathbf{X}^{(t)}$:

$$\boldsymbol{\mu}^{(t|s)} = \mathbf{K}_{st}^{\mathrm{T}(s)} \left(\mathbf{K}^{(s)} + \sigma_{n_s}^2 \mathbf{I}\right)^{-1} \mathbf{Y}^{(s)}, \tag{11}$$

$$\mathbf{V}^{(t|s)} = \mathbf{K}_{tt}^{(s)} - \mathbf{K}_{st}^{\mathrm{T}(s)} \left[\mathbf{K}^{(s)} + \sigma_{n_s}^2 \mathbf{I}\right]^{-1} \mathbf{K}_{st}^{(s)}, \tag{12}$$

where $\mathbf{K}_{st}^{(s)} = k^{(s)}(\mathbf{X}^{(t)}, \mathbf{X}^{(t)})$ and $\mathbf{K}_{tt}^{(s)} = \mathbf{k}^{(s)}(\mathbf{X}^{(t)}, \mathbf{X}^{(t)})$. Given the above prior, the adapted predictive distribution for a test input $\mathbf{x}_*^{(t)}$ after observing the target data is given by [9]

$$\begin{aligned} \mu_{ad}^{(t|s)}(\mathbf{x}_*^{(t)}) &= \mu^{(s)}(\mathbf{x}_*^{(t)}) \\ &+ \mathbf{V}_*^{\mathrm{T}(t|s)} \left(\mathbf{V}^{(t|s)} + \sigma_{n_s}^2 \mathbf{I}\right)^{-1} \left(\mathbf{Y}^{(t)} - \boldsymbol{\mu}^{(t|s)}\right), \end{aligned} \tag{13}$$

$$\begin{aligned} V_{ad}^{(t|s)}(\mathbf{x}_*^{(t)}) &= V^{(s)}(\mathbf{x}_*^{(t)}) \\ &- \mathbf{V}_*^{\mathrm{T}(t|s)} \left(\mathbf{V}^{(t|s)} + \sigma_{n_s}^2 \mathbf{I}\right)^{-1} \mathbf{V}_*^{(t|s)}, \end{aligned} \tag{14}$$

with

$$\begin{aligned} \mathbf{V}_*^{(t|s)} &= k^{(s)}(\mathbf{X}^{(t)}, \mathbf{x}_*^{(t)}) \\ &- K_{st}^{\mathrm{T}(s)} \left(\mathbf{K}^{(s)} + \sigma_{n_s}^2 \mathbf{I}\right)^{-1} k^{(s)}(\mathbf{X}^{(s)}, X_*^{(t)}) \end{aligned} \tag{15}$$

It is clear from (13) that the final prediction shifts the prior mean obtained on the source data towards the distribution of the target data. Also, from (14) we notice that the models confidence is improved by reducing the prior variance.

C. Gaussian Processes Domain Experts (GPDE)

For our training set we have both the source and target data $\mathcal{D} = \left(\mathcal{D}^{(s)}, \mathcal{D}^{(t)}\right)$. We will assume that the source data is a combination of multiple smaller source datasets $\mathcal{D}^{(s)} = \{\mathcal{D}^{(s_1)}, \mathcal{D}^{(s_2)}, ..., \mathcal{D}^{(s_M)}\}$, where M is a total number of these datasets. A GP expert will be trained on each dataset and their results will be combined into a final prediction. An overall computation will be performed by combining many independent smaller computations performed by experts. This approach enables parallelisation and distributed computing.

Given the above mentioned data split and assuming conditional independence of the labels from each domain given their corresponding input features, the marginal likelihood can be approximated by

$$\begin{aligned} &p(\mathbf{Y}^{(s,t)}|\mathbf{X}^{(s,t)}, \boldsymbol{\theta}^{(s,t)}) \approx \\ &p(\mathbf{Y}^{(t)}|\mathbf{X}^{(t)}, \boldsymbol{\theta}^{(t)}) \prod_{k=1}^{M} p_k(\mathbf{Y}^{(s_k)}|\mathbf{X}^{(s_k)}, \boldsymbol{\theta}^{(s)}). \end{aligned} \tag{16}$$

We assume that a standard GP is sufficient to model the source data, so all source domains share a set of hyper-parameters $\boldsymbol{\theta}^{(s)}$. Each factor from (16) is determined by a GP expert. For training the GPDE, we find the GP hyperparameters $\boldsymbol{\theta}^{(s)}$ and $\boldsymbol{\theta}^{(t)}$ that maximize the corresponding log-marginal likelihood

$$\begin{aligned} &\log p(\mathbf{Y}^{(s,t)}|\mathbf{X}^{(s,t)}, \boldsymbol{\theta}^{(s,t)}) = \\ &\log p(\mathbf{Y}^{(t)}|\mathbf{X}^{(t)}, \boldsymbol{\theta}^{(t)}) + \sum_{k=1}^{M} \log p_k(\mathbf{Y}^{(s_k)}|\mathbf{X}^{(s_k)}, \boldsymbol{\theta}^{(s)}). \end{aligned} \tag{17}$$

Each term in (17) is independently computed and given by (10), for $v = t, s_1, s_2, ..., s_M$. This means that the complexity of calculating (17) is much lower than using (10) on the source and target data combined, or just on the source data as is needed in the pGP technique. This is because inverting one big matrix is considerably more expensive than inverting $M + 1$ much smaller ones. Also, in equation (17) we have a sum of two terms where one depends on $\boldsymbol{\theta}^{(s)}$ and the other depends on $\boldsymbol{\theta}^{(t)}$. This means that we can learn these two sets of hyper-parameters independently from one another. If we have a fixed source set, than for each new target that we need to do predictions on, we only need to retrain $\boldsymbol{\theta}^{(t)}$, which can bring even more computational savings.

When the GPDE is trained, we need to combine the prediction of each expert into an overall prediction. The predictive distribution is given by

$$\begin{aligned} p(f_*|\mathbf{x}_*, \mathcal{D}, \boldsymbol{\theta}^{(s,t)}) &= p(f_*|\mathbf{x}_*, \mathcal{D}^{(t)}, \boldsymbol{\theta}^{(t)}) \\ &\prod_{k=1}^{M} p(f_*|\mathbf{x}_*, \mathcal{D}^{(s_k)}, \boldsymbol{\theta}^{(s)}). \end{aligned} \tag{18}$$

Instead of a regular GP for each source expert, we can adapt it to the target data using the pGP technique. Then, the predictive distribution is given by

$$\begin{aligned} p(f_*|\mathbf{x}_*, \mathcal{D}, \boldsymbol{\theta}^{(s,t)}) &= p(f_*|\mathbf{x}_*, \mathcal{D}^{(t)}, \boldsymbol{\theta}^{(t)}) \\ &\prod_{k=1}^{M} p(f_*|\mathbf{x}_*, \mathcal{D}^{(s_k)}, \mathcal{D}^{(t)}, \boldsymbol{\theta}^{(s)}). \end{aligned} \tag{19}$$

In both of these approaches, when predicting unseen target inputs $\mathbf{x}_*^{(t)}$, each expert predicts the mean $\mu_*^{(v)} = \mu^{(v)}(\mathbf{x}_*^{(t)})$ and variance $V_*^{(v)} = V^{(v)}(\mathbf{x}_*^{(t)})$ independently, where $v = t, s_1, s_2, ..., s_M$. The joint prediction is obtained as the product of all experts predictions, which is proportional to a Gaussian distribution with variance and mean [10]

$$V_*^{(gpde)} = \left[(V_*^{(t)})^{-1} + \sum_{k=1}^{M} (V_*^{(s_k)})^{-1}\right]^{-1}, \tag{20}$$

$$\mu_*^{(gpde)} = (V_*^{(gpde)}) \left[(V_*^{(t)})^{-1} \mu_*^{(t)} + \sum_{k=1}^{M} (V_*^{(s_k)})^{-1} \mu_*^{(s_k)} \right]. \quad (21)$$

The strength of this model is that the overall prediction $p(f_*|\mathbf{x}_*, \mathcal{D}, \boldsymbol{\theta}^{(s,t)})$ is straightforward to compute. A shortcoming is that with an increasing number of experts the predictive variances vanish (the precisions add up), which leads to overconfident predictions, especially in regions without data [10]. Thus this model is inconsistent in the sense that it doesn't fall back to the prior far from the training data.

The importance of experts can be increased as

$$p(f_*|\mathbf{x}_*, \mathcal{D}, \boldsymbol{\theta}^{(s,t)}) = p(f_*|\mathbf{x}_*, \mathcal{D}^{(t)}, \boldsymbol{\theta}^{(t)})^{\beta_t} \prod_{k=1}^{M} p(f_*|\mathbf{x}_*, \mathcal{D}^{(s_k)}, \mathcal{D}^{(t)}, \boldsymbol{\theta}^{(s)})^{\beta_{s_k}}. \quad (22)$$

where β_k is the weight (or the contribution) of the k-th expert. The predictive variance and mean are, therefore,

$$V_*^{(gpde)} = \left[\beta_t (V_*^{(t)})^{-1} + \sum_{k=1}^{M} \beta_{s_k} (V_*^{(s_k)})^{-1} \right]^{-1}, \quad (23)$$

$$\mu_*^{(gpde)} = (V_*^{(gpde)}) \left[\beta_t (V_*^{(t)})^{-1} \mu_*^{(t)} + \sum_{k=1}^{M} \beta_{s_k} (V_*^{(s_k)})^{-1} \mu_*^{(s_k)} \right]. \quad (24)$$

The strength of this model is that if $\beta_t + \sum_{k=1}^{M} \beta_{s_k} = 1$ the predictive distribution falls back to the prior outside the range of the training data [10], so the model doesn't make overconfident predictions in that region anymore. A weakness is that inside the range of the data the model tends to overestimate the variances, especially with an increasing number of GP experts.

In this paper, we will not consider the effect of tuning the parameters β_k. They will be set to $\beta_k = \frac{1}{M+1}$, where $M+1$ is the number of experts. In this setting, the prediction means (21) and (24) are identical, but the variance differs in a sense that the variance in equation (20) starts to resemble the prior variance outside the training data range, as mentioned above.

III. Experiments

A. Data

For this article we collected data from the ADNI database (adni.loni.usc.edu). We used the standard dataset processed for TADPOLE challange [5]. This dataset contains 1737 unique subjects and was created from the ADNIMERGE spreadsheet, to which regional MRI (volumes, cortical thickness, surface area), PET (FDG, AV45, AV1451), DTI (regional means of standard indices) and cerebrospinal fluid (CSF) biomarkers were added. We used all the features from this data, except for the normalized verticle volumens (ICVn) and all cognitive scores except ADAS-Cog13. The task was to predict the Alzheimer's Disease Assessment Scale-Cognition Sub-Scale (ADAS-Cog13) score. The subjects clinical status (CS) was used to divide the source dataset into subsets. With that data, multi-modal feature set was constructed from six modalities: demographics (6 features), genetics (3 features), cognitive tests (9 features), CSF (3 features), MRI (365 features), and DTI (229 features). Due to sparseness, we excluded PET data entirely. We used a cohort of 100 subjects with each having 21 planned visits every 6 months. Each subject has no more than 10 visits missing and no more than 82.5% features missing for the experiments.

B. Models

All the models are first-order autoregressive, meaning that they are using the features of the current visit, along with the ADAS-Cog13 current cognitive score $(\mathbf{x}_t, y_t)$ as an input to the model. The goal is to predict the ADAS-Cog13 cognitive score y_{t+1} for the next visit. When we are predicting y_{t+1} for a specific patient, the source data will consist of all other patients we have in the test set, while the target data will consist of all the visits for that patient up to visit $t-1$, and the input to the model will be $(\mathbf{x}_t, y_t)$. We evaluated five models on this dataset:

- **Source GP (sGP)**: A regular GP, as mentioned in section II-A. The training set will consist of the source data, and the predictive distribution is given with equations (7) and (8).
- **Personalized GP (pGP)**: This is the GP adaptation technique explained in II-B. The predictive distribution is given with equations (13) and (14).
- **Target GP (tGP)**: A regular GP, as mentioned in section II-A. The training set will consist of the target data, and the predictive distribution is given with equations (7) and (8). Because of the small amount of subject visits and common missing data, the hyper-parameters for this model are learned on the source data.
- **GP Domain Experts (GPDE)**: This is the technique explained in II-C. This model uses the expert contribution parameters β_i (all experts have the same contribution, and they all sum to 1) and breaks the source data into three subgroups based on the patients clinical status:
  - Subjects who are cognitively normal (CN).
  - Subjects who are mildly cognitively impaired (MCI) or have converted from CN to MCI.
  - Subjects who have Alzheimer disease (AD), or have converted from CN/MCI to AD.

  The predictive distribution is given by equations (23) and (24). The target expert hyper-parameters are taken to be $\boldsymbol{\theta}^{(s)}$, which are trained only on the second term of equation (17).
- **Personalized GP Domain Experts (pGPDE)**: This is the technique explained in II-C. This model uses the expert contribution parameters β_i (all experts have the same contribution, and they all sum to 1) and breaks the source data into three subgroups based on the patients clinical status. First, the model adapts each source expert using equations (13) and (14). Then, it computes the predictive distribution using equations (23) and (24). The

target expert hyper-parameters are taken to be $\boldsymbol{\theta}^{(s)}$, which are trained only on the second term of equation (17).

The sGP, pGP and tGP models are used in [2] and [8] to predict future cognitive scores. We propose using GPDE and pGPDE for the same task, and compare the results. All of the models use the squared-exponential kernel.

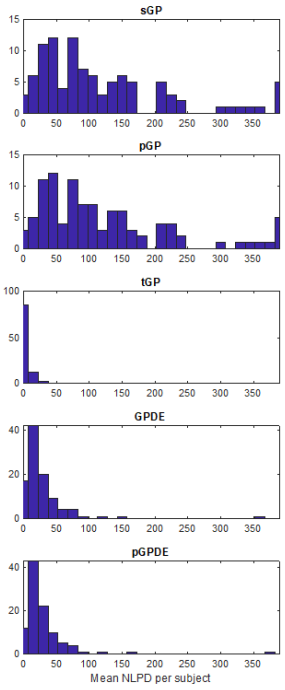


Fig. 1. The histograms of the NLPD metric averaged per subject for all of the considered models during the leave-one-subject-out validation procedure.

C. Evaluation

All models are evaluated using the leave-one-out validation method—each subject is selected for predictions, while the other 99 make up the source dataset and all the results are averaged at the end. Each subject that is selected for prediction has 21 doctor visits planned. When he missed a visit, the last available ADAS-Cog13 score was put in the database. Missing features were imputed as zero values. When computing the final metrics, the predictions which have no true label (due to missed visits) are not used. As the evaluation metric, we use the mean absolute error (MAE) defined as:

$$\frac{1}{N}\sum_i |y^{(i)}_{\text{pred}} - y^{(i)}_{\text{true}}|, \tag{25}$$

which encodes the error in the point estimates by the model. However, it fails to encode the model's uncertainty in its prediction. For this, we compute the mean negative log predictive density (MNLPD), defined as:

$$\begin{aligned}&\frac{1}{N}\sum_i -\log p(y^{(i)}_{\text{true}}|\mathbf{x}^{(i)}, \mathcal{D}, \boldsymbol{\theta}) \\ =&\frac{1}{N}\sum_i \left(\frac{1}{2}\log {\sigma^{2}}^{(i)}_{\text{pred}} + \frac{(y^{(i)}_{true} - \mu^{(i)}_{\text{pred}})^2}{2{\sigma^{2}}^{(i)}_{\text{pred}}} + \frac{1}{2}\log 2\pi\right),\end{aligned} \tag{26}$$

which, in addition to the point estimates, also accounts for the predictive variance. Note that if the prediction is close to the true value, but the model is not confident in its prediction, it receives a lower MNLPD score than a model that has the same good prediction but a lower variance. Similarly, if the prediction is less accurate but the model is overconfident, that is rated worse than when the model is uncertain about its prediction.

D. Results

Experimental results are shown in Table I. By looking into the performance measures for all subjects (all theree subgroups trained together), we note that while pGP and pGPDE models have similar MAE for their point estimates, the MNLPD of the latter is much lower. This can be noticed more clearly by looking at Fig. 1 and it signals that pGPDE provides better estimates of prediction uncertainty (i.e. higher variance when the point estimates are further away from the true label). Also, by looking at Fig. 1 we notice that the GPDE model, which is also based on the idea of combining experts, achieves a relatively low MNLPD per subject. Correctly encoding the model's uncertainty is important for the interpretation of and confidence in the predicted target scores. While looking at the performance measures of the CN subjects, it turns out that the tGP model is best both in terms of MAE and MNLPD. This is because the ADAS-Cog13 score changes very slowly for these subjects, and the tGP model looks only at the previous visits for the observed subject. Thus, a simple smoothing of the subject's previous scores leads to reasonably accurate estimates. When we look at the MCI subjects we notice that the proposed GPDE and pGPDE models achieve the lowest MAE, with a relatively low MNLPD. It is interesting to notice that both GPDE and pGPDE have almost the same MAE. It seems that the effects of the model personalization via the GP posteriors are less pronpunced here. This may be due to the fact that the GPDE model is already performing another type of the model-level personalization by different weighting of the expert models for each test data. Among the expert models, there is an expert that uses the MCI dataset subgroup and that expert could have the largest contribution in the prediction. Finally, when we look at the AD subjects, the pGP model has the best MAE, but its MNLPD is quite poor. The pGPDE achieves a slightly higher MAE, but with a much lower MNLPD, so it can be more useful in real-world applications. All the models evaluated on the AD subjects have considerably larger MNLPD. This tells us that it is an extremely difficult task to predict accurately the scores of the AD subjects, where the changes in their individual scores vary a lot. We plan to investigate alternative strategies for tackling this in our future work.

The tGP model always achieves the lowest MNLPD, but has a relatively high MAE. This is because it makes predictions using a really small target dataset as its training set, so his predictive variance is almost always high. This model makes mistakes often, but it isn't certain in those predictions so the MNLPD doesn't penalize that. Looking only at the MNLPD this model could be declared as the best, but it isn't because

TABLE I

THIS TABLE PRESENTS THE LEAVE-ONE-SUBJECT-OUT VALIDATION RESULTS OF ALL THE MODELS. PERFORMANCE MEASURES WERE FIRST AVERAGED FOR EACH SUBJECT, AND THE MEAN VALUE OF THOSE MEASURES IS PRESENTED IN THIS TABLE ALONG WITH ONE STANDARD DEVIATION. THE METRICS ARE FIRST AVERAGED ON ALL THE SUBJECTS, THEN ON THE SUBJECTS FROM ALL OF THE THREE CLINICAL STATUS GROUPS. DATAPOINTS WITH THE MISSING VISIT y_{t+1} WERE DISCARDED FROM THE EVALUATION.

Models	**All subjects**		**CN subjects**		**MCI subjects**		**AD subjects**	
	MAE	**MNLPD**	**MAE**	**MNLPD**	**MAE**	**MNLPD**	**MAE**	**MNLPD**
sGP	4.01 ± 1.7	142.67 ± 227.54	2.96 ± 1.14	72.9 ± 75.18	3.53 ± 0.99	82.8 ± 55.83	4.82 ± 1.85	212.68 ± 306.13
pGP	$\mathbf{3.77 \pm 1.61}$	145.21 ± 229.42	2.75 ± 1.08	74.60 ± 76.65	3.28 ± 0.84	84.71 ± 56.45	$\mathbf{4.57 \pm 1.76}$	216.02 ± 308.44
tGP	4.32 ± 2.44	5.4 ± 4.81	$\mathbf{2.41 \pm 0.8}$	$\mathbf{2.77 \pm 0.45}$	3.34 ± 0.91	3.34 ± 0.85	5.87 ± 2.61	7.93 ± 5.95
GPDE	3.92 ± 1.94	$\mathbf{29.71 \pm 41.88}$	2.75 ± 1.02	15.44 ± 13.25	$\mathbf{3.08 \pm 0.87}$	$\mathbf{15.45 \pm 10.1}$	5.01 ± 2.15	$\mathbf{45.16 \pm 55.19}$
pGPDE	$\mathbf{3.79 \pm 1.83}$	$\mathbf{30.79 \pm 43.11}$	2.6 ± 1	16.02 ± 13.77	$\mathbf{3.07 \pm 0.86}$	$\mathbf{16.27 \pm 10.46}$	4.82 ± 1.99	$\mathbf{46.65 \pm 56.81}$

TABLE II

AVERAGE TRAINING TIME, ALONG WITH ONE STANDARD DEVIATION, FOR MODELS THAT USE THE WHOLE SOURCE SET VERSUS EXPERT MODELS THAT USE THE SOURCE SUBGROUPS (CN, MCI, AD). THIS IS OBTAINED BY THE LEAVE-ONE-SUBJECT-OUT VALIDATION PROCEDURE.

Models	**Average training time**
Single expert models (sGP, pGP, tGP)	54.78 ± 2.72s
Domain expert models (GPDE, pGPDE)	17.5 ± 0.94s

it is very often uncertain in his predictions, while also having a relatively high error.

The average training time during the leave-one-out validation is shown in II. We can see that the expert models, GPDE and pGPDE, which break the source dataset of size N into three subgroups ($\sim \mathcal{O}\left(3(\frac{N}{3})^3\right)$ complexity), need considerably less time to learn the hyper-parameters than other models that learn hyper-parameters on the whole source dataset ($\mathcal{O}(N^3)$ complexity). In this article we used a relatively small cohort of 100 subjects. In practical application, the cohort could be much larger and the difference in the training times could be much bigger. Also, if the datasets were too big for matrix inversions and storage, using the GP experts could make the problem more computationally tractable.

IV. CONCLUSION

We proposed using the GPDE and pGPDE models for predicting a subject's next visit ADAS-Cog13 cognitive score, with the current score and current genetic, demographic, DTI, MRI and CSF features as inputs to the models. We compared their results with the standard and domain adaptive GP models which were used on the same dataset in previous research [2], [8]. We show that the expert models tend to be more confident in more accurate point estimates and less confident in less accurate point estimates than the previously proposed domain adaptive GPs that employ a single expert (learn using the whole training population data), while achieving a relatively good mean absolute error. Also, we show that the expert models have a clear advantage in terms of the required training time and point out that this would have a much larger impact in practical settings, where the datasets would be much larger.

ACKNOWLEDGMENT

This research was supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia, Contracts No. TR32038 and III42007. The work of O. Rudović is funded by the European Union H2020, Marie Curie Action - Individual Fellowship no. 701236 (EngageMe).

REFERENCES

[1] M. W. Weiner, D. P. Veitch, P. S. Aisen, L. A. Beckett, N. J. Cairns, R. C. Green, D. Harvey, C. R. Jack, W. Jagust, J. C. Morris *et al.*, "Recent publications from the alzheimer's disease neuroimaging initiative: Reviewing progress toward improved ad clinical trials," *Alzheimer's & Dementia*, 2017.

[2] Y. Utsumi, O. Rudovic, K. Peterson, R. Guerrero, and R. W. Picard, "Personalized gaussian processes for forecasting of alzheimer's disease assessment scale-cognition sub-scale (adas-cog13)," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 4007–4011.

[3] J. Skinner, J. O. Carvalho, G. G. Potter, A. Thames, E. Zelinski, P. K. Crane, L. E. Gibbons, A. D. N. Initiative *et al.*, "The alzheimer's disease assessment scale-cognitive-plus (adas-cog-plus): an expansion of the adas-cog to improve responsiveness in mci," *Brain imaging and behavior*, vol. 6, no. 4, pp. 489–501, 2012.

[4] R. C. Mohs, D. Knopman, R. C. Petersen, S. H. Ferris, C. Ernesto, M. Grundman, M. Sano, L. Bieliauskas, D. Geldmacher, C. Clark *et al.*, "Development of cognitive instruments for use in clinical trials of antidementia drugs: additions to the alzheimer's disease assessment scale that broaden its scope." *Alzheimer disease and associated disorders*, 1997.

[5] D. Alexander, F. Barkhof, F. Nick, E. Bron, and A. Toga, "The alzheimer's disease prediction of longitudinal evolution (tadpole) challenge," May 2017. [Online]. Available: https://tadpole.grand-challenge.org/home/

[6] J. L. Cummings, "Challenges to demonstrating disease-modifying effects in alzheimer's disease clinical trials," *Alzheimer's & Dementia*, vol. 2, no. 4, pp. 263–271, 2006.

[7] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[8] K. Peterson, O. Rudovic, R. Guerrero, and R. Picard, "Personalized gaussian processes for future prediction of alzheimer's disease progression," *International Conference on Neural Information Processing Systems (Machine Learning for Health Workshop)*, 2017.

[9] S. Eleftheriadis, O. Rudovic, M. P. Deisenroth, and M. Pantic, "Gaussian process domain experts for modeling of facial affect," *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4697–4711, 2017.

[10] M. Deisenroth and J. W. Ng, "Distributed gaussian processes," in *International Conference on Machine Learning*, 2015, pp. 1481–1490.

Estimacija trajektorije i kinematičkih parametara pasivnom monosenzornom kamerom

Marko Antonijević, Filip Ilić, *Vojna Akademija, Univerzitet Odbrane, Beograd*

***Apstrakt*— U ovom radu je analizirana procena daljine i kinematičkih parametara cilja na osnovu površine cilja, koja je izdvojena osnovnim metodama obrade slike. Korišćeni su rezultati eksperimenta u kojima je simulirano kretanje cilja. Model Sistema za pasivnu osmatranje i praćenje činili su VC-C50i kamera i PC računar. Za digitalnu obradu slike korišćen je MATLAB softver. Dobijeni rezultati prikazuju trajektoriju i osnovne kinematičke parametre cilja kao što su pozicija, brzina i ugaona brzina.**

***Ključne reči*—Estimacija trajektorije, praćenje cilja, upravljanje kamerom, estimacija kinematičkih parametara**

I. UVOD

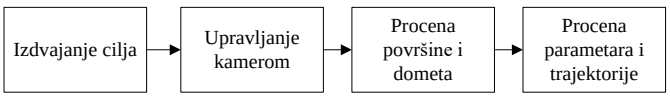
Sistemi sa pasivnim senzorima imaju mnoge primene, najčešće za video nadzor, bezbednost, kontrolu vazdušnog prostora, navođenje raketa, u robotici. Zbog povećane upotrebe bespilotnih letelica ,dronova i drugih letelica male brzine u vojne svrhe,njihova detekcija konvencionalnim radarskim sistemima je otežana. Sistemi sa pasivnim pasivnim senzorima u ovim primenama je pogotovo izražena, a najviše u sistemima za proenu daljine i praćenje. Druga važna prednost primene sistema sa pasivnim senzorom u vojne svrhe je što tokom osmatranja ne emituju zračenje tako da ne odaju položaj i time pružaju bezbednost operaterima.

Rad [1] analizira procenu daljine bespilotne letelice, a radovi [2,3] pokazuju procenu daljine digitalnom obradom slike. Takođe, čest problem koji se rešava pasivnim monosenzornim sistemima je procena daljine u robotici [4].

Cilj ovog rada je da se na temelju znanja iz obrade slike pokaže moderan algoritam za obradu slike i procenu daljine, a da se zatim to znanje potkrepi eksperimentalnim rezultatima dobijenim kontrolisanom kamerom

II. BLOK DIJAGRAM SISTEMA

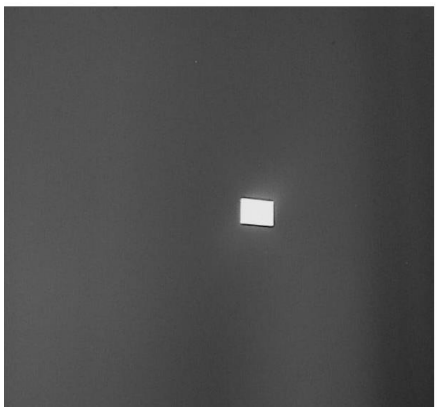
Sistem za praćenje sa video senzorom može se opisati blok šemom sa slike 1. Prvi blok služi za obradu slike koja odvaja cilj od okruženja u svakom frejmu. Drugi blok služi za kontrolu step motora kamere na osnovu pozicije cilja na slici. Tu se računaju pomeraji uglova kamere i komunicira sa kamerom. Sledeći blok procenjuje daljinu na osnovu površine cilja. Poslednji blok iscrtava trajektoriju clja i daje informacju o kinematičkim i drugim značajnim parametrima cilja .



Slika 1. Blok šema sistema

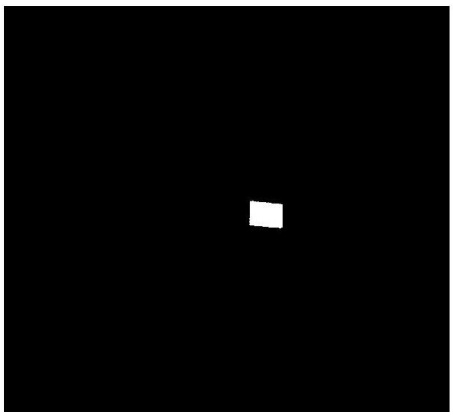
III. BLOK ZA IZDVAJANJE CIJJA

Da bi se izračunala trajektorija cilja prvo je potrebno da se cilj odvoji od okoline. Početna slika sadrži cilj sa okruženjem (Slika 2.).



Slika 2. Cilj sa okruženjem

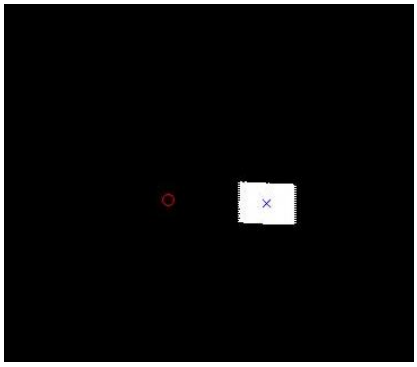
Slika se zatim obrađuje Kapur metodom. Ova metoda izračunava prag . Time se vrši binarizacija slike i svi pikseli koji su prešli izračunati prag dodeljuju se cilju.



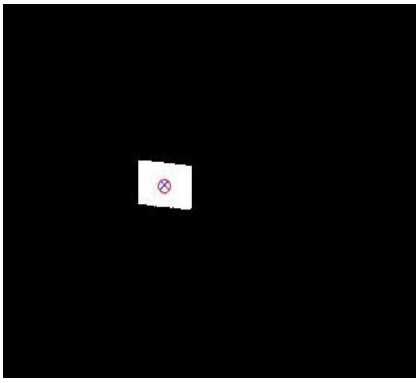
Slika 3. Cilj nakon binarizacije

IV. BLOK ZA KONTROLU KAMERE

Kada se cilj izdvoji računa se njegov centar.U svakom frejmu računa se daljina u pikselima između centra cilja i centra slike. Ta razlika se zatim pretvara u u uglove po azimutu i elevaciji. Izračunati uglovi se pretvaraju u ASCII kod koji predstavlja pomeraje step motora koji kontrolišu kameru. VC-C50i kamera sa računarom komunicira preko serijskog porta. Centar cilja i centar kamere se zatim poravnavaju u svakom frejmu i kamera prati cilj (slika 4, 5)



Slika 4. Cilj je detektovan



Slika 5. Cilj je centriran

Step motori korišćene kamere dozvoljavaju maksimalni pomeraj od +/− 90° po azimutu i +/− 10° po elevaciji. Takođe maksimalna brzina okretanja kamere je 90°/s što može predstavljati problem ako se cilj kreće brže ili izvan zadatih granica.

V. BLOK PROCENU POVRŠINE I DALJINE

Kada je cilj izdvojen i centriran moguće je izračunati površinu cilja a na osnovu površine proceniti daljinu do cilja .Korišćene metode date su u [5-8]. Matematička relacija koja opisuje zavisnost površine i daljine je:

$$D_e(k) = D_0\sqrt{\frac{S_0}{S_e(k)}}. \qquad (1)$$

gde su: $D_e(k)$ procenjena daljina do cilja, $S_e(k)$ procenjena površina cilja na k-toj slici, a D_0 i S_0 daljina do cilja i površina na prvoj slici. Što znači da je za estimaciju potebno poznavati unapred samo daljinu od cilja u početnom trenutku.

VI. BLOK ZA PROCENU TRAJEKTORIJE I KINEMATIČKIH PARAMETARA CILJA

Koriseteći procenjenu daljinu i poznate uglove azimuta i elevacije koje kamera vraća u svakom frejmu moguće je estimirati i iscrtati trajektoriju.Koriste se poznate relacije iz fizike:

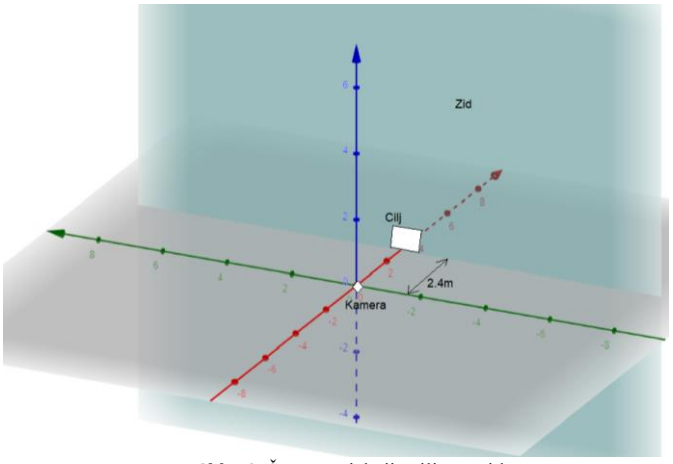
$$v_x = \frac{dx(t)}{dt}, v_y = \frac{dy(t)}{dt}, v_z = \frac{dz(t)}{dt}. \qquad (2)$$

$$\omega_a = \frac{d\varphi(t)}{dt}, \omega_e = \frac{d\phi(t)}{dt}. \qquad (3)$$

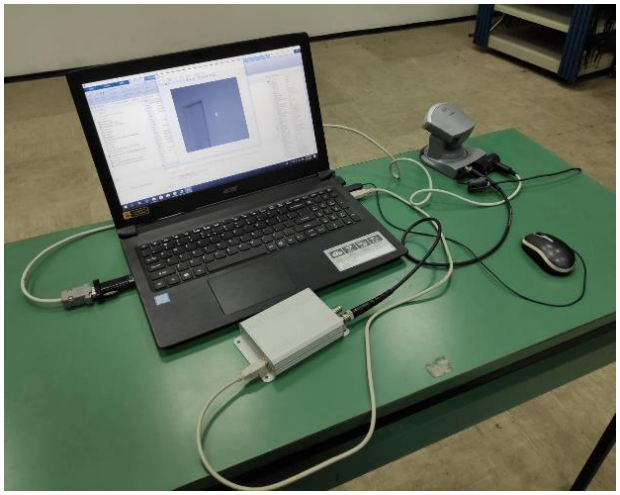
Pri čemu su v_x, v_y i v_z komponente linijske brzine, φ ugao azimuta , ϕ ugao elevacije ,a ω_a i ω_e ugaone brzine azimuta i elevacije respektivno.Na osnovu relacija (2,3) moguće je proceniti osnovne kinematičke parametre cilja kao što su linijska i ugaona brzina.

VII. POSTAVKA EKSPERIMENTA

Cilj je u ovom eksperimentu bio je pravougaonik projektovan na zid (slika 6.) čije je rastojanje od kamera unapred izračunato.Kamera je priključena na računar preko COM porta i komunicirala je RS-232 komunikacijom sa zadatkom da prati cilj (slika 7.), dok se na računaru obrađuje slika. Cilj se kretao po zadatoj trajektoriji.



Slika 6. Šema projekcije cilja na zid

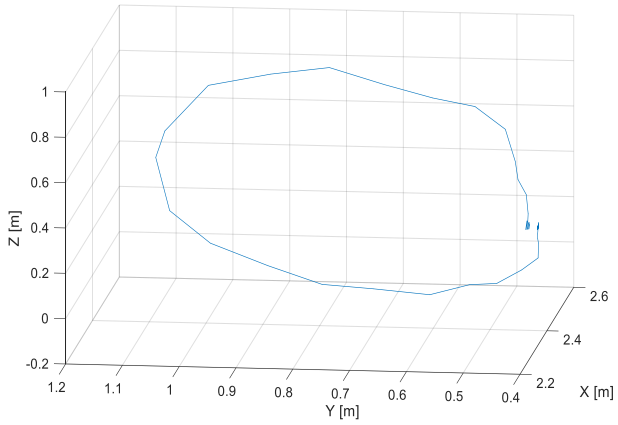


Slika 7. Kamera povezana sa računarom na kome se vrši obrada slike

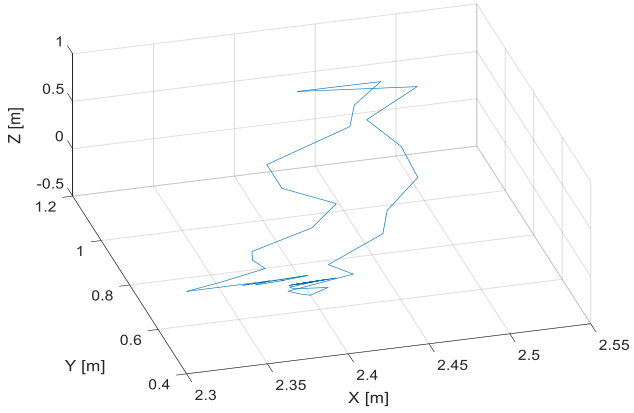
Korišćene su dve vrste trajektorija. Prvu vrstu su činile kružna i trougaona trajektorija sa ciljem da se na osnovnim trajektorijama kao što su gore navedene izračuna greška koju kamera pravi u proceni daljine do cilja. Druga vrsta trajektorija je bila iscrtana slobodnom rukom sa ciljem da se simulira let cilja što verniji realističnoj situaciji.

VIII. REZULTATI PROCENE TRAJEKTORIJE

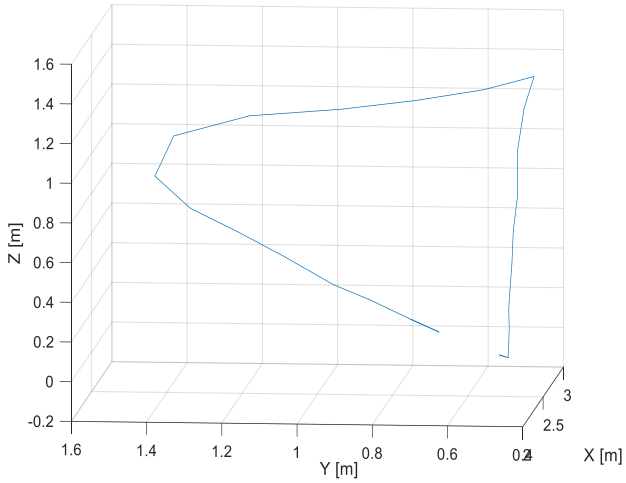
Merenjem kamerom dobijeni su sledeći rezultati za procenu trajektorije:



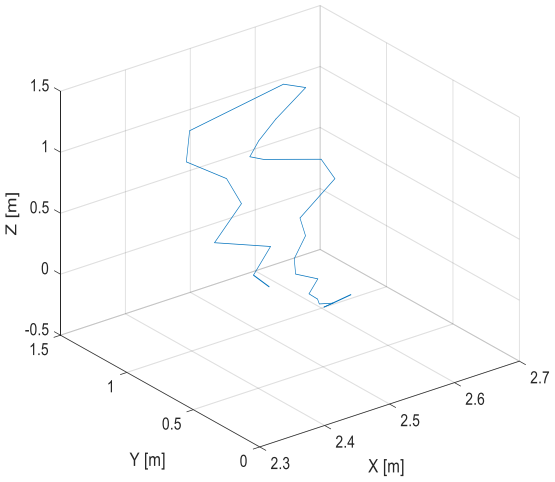
Slika 8. Kružna trajektorija (pogled iz ugla kamere)



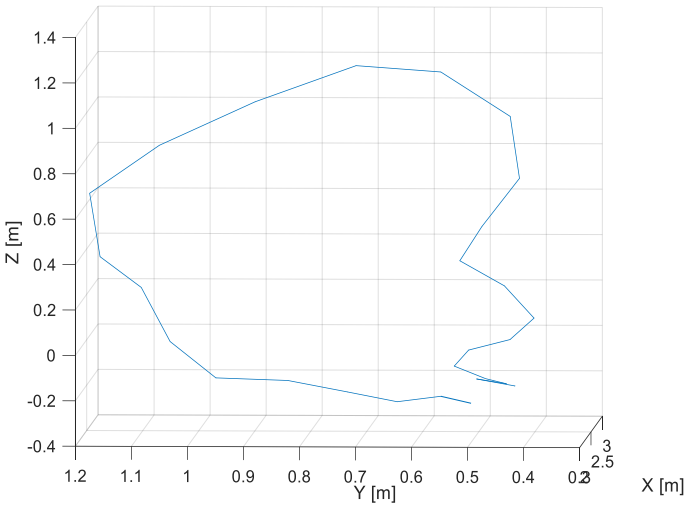
Slika 9. Kružna trajektorija (pogled sa strane)



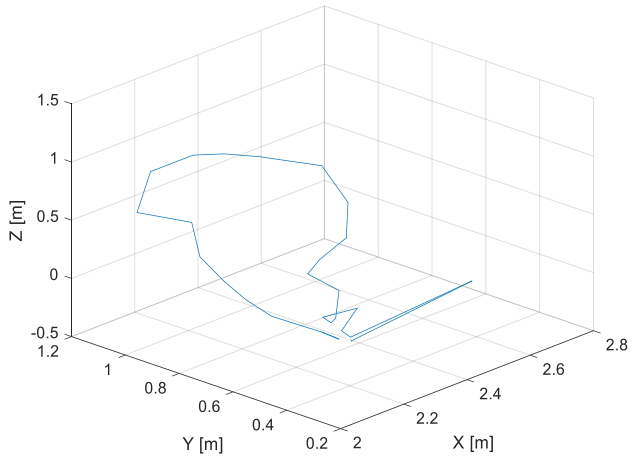
Slika 10. Trougaona trajektorija (pogled iz ugla kamere)



Slika 11. Trougaona trajektorija (pogled sa strane)



Slika 12. Trajektorija nacrtana slobodnom rukom (iz ugla kamere)

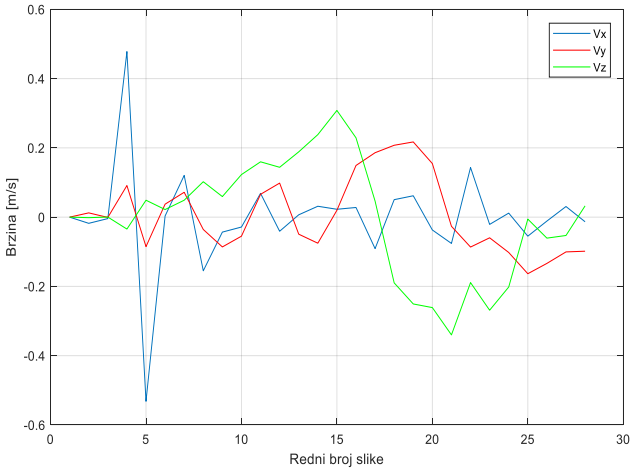


Slika 13. Trajektorija nacrtana slobodnom rukom (pogled sa strane)

Sa slika (8-13) uočava se da postoji greška u proceni trajektroije. Grafik priakzuje kretanje i po x-osi iako ono postoji samo u yOz ravni.Ovo je posledica nemogućnosti kamere da trenutno izoštri sliku. Zato što je cilj zamućen pravi se procena površine cilja, a samim tim i greška procene daljine. Ova greška raste sa porastom brzine cilja ,a manja je ukoliko je cilj dalje od kamere. Greška proocenene ne prelazi 6.4% za trougaonu i kružnu trajektroiju ,a 4.8% za trajektoriju nacrtanu slobodnom rukom što je prihvatljivo.

IX. PROCENJENI KINEMATIČKI PARAMETRI

S obzirom da imamo poznatu poziciju cilja u svakom trenutku, možemo odrediti brzinu i koristeći (2-6). Razmatrana je samo trajektorija crtana slobodnom rukom zato što za tu trajektoriju kretanje cilja najviše odgovara realističnom. Dobijeni su sledeći rezultati:

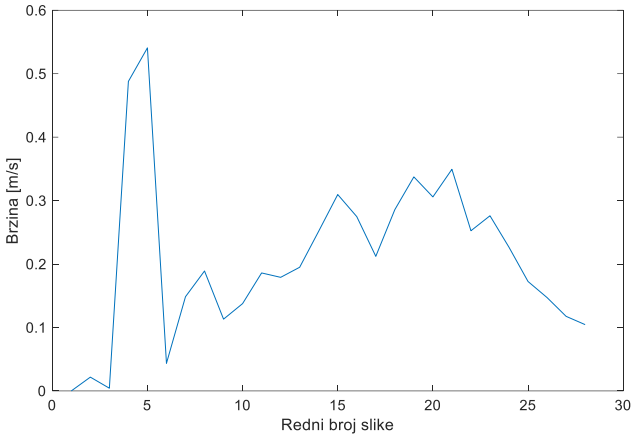


Slika 14. Promena komponenti linijske brzine po svakoj od koordinata

Kada su poznate komponente brzine linijsku brzinu moguće je odrediti prema formuli:

$$v = \sqrt{v_x^{\,2} + v_y^{\,2} + v_z^{\,2}}\,. \tag{7}$$

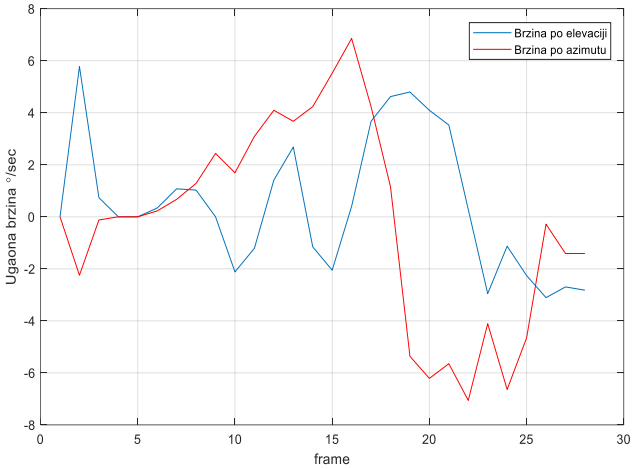
Dobijeni su rezultati:



Slika 15. Promena linijske brzine

Sa slika 14 i 15 može se primetiti da veliki skok brzine po x-osi , uzrokovan lošom procene daljine, izazvao i skok linijske brzine u 4 i 5 frejmu. U narednim frejmovima greška se smanjuje, rad kamere je pravilniji i brzina se kreće oko srednje vrednosti od 2.8 m/s bez značajnijih odstupanja.

Kamera u svakom trenutku vraća i vrednosti uglova na osnovu kojih se procenjuje ugaona brzina:

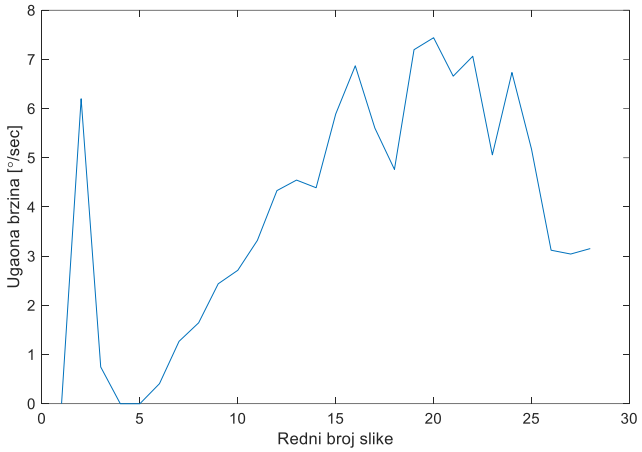


Slika 16. Promena komponenti ugaone brzine

Na osnovu vrednosti komponenti moguće je izračunati vredost ugaone brzine koristeći relaciju:

$$w = \sqrt{w_e^{\,2} + w_a^{\,2}}\,. \tag{8}$$

Dobijene vredosti ugaone brzine su:



Slika 17. Promena ugaone brzine

Ugaona brzina ima veliki skok samo između 2 i 4 frejma kada cilj pravi najbrži manevar. Uglovi na osnovu kojih je izračunata ugaona brzina dobijaju se sa kamere i ne zavise od obrade slike. Poređenjem slike 17. sa slikom 15. dolazi se do zaključka da je veliki skok u 4 i 5 frejmu zaista posledica pogrešne procene daljine.

X. ZAKLJUČAK

Ovim radom eksperimentalno je pokazano da je moguće uspešno detektocati i pratiti cilj koristeći VC-C50i kameru, računar i softverski paket MATLAB. Praćenje kamere je funkcionisalo na osnovu obrade slike i računanju razlike centra slike i centra cilja. Obrada slike omogućavala procenu daljine, samim tim i računanje kinematičkih parametara. Procenjena daljina omogućava proračun trajektorije cilja ,a izračunati parametri omogućavaju klasifikaciju ciljeva prema brzini.

U budućim radovima biće implementiran i Klamanov filtar koji omogućuje predikciju kretanja cilja i povećava tačnost estimacije trajektorije i parametara.

LITERATURA

[1] Shakernia, Omid, Won-Zon Chen, and Vince Raska. "Passive ranging for UAV sense and avoid applications." *Infotech@ Aerospace*. 2005. 7179.

[2] Atherton, Tim J., Darren J. Kerbyson, and Graham R. Nudd, Passive estimation of range to objects from image sequences, BMVC91. Springer London, 1991. 343-346.

[3] Anderson, Joel R. Monocular passive ranging by an optical system with band pass filtering. No. AFIT/GAP/ENP/10-M01. Air force inst of tech wright-patterson afb oh graduate school of engineering and management, 2010.

[4] Wahab, M. N. A., N. Sivadev, and K. Sundaraj. "Target distance estimation using monocular vision system for mobile robot." *2011 IEEE Conference on Open Systems*. IEEE, 2011.

[5] Z. P. Barbaric, B. P. Bondzulic, S. T. Mitrovic, Passive ranging using image intensity and contrast measurements, Electronics Letters, 48(18), pp. 1122– 1123, (2012), doi:10.1049/el.2012.0632.

[6] B. P. Bondžulić, S. T. Mitrović, Ž. P. Barbarić, M. S. Andrić, "A comparative analysis of three monocular passive ranging methods on real infrared sequences", Journal of Electrical Engineering 64(5), pp. 305–310, (2013), doi:10.2478/jee-2013-0044.

[7] Mikluc, D. L., Andrić, M. S., Mitrović, S. T., & Bondžulić, B. P. (2017). Improved method for passive ranging based on surface estimation of an airborne object using an infrared image sensor. Optica Applicata, 47(3), 383-394

[8] Kapur, J. N., Sahoo, P. K., & Wong, A. K. (1985). A new method for gray-level picture thresholding using the entropy of the histogram. Computer vision, graphics, and image processing, 29(3), 273-285.

ABSTRACT

In this paper distance and kinematic parameter estimation based on the area of the target which has been extracted using basic image processing methods has been analyzed. The results from the experiment designed to simulate target movement were used. The model for target tracking was made of the controlled camera, VC-C50i and a personal computer. The results show the calculated trajectory and basic kinematic parameters such as speed and angular speed.

Trajectory and kinematic parameter estimation using passive monosensor camera

Marko Antonijević, Filip Ilić

Analiza kvaliteta estimacije u sistemu za praćenje više ciljeva sa video senzorom u zavisnosti od šuma merenja

Filip Ilić, Marko Antonijević, *Vojna Akademija, Univerzitet Odbrane, Beograd*

***Apstrakt*—U ovom radu je izvršena višestruka analiza kvaliteta estimacije u sistemu za praćenje više ciljeva. Sistem za praćenje više ciljeva sadrži senzor slike na osnovu kojeg su formiranja merenja. Gausov šum merenja različitih vrednosti varijansi je dodat generisanim trajektorijama četiri bliska cilja. Upoređeni su rezultati procene pozicije ciljeva na slici u zavisnosti od primene algoritama za pridruživanje podataka: globalno najbliži sused, najboljih *n* rešenja i Munkresov algoritam. Druga analiza je izvršena kroz uticaj proračuna ukupne estimacije stanja, dobijena na osnovu težinskih koeficijenata u interaktivnom višestrukom modelu za estimaciju.**

***Ključne reči*— Pridruživanje podataka; Praćenje više ciljeva; Video senzor; Estimacija.**

I. UVOD

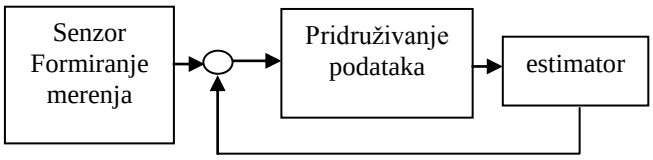
Kada se prikupljaju podaci sa nekog senzora, veoma je bitno imati na umu greške koje taj sensor unosi prilikom merenja. Unešene greške tokom merenja mogu dovesti do toga da obrađeni rezultati budu neupotrebljivi u određene svrhe. Shodno tome treba ispitati do kojih granica i pod kojim uslovima su greške dozvoljene a da rezultati budu merodavni.

Današnji sistemi, čiji su osnovni zadaci praćenje više ciljeva, su veoma rasprostranjeni. Njihova kompleksnost se sastoji u složenim matematičkim algoritmima i operacijama, koje su potrebne u cilju što kvalitetnije procene obeležja ciljeva koji se prate, neki od njih su dati u [1]. Najpre se treba fokusirati na senzor koji formira merenje na osnovu kojeg se može govoriti o sistemima sa pasivnim ili aktivnim senzorima koji su opisani u [2]. Sistemi sa aktivnim senzorima, često imaju male šumove merenja što ima za posledicu nižu grešku procene. S druge strane, sistemi za praćenje sa pasivnim senzorima su u prednosti jer je teško odrediti njihovu lokaciju, ali su metode za procenu često nelinearne i kompleksne. Drugi bitan blok ovakvih sistema je estimator. Kada se govori o linearnim modelima procesa koristi se optimalni estimator, a to je Kalmanov filter, dok se u nelinearnim procesima upotrebljeva prošireni Kalmanov filter, čestični filter ili neki od adaptivnih algoritama estimacije koji su predloženi u [3].
Problem opisanih sistema je dodatno usložen ukoliko se razmatra problem pridruživanja podataka sto je predstavljeno u [4]. Estimator ima osnovnu ulogu da proceni stanja slučajnog procesa, odnosno kada se govori o sistemima za praćenje tada je uloga u proceni položaja cilja koji se prati. Informacije koje su potrebne estimatoru su merenje i prethodna estimacija kako bi formirao novu procenu. Sada se može videti problem koji nastaje u sistemima za praćenje više ciljeva a to je odluka o dodeljivanju merenja prethodno formiranim procenama. Kvalitet odluke o dodeljivanju merenja značajno utiče ne samo na kvalitet procene već i na to da li će se procenjeni ciljevi međusobno zameniti u daljoj obradi.

Kao što je navedeno u [5], kvalitet odluke između ostalog zavisi od šuma procesa merenja. U ovom radu smo pronašli granice do kojih šum procesa merenja može da varira kako bi dao merodavne rezultate. Osnovni kriterijum procene je srednja kvadratna greška.

II. SISTEM ZA PRAĆENJE VIŠE CILJEVA

Sistem za praćenje više ciljeva je složen sistem, koji se sastoji od nekoliko blokova. Opšta blok šema sistema se može predstaviti kao na slici 1.



Sl. 1. Opšta šema sistema za praćenje više ciljeva.

A. Blok za formiranje merenja

Blok za formiranje merenja ima osnovnu ulogu da obradi prikupljene podatke i generiše informacije o cilju koji se prati. Navedeni blok može da sadrži jedan ili više aktivnih ili pasivnih senzora. U ovom radu je fokus na video senzoru, i odabrana je kamera VC-C50i u toj ulozi. Kvalitet sistema za praćenje umnogome zavisi i od kvaliteta senzora kao i od drugih blokova, ali u ovom radu je izvršena jedna od analiza a to je uticaj šuma merenja. Analiza se zasniva na proceni kinematičkih stanja cilja u zavisnosti od dodatog Gausovog šuma generisanim trajektorijama, čime je modelovan šum merenja senzora.

Kamera snima scenario koji je generisan i objašnjen u daljem tekstu. Sa senzora prikupljamo podatke u vidu slike koju konvertujemo u sive slike, a zatim primenjujemo Kittlerovu metodu za određivanje adaptivnog praga inteziteta sive slike kao što je opisano u [6].

B. Blok za estimaciju

U modernim sistemima za praćenje više ciljeva obično se koristi nekoliko Kalmanovih filtera sa različitim modelima kretanja ciljeva koji rade paralelno. Najpoznatiji i danas najviše primenjivan je interaktivni višestruki model (IMM), koji daje dobre rezultate posebno za praćenje ciljeva koji prave složene manevre. IMM algoritam je metoda za kombinovanje stanja hipoteze iz više modela filtera kako bi se poboljšala procena stanja ciljeva sa promenljivom dinamikom kretanja. Osnove IMM algoritma su date u [7], gde je detaljno opisan rad navedenog modela ali i njegove prednosti u odnosu na jednostruke klasične modele.

Blok za estimaciju se sastoji iz algoritma sa interaktivnim višestrukim modelima. IMM algoritam je realizovan pomoću dva Kalmanova filtera u paraleli, prvi je model kretanja sa manjim manevrom a drugi model kretanja sa oštrijim manevrom. Jedan od problema koji se rešava primenom IMM algoritma je proračun estimacije kinematičkih stanja. Međutim kako se za svaki od modela takođe proračunavaju težinski koeficijenti, tako se oni mogu iskoristiti u bloku za pridruživanje podataka. Dakle, formiranje estimacija za svaki model se koristi u bloku za pridruživanje podataka da bi se izračunala statistička distanca, a zatim uz primenu težinskih koeficijenata dobija se konačna statistička distanca. Ovo se može uraditi na dva načina i to je takođe bila analiza u ovom radu. Prvi metod se naziva kumulativni metod i izračunava se prema sledećem izrazu:

$$d^2(k) = \sum_{i=1}^{r} \mu_i d_i^2(k), \tag{1}$$

gde su μ_i verovatnoća i -tog modela i $d_i^2(k)$ distanca u k-tom trenutku i -tog modela.

Drugi metod se zasniva na određivanju konačne statističke distance na osnovu maksimalne vrednosti težinskih koeficijenata IMM algoritma:

$$d^2(k) = d_i^2(k), \text{gde je } i \text{ za koje je } \max_{i=1,\ldots,r}(\mu_i). \tag{2}$$

Za praćenje ciljeva odabran je model konstantnog kretanja (*Constant Velocity* - CV) sa dva stanja po x i y koordinati i njihovim brzinama, tako da je ukupan vektor stanja definisan kao:

$$X = \begin{bmatrix} x & \dot{x} & y & \dot{y} \end{bmatrix}. \tag{3}$$

Polazeći od dinamičkih jednačina stanja, koje se mogu modelovati izrazom:

$$X(k+1) = F(k)X(k) + G(k)v(k), \tag{4}$$

i jednačine merenja koja je data kao:

$$Y_j(k) = H(k)X(k) + w(k),\ j = 1, 2, \ldots, m(k). \tag{5}$$

Pri čemu je:

$X(k)$ – vektor stanja cilja u k-tom skenu,

$Y_j(k)$ – j-ta opservacija koja je primljena u k-tom skenu,

$F(k)$ – Tranziciona matrica (matrica prelaza iz stanja u stanje),

$H(k)$ – matrica merenja (opservacija),

$v(k)$ – šum procesa,

$w(k)$ – šum merenja, tj. nekorelisani beli Gausov šum sa poznatom kovarijacionom matricom R,

$G(k)$ – matrica šuma procesa $v(k)$,

$m(k)$ – broj merenja pristiglih u k-tom skenu.

Matrice prelaza stanja i uticaja šuma procesa na vektor stanja za oba modela su:

$$F_1 = F_2 = \begin{bmatrix} 1 & T & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & T \\ 0 & 0 & 0 & 1 \end{bmatrix}; G_1 = G_2 = \begin{bmatrix} \frac{T^2}{2} & 0 \\ T & 0 \\ 0 & \frac{T^2}{2} \\ 0 & T \end{bmatrix}. \tag{6}$$

Gde je sa T označena perioda skeniranja. Kovarijaciona matrica šuma procesa i merenja su:

$$Q^j = \sigma_j^2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; H = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, j = 1, 2. \tag{7}$$

Gde σ_j^2 predstavlja varijanse šumova procesa i iznose $\sigma_1^2 = 1^2\ piksel/s^2$ i $\sigma_2^2 = 5^2\ piksel/s^2$.

IMM algoritam je projektovan sa sledećim verovatnoćama za prelazak na novi model:

$p_{11} = 0.95, p_{12} = 0.05, p_{22} = 0.1, p_{21} = 0.9$.

C. Blok za pridruživanje podataka

Sledeći blok je blok za pridruživanje podataka, koji se koristi nakon formiranih merenja, a nakon što su izvršene procene stanja ciljeva u prethodnom diskretnom trenutku. U ovom blok su implementiraju matematički metode za pridruživanje podataka i u ovom radu su odabrane globalno najbliži sused (*Global Nearest Neighbour* - GNN), najboljih *n* rešenja i Munkresov algoritam.

1) GNN

Kada se odrede distance matrice asocijacije, GNN algoritam postupa po sledećim pravilima:

- Dodeliti opservaciju tragu, ako je jedina u prozoru tog traga
- Dodeliti trag opservaciji, ako prozor tog traga obuhvata jednu opservaciju
- Dodeliti *j*-tu opservaciju *i*-tom tragu, čija je distanca najmanja u matrici asocijacije

Neke od primene ovog algoritma su date u [8].

2) Najboljih n rešenja

Algoritam određuje sve moguće dodele na osnovu distanci iz matrice asocijacije. Zatim sortira sve moguće dodele po kritrijumu minimalne distance i uzima u obzir *n* najboljih dodela.

3) Munkresov algoritam

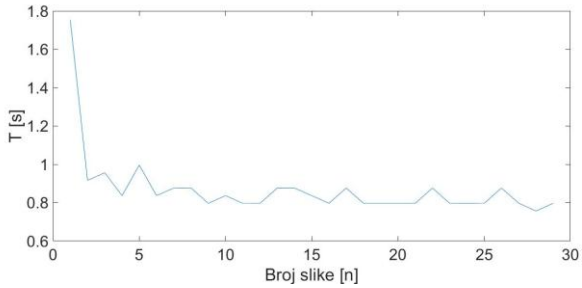
Algoritam se zasniva na obradi matrice asocijacije čiji su elementi distance. Princip rada algoritma je opisan u [9], neke od primena algoritma se mogu naći u [10].

III. REZULTATI

Ovo poglavlje će biti podeljeno u tri celine. Prva celina je opis generisanog scenarija i formiranje merenja video senzorom, dok druga i treća celina sadrže objašnjenja, prikaz i komentare rezultata estimacija na osnovu izvršenih analiza i implementacija.

A. Opis generisanih trajektorija

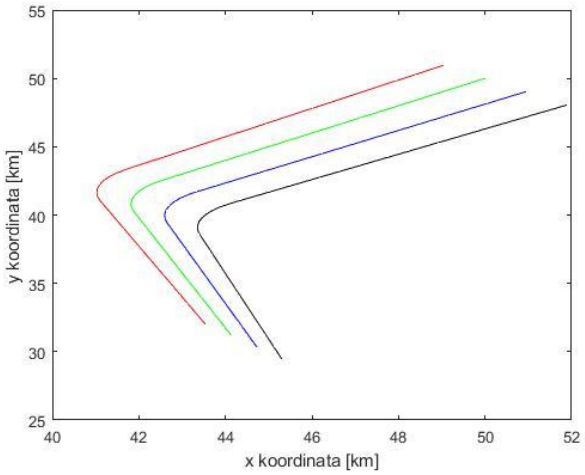
U ovom radu je generisan scenario koji sadrži trajektorije 4 cilja. Vrednost periode diskretizacije je usvojena na osnovu diskretnih vremenskih trenutaka koji su dobijeni pri preuzimanju slike sa video senzora, čije vrednosti su date na Sl. 2.



Sl. 2. Vrednost periode diskretizacije

Na osnovu prikazanih vrednosti prvih 30 slika koje se dobijaju sa kamere, može se usvojiti srednja vrednost i ona iznosi T= 0.9 s. Naravno mora se napomenuti da će se usvajanjem konstantne vrednosti za periodu diskretizacije uticati na kvalitet estimacije, ali obzirom da su merenja zašumljenja, navedeni uticaj neće doći do izražaja.

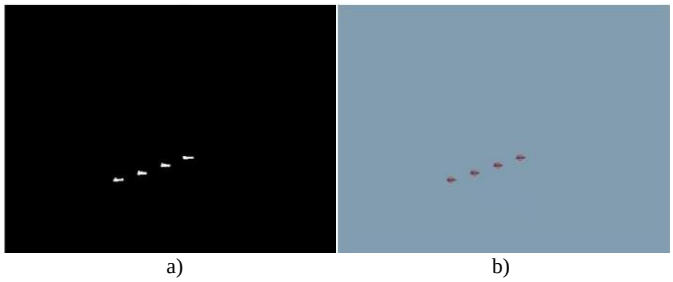
Referentne trajektorije su definisane na osnovu [2], gde su formirana 4 cilja sa brzinom kretanja od V=311 m/s, pri čemu su vršeni manevri inteziteta od 4g u periodu od 38·T do 42·T, pri čemu je ukupno trajanje od 72·T, Sl. 3.



Sl. 3. Trajektorije ciljeva

Ovako definisane trajektorije su prikazane u ravni prostora dimenzija [40 km – 52 km] × [25 km – 55 km] i projektovane na platno, tako da video senzor VC-C50i vidi kompletan prostor i na taj način se konvertuje u sliku dimenzija 704×576 piksela.

Slike koje se dobijaju sa video senzora se konvertuju u sive slike a zatim po Kittlerovoj metodi određen je prag inteziteta sive slike nakon čega se slika binarizuje tako što svaki piksel čiji je intezitet ispod praga dobija vrednost 0, dok pikseli čiji su inteziteti veći od praga dobijaju vrednost 1, kao što je prikazano na slici 4. a). Na taj način se dobija maska pomoću koje se izdvajaju ciljevi i određuju se njihovi centroidi, Sl 4.b, koji predstavljaju koordinate ciljeva koje dalje obrađuje estimator. Obrada slike koja je pomenuta data je u [11].



a) b)

Sl. 4. a) Segmentacija slike primenom adaptivnog praga, b) Detektovani ciljevi i prikaz njihovih centroida.

Na osnovu formirana merenja sledi analiza uticaja varijanse šuma merenja i proračuna ukupne distance na kvalitet estimacije primenom različitih metoda za pridruživanje podataka.

B. Uticaj šuma merenja na kvalitet estimacije

Sistemi za praćenje pokretnih ciljeva koriste estimatore stanja u situacijama kada se prati jedan cilj. Međutim, proces praćenja se komplikuje kada se u gejtu posmatranog traga nađu dva ili više ciljeva kao što je objašnjeno u [12]. Tada je neophodno koristiti algoritme za pridruživanje podataka. Algoritmi za pridruživanje podataka, po pravilu, sadrže neki od filtera za estimaciju stanja. Na taj način moguće je

napraviti poređenje ovih algoritama sa stanovišta srednje kvadratne greške praćenja po poziciji.

Osnovni kriterijum za uporednu analizu bila je srednja kvadratna greška po poziciji (RMSE), koja se izračunava kao:

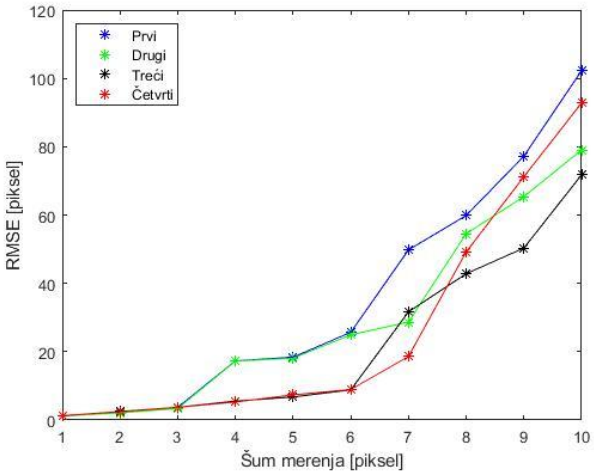
$$RMSE(k) = \sqrt{(x\ (k) - x_e(k))^2 + (y\ (k) - y_e(k))^2}\ , \qquad (8)$$

gde su x i y vrednosti koordinata cilja bez merenog šuma, dok su x_e i y_e estimacije po koordinatama u *k*-tom trenutku.

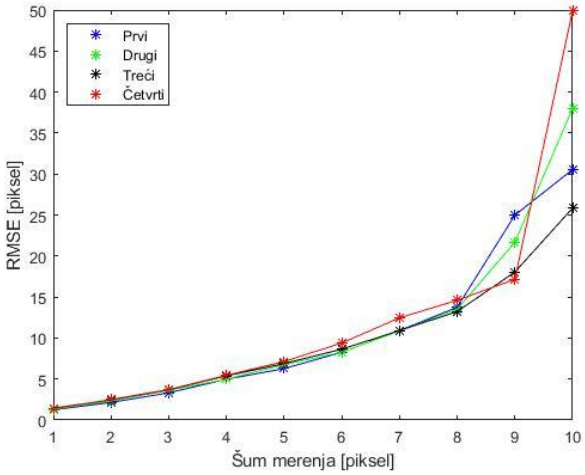
Kako je potrebno odrediti do kojih granica možemo verovati algoritmima za dodelu podataka, trajektorijama ciljeva dodat je šum merenja procesa sa Gausovom raspodelom sa nultom srednjom vrednosti. Standardna devijacija merenja σ_m varira od jednog piksela do 10 piksela × 10 piksela.

Na slici 5. prikazana je RMSE po poziciji sva četiri cilja u odnosu na prave pozicije, odnosno pozicije bez dodatog šuma u zavisnosti od šuma merenja koji varira. RMSE predstavlja srednju vrednost trenutnih grešaka tokom scenarija i izračunava se na sledeći način:

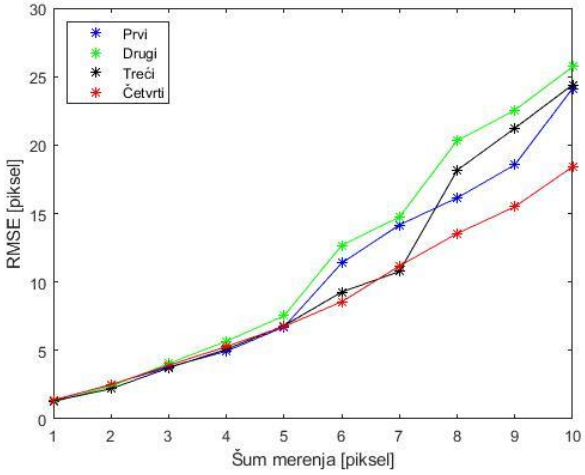
$$RMSE = n^{-1} \cdot \sum_{k=1}^{n} RMSE(k) \qquad (9)$$



a)



b)

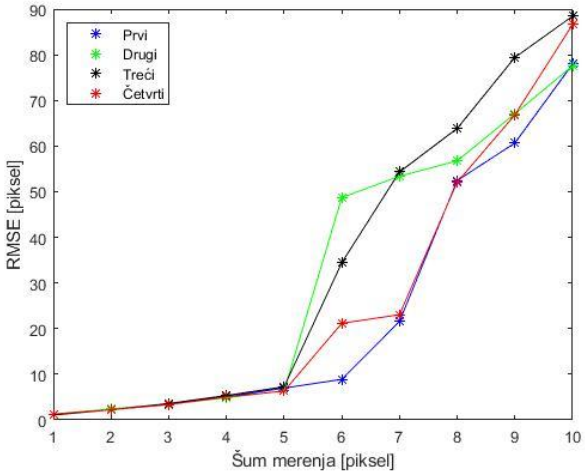


c)

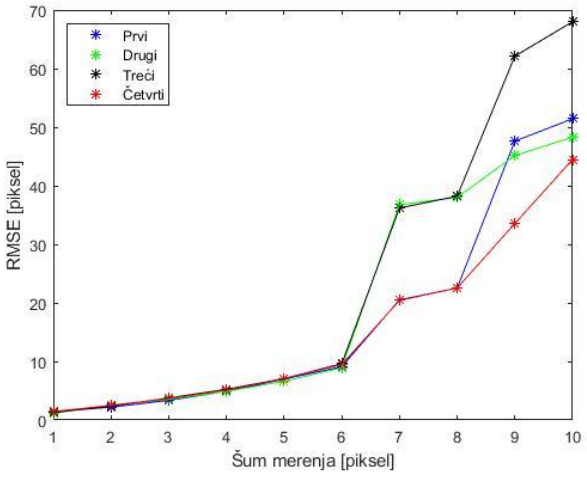
Sl. 5. RMSE a) primenom GNN, b) primenom najboljih *n* rešenja, c) primenom Munkres algoritma.

C. Uticaj proračuna ukupne statističke distance na kvalitet estimacije

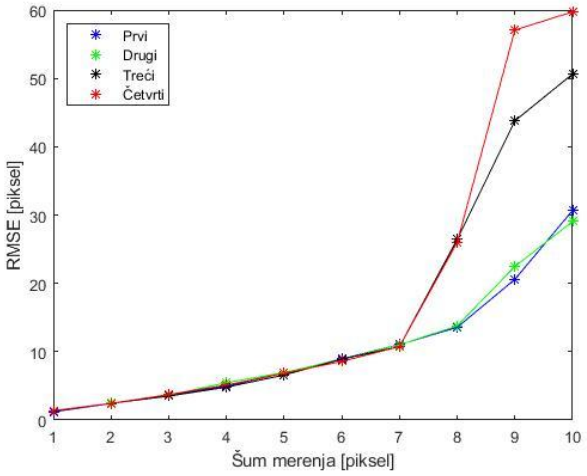
U ovom delu će biti razmatran i uticaj šuma na kvalitet estimacije ukoliko se primeni kumulativna metoda proračuna ukupne statističke distance, što je prikazano na slici 6.



a)



b)



c)

Sl. 5. RMSE kumulativnom metodom a) primenom GNN, b) primenom najboljih *n* rešenja, c) primenom Munkres algoritma.

IV. ZAKLJUČAK

Rezultati sa Sl. 5. i 6. prikazuju da sa povećanjem šuma merenja, povećava se i RMSE. Kada uporedimo rezultate jednostruke metode, vidimo da za GNN algoritam RMSE prvog i drugog cilja drastično raste sa 3 piksela na 20 za šum merenja od 4^2 piksela^2. Algoritam sa Najboljih *n* rešenja daje malo bolje rezultate. Kod njega RMSE raste linearno do šuma merenja od 8^2 piksela^2 nakon čega RMSE naglo raste. Najbolje rezultate daje Munkresov algoritam, gde RMSE raste linearno gde za vrednost varijanse šuma merenja od 10^2 piksela^2 i iznosi svega 30 piksela.

Što se tiče uporedne analize kumulativne metode vidimo da se nagli skokovi za algoritme GNN, Najboljih *n* rešenja i Munkres algoritma javljaju na varijanse šuma merenja od 5^2, 6^2 i 7^2 piksela^2, respektivno.

Na osnovu ove analize može se zaključiti da jednostruki metod daje nešto bolje rezultate od kumulativne metode kada su u pitanju veliki šumovi merenja. Za male šumove merenja, kumulativna metoda daje bolje rezultate jer je rast RMSE u odnosu na šum merenja linearan.

Ono što se još može zaključiti na osnovu komparativne analize, jeste da Munkresov algoritam daje najbolja rešenja i po jednoj i po drugoj metodi, stoga se preporučuje njegovo implementiranje u aplikacijama koje se bave praćenjem više ciljeva.

LITERATURA

[1] Blackman, S., Populi, R.: Design and Analysis of Modern Tracking Systems, Artech House, 1999.

[2] Shalom, Y. Bar, Blair, W. D.: Multitarget-Multisensor Tracking: Applications and Advances-Volume III, Artech House, Norwood, MA 02062, 2000.

[3] Blackman, S.: Multiple-Target Tracking with Radar Applications, Artech House, Dedham, 1986.

[4] Bar-Shalom, Yaakov, and Xiao-Rong Li. *Multitarget-multisensor tracking: principles and techniques*. Vol. 19. Storrs, CT: YBs, 1995.

[5] Reid, Donald. "An algorithm for tracking multiple targets." *IEEE transactions on Automatic Control* 24.6 (1979): 843-854.

[6] Kittler, Josef, and John Illingworth. "Minimum error thresholding." *Pattern recognition* 19.1 (1986): 41-47.

[7] Kirubarajan, Thiagalingam, and Yaakov Bar-Shalom. "Kalman filter versus IMM estimator: when do we need the latter?." *IEEE Transactions on Aerospace and Electronic Systems* 39.4 (2003): 1452-1457.

[8] Konstantinova, Pavlina, Alexander Udvarev, and Tzvetan Semerdjiev. "A study of a target tracking algorithm using global nearest neighbor approach." *Proceedings of the International Conference on Computer Systems and Technologies (CompSysTech'03)*. 2003.

[9] Bourgeois, François, and Jean-Claude Lassalle. "An extension of the Munkres algorithm for the assignment problem to rectangular matrices." *Communications of the ACM* 14.12 (1971): 802-804.

[10] Hwang, Inseok, et al. "Multiple-target tracking and identity management in clutter, with application to aircraft tracking." *Proceedings of the 2004 American Control Conference*. Vol. 4. IEEE, 2004.

[11] Gonzalez, Rafael C., and Richard E. Woods. "Digital image processing [M]." *Publishing house of electronics industry* 141.7 (2002).

[12] Collins, Robert T. "Multitarget data association with higher-order motion models." *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012.

ABSTRACT

In this paper, the quality estimation in multiple target tracking system was analysed. Multiple target tracking system has a video sensor which is used for recording. Gauss noise of different variances was added to the generated trajectories of four imminent targets. The position estimation results were compared based on the algorithm that was used: Global nearest neighbor, N best solution and Munkres algorithm. The second analysis was made trough the influence of the calculation of the total estimation, that was calculated based on weight coefficients in interactive multiple model.

Analysis of the quality of the estimation in the multi target tracking system using one video sensor depending on the measurment noise

Ilić Filip, Marko Antonijević

A 5 GHz Low-Noise Amplifier with Sliding Mode Based Phase Control Loop

Darko Mitić, *IEEE*, Goran Jovanović, Tatjana Nikolić and Dragan Antić, *Member, IEEE*

***Abstract*— The paper considers a tunable Low-Noise Amplifier (LNA) incorporating a phase loop with sliding mode control (SMC). Phase control loop forces a resonant frequency to be equal to an input signal frequency by tuning the amplifier resonant constituents. Thanks to the sliding mode control, LNA is robust to parameter perturbations in the full operating range, possesses maximum gain at the resonant frequency and attain input signal frequency faster. The 0.13 μm SiGe BiCMOS technology was used for LNA design and validation. LNA has ~30 dB gain, quality factor *Q* ~ 41 and resonant frequency from 5133 up to 5783 MHz.**

***Index Terms*— Low-Noise Amplifier, Self-tuning, Phase Control Loop, Resonant circuit, Sliding Mode Control.**

I. Introduction

Low-Noise Amplifier (LNA) is basically used in the receiving end of wireless communication system to amplify very weak received signal to the acceptable level with adding as little additional noise as possible and providing high linearity, sufficient gain, low power consumption and well defined resistive input impedance [1, 2]. With the aforementioned characteristics, LNA can be considered as band-pass filter as well.

The values of manufactured analog circuit components often differ from the design specifications because of process parameters, supply voltage, and temperature (PVT) variations, so that the practically obtained results are not optimal. In order to compensate frequency characteristic variations caused by these PVT perturbations, tunable filters, master-slave filter tuning schemes and self-tuning filters [3-6] are proposed in the literature. The latter approach, based on using of the phase loop with sliding mode control (SMC), is suggested in this paper.

Tunable selective amplifiers can be realized by using digitally controlled binary-weighted capacitor array, or current mirror array [4]. The frequency accuracy depends on the number of tuning bits used. Analog filters and selective amplifier are also adjusted by master–slave tuning schemes [5, 6] with a master-circuit designed as the voltage-controlled oscillator (VCO), and a slave filter built with identical integrators. By comparing the phases of input and output signals, it is possible to detect the mismatch of LNA characteristics [7]. To obtain the desired phase transfer function, self-tuning filters use the estimated phase error as correction factor and a phase control loop similar to one used in tuning oscillators with phase locked loop (PLL), or delay lines with delay locked loop (DLL) [8, 9].

In this paper, we propose cascade common source LNA with a sliding mode based phase control loop forming a self-tuning LNA (STLNA) with SMC. STLNA with SMC has a defined phase shift at resonant frequency, so the phase comparison of input and output signals detects amplifier characteristics mismatch. The phase loop with SMC utilizes a phase shift value to adjust LNA to maximal gain by using MOS varicap for frequency tuning. STLNA with SMC will be always tuned to the input signal frequency even in the presence of parameter variations, and will have the maximal gain at input signal frequency, acting as a selective amplifier in a wide frequency range. The similar sliding mode based phase control loop is considered for design of phase-synchronizer in [10].

The paper is organized as follows. Section II concentrates on realization of the tunable LNA. The structure of STLNA with sliding mode based phase control loop is defined and described in the Section III. Simulation results are shown in Section IV and the example of STLNA application is presented in Section V. Section VI contains some concluding remarks.

II. A Tunable Narrowband Low-Noise Amplifier

The LNA schematic is based on the inductively degenerated common source LNA topology [11] and it is shown in Fig. 1. It consists of two MOS transistors, M_1 and M_2, where M_1 operate as common-source. The inductor L_{pr} is connected to a DC biasing node. To decrease the Miller effect for LNA that strongly limits its frequency characteristics and provides poor reverse isolation, the transistor M_2 operates in common-gate mode. Thanks to two active elements, M_1 and M_2, a high LNA gain is obtained.

The output load is implemented as resonant circuit consisting of L_1, C_1 and MVaricap. The voltage drop over the load is lower than the drop over a resistive load of the same impedance. This solution provides correct circuit operation at lower power supply voltage level (V_{dd} = 1.8 V), and has lower power dissipation.

MVaricap element corresponds to the voltage controlled capacitor. In IHP BiCMOS technology, it is implemented as a modified PMOS transistor [12]. By controlling polarization voltage of a N-well, the capacitance between gate and channel of PMOS transistor varies. The value of a control voltage, V_{ctrl}, determines a capacitance of MVaricap element [12]

$$C_{MVar} = \frac{C^*_{MVar}}{\sqrt{1+V_{ctrl}/V_\rho}}, \quad (1)$$

All authors are with the Faculty of Electronic Engineering, University of Niš, 14 Aleksandra Medvedeva, 18000 Niš, Serbia (e-mail: {darko.mitic, goran.jovanovic, tatjana.nikolic, dragan.antic}@elfak.ni.ac.rs).

and the resonant frequency f_r is given by

$$f_r = \frac{1}{2\pi\sqrt{L_1(C_1 + C_{MVar})}}. \quad (2)$$

The dynamic impedance of a resonant circuits, at f_r, is very high so that the LNA has very high gain, as well. The phase shift of LNA is

$$\theta = -180^o - \text{atan}\left(2Q\frac{f_s - f_r}{f_r}\right), \quad (3)$$

where f_s is a frequency of input signal V_{in} and Q is a quality factor.

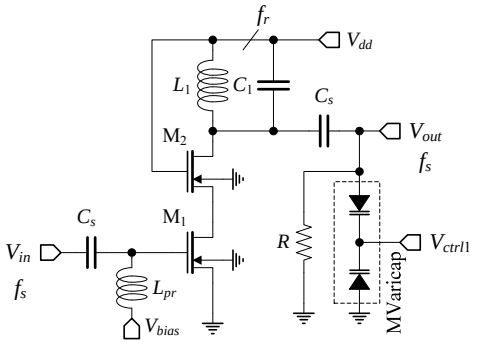


Fig. 1. Low-noise amplifier scheme

Elements of resonant circuit are also chosen in order to provide high quality factor Q and, consequently, narrow LNA bandwidth $BW = f_r/Q$. When BW is narrow then LNA gain is high, the receiver selectivity is good, the attenuation of symmetrical signals in heterodyne receiver is high, and noise level is low. The amplitude and phase characteristics of proposed LNA are given in Fig. 2. As one can see, the maximal gain is obtained at the resonant frequency for the phase shift of –180°.

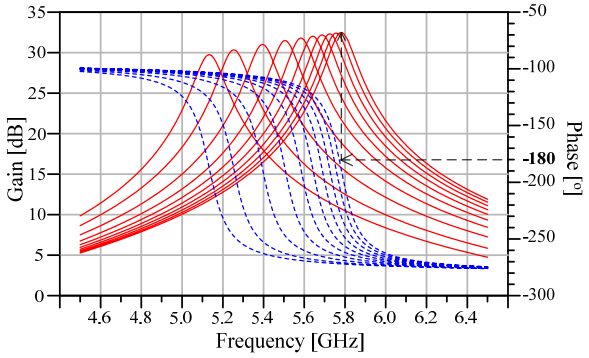


Fig. 2. Gain magnitude and phase characteristics

III. Architecture of Self-Tuning LNA with Sliding Mode Control

The design concept of STLNA employs phase control strategy where the phase shift between input and output signals is used to generate control voltage, V_{ctrl}. The control voltage defines the resonant frequency f_r, which determines the LNA phase characteristic (3). When the filter phase shift is –180°, then f_r is tuned to the frequency of the input signal f_s. In that manner, we control the phase shift of the output signal similar to the pulse delay control that can be found in DLL circuits [8, 9]. The block diagram of the STLNA with SMC is depicted in Fig. 3. The LNA, introduced in Sec. II, represents a building block of the proposed architecture (see Fig. 3).

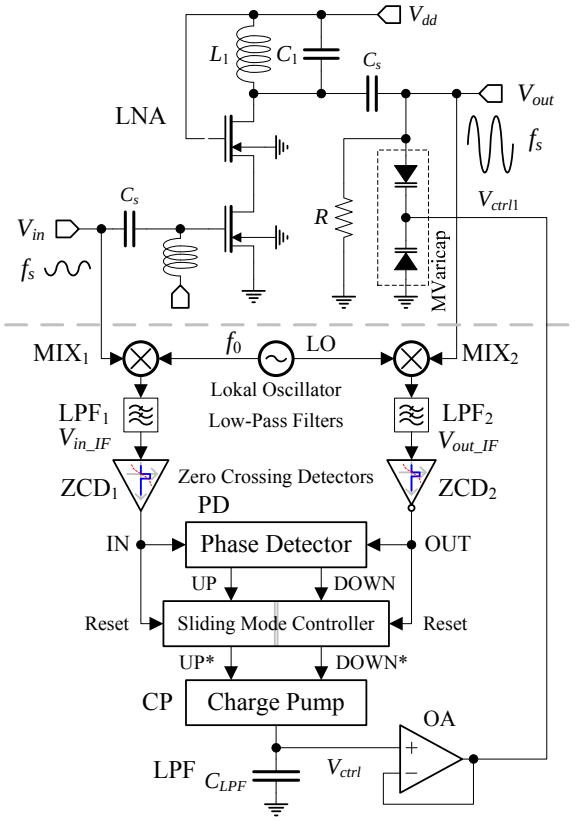


Fig. 3. Block diagram of STLNA with SMC.

Correct operation of a phase detector is difficult to realize at high RF/MW frequency so the frequency down conversion of signals V_{in} and V_{out} is performed by using two mixers, MIX_1 and MIX_2, local oscillator (LO), as generator of frequency f_0, and two low-pass filters, LPF_1 and LPF_2 (see Fig. 3.). The outputs of frequency down converter are two signals, V_{in_IF} and V_{out_IF}, at lower frequency, $f_0 - f_s$. Then signals V_{in_IF} and V_{out_IF}, is amplified first and shaped to rectangular forms by using two zero crossing detectors, ZCD1 and ZCD2 presented in Fig 4. ZCD_2 has one inverter stage more than ZCD_1, introducing an additional -180° phase shift to the output signal. In that way, the phase error, estimated by phase detector, PD, is

$$\theta_{PD} = -\text{atan}\left(2Q\frac{f_s - f_r}{f_r}\right). \quad (4)$$

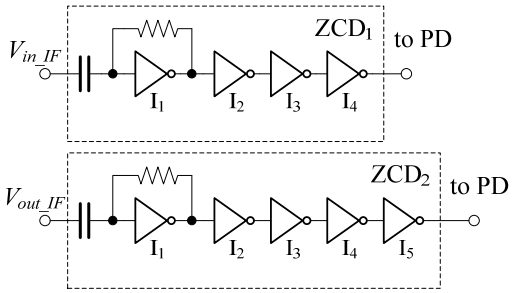


Fig. 4. Zero crossing detector

Phase comparison of input IN and output OUT signals is performed by a PD, which generates UP and DOWN signals. DOWN signal is on when the OUT signal phase leads in respect to the IN signal phase, while in the opposite, UP signal is active. Time durations of UP and DOWN signals are proportional to the phase shift. UP and DOWN signals represent error signal used in the sliding mode controller to generate the UP* and DOWN* control signal that drive a charge pump (CP). CP charges and discharges the load capacitor C_{LPF} providing V_{ctrl} that is used as control voltage for LNA. CP acts as an integrator and, therefore, V_{ctrl} can be written as

$$V_{ctrl} = -k_{cp}\int_{-\infty}^{t} V_{smc}(\theta_{PD})\mathrm{d}\tau\,, \quad (5)$$

where k_{cp} is a CP positive gain depending on a constant current of charge pump I_{CP} and the capacitor C_{LPF}. $V_{smc}(\theta_{PD})$ is a SMC law. In stable-state the phase shift between V_{out} and V_{in} signals is 0°. The detailed descriptions of all these elements from Fig. 3 are presented in [7]. This paper would focus only on design of the sliding mode controller.

SMC belongs to the class of nonlinear control techniques known as variable structure control [13], which has been studied and implemented worldwide since 1950's. A sliding mode exists in control system when a system state is forced to move along a predefined sliding manifold by discontinuous control with infinite switching, providing system robustness to parameter perturbations and external disturbances. Such control action has to satisfy the reaching and existence conditions of sliding mode for all time, i.e. to ensure that the velocity vector of the system state trajectories always points toward the sliding manifold. The motion of a system with SMC consists of a reaching phase, where the system state is driven to the sliding manifold from any initial state, and a sliding mode phase, where the system state slides along predefined switching manifold having predetermined stable dynamics. Therefore, SMC design procedure has two steps. The first step is to select the sliding manifold so that the system has desirable dynamical behavior in the sliding mode. Secondly, a discontinuous control algorithm should be found to ensure that the system state reaches the sliding manifold in finite time. In order to describe the SMC design procedure briefly, let us consider the system

$$\dot{\mathbf{x}} = \varphi(\mathbf{x},t) + \gamma(\mathbf{x},t)u\,, \quad (6)$$

where $\mathbf{x} \in R^n$ is a system state vector, u is a scalar control input, $\varphi(\mathbf{x},t)$ and $\gamma(\mathbf{x},t)$ are nonlinear functions, and n is a system order. Following SMC design steps, the sliding manifolds with desirable dynamics is defined as

$$s(\mathbf{x}) = 0\,, \quad (7)$$

where $s(\mathbf{x})$ also represents the switching function. Then SMC u is selected as

$$u = \begin{cases} u^+(\mathbf{x}) & \text{for } s(\mathbf{x}) > 0 \\ u^-(\mathbf{x}) & \text{for } s(\mathbf{x}) < 0 \end{cases}, \quad (8)$$

to meet the reaching and existence conditions of sliding mode

$$s(\mathbf{x})\dot{s}(\mathbf{x}) < 0\,. \quad (9)$$

Taking into account (1), (2) and (5), the LNA phase loop dynamics can be obtained by time-differencing (4)

$$\frac{d\theta_{PD}}{dt} = -\frac{2k_{cp}\pi^2 QL_1 C_{MVar}^3 f_s f_r^3}{V_\rho\left(C_{MVar}^*\right)^2\left(f_r^2 + 4Q^2\left(f_s - f_r\right)^2\right)}V_{SMC}\left(\theta_{PD}\right). \quad (10)$$

Since the LNA phase loop dynamics (10) is of the first order, the switching function is chosen according to

$$s(\theta_{PD}) = \theta_{PD}\,, \quad (11)$$

and SMC law is selected as

$$V_{smc}(\theta_{PD}) = \text{sgn}(\theta_{PD})\,. \quad (12)$$

One can see that by substituting (12) in (10), the reaching and existence condition of sliding mode (9)

$$s(\theta_{PD})\dot{s}(\theta_{PD}) = \theta_{PD}\dot{\theta}_{PD} = -\zeta\left|\theta_{PD}\right| < 0 \quad (13)$$

is always satisfied since

$$\zeta = \frac{2k_{cp}\pi^2 QL_1 C_{MVar}^3 f_s f_r^3}{V_\rho\left(C_{MVar}^*\right)^2\left(f_r^2 + 4Q^2\left(f_s - f_r\right)^2\right)} > 0\,. \quad (14)$$

The relay type SMC law (12) produces the chattering, the undesirable phenomenon [14] involving high control activities that may excite non-modeled high frequency dynamics and oscillations of phase in vicinity of $\theta_{PD} = 0$. To alleviate the chattering, the boundary-layer approach [15] is suggested with the control

$$V_{smc}(\theta_{PD}) = \begin{cases} +1 & \theta_{PD} > \delta \\ \theta_{PD} & \left|\theta_{PD}\right| \le \delta \\ -1 & \theta_{PD} < -\delta \end{cases}, \quad (15)$$

where δ is a positive real constant defining the boundary-layer width around $\theta_{PD} = 0$. Note that the control law $V_{smc}(\theta_{PD})$ is linear in δ - neighborhood of $\theta_{PD} = 0$, so by the appropriate choice of δ, the chattering could be eliminated. The scheme of sliding mode controller with control law (15) is shown in Fig. 5.

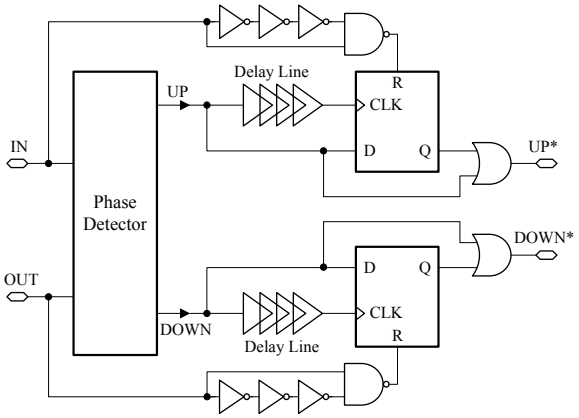


Fig. 5. Scheme of sliding mode controller.

IV. Simulation Results

The proposed solution, which relates to design of STLNA with sliding mode based phase control loop, is verified by Spice simulation. The IHP design kit for 0.13 µm SiGe BiCMOS technology was used [12]. The supply voltage V_{dd} was chosen to be 1.8 V. The characteristics of LNA are shown in Table I.

TABLE I
LNA CHARACTERISTICS

Gain	29.7 – 32.5 dB
Resonant frequency range	5133 –5783 MHz
Bandwidth	126 – 136 MHz
Quality factor	37 – 45
P_{total} (LNA core)	2.1 mW

First we have realized the selective LNA whose resonant frequency is tunable. The tuning process, in our design, was performed by varying the capacitance MVaricap, what was achieved by adjusting a DC biasing point. Consequently, the resonant f_r frequency was subject to change.

Fig. 6 presents time responses of STLNA with SMC. The waveforms of the down-converted input V_{in_IF} and output V_{out_IF} signals are given in Fig. 6(a). IN and OUT signals at the outputs of ZCD1 and ZCD2 respectively, UP and DOWN control signals, obtained at the outputs of PD, UP* and DOWN* control signals, derived from sliding mode controller, as well as V_{ctrl}, are presented in Fig. 6(b). The steady-state state is reached, i.e. the phase loop is locked, at the moment when: i) the signals V_{in} and V_{out} , i.e V_{in_IF} and V_{out_IF} are of opposite phases, ii) UP, DOWN, UP*, and DOWN* signals disappear, and iii) the control voltage V_{ctrl} has a constant value. The settling time of a system is approximately 50 ns - 5 times faster than in the case without sliding mode controller in the phase control loop. Settling time of the LNA input V_{in} and output V_{out} signal is presented in Fig. 6(c). The sliding mode based phase control loop changes the referent resonant frequency of resonant circuits until a condition $f_r = f_s$ is reached. At the resonant frequency STLNA has its maximal gain.

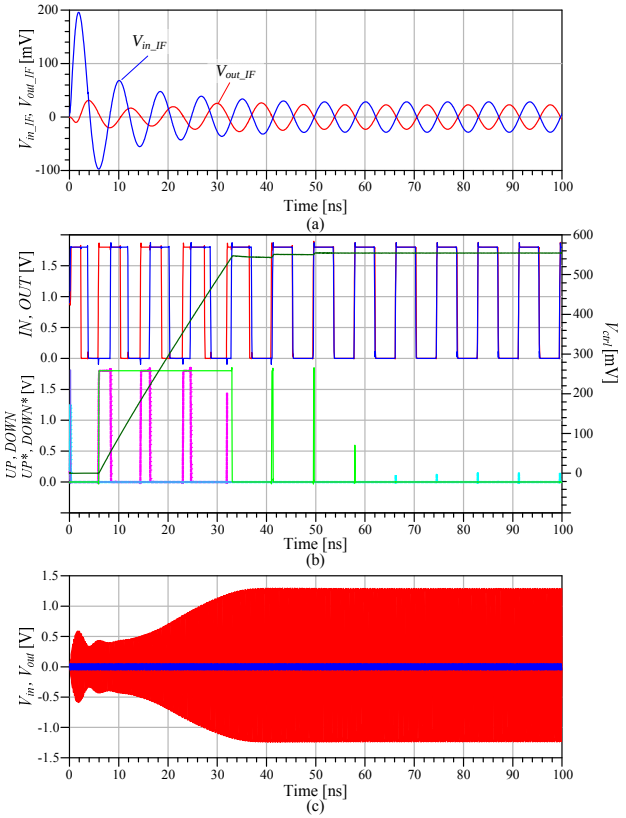


Fig. 6. Time response of STLNA with SMC: (a) V_{in_IF} and V_{out_IF} and (b) IN, OUT, UP, DOWN, UP*, DOWN* and V_{ctrl}, (c) V_{in}, V_{out}

V. Application of STLNA with SMC

The application of STLNA with SMC is shown in Fig. 7. The structure consists of two STLNAs, master and slave. The master STLNA is excited by the referent frequency source f_s and it generates control voltage, V_{ctrl}. The slave STLNA is of identical structure as the master. It is driven with the same control voltage V_{ctrl}, generated by the master STLNA. The slave is used for filtering and amplification of the input signal.

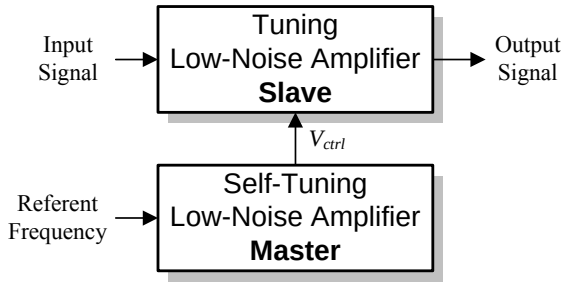


Fig. 7. Typical application for self-tuning band-pass filter

VI. Conclusion

Selective self-tuning low-noise amplifier (STLNA) with sliding mode based phase control loop is considered in this paper. The proposed solution is suitable for VLSI implementation. Unlike self-tuning solutions based on sensing of power [16] or voltage [17] level, the phase loop with sliding mode control (SMC) is introduced to tune the circuit resonant frequency to the frequency of input signal in the presence of PVT variations. This is performed by changing MOS varicap capacitance with control voltage, which is designed according to SMC theory. STLNA is working within a frequency range from 5133 up to 5783 MHz, and has high quality factor Q (37-45), and high gain (29.7 – 32.5 dB). The duration of the entire self-tuning process is less than 50 ns in full frequency range of STLNA operation. The proposed circuit is useful for realization of narrow-band amplifiers implemented in heterodyne receivers. Simulation results are promising for practical implementation of the described solution.

Acknowledgment

This work was supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia, as a part of the projects TR 32009, III 43007, and TR 35005.

References

[1] D. K. Shaeffer, T. H. Lee, "A 1.5-V, 1.5-GHz CMOS low noise amplifier", IEEE J. Solid-State Circuit, vol. 32, no. 5, pp. 745–759, 1997.

[2] D. J. Cassan, J. R. Long, "A 1-V transformer-feedback low-noise amplifier for 5-GHz wireless LAN in 0.18-μm CMOS", IEEE J. Solid-State Circuits, vol. 38, no. 3, pp. 427–435, 2003.

[3] Gh. Z. Fatin, and Z. D. K. Kanani, "A very low power band-pass filter for low-IF applications", J. Circuits, Systems and Computers, vol. 17, no. 4, pp. 685-701, 2008.

[4] Z. Y. Chang, D. Haspeslagh, and J. Verfaillie, "A highly linear CMOS Gm-C band-pass filter with on-chip frequency tuning", IEEE J. of Solid-State Circuits, vol. 32, no. 3, pp. 388-397, 1997.

[5] C. Yoo, S.-W. Lee and W. Kim, "A ±1.5-V, 4-MHz CMOS continuous-time filter with a single-integrator based tuning", IEEE J. Solid-State Circuits, vol. 33, no. 1, pp. 18-27, 1998.

[6] G. Jovanović, D. Mitić, M. Stojčev and D. Antić, "Self-tuning biquad band-pass filter", J. of Circuits, Systems and Computers, vol. 22, no. 3, pp. 1-19, 2013.

[7] G. Jovanović, D. Mitić, M. Stojčev, D. Antić: "Self-tuning low-noise amplifier", *Facta Universitatis, Series: Automatic Control and Robotics*, vol. 12, no. 2, pp. 139-145, 2013.

[8] J. Maneatis, "Low-jitter process-independent DLL and PLL based on self-biased techniques", IEEE J. Solid-State Circuits vol. 31, no. 11, pp. 1723-1732, 1996.

[9] M. Stojčev and G. Jovanović, "Clock aligner based on delay locked loop with double edge synchronization", Microelectronics Reliability, vol. 48, pp. 158–166, 2008.

[10] D. B. Mitić**,** G. S. Jovanović, M. K. Stojčev, D. S. Antić: "Phase-synchroniser based on gm-C all-pass filter chain with sliding mode control", *Int. J. of Electronics*, vol. 102, no. 3, pp. 362-375, 2015.

[11] P. Leroux and M. Steyaert, *LNA-ESD Co-Design for Fully Integrated CMOS Wireless Receivers*, Dordrecht, Springer, 2005.

[12] IHP-Microelectronics, *SiGe:C BiCMOS technologies for MPW & prototyping*. [Online] Cited 2013-08-30. Available at: http://www.ihp-microelectronics.com/en/services/mpw-prototyping/sigec-bicmos-technologies.html.

[13] X. Yu and O. Kaynak, "Sliding-mode control with soft computing: A survey", *IEEE Trans. on Ind. Elec.*, vol. 56, pp. 3275-3285, 2009.

[14] V. I. Utkin, "Chattering problem", *Preprints of the 18th IFAC World Congress*, pp. 13374-13379, 2011.

[15] J-J. E. Slotine and S. S. Sastry, "Sliding controller design for non-linear systems", *Int. J. of Control*, vol. 40, pp. 421-434, 1984.

[16] J. Choi, D. Im, K. Lee, "A self-tuned balun-LNA with differential imbalance correction and blocker filtering", IEEE Microwave and Wireless Components Letters, vol. 21, no. 12, pp. 673-675, 2011.

[17] N. Ahsan, J. Dabrowski, A. Ouacha, "A self-tuning technique for optimization of dual band LNA", in *Proc. of the 1st European Wireless Technology Conference*. Amsterdam (Nederland), pp. 178-181, 2008.

FPGA-based quadrotor attitude estimation using experimental results from 9DOF IMU sensor

Taki Eddine Lechekhab, Stojadin Manojlović, Slobodan Simić, Davorin Mikluc

Abstract— **In order to build quadrotor prototype and apply different control algorithms, in this paper, experimental results from Inertial Measurement Unit (IMU) sensor fixed on a quadrotor test bench, mounted on the 3-axis platform, are analyzed and discussed. A Field Programmable Gate Array (FPGA) platform is used to control the 3-axis platform and process the measurements. The raw data are analyzed and processed by two algorithms: Kalman filter (KF) and Complimentary Filter (CF) to estimate the attitude angles of quadrotor, combining the measurements from the gyroscopes and the accelerometers.**

***Index Terms*— Quadrotor, FPGA, Attitude estimation, Kalman Filter , Complimentary Filter, MPU9250 IMU sensor.**

I. INTRODUCTION

Recently the interest in the unmanned aerial vehicles (UAVs) is increasing, and they have been used for many applications as: military operations, aerial photography for mapping, news coverage, inspection of power lines, atmospheric analysis for weather forecasts, traffic monitoring in urban areas, crop monitoring and spraying, border patrol, surveillance for illegal imports and exports, fire detection and control, agriculture, search and rescue operations for missing persons, natural disasters, etc [1] [2].

The quadrotors make up an easy-to-use and challenging platform for students and researchers from control systems area, because of its low-cost, the nonlinear nature and under-actuated configuration, and make it ideal to synthesize and analyze different control algorithms [1].

The quadrotor has a cross design with two pairs of opposite rotors rotating clockwise, while the other rotor pair rotating counter-clockwise to balance the torque. The attitude of the quadrotor and up-thrust actions are controlled by changing the angular velocities of the actuator rotors using pulse width modulation (PWM) to give the desired output. The current control and PWM control can be generated using different types of platforms as Arduino, FPGA, Raspberry pi etc. In this work a FPGA BASYS 2 platform is used. Nowadays the FPGAs are increasingly used as embedded systems in this field, which gives a motivation for research and using them to integrate different control algorithms on quadrotors.

The advances in electronics' field allowed producing cheap lightweight flight controllers, IMUs, actuators, positioning system (GPS) and other sensors [3]. This resulted in the quadrotor configuration becoming popular for small UAVs. With their small size and maneuverability, these quadrotors can be flown indoors and outdoors as well for different missions.

The aim of this work is to study the reliability of a low cost 9DOF MPU9250 IMU sensor for the stability and orientation feedback, in order to build a quadrotor test bench for application of different controllers using FPGA platform. It receives the raw measurements from the gyroscopes and the accelerometers, and compare them to the reference angles translated from the potentiometers placed on each axis (X, Y, Z) of the 3-axis platform equipped with the DC motors that control the pan, tilt and roll motions.

Among the plenty approaches to solve the attitude estimation problem, the Kalman filter took the main interest [4]. In our work the measurement results from the gyroscopes and the accelerometers are used to estimate the quadrotor attitude angles. Therefore two filters are used and compared: Kalman filter and complimentary filter, which combine the accelerometers and the gyroscopes measurements for estimating the roll and pitch angles. These methods are taking on consideration the measurement noise from each, and the biases from the gyroscopes that causes the drifting .

This paper is organized in six sections. The hardware used in the experiments is described in the second section, while the measurement models and proposed filters are explained in the third and fourth sections, respectively. After the experimental results presented in the fifth section, the paper ends with conclusions and future perspectives.

II. HARDWARE DESCRIPTION

A. Three-axis platform

The 3-axis platform (pan, tilt and roll) used in this work is presented in Fig. 1. It consists of an outer gimbal, inner gimbal, test table, three DC motors as axis actuators and three axis positional sensors, all mounted on the fixed platform base. The quadrotor prototype is mounted on the test table. The outer gimbal rotates around a pan axis and it is driven by a motor located inside the base. The inner gimbal is driven by a DC motor, which is fixed on the outer gimbal and it moves around the polarization axis. The third motor drives the

Taki Eddine Lechekhab is with the Military academy, University of Defense, Pavla Jurisica Sturma 33, 11000 Belgrade, Serbia (e-mail: Boughdiri.taki@gmail.com)

Stojadin Manojlovic is with the Military academy, University of Defense, Pavla Jurisica Sturma 33, 11000 Belgrade, Serbia (e-mail: colemanojle@yahoo.com).

Slobodan Simic is with the Military academy, University of Defense, Pavla Jurisica Sturma 33, 11000 Belgrade, Serbia (e-mail: simasimic01@gmail.com).

Dovorin Mikluc is with the Military academy, University of Defense, Pavla Jurisica Sturma 33, 11000 Belgrade, Serbia (e-mail: miklucd@yahoo.com)

platform around the tilt axis (yaw angle of the quadrotor), and it is fixed inside the inner gimbal [5] .

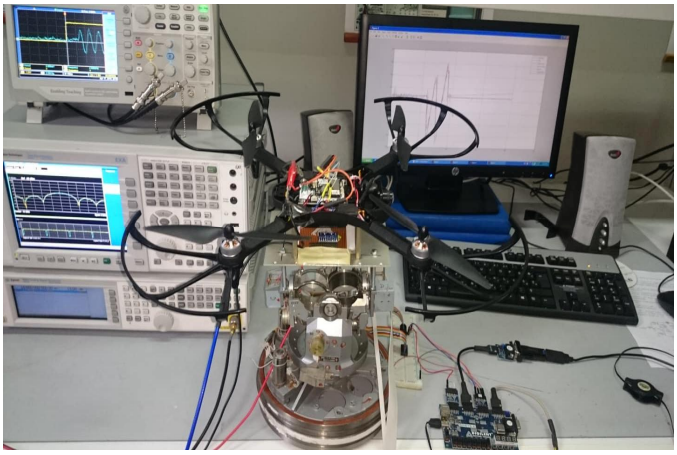


Fig. 1. Quadrotor prototype mounted on the three-axis platform.

B. *Field Programmable Gate Array platform*

In this work a Digilent's BASYS 2 FPGA board, shown in Fig. 2 (d), is used, which is an circuit design and implementation platform. The specifications of this platform provides us a complete and easy to use hardware for hosting circuits ranging from basic logic devices to complex controllers. The considerable number of I/O, make it ideal to design the suitable filters and control algorithms, and reading the measurements from different sensors and modules, simultaneously.

The figure. 2.(a) shows PmodRS232X a serial converter and interface that connects the FPGA platform with the computer for data processing and analyzing using different software like MATLAB. The Digilent's Pmod AD1, represented in Fig. 2.(c), is a two channel, 12-bit analog-to-digital converter that features Analog Devices AD7476A. With a sampling rate of up to 1 million samples per second, it is used to link and calibrate the potentiometers, from which the raw measurements are used as referent angles, because they give the platform attitude angles in the inertial coordinate system.

C. *MPU9250/6500 IMU module*

The MPU-9250 is a 9-axis (gyro+accelerometer+ compass) MEMS motion tracking device. This multi-chip module (MCM), shown in Fig. 2.(b), consists of two dies integrated into a single QFN package. One die houses the 3-axis gyroscope and the 3-axis accelerometer. The other die contains the AK8963 3-axis magnetometer. These integrated dies are controlled by a Digital Motion Processor (DMP) , all combined and boxed in a small 3x3x1mm package. This module has an embedded temperature sensor and an on-chip oscillator with ±1% variation over the operating temperature range. This tracking module can be fined in many devices as mobile phones, virtual reality devices and UAVs. The characteristics of this module are given in Table. I, that can be found in [6].

TABLE I
MPU9250 IMU SENSOR MODULE SPECIFICATIONS

Specifications	MPU9250/6500
Size	25.5 × 15.4 × 3 mm
Weight	2.72 g
Gyro range	±250, ±500, ±1000, ±2000 °/s
Gyro bias	2 °/s
Gyro nonlinearity	±0.1 %
Gyro noise performance	0.1 (rms)
Accel. range	±2g, ±4g, ±8g, ±16 g
Accel. bias	±60 mg
Accel. nonlinearity	±0.5 %
Accel. noise performance	8 mg(rms)
Magnetometer range	± 4800 µT

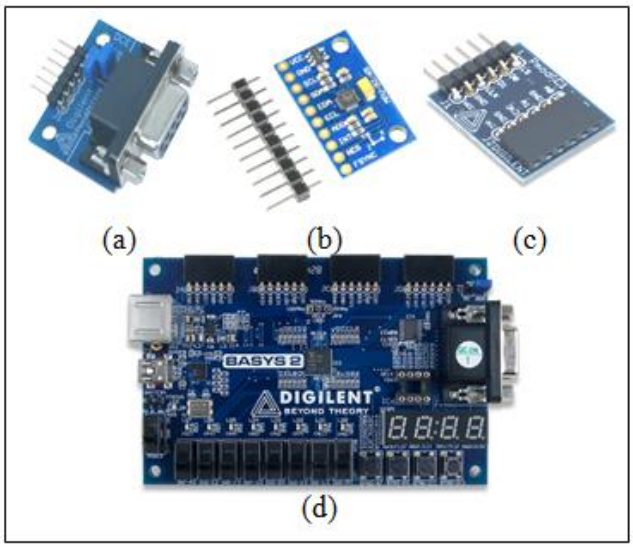


Fig. 2. The hardware used for the measurements .

III. MEASUREMENT MODELS

The mathematical model of the quadrotor is based on the Newton-Euler equations which describe its translational and angular dynamics. The translational dynamics is defined in the Earth-centred-Earth-fixed (ECEF) reference frame and the angular dynamics is represented in the body-fixed reference frame [7]. The rotation matrix $\mathbf{T}_{B/E}$ that define the orientation of the body frame (B) with respect to the Earth reference frame (E), using the sequence of yaw-pitch-roll elementary transformations, is expressed as:

$$\mathbf{T}_{B/E} = \begin{bmatrix} c\theta c\psi & c\theta s\psi & -s\theta \\ s\phi s\theta c\psi - c\phi s\psi & s\phi s\theta s\psi + c\phi c\psi & s\phi c\theta \\ c\phi s\theta c\psi + s\phi s\psi & c\phi s\theta s\psi - c\phi c\psi & c\phi s\theta \end{bmatrix} \quad (1)$$

with 'c' and 's' denoting cos and sin, respectively. The components of the quadrotor body absolute angular velocity,

expressed in the body frame $\omega_B = \begin{bmatrix} p & q & r \end{bmatrix}^T$, are related with Euler angles $\xi = \begin{bmatrix} \phi & \theta & \psi \end{bmatrix}^T$ as following:

$$\dot{\xi} = \begin{bmatrix} 1 & \sin\phi\tan\theta & \cos\phi\tan\theta \\ 0 & \cos\phi & -\sin\phi \\ 0 & \dfrac{\sin\phi}{\cos\theta} & \dfrac{\cos\phi}{\cos\theta} \end{bmatrix} \omega_B \qquad (2)$$

The stabilization of the quadrotor at hover mostly depends on the control of the roll and pitch angles, so in this work the magnetometer measurements are not used, and the yaw angle is not considered in the experimental results.

The accelerometers measure the quadrotor body translational accelerations $\mathbf{a}_B$ in the body-fixed frame. According to [8] and [9] the approximated relation of the components of the normalized acceleration vector in the body frame and the pitch and roll angles is:

$$\mathbf{a} = \frac{\mathbf{a}_B}{|\mathbf{a}_B|} \approx \begin{bmatrix} \sin\theta \\ -\sin\phi\cos\theta \\ \cos\phi\cos\theta \end{bmatrix} \qquad (3)$$

The measurement models of the IMU sensors include accelerometer noise n_a, gyroscope noise n_g and gyroscope bias β_ω:

$$\begin{aligned} \bar{\omega} &= \omega + \beta_\omega + n_g \\ \bar{a} &= a + n_a \end{aligned} \qquad (4)$$

where ω and $\bar{\omega}$ are true and measured gyroscopes data, and a and $\bar{a}$ are true and measured accelerometers data.

IV. Attitude Estimation Algorithms

A. *Kalman filter*

In order to estimate the attitude angles and the gyroscopes biases the state vector is chosen as $x = \begin{bmatrix} \phi & \beta_x & \theta & \beta_y \end{bmatrix}$,where β_x and β_y are the biases along X and Y axis. For the implementation of the Kalman algorithm, based on the equations (2), (3) and (4), the state space model can be expressed as following [8]:

$$\begin{aligned} \dot{x} &= \begin{bmatrix} (\bar{\omega}_x - \beta_x) + (\bar{\omega}_y - \beta_y)\sin\phi\tan\theta \\ 0 \\ (\bar{\omega}_y - \beta_y)\cos\phi \\ 0 \end{bmatrix} + \mathbf{w} \\ y &= \begin{bmatrix} \sin\theta \\ -\sin\phi\cos\theta \\ -\cos\phi\cos\theta \end{bmatrix} + \mathbf{v} \end{aligned} \qquad (5)$$

In the nonlinear model (5) with $\bar{\omega}_x$ and $\bar{\omega}_y$ are denoted gyroscope measurements along X and Y axis respectively. The process noise **w** and measurement noise **v** are both assumed to be Gaussian with covariance matrices **Q** and **R**:

$$\begin{aligned} \mathrm{E}\left[\mathbf{w}\mathbf{w}^T\right] &= \mathbf{Q} \\ \mathrm{E}\left[\mathbf{v}\mathbf{v}^T\right] &= \mathbf{R} \end{aligned} \qquad (6)$$

Using small angles approximations: $\sin\alpha \approx \alpha, \cos\alpha \approx 1$, where $\alpha = \{\phi, \theta\}$ the equations (2) and (3) can be linearized, and the model (5) can be represented in the following linear form :

$$\begin{aligned} \dot{x} &= \begin{bmatrix} 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \end{bmatrix} x + \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} u + \mathbf{w} \\ y &= \begin{bmatrix} 0 & 0 & 1 & 0 \\ -1 & 0 & 0 & 0 \end{bmatrix} x + \mathbf{v} \end{aligned} \qquad (7)$$

The input vector of this system represents the angular velocity measurements $u = \begin{bmatrix} \bar{\omega}_x, \bar{\omega}_y \end{bmatrix}$ and $y = \begin{bmatrix} \bar{a}_x, \bar{a}_y \end{bmatrix}$ represents the acceleration measurements.

The discretization of the linear system (7) can be made with the sampling time *T*, assuming zero-order hold of the inputs:

$$\begin{aligned} x_{k+1} &= \mathbf{A}_k x_k + \mathbf{B}_k u_k & \mathbf{E}\left[\mathbf{w}_k \mathbf{w}_k^T\right] &= \mathbf{Q}_k \\ y_k &= \mathbf{H}_k x_k + v_k & \mathbf{E}\left[\mathbf{v}_k \mathbf{v}_k^T\right] &= \mathbf{R}_k \end{aligned} \qquad (8)$$

where the matrices $\mathbf{A}_k$, $\mathbf{B}_k$ and $\mathbf{H}_k$ are defined as :

$$\mathbf{A}_k = e^{\mathbf{A}T} = \begin{bmatrix} 1 & -T & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -T \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{B}_k = \int_0^T e^{\mathbf{A}\tau} d\tau = \begin{bmatrix} T & 0 \\ 0 & 0 \\ 0 & T \\ 0 & 0 \end{bmatrix}$$

$$\mathbf{H}_k = \begin{bmatrix} 0 & 0 & 1 & 0 \\ -1 & 0 & 0 & 0 \end{bmatrix}$$

The discrete model (8) is used for Kalman filter implementation. As denoted in [8], the model is decoupled, and it is possible to implement two Kalman filters for each angle estimation.

The equations of KF are summarized in two equation systems defined in the state space form, the time update system (prediction) in (9), and measurement update system in (10):

$$\begin{aligned} \bar{x}_k &= \mathbf{A}_{k-1}x_{k-1} + \mathbf{B}_{k-1}u_{k-1} \\ \bar{\mathbf{P}}_k &= \mathbf{A}_{k-1}\mathbf{P}_{k-1}\mathbf{A}_{k-1}^T + \mathbf{Q}_k \end{aligned} \tag{9}$$

$$\begin{aligned} \mathbf{S}_k &= \mathbf{H}_k\mathbf{P}_k^-\mathbf{H}_k^T + \mathbf{R}_k \\ \mathbf{K}_k &= \mathbf{P}_k^-\mathbf{H}_k^T\mathbf{S}_k^{-1} \\ x_k^+ &= x_k^- + \mathbf{K}_k\left(y_k - \mathbf{H}_k x_k^-\right) \\ \mathbf{P}_k^+ &= \left(\mathbf{I} - \mathbf{K}_k\mathbf{H}_k\right)\mathbf{P}_k^- \end{aligned} \tag{10}$$

B. Complimentary filter

Considering that the raw acceleration measurements along each axis are denoted as $\bar{a}_x$, $\bar{a}_y$ and $\bar{a}_z$, the estimations of the roll and pitch angles can be formed by resolving forces using basic trigonometry:

$$\hat{\phi}_{Acc} = \arctan\left(\frac{\bar{a}_y}{\sqrt{\bar{a}_x^2 + \bar{a}_z^2}}\right) \tag{11}$$

$$\hat{\theta}_{Acc} = \arctan\left(\frac{\bar{a}_x}{\sqrt{\bar{a}_y^2 + \bar{a}_z^2}}\right) \tag{12}$$

Using the equation (2), the roll and pitch angular velocities $\dot{\phi}_{gyro}$ and $\dot{\theta}_{gyro}$ can be calculated based on the gyroscope measurements.

The high frequency noise overlaying the signal disturbs the accelerometer measurement, and effecting on the estimation of the attitude angles. Therefore, some type of the law-pass filter should be added. Complimentary filter reduces the gyroscope drift and the impact of accelerometer noise during the measurements [10]. This filter is based on the sensor data fusion by choosing a constant α where: $0 < \alpha < 1$ and it raises when the accelerometer measurements are showing more reliability. The complimentary filter estimations of the roll and the pitch angle are given with:

$$\begin{aligned} \hat{\phi}(k+1) &= (1-\alpha)(\hat{\phi}(k) + \dot{\phi}_{gyro}(k)T) + \alpha\hat{\phi}_{Acc}(k) \\ \hat{\theta}(k+1) &= (1-\alpha)(\hat{\theta}(k) + \dot{\theta}_{gyro}(k)T) + \alpha\hat{\theta}_{Acc}(k) \end{aligned} \tag{13}$$

V. Experimental Results

Using the hardware explained in Section II, and XILINX System Generator in MATLAB, a number of scenarios are made in order to extract and process the data from the IMU sensor and potentiometer. The SIMULINK/System generator blocks are used to read the data from the gyroscope, accelerometer and potentiometer and are shown in Fig. 3.

The first scenario is rotation along X axis (roll angle only), the second scenario is testing the rotation along Y axis (pitch angle only), while the third scenario is combination of rolling and pitching .

The FPGA platform generates PWM signal to control the 3-axis platform's motors to follow a desired trajectory. The first scenario aims to save the data from the IMU module while the quadrotor test bench changes only the roll angle. Considering that angle measured from potentiometers as a referent angle, the results without filters are shown in Fig. 4.

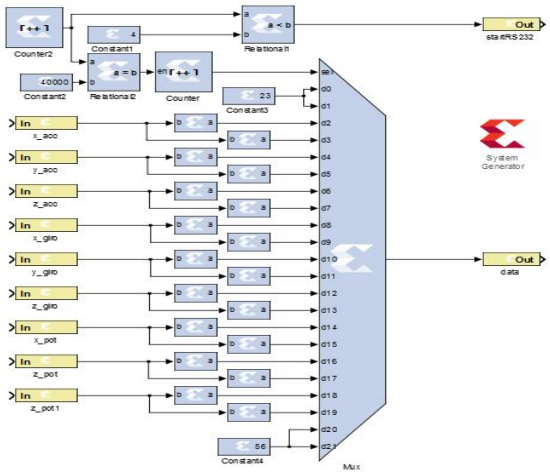


Fig. 3. System Generator blocks to read the measurements from IMU sensor.

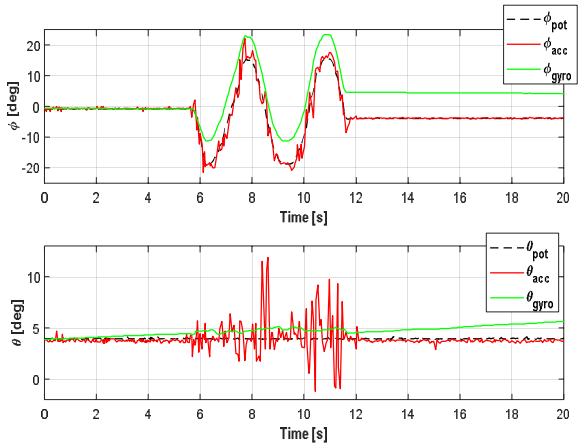


Fig. 4. Attitude angles based on the raw measurements from the first scenario

The accelerometer estimation of the roll angle in Fig. 4 shows a high sensitivity compared to the gyroscope data , and we can notice as well the drift caused by integrating gyroscope noise during the experiment time even along the Y axis. This gives an idea which sensor is more reliable for the complimentary filter method.

The second experiment is made by fixing the role angle while changing the pitch angle and the results are presented in Fig. 5, without any filter, as in the previous experiment.

Unlike the first measurement, the acceleration along the fixed axis shows less sensibility, while the gyroscopic drift still existing caused by integration the bias. The estimated measurements from accelerometer and gyroscope appear to be identical with the reference angle from the potentiometer except the drift of gyroscope angle and accelerometer's noise.

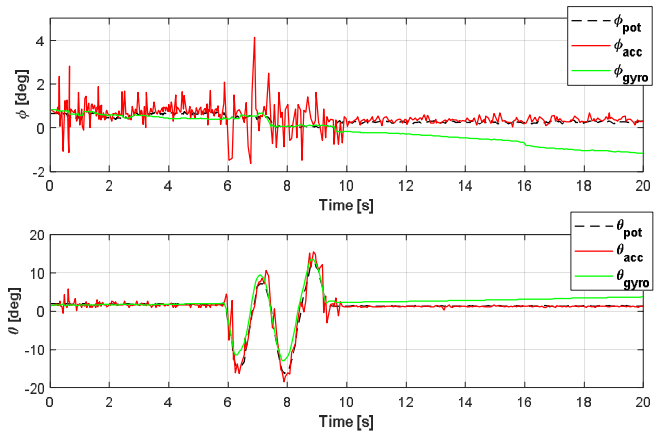


Fig. 5. Attitude angles based on the raw measurements from the second scenario

The third measurement scenario is considered the combined roll-pitch movement and results are presented in Fig. 6.

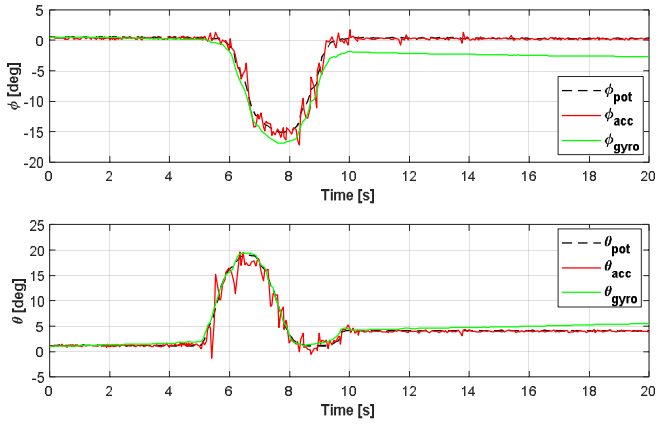


Fig. 6. Attitude angles based on the raw measurements from the third scenario.

By using the equations explained in the fourth section, the Kalman filter has been applied to estimate the attitude and reduce the noise that have been generated by the accelerometer and to eliminate the bias from the gyroscope.

The results are compared with the angles measured by the potentiometer. The values of the covariance matrices elements are chosen based on a number of simulations. For each angle (roll and pitch), two Kalman filters are separately applied. The estimated angles are compared to the referent angles from the potentiometers and illustrated in Figs. 7. – 9.

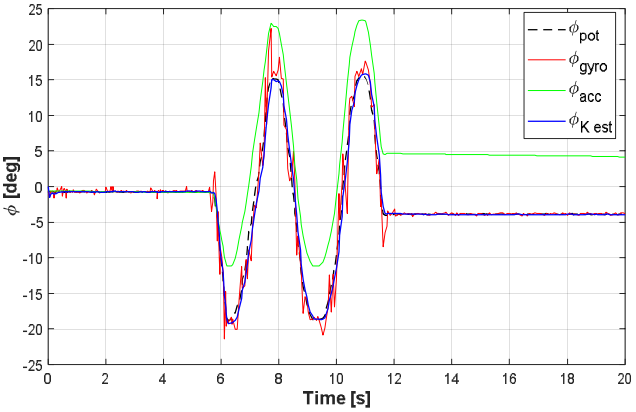


Fig. 7. KF estimation of roll angle (the first scenario).

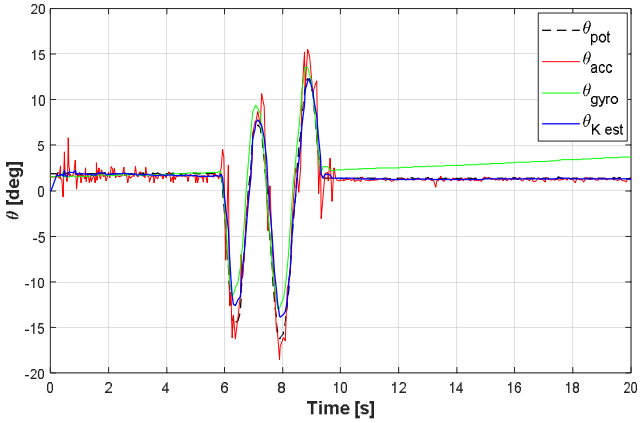


Fig. 8. KF estimation of pitch angle (the second scenario) .

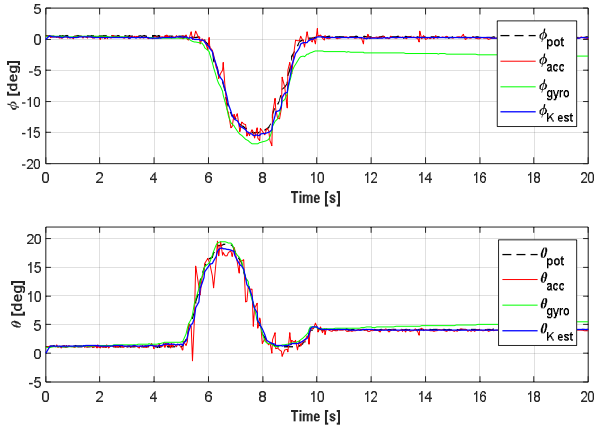


Fig. 9. KF estimation for both angles (the third scenario).

The following Fig. 10. and Fig. 11. show the estimated bias along X and Y axis from the third scenario, where we can notice that when the 3-axis platform is stabilized after 10 sec, the gyroscope biases converge to the constant values approximately $\beta_x = -0.11[\text{deg}/s]$ and $\beta_y = 0.15[\text{deg}/s]$.

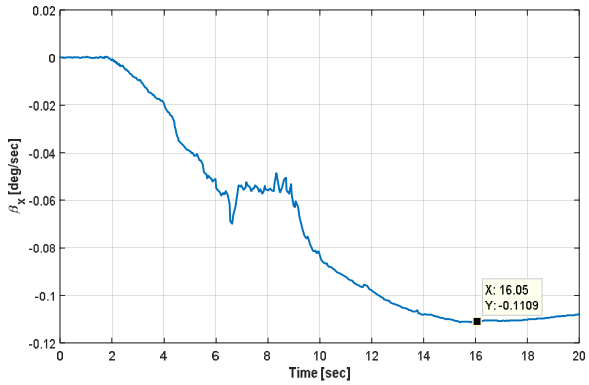


Fig. 10. Bias estimation along X axis

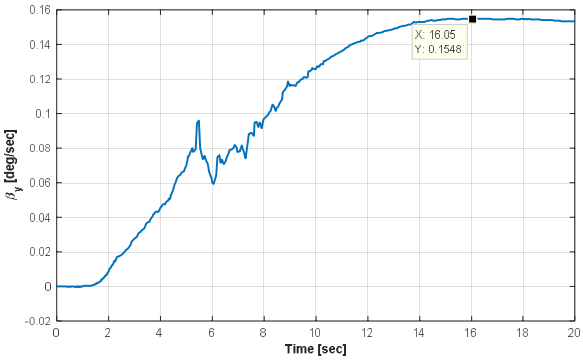


Fig. 11. Bias estimation along Y axis

The complimentary filter needs only one tuning parameter and a simple code, so it is optimal for energy saving, which is the actual UAVs' problem. In all scenarios, the parameter $\alpha = 0.4$ have been chosen after a number of tests that gives the appropriate results. By raising this parameter we more rely on the accelerometer measurements and a higher frequency noise is added. The following figure shows the results by applying this filter during the third scenario, compared with the Kalman filter estimations. As can be seen, the noise from the accelerometer influences the estimated angles, while the gyroscope drift is reduced.

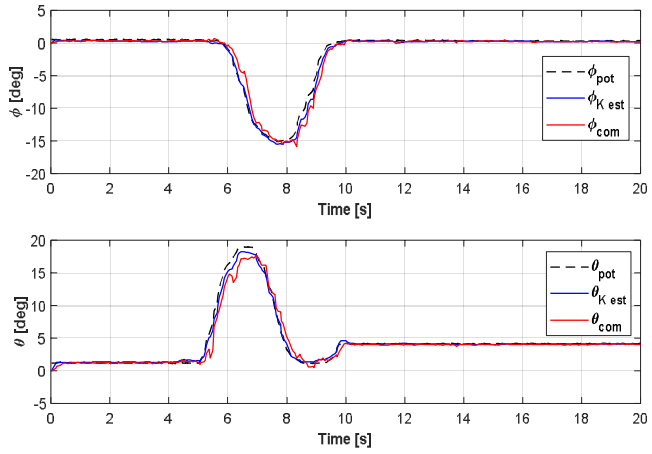


Fig. 12. Comparison of the different estimations results.

VI. Conclusion and Future Works

In this paper the hardware configuration, including 3-axis motion platform and FPGA processing board, for testing the main characteristics of the low-cost MPU9250 IMU module, is evaluated.

The analysis of the raw measurements from the MPU9250 module, allowed us to determinate the main characteristics of this sensor, as the biases and measurement noises, and lead us to apply the suitable filter. The Kalman filter gave better estimation compared to the complimentary filter results. The results show that appropriately filtered measurements from this low-cost module can be used in the next step of building the quadrotor prototype.

The future work will be the FPGA-based implementation of the robust controller for the quadrotor attitude stabilization and trajectory tracking, including the filtering of the vibrations caused by the propellers and testing the energy efficiency of different algorithms.

References

[1] Zulu Andrew, and Samuel John. "A review of control algorithms for autonomous quadrotors." *arXiv preprint arXiv:1602.02622* (2016).

[2] Premkumar, G., R. Jayalakshmi, and Md Akramuddin. "Design and Implementation of FPGA Based Quadcopter." (2018).

[3] Leong, Bernard Tat Meng, Sew Ming Low, and Melanie Po-Leen Ooi. "Low-cost microcontroller-based hover control design of a quadcopter." *Procedia Engineering* 41 (2012): 458-464.

[4] Farrell, James Lawrence. "Attitude determination by Kalman filtering." *Automatica* 6, no. 3 (1970): 419-430.

[5] Stanković, Momir R., Stojadin M. Manojlović, Slobodan M. Simić, Srđan T. Mitrović, and Milica B. Naumović. "FPGA system-level based design of multi-axis ADRC controller." *Mechatronics* 40 (2016): 146-155.

[6] Premkumar, G., R. Jayalakshmi, and Md Akramuddin. "Design and Implementation of FPGA Based Quadcopter." (2018)..

[7] Bouabdallah, Samir. *Design and control of quadrotors with application to autonomous flying*. No. THESIS. Epfl, 2007..

[8] Ricardo S, Luis R, Pedro G and Pedro C. "Improving Attitude Estimation Using Inertial Sensors for Quadrotor Control Systems " In *2014 International Conference on Unmanned Aircraft systems* , pp. 895-901. Orlando FL, USA. May 27-30, 2014

[9] Martin, Philippe, and Erwan Salaün. "The true role of accelerometer feedback in quadrotor control." In *2010 IEEE International Conference on Robotics and Automation*, pp. 1623-1629. IEEE, 2010.

[10] Chan, Ai-Ling, Su-Lim Tan, and Chu-Lih Kwek. "Sensor data fusion for attitude stabilization in a low cost Quadrotor system." In *2011 IEEE 15th International Symposium on Consumer Electronics (ISCE) Harvard*, pp. 34-39. IEEE, 2011.

Primena nelinearnog ADRC algoritma za upravljanje planarnim manipulatorom

Milan Svetozarević, Momir Stanković

***Apstrakt*— U radu je predložena primena nelinearnog regulatora sa aktivnim potiskivanjem poremećaja (*Nonlinear Active Disturbance Rejection Control*-NADRC) za upravljanje planarnim robotskim manipulatorom, koji predstavlja složeni nelinearni sistem sa dva ulaza i dva izlaza. Pri realizaciji upravljanja unakrsna dinamika sistema je uključena u totalne poremećaje po kanalima upravljanja, tako da su projektovani nezavisni regulatori za oba kraka manipulatora. Razmatrani problem praćenja zadate reference je preformulisan u problem regulacije (*error-based structure*), čime je pojednostavljena standardna struktura ADRC regulatora. Detaljna simulaciona analiza je pokazala uticaj izbora parametara nelinearnog regulatora na performanse upravljanja.**

***Ključne reči*— Nelinearno upravljanje sa aktivnim potiskivanjem poremećaja (ADRC); Planarni robotski manipulator; Prošireni opserver stanja (ESO).**

I. UVOD

Opis kinematike kretanja robotskih manipulatora najčešće zahteva kompleksne matematičke modele sa velikim brojem parametara i više ulaznih i izlaznih veličina, koje su međusobno spregnute, i to najčešće nelineranim vezama. Shodno tome, precizna identifikacija parametara ovakvih sistema je prilično otežana i ograničena. U ovom radu je razmatran problem upravljanja planarnog manipulatora sa dva kraka [1], odnosno dva stepena slobode (sistem sa dva ulaza i dva izlaza), koji nalazi široku primenu u različitim industrijskim postrojenjima [2], humanoidnim robotima [3], kao i fizioterapeutskim rehabilitacionim uređajima [4]. Imajući u vidu navedeni problem neodređenosti parametara (*parametar uncertanties*) manipulatora, konvencionalni upravljački algoritmi poput PI/PID regulatora [5], kao i algoritmi koji se baziraju na poznavanju preciznog modela objekta upravljanja [5], uglavnom ne daju zadovoljavajuće performanse. Shodno tome, u radu je analizirana primena regulatora na bazi upravljanja sa aktivnim potiskivanjem poremećaja (*Active Disturbance Rejection Control*-ADRC), koji omogućuje visoke performanse sistema uz minimalnu zavisnost od poznavanja modela objekta upravljanja [6].

Osnovna ideja ADRC-a je primena proširenog opservera stanja (*Extended State Observer*-ESO) pomoću koga se vrši estimacija i nakon toga potiskivanje generalizovanog (totalnog) poremećaja, koji obuhvata sve unutrašnje i spoljašnje poremećaje sistema. Na ovaj način, teoretski je pokazano da se bilo koji složeni sistem *n*-tog reda svodi na model redne veze *n* integratora bez poremećaja (*disturbance free model*), kojom se relativno lako može upravljati.

Primenom ADRC tehnike omogućeno je razdvajanje kanala upravljanja jednog i drugog kraka manipulatora. Naime, sistem sa dva ulaza i dva izlaza je raspregnut na dva sistema sa jednim ulazom i jednim izlazom, gde je celokupna unakrsna dinamika uključena u totalne poremećaje po jednom i drugom kraku.

Imajući u vidu da se upravljanje robotkskih manipulatora uglavnom svodi na problem praćenja zadate referentne pozicije krakova, standradni ADRC algoritam se najčešće analizira kao algoritam sa dva stepena slobode (*two-degree-of-freedom*-2DOF) [7] koji kao ulaze koristi referentni signal i izlazni signal objekta upravljanja. Međutim, kako je većina industrijskih sistema projektovana za upravljačke algoritme sa jednim stepenom slobode (*one-degree-of-freedom*-1DOF), kao što su PI/PID regulatori, jasno je da 2DOF topologija ADRC može predstavljati ograničenje u njegovoj praktičnoj primeni.

Imajući u vidu navedeno, u ovom radu je predložena reformulisana ADRC struktura koja ima za cilj predstavljanje algoritma kao 1DOF strukture, koja kao ulaz koristi grešku praćenja referentnog signala (*error-based structure*). Pored toga, za razliku od [7] razmotrena je primena nelinearnog ADRC (NADRC) algoritma sa različitim parametrima nelinearnih funkcija. Detaljna simulaciona analiza, sprovedena kroz više upravljačkih scenarija, je pokazala prednosti i nedostatke predloženih upravljačkih tehnika.

II. MODEL PLANARNOG MANIPULATORA

Planarni manipulator je sistem koji se sastoji iz dva kraka koja se kreću u istoj ravni i njegova struktura je prikazana na Sl. 1.

Ulazni signali sistema su naponi aktuatora u_{m1} i u_{m2}, koji omogućuju ugaono pokretanje kraka. Izlazni signali sistema su ugaone pozicije θ_1 i θ_2, prvog i drugog kraka planarnog manipulatora, respektivno. Kinematičko kretanje kraka planarnog manipulatora može se opisati kao ,

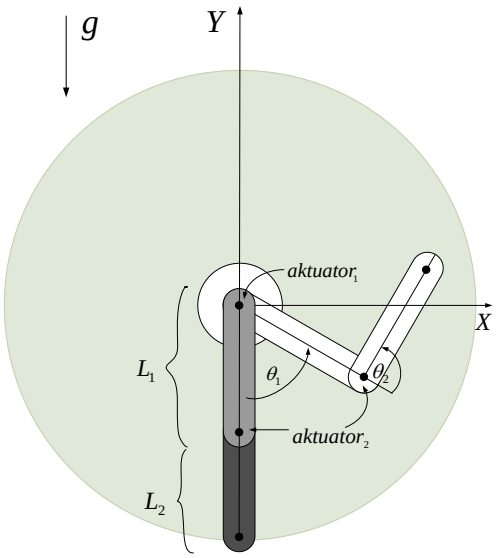
$$I_i\ddot{\theta}_i + F_i\dot{\theta}_i = \tau_{mi} + \tau_{zi} - \tau_{ci}\,, \tag{1}$$

gde se indeks *i*=1, 2 odnose na prvi, odnosno drugi krak, I_i [kgm^2] je ukupni moment inercije na osovini *i* -tog aktuatora, F_i [Nms/rad] predstavlja koeficijent viskoznog trenja, τ_{mi} [Nm], τ_{zi} [Nm], τ_{ci} [Nm], su obrtni moment aktuatora, poremećajni moment i moment unakrsnih veza, respektivno.

Milan Svetozarević – Vojna akademija, Univerzitet odbrane u Beogradu, Pavla Jurišića 31, 11000 Beograd, (e-mail: svetozarevicrks@ gmail.com).

Momir Stanković – Vojna akademija, Univerzitet odbrane u Beogradu, Pavla Jurišića 31, 11000 Beograd, (e-mail: momir_stankovic@ yahoo.com).

Da bismo detaljnije opisali model razmatranog sistema uvedeni su dinamički parametri p_1- p_5 dati u Tabeli I, gde su m_i, L_i, J_i masa, dužina i moment inercije i-tog kraka respektivno, a g je gravitaciono ubrzanje.



Sl. 1. Skica planarnog manipulatora [7]

TABELA I
DINAMIČKI PARAMETRI PLANARNOG MANIPULATORA

p_1 [kg m^2]	$m_2L_2^2 + J_2$
p_2 [kg m^2]	$2m_2L_1L_2$
p_3 [kg m^2]	$J_1 + m_2L_1^2 + 4m_2L_1^2$
p_4 [Nm]	m_2L_2g
p_5 [Nm]	$(m_1L_1 + 2m_2L_1)g$

Na osnovu uvedenih parametara možemo zapisati da je:

$$\begin{aligned} I_1 &= J_{m1} + \eta_1^2(p_1 + p_5), \\ I_2 &= J_{m2} + \eta_2^2 p_4, \end{aligned} \tag{2}$$

gde J_{mi} predstavljaju momente inercije i-tog aktuatora, dok je η_i odnos redukcije preko koje su povezani *i*-ti aktuator i *i*-ti krak . Uticaj unakrsnih veza može se opisati kao:

$$\begin{aligned} \tau_{c1} &= \eta_1\ddot{\theta}_1 p_2 c_{m1} + \eta_2\ddot{\theta}_2 p_2 c_{m2} - \eta_1\dot{\theta}_1\eta_2\dot{\theta}_2 p_2 s_{m2} \\ &- \eta_2\ddot{\theta}_2 p_2(\eta_1\dot{\theta}_1 + \eta_2\dot{\theta}_2)s_{m2} + p_5 c_{m1} + p_4 c_{m12}, \end{aligned} \tag{3}$$

$$\tau_{c2} = (p_2 + p_2 c_{m2})\eta_2\ddot{\theta}_2 + \eta_2^2\dot{\theta}_1^2 p_2 s_{m2} + p_4 c_{m12},$$

gde je $s_{mi} \equiv \sin(\theta_i)$, $c_{mi} \equiv \cos(\theta_i)$, $c_{m12} \equiv \cos(\theta_1 + \theta_2)$.
U ovom slučaju pretpostavljeno je da su aktuatori motori jednosmerne struje pa je obrtni moment definisan kao

$$\tau_{mi} = \frac{k_{Ii}}{R_i}(U_{mi} - k_{ei}\dot{\theta}_i) \tag{4}$$

gde je k_{Ii} [Nm/A] eletktromehanička konstanta, k_{ei} [Vs/rad] mehaničko-električna konstanta i R_i [Ω] otpornost namotaja motora. Treba napomenuti da je u (6) induktivnost rotorskog namotaja motora zanemarena zbog njenog malog uticaja u realnim sistemima.

III. PROJEKTOVANJE NELINEARNOG ADRC ALGORITMA

Opisani model planarnog manipulatora sa dva ulaza i dva izlaza možemo predstaviti kao dva raspregnuta sistema sa jednim ulazom i jednim izlazom

$$\begin{aligned} \ddot{\theta}_1 &= f_1 + b_{01}u_{m1}, \\ \ddot{\theta}_2 &= f_2 + b_{02}u_{m2}, \end{aligned} \tag{5}$$

gde su b_{01} i b_{02} najbolje aproksimacije parametara sistema $b_1 = k_{I1}/(R_1I_1)$ i $b_2 = k_{I2}/(R_2I_2)$, respektivno, dok f_1 i f_2 predstavljaju totalne poremećaje po prvom i drugom kraku:

$$\begin{aligned} f_1 &= -\frac{F_1}{I_1}\dot{\theta}_1 - \frac{k_{I1}k_{e1}}{R_1I_1}\theta_i + \frac{\tau_{z1}}{I_1} - \frac{\tau_{c1}}{I_1} + (\frac{k_{I1}}{R_1I_1} - b_{01})u_{m1}, \\ f_2 &= -\frac{F_2}{I_2}\dot{\theta}_2 - \frac{k_{I2}k_{e2}}{R_2I_2}\theta_2 + \frac{\tau_{z2}}{I_2} - \frac{\tau_{c2}}{I_2} + (\frac{k_{I2}}{R_2I_2} - b_{02})u_{m2}. \end{aligned} \tag{6}$$

Imajući vidu da je predložena identična struktura ADRC regulatora za upravljanje prvim i drugim krakom, u nastavku će njegovo projektovanje biti opisano za i-ti krak (i=1, 2).

Sistem (5) se može predstaviti u formi modela u prostoru stanja

$$\begin{aligned} \dot{x}_{1i} &= x_{2i}, \\ \dot{x}_{2i} &= f_i + b_{0i}u_{mi}, \\ x_{1i} &= \theta_i \end{aligned} \tag{7}$$

Kako razmatramo problem praćenja zadate referentne pozicije *i*-tog kraka θ_{ri}, (7) možemo zapisati u formi greške praćenja $e_{1i} = \theta_{ri} - \theta_i$ [8]:

$$\begin{aligned} \dot{e}_{1i} &= e_{2i}, \\ \dot{e}_{2i} &= f_{ei} - b_{0i}u_{mi}, \end{aligned} \qquad \text{za } i=\{1, 2\}, \tag{8}$$

gde je f_{ei} novi totalni poremećaj koji ima oblik:

$$f_{ei} = \ddot{\theta}_{ri} + (\frac{F_i}{I_i} + \frac{k_{Ii}k_{ei}}{R_iI_i})(\dot{\theta}_{ri} - \dot{e}_i) - \frac{\tau_{zi}}{I_i} + \frac{\tau_{ci}}{I_i} - (\frac{k_{Ii}}{R_iI_i} - b_{0i})u_{mi}, \tag{9}$$

Na ovaj način problem praćenja je preveden u problem regulacije, gde referentni signal možemo okarakterisati kao dodatni poremećaj odnosno kao deo totalnog poremećaja (9), a grešku praćenja kao izlazni signal koji treba minimizovati. Primenom ADRC tehnike upravljanja model (8) se može zapisati u formi sa proširenim stanjem koje predstavlja totalni poremećaj $x_{3ei} = f_{ei}$.

$$\begin{aligned} \dot{x}_{1ei} &= x_{2ei}, \\ \dot{x}_{2ei} &= x_{3ei} - b_{0i}u_{mi}, \\ \dot{x}_{3ei} &= \dot{f}_{ei} \end{aligned} \tag{10}$$

gde je $x_{1ei} = e_{1ei}$ i $x_{2ei} = e_{2ei}$. Pretpostavljajući da u razmatranim sistemima totalni poremećaj najčešće ima sporopromenljivu dinamiku ($\dot{f}_{ei} \approx 0$) za prošireni sistem (10) možemo projektovati klasičan *Luenberger*-ov opserver.

$$\begin{aligned} \dot{\hat{e}}_{1i} &= \hat{e}_{2i} + \beta_{1i}(e - \hat{e}_{1i}), \\ \dot{\hat{e}}_{2i} &= \hat{e}_{3i} + \beta_{2i}(e - \hat{e}_{1i}) - b_{0i}u_{mi}, \\ \dot{\hat{e}}_{3i} &= \beta_{3i}(e - \hat{e}_{1i}), \end{aligned} \tag{11}$$

gde su $\hat{e}_{1i}$ i $\hat{e}_{2i}$ estimacije stanja e_1 i e_2, a $\hat{e}_{3i}$ predstavlja estimaciju dodatnog stanja sistema (8), odnosno totalnog poremećaja f_{ei}, dok su sa β_{1i}, β_{2i} i β_{3i} označena pojačanja opservera.

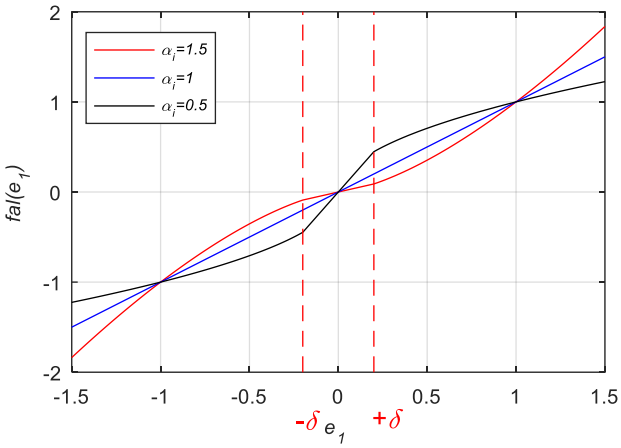
Na osnovu dobijenih estimiranih stanja sistema možemo realizovati nelinearni upravljački zakon sa aktivnim potiskivanjem totalnog poremećaja:

$$u_i = \frac{\hat{e}_{3i}}{b_{0i}} + \frac{k_{1i}\, fal_1(e_{1i}, \alpha_{1i}, \delta) + k_{2i}\, fal_2(e_{2i}, \alpha_{2i}, \delta)}{b_{0i}}, \tag{12}$$

gde su k_{1i} i k_{2i} i-ti koeficijenti upravljanja kojim podešavamo željenu dinamiku sistema u zatvorenoj povratnoj sprezi, Dok je $fal(e_{1i}, \alpha_{1i}, \delta)$ nelinearna funkcija definisana izrazom [1]:

$$fal(e_1, \alpha_i, \delta) = \begin{cases} \dfrac{e_1}{\delta^{1-\alpha_i}}, & |e_1| \le \delta \\ |e_1|^{\alpha} sign(e_1), & |e_1| > \delta \end{cases}, \tag{13}$$

Da bismo pojasnili razlog uvođenja nelinearne funkcije (13) njen grafički prikaz za δ=0.2 i različite vrednosti α_i dat je na Sl. 2.



Sl. 2. Grafički prikaz *fal* funkcije

Kao što možemo uočiti, izborom parametra δ menja se širina linearnog regiona oko nulte vrednosti e_1 i najčešće se δ usvaja tako da bude manje od greške praćenja u stacionarnom stanju [9]. Takođe vidimo da se izborom α <1 vrši smanjenje uticaja pojačanja regulatora kada je greška praćenja velika, čime se smanjuju oscilacije sistema u toku prelaznog perioda, pa se α <1 najčešće usvaja za funkciju fal_1. U drugu ruku, za α >1 imamo suprotan efekat (veće pojačanje pri velikim greškama praćenja), pa je iz tog razloga preporučen izbor α >1 za funkciju fal_2, imajući u vidu da ona utiče na diferencijalno dejstvo regulatora i tako smanjuje preskok u odzivu sistema sa zatvorenom povratnom spregom. Treba primetiti da se izborom α =1 (12) praktično svodi na linearnu funkciju.

IV. Simulaciona analiza

Na osnovu matematičkog modela (1) u programskom paketu Matlab/Simulink je formiran simulacioni model razmatranog planarnog manipulatora, dok je projektovanje ADRC regulatora realizovano na osnovu uprošćenog modela sistema (5). Usvojene vrednosti parametara manipulatora i aktuatora su date u Dodatku rada [7].

Uporedna simulaciona analiza performansi upravljanja je izvršena za sisteme sa različitim vrednostima parametara nelinearne *fal* funkcije. U prvom slučaju usvojeno je $\alpha_1 = 1$ i $\alpha_2 = 1$ što praktično predstavlja linearni ADRC regulator (u nastavku nazvan "LADRC"). Drugi regulator je projektovan sa usvojenim vrednostima $\alpha_1 = 0.5$ i $\alpha_2 = 1$ (uvođene nelinearnosti samo u proporcionalno pojačanje, u nastavku nazvan "NADRC$_1$"), dok je za treći regulator uzeto $\alpha_1 = 0.5$ i $\alpha_2 = 1.5$ (u nastavku nazvan "NADRC$_2$").

Ostali parametri regulatora su bili identični, odnosno usvojeno je δ=0.1. Dok su pojačanja observera definisana izrazima:

$$\beta_1 = 3\omega_o,\ \beta_2 = 3\omega_o^2,\ \beta_2 = \omega_o^3, \tag{14}$$

gde ω_o=80 *rad/s* predstavlja usvojeni propusni opseg

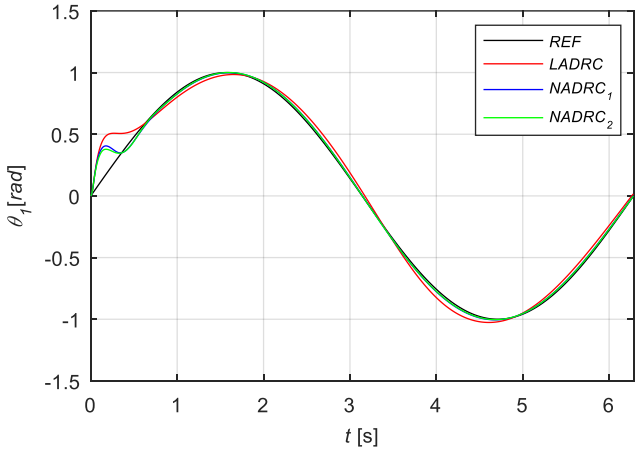
opservera, a pojačanja regulatora izrazima

$$k_1 = \omega_c,\ k_2 = 2\omega_c, \tag{15}$$

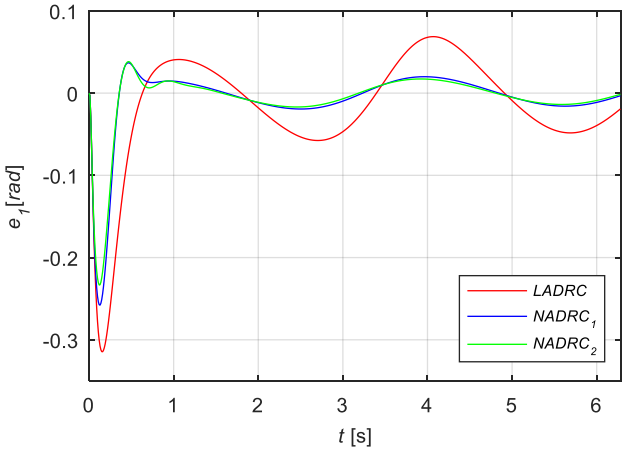
gde ω_c=10 *rad/s* predstavlja željeni propusni opseg sistema sa zatvorenom povratnom spregom.

Simulacije su sprovedene kroz tri različita scenarija koja obuhvataju ugaono pozicioniranje kraka planarnog manipulatora, u prisustvu spoljašnjih poremećaja, neodređenosti parametara, šuma merenja i ograničenja maksimalnih vrednosti upravljačkih signala.

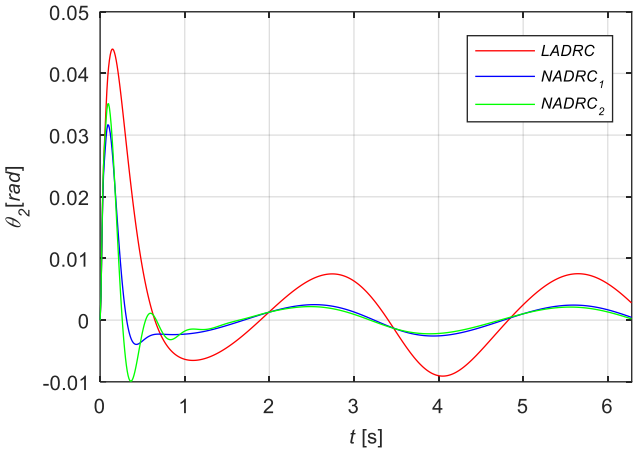
U okviru prvog scenarija je pretpostavljeno da je u potpunosti poznata vrednost parametara b_i, odnosno $\hat{b}_{o1} = b_{o1}$ i $\hat{b}_{o2} = b_{o2}$. Simulirani su odzivi sistema na referentnu pobudu $\theta_{r1} = \sin(2\pi t)$ dovedenu na prvi krak manipulatora. Dok je na drugi krak dovedena pobuda $\theta_{r2} = 0$. Performanse praćenja za sva tri tipa regulatora prikazane su na Sl. 3, dok su greške praćenja date na Sl. 4. Uticaj poremećaja na drugom kraku, koji je posledica unakrsnih veza prikazan je na Sl. 5.



Sl. 3. Performanse praćenja referentnog signala prvog kraka za scenario 1



Sl. 4. Greške praćenja referentnog signala prvog kraka za scenario 1



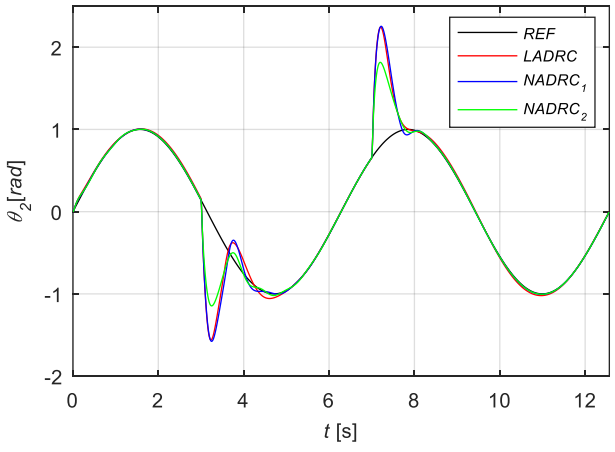
Sl. 5. Performanse praćenja referentnog signala drugog kraka za scenario 1

Sa slika se može uočiti da kod praćenja reference NADRC$_1$ i NADRC$_2$ imaju približno iste performanse sa aspekta brzine odziva, preskoka i vremena smirenja, dok LADRC ima lošije performanse. Takođe, može se uočiti da NADRC$_2$ ima nešto lošije performanse od NADRC$_1$ u toku prelaznog perioda kod potiskivanja poremećaja na drugom kraku manipulatora.

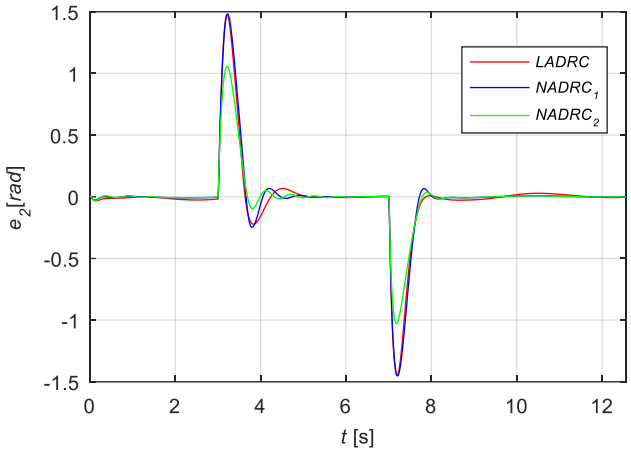
U okviru drugog scenarija je pretpostavljeno da je nepoznat model parametara b_i ($\hat{b}_{o1} = 2b_{o1}$ i $\hat{b}_{o2} = 2b_{o2}$) i simulirani odzivi na referentnu pobudu $\theta_{r2} = \sin(2\pi t)$ dovedenu na drugi krak manipulatora, a za prvi krak je podešeno $\theta_{r1} = 0$. Pored toga na oba kraka planarnog manipulatora je dodat poremećajni moment τ_{zi}

$$\tau_{zi} = \begin{cases} 5, & 3s \le t \le 7s \\ -1, & 7s < t \end{cases}. \tag{16}$$

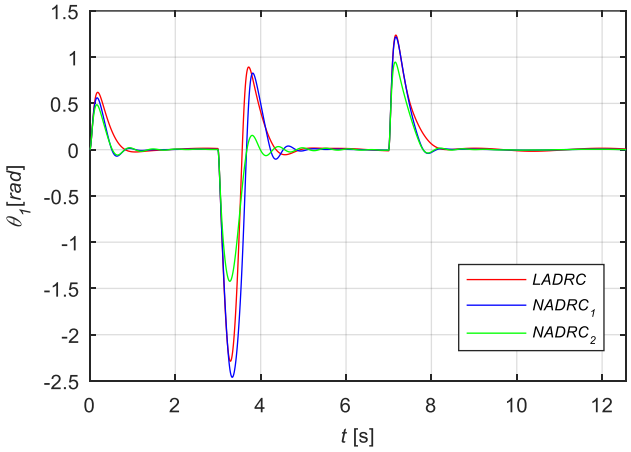
Performanse praćenja i greške praćenja date su na Sl. 6 i Sl. 7 respektivno, dok je uticaj poremećaja na prvom kraku prikazan na Sl. 8.



Sl. 6. Performanse praćenja referentnog signala drugog kraka za scenario 2



Sl. 7. Greške praćenja referentnog signala drugog kraka za scenario 2



Sl. 8. Performanse praćenja referentnog signala prvogog kraka za scenario 2

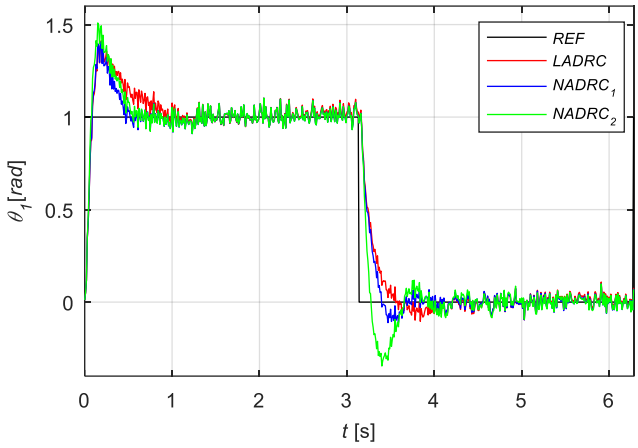
Sa slika se može uočiti da sva tri tipa regulatora zadržavaju zadovoljavajuće performanse uprkos značajnoj varijaciji parametara sistema. Takođe primećujemo da NADRC$_2$ sistem obezbeđuje najbolje potiskivanje poremećaja odskočnog tipa (najmanji preskok i najkraće vreme smirenja), što je posledica uvedenih nelinearnosti u diferencijalnom dejstvu regulatora.

Treći scenario je koncipiran tako da je zadržano nepoznavanje modela parametara b_i ($\hat{b}_{o1} = 2b_{o1}$ i $\hat{b}_{o2} = 2b_{o2}$), a simulirano je ponašanje sistema na referentne pobude

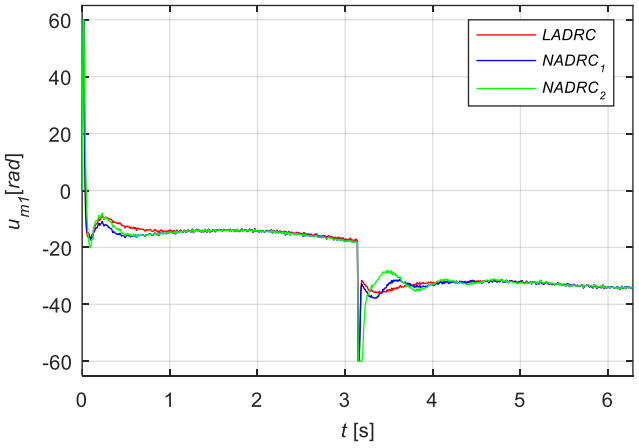
$$\theta_{r1} = \begin{cases} 1, & t \le T/2 \\ 0, & T/2 < t \le T \end{cases}, \; gde\ je\ T = 2\pi, \qquad (17)$$
$$\theta_{r2} = \sin(2\pi t).$$

Pored toga da bi uslove što više približili realnom sistemu dodat je šum merenja sa varijansom $\sigma_v^2 = 10^{-7}\, rad/s$ na oba kraka planarnog manipulatora i zasićenje upravljačkih signala aktuatora $\max(|u_{m1}|) = \max(|u_{m2}|) = \pm 60V$.

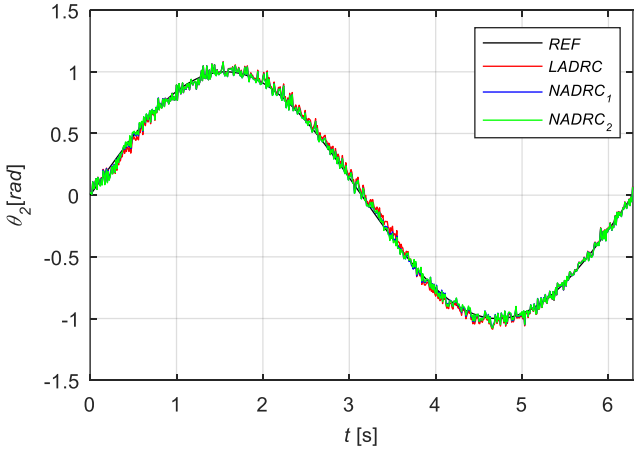
Karakteristike praćenja referentne pozicije prvog kraka date su na Sl. 9, a vrednosti upravljačkih signal odgovarajućih aktuatora na Sl. 10. Performanse praćenja i upravljački signali za drugi krak predstavljeni su na Sl. 11 i Sl. 12, respektivno.



Sl. 9. Performanse praćenja referentnog signala prvog kraka za scenario 3



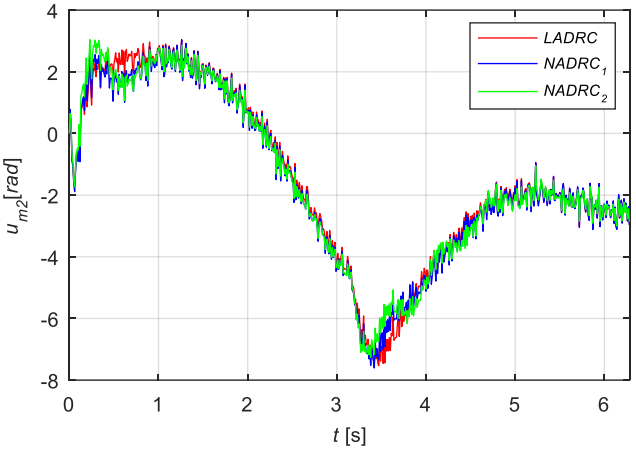
Sl. 10. Upravljački signal prvog kraka za scenario 3



Sl. 11. Performanse praćenja referentnog signala drugog kraka za scenario 3

Na osnovu izgleda upravljačkih signala (Sl. 10 i Sl. 12) se može zaključiti da regulatori imaju približno jednaku

osetljivost na šum merenja. U drugu ruku, ograničenje upravljačkih signala aktuatora ostvarilo je najveći uticaj na $NADRC_2$ (povećan preskok, duže vreme smirenja) i to u slučaju spoljašnjih odskočnih poremećaja, koji zahtevaju velike vrednosti upravljačkih signala.



Sl. 12. Upravljački signal drugog kraka za scenario 3

V. ZAKLJUČAK

U radu je izvršena analiza mogućnosti upravljanja planarnim manipulatorom primenom nelinearnog ADRC regulatora. Projektovanje upravljanja je realizovano pod pretpostavkom minimalnog poznavanja modela sistema. Detaljna simulaciona analiza je sprovedena kroz više upravljačkih scenarija za različite vrednosti parametara uvedenih nelinearnih funkcija. Rezultati su potvrdili prednosti nelinearnih algoritama u odnosu na linearne ADRC algoritme, kako u praćenju željene pozicije, tako i pri potiskivanju spoljašnjih poremećaja. Buduća istraživanja biće usmerena ka praktičnoj realizaciji razmatranih algoritama upravljanja, kao i na projektovanju složenijih proširenih opservera kako bi omogućilo potpuno potiskivanje šire klase totalnih poremećaja.

DODATAK

Parametri planarnog manipulatora:

L_1 =0.25 m L_2 =0.18 m
m_1 =0. 5 kg m_2 =0. 2 kg
η_1 =1/36 η_2 =1/20.25
J_1 =0.1 kg m^2 J_2 =0.1 kg m^2
F_1 =0.05 Nms/rad F_2 =0.05 Nms/rad

Parametri aktuatora:

R_i =20 Ω J_{mi} =0.01kg m^2 kg
k_{Ii} =1.5 Nm/A k_{ei} =1.5 V/rad/s

LITERATURA

[1] M. Przybyła, R. Madoński, M. Kodrasz and P. Herman, "An experimental comparison of model-free control methods in a nonlinear manipulator", Proc. International Conference on Intelligent Robotics and Applications, Berlin, Germany, pp. 53-62, December, 2011.

[2] S. Wolf and G. Hirzinger, "A new variable stiffness design: Maching requirements of the next robot generation," Proc. IEEE International Conference on Robotics and Automation, Pasadena, California, USA, pp. 1741-1746, May 2008.

[3] A. De Santis, B. Siciliano, A. De Luca and A. Bicchi, "An atlas of physical human-robot interaction," in Mechanism and Machine theory, vol. 43, no. 3, pp. *253*–270, 2008.

[4] R. Colombo, F. Pisano, A. S. Micera, Mazzone, C. Delconte, C. Carrozza, P. Dario and G. Micuno, "Robotic techniques for upper limb evaluation and rehabilitation of stroke patients," Proc. IEEE transaction on neural and rehabilitation engineering, vol. 13, no. 3, pp. 311-324, September 2005.

[5] K. H. Ang, G. Chong and Y. Li "PID control system analysis, designs and technology" Proc. IEEE transaction on control and control systems technology, vol. 13, no. 4, pp. 559-576, 2005.

[6] J. Han, "From PID to active disturbance rejection control", IEEE Trans. Ind. Electron., vol. 56, no. 3, pp. 900-906, 2009.

[7] M. Przybyła, M. Kodrasz, R. Madoński, P. Herman and P Sauer, "Active Disturbance Rejection Control of a 2DOF manipulator with significant modeling uncertaity", Bulletin of Polish Academy of Sciences Tehnical Sciences, Vol. 60(3), pp.509-520, . doi:10.2478/v10175-012-0064-z

[8] S. Chiaverini, S. Siciliano and O. Egeland, "Review of the damped least-squares inverse with experiments on industrial robot manipulator," Proc. IEEE Transaction on control and control systems technology, vol. 2, no. 2, pp. 123-134, 1994.

[9] J. Li, Y. Xia, X. Qi and Z. Gao, "On the necessity, scheme and basis of the linear-nonlinear switching in disturbance rejection control," Proc. IEEE Transaction on Industrial Electronics, vol. 64, no. 2, pp. 1425-1435, February 2017.

ABSTRACT

In this paper, Nonlinear Active Disturbance Rejection Control (NADRC) of planar robotic manipulator is proposed. Appalling ADRC approach, the complex nonlinear model of two inputs-two outputs planar manipulator system is transformed into two single input-single output (SISO) systems, including the cross-coupled dynamics in the total disturbance of two independent SISO systems. Further, the considered tracking problem is reformulated to regulation problem, and it is enabled the simpler structure of error-based NADRC controller. The detailed simulation analysis is shown the influence of the controller parameter setup on control performances.

Nonlinear Active Disturbance Rejection Control of Planar Robot Manipulator

Milan Svetozarević, Momir Stanković

Stabilnost linearnih dinamičkih sistema sa vremenskim kašnjenjem

Vukan Turkulov, Milan R. Rapaić, Rachid Malti

***Apstrakt*—U ovom radu bavimo se problemom određivanja oblasti stabilnosti linearnih, vremenski invarijantnih sistema sa višestrukim vremenskim kašnjenjima. Prikazani metod je iterativan, a zasniva se na primeni Rošeove teoreme i fundamentalne teoreme matematičke analize. Izlaganja su ilustrovana primerom.**

***Ključne reči*: stabilnost; analiza sistema; vremensko kašnjenje.**

I. Uvod

Stabilnost je jedna od osnovnih karakteristika dinamičkih sistema. Stabilnost linearnih, vremenski invarijantnih (Linear Time Invariant, LTI) sistema konačne dimenzije može se odrediti pomoću nula karakteristične jednačine sistema. Sistemi beskonačne dimenzije, kako im i samo ime kaže, imaju neograničen broj rešenja karakteristične jednačine. Otuda je za ovakve sisteme nepogodno, a najčešće i nemoguće, ispitivati stabilnosti direktnom proverom položaja svakog pola.

Među najčešće sretanim sistemima beskonačne dimenzije su sistemi sa vremenskim kašnjenjem. Karakteristična funkcija ovakvih sistema može se u opštem obliku zapisati kao

$$f(s, e^{-s\tau_1}, e^{-s\tau_2}, \ldots, e^{-s\tau_n}), \tag{1}$$

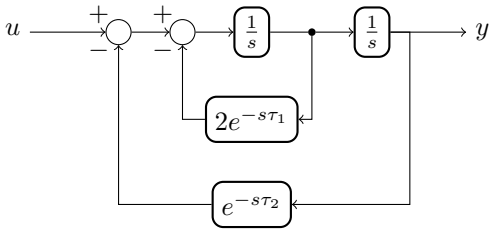
gde vrednosti $\tau_1, \tau_2, \ldots, \tau_n$ predstavljaju čista vremenska kašnjenja prisutna u sistemu. Pogodno je sva kašnjenja predstaviti jednim vektorom $\boldsymbol{\tau} = [\tau_1, \tau_2, \ldots, \tau_n]$. Od značaja je ispitati stabilnost sistema u zavisnosti od parametara $\boldsymbol{\tau}$, odnosno odrediti oblast stabilnosti sistema u prostoru parametara $\boldsymbol{\tau}$.

Ukoliko sistem poseduje vremensko kašnjenje isključivo na ulazu i/ili izlazu, funkcija prenosa sistema je oblika $G_0(s) = G(s)e^{-s\tau}$, gde je $G(s)$ racionalna funkcija prenosa. Stabilnost takvih sistema konvencionalno se ispituje Nikvistovim kriterijumom [1]. Metode za analizu nekih klasa sistema sa kašnjenjem predstavljene su u [2] [3] [4] i [5]. Interesantan alternativan pristup koji se zasniva na GMK prikazan je u [6].

Cilj ovog rada je uvođenje opšteg iterativnog postupka (algoritma) za analizu stabilnosti LTI sistema sa višestrukim vremenskim kašnjenjima. Bez gubitka opštosti, u ovom radu ćemo postupak primeniti na primer sistema sa višestrukim kašnjenjem opisan sledećim matematičkim modelom u prostoru stanja

$$\begin{aligned} \dot{x_1} &= -2x_1(t-\tau_1) - x_2(t-\tau_2) + u \\ \dot{x_2} &= x_1 \\ y &= x_2 \end{aligned} \tag{2}$$

gde je t vreme, x_1 i x_2 promenljive stanja sistema, u ulazni signal, y izlazni signal, a τ_1 i τ_2 čista vremenska kašnjenja. Blok dijagram sistema opisanog ovim modelom prikazan je na slici 1.



Slika 1. Blok dijagram sistema opisanog modelom (2).

Prilagodićemo pristup razvijen za analizu stabilnosti frakcionih sistema predložen u [7] i [8]. Postupak se zasniva na poznavanju neke određene vrednosti parametara $\boldsymbol{\tau^0}$ za koje je sistem stabilan. Počevši od date vrednosti $\boldsymbol{\tau^0}$, iterativno određujemo vrednosti $\boldsymbol{\tau}$ za koje je sistem takođe stabilan. Na ovaj način pravimo "skokove" po određenoj pravoj u prostoru parametara $\boldsymbol{\tau}$, garantujući da je sistem stabilan na svakom "preskočenom" intervalu.

Rad je podeljen u sledeće celine. U poglavlju II prikazana je matematička osnova predloženog postupka. U poglavlju III nalazi se primena predloženog postupka na sistem opisan modelom 2. Zaključak je dat u poglavlju IV.

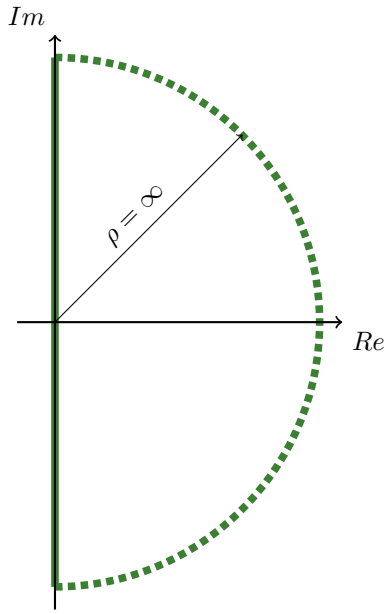
II. Opis algoritma

Predloženi postupak vrši analizu stabilnosti sistema duž odabrane prave u prostoru parametara $\boldsymbol{\tau}$. Pogodno je kretanje po odabranoj pravoj opisati jednim realnim parametrom θ kao

$$\boldsymbol{\tau} = \boldsymbol{\tau_0} + \theta\boldsymbol{\tau_p} \tag{3}$$

gde je $\boldsymbol{\tau_0}$ tačka kroz koju prolazi odabrana prava, a $\boldsymbol{\tau_p}$ jedinični vektor paralelan sa odabranom pravom. Stabilnost sistema dalje možemo diskutovati u zavisnosti od parametra θ. Postupak možemo potom primenjivati na proizvoljan broj pravih u prostoru parametara $\boldsymbol{\tau}$, analizirajući stabilnost duž svake od njih.

V. Turkulov (vukan_turkulov@uns.ac.rs), M. R. Rapaić (rapaja@uns.ac.rs) – Univerzitet u Novom Sadu, Fakultet tehničkih nauka, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija
R. Malti({firstname.lastname}@ims-bordeaux.fr) – Univ. Bordeaux, IMS – UMR 5218 CNRS, Francuska



Slika 2. Grafički prikaz konture $\mathcal{C}$ u kompleksnoj s ravni

Na odabranoj pravoj, karakterističnu funkciju sistema možemo preko novouvedenog parametra θ izraziti kao $f(s,\theta)$. Pretpostavka postupka je poznavanje neke vrednosti θ_0 za koju je sistem stabilan. Za takvo θ_0, funkcija $f(s,\theta_0)$ nema nijednu nulu u desnoj kompleksnoj poluravni. Takođe, pretpostavljamo da je funkcija $f(s,\theta)$ analitička u desnoj kompleksnoj poluravni za svaku vrednost parametra θ.

Predloženi postupak se oslanja na Rošeovu teoremu [9], koja je formulisana na sledeći način:

Teorema 1. Ukoliko za dve kompleksne funkcije f i g koje su holomorfne unutar zatvorene konture $\mathcal{K}$ važi $|g(s)| < |f(s)|$ za svako s na konturi $\mathcal{K}$, onda funkcije f i $f+g$ imaju isti broj nula unutar konture $\mathcal{K}$.

Rošeovu teoremu možemo primeniti na karakterističnu funkciju $f(s,\theta)$ i konturu $\mathcal{C}$ koja obuhvata celu desnu poluravan, kao što je prikazano na slici 2. Kontura $\mathcal{C}$ se sastoji iz dva segmenta - imaginarne ose gde je $s = j\omega$ za $\omega \in \mathbb{R}$, i polukruga beskonačnog poluprečnika gde je $s = \rho e^{j\varphi}$ za $\rho \to \infty, \varphi \in [-\frac{\pi}{2}, \frac{\pi}{2}]$.

Ukoliko je $|f(s,\theta) - f(s,\theta_0)| < |f(s,\theta_0)|$ za svako $s \in \mathcal{C}$, onda $f(s,\theta)$ i $f(s,\theta_0)$ imaju isti broj nula u desnoj poluravni. S obzirom na to da $f(s,\theta_0)$ nema nijednu nulu u desnoj poluravni, ovaj uslov bi garantovao da ni $f(s,\theta)$ nema nijednu nulu u desnoj poluravni. Dakle, sistem je sigurno stabilan za one vrednosti parametra θ za koje važi

$$|f(s,\theta) - f(s,\theta_0)| < |f(s,\theta_0)|, \forall s \in \mathcal{C}. \tag{4}$$

U interesu nam je da pronađemo što veću vrednost θ za koju važi uslov (4). Pored toga, uslov (4) predstavlja konzervativnu granicu stabilnosti, što ćemo ilustrovati sledećim primerom: recimo da postoji θ_1 takvo da nejednakost (4) važi za $\theta \in [\theta_0, \theta_1]$, a ne važi za $\theta > \theta_1$. Drugim rečima, θ_1 je najveća vrednost parametra θ za koju je uslov Rošeove teoreme ispunjen. Na osnovu te činjenice možemo zaključiti da je sistem sigurno stabilan za $\theta \in [\theta_0, \theta_1]$. Međutim, moguće je da postoji neko $\theta > \theta_1$ za koju je sistem takođe stabilan, iako uslov (4) nije ispunjen. Upravo je ovo razlog zbog kog je predloženi postupak iterativan. Kada pronađemo vrednost θ_1 za koju je sistem sigurno stabilan, ponavljamo postupak tražeći novu vrednost θ_2 za koju važi nejednakost $|f(s,\theta_2) - f(s,\theta_1)| < |f(s,\theta_1)|, \forall s \in \mathcal{C}$. Na ovaj način iterativno pravimo korake kroz prostor parametara, pri čemu možemo sa sigurnošću da tvrdimo da je sistem stabilan u svakom pređenom koraku.

Posmatrajmo sada uslov (4). Jednostavno je dokazati da data nejednakost važi na kružnom luku konture $\mathcal{C}$ ukoliko koeficijent uz najveći stepen promenljive s karakteristične funkcije sistema ne zavisi od parametra θ. Izazov je pokazati za koju vrednost θ data nejednakost važi na segmentu konture koji se poklapa sa imaginarnom osom. Primetimo da je funkcija $f(s,\theta)$ simetrična u odnosu na realnu osu za sve sisteme sa realnim koeficijentima. Zbog toga je dovoljno proveriti da li je uslov Rošeove teoreme ispunjen na gornjoj polovini imaginarne ose. Ovime se problem svodi na određivanje vrednosti θ za koje važi

$$|f(j\omega,\theta) - f(j\omega,\theta_0)| < |f(j\omega,\theta_0)|, \forall \omega \in \mathbb{R}^+. \tag{5}$$

Nažalost, kao što ćemo pokazati u poglavlju III, nejednakosti ovog oblika uglavnom nisu pogodne za rad. Iskoristićemo fundamentalnu teoremu analize [10] kako bismo dobili nejednakosti koje su konzervativnije, ali pogodnije za rad.

Teorema 2. Neka su $f(x)$ i $F(x)$ funkcije realne promenljive na zatvorenom intervalu $[a,b]$, tako da važi

$$F'(x) = f(x). \tag{6}$$

Ukoliko je funkcija f integrabilna na intervalu $[a,b]$, onda važi

$$\int_a^b f(x)dx = F(b) - F(a). \tag{7}$$

Fundamentalnu teoremu analize ćemo primeniti na levu stranu nejednakosti (5). Iako je f kompleksna funkcija, duž imaginarne ose je možemo posmatrati kao funkciju realne promenljive ω, koristeći zapis

$$f(j\omega,\theta) = f(\omega,\theta) = f_R(\omega,\theta) + jf_I(\omega,\theta) \tag{8}$$

gde su $f_R(\omega,\theta)$ i $f_I(\omega,\theta)$ realni i imaginarni delovi funkcije $f(\omega,\theta)$. Primenom fundamentalne teoreme na izraz sa leve strane nejednakosti (5) dobijamo

$$f(j\omega,\theta) - f(j\omega,\theta_0) = \int_{\theta_0}^{\theta} \frac{\partial f_R}{\partial \theta}(\omega,\beta)d\beta + j\int_{\theta_0}^{\theta} \frac{\partial f_I}{\partial \theta}(\omega,\beta)d\beta = \int_{\theta_0}^{\theta} \frac{\partial f}{\partial \theta}(\omega,\beta)d\beta \tag{9}$$

Uvrštavanjem dobijenog izraza, nejednakost (5) postaje

$$\left| \int_{\theta_0}^{\theta} \frac{\partial f}{\partial \theta}(j\omega,\beta)d\beta \right| < |f(j\omega,\theta_0)|, \forall \omega \in \mathbb{R}^+. \tag{10}$$

Iako nam je u interesu da pronađemo što veće θ za koje nejednakost važi, možemo radi jednostavnosti uzeti konzervativniju vrednost θ za koju možemo da tvrdimo da nejednakost važi. Kao posledica korišćenja konzervativnih vrednosti i nejednakosti, "skokovi" koje vršimo duž posmatrane prave biće manji. Drugim rečima, što konzervativnije vrednosti uzimamo, biće potrebno više iteracija da "pređemo" isti interval duž posmatrane prave.

Uzimajući to u obzir, levu stranu nejednakosti (10) možemo zameniti konzervativnijim izrazima

$$\left| \int_{\theta_0}^{\theta} \frac{\partial f}{\partial \theta}(j\omega, \beta) d\beta \right| \leq \int_{\theta_0}^{\theta} \left| \frac{\partial f}{\partial \theta}(j\omega, \beta) \right| d\beta \tag{11}$$

$$\int_{\theta_0}^{\theta} \left| \frac{\partial f}{\partial \theta}(j\omega, \beta) \right| d\beta \leq \left(\max_{\theta_0 \leq \beta \leq \theta} \left| \frac{\partial f}{\partial \theta}(j\omega, \beta) \right| \right) (\theta - \theta_0). \tag{12}$$

Uvođenjem smene $\Delta\theta = \theta - \theta_0$ koja predstavlja promenu ili "pomeraj" parametra θ i primenom konzervativnijih granica, nejednakost (10) postaje

$$\Delta\theta < \frac{|f(j\omega, \theta_0)|}{\max\limits_{\theta_0 \leq \beta \leq \theta} \left| \frac{\partial f}{\partial \theta}(j\omega, \beta) \right|}, \forall \omega \in \mathbb{R}^+. \tag{13}$$

Dakle, ako nejednakost (13) važi za neki pomeraj $\Delta\theta$, sistem će biti stabilan za $\theta = \theta_0 + \Delta\theta$. Pošto nejednakost mora da važi za svako $\omega \in \mathbb{R}^+$, "dozvoljene" korake možemo izraziti uzimajući u obzir najmanju moguću vrednost izraza sa desne strane nejednakosti:

$$\Delta\theta < \min_{\omega} \frac{|f(j\omega, \theta_0)|}{\max\limits_{\theta_0 \leq \beta \leq \theta} \left| \frac{\partial f}{\partial \theta}(j\omega, \beta) \right|} \tag{14}$$

Opšti pregled postupka za analizu stabilnosti sistema sa karakterističnom funkcijom $f(s, \theta)$ prikazan je kao Algoritam 1. Promenljiva ε koristi se kao kriterijum zaustavljanja algoritma.

Algoritam 1: Opšti pregled algoritma

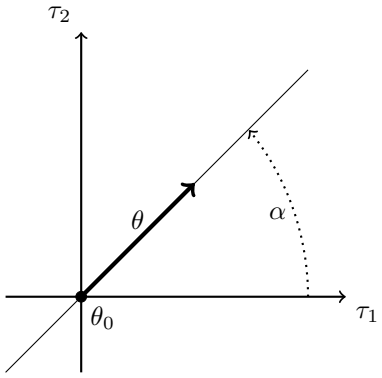
$\theta_k = \theta_0$;
$\Delta\theta = \infty$;
while $\Delta\theta > \varepsilon$ **do**
  $\Delta\theta = \min\limits_{\omega} \dfrac{|f(j\omega, \theta_k)|}{\max_{\theta_k \leq \beta \leq \theta} \left| \frac{\partial f}{\partial \theta}(j\omega, \beta) \right|}$;
  $\theta_k = \theta_k + \Delta\theta$;

III. Primer

Postupak ćemo primeniti na sistem sa karakterističnom funkcijom

$$f(s, \tau_1, \tau_2) = s^2 + 2se^{-s\tau_1} + e^{-s\tau_2}. \tag{15}$$

Stabilnost ispitujemo duž jedne prave u ravni parametara $\tau_1 O \tau_2$, kao što je prikazano na slici 3. Prava prolazi kroz



Slika 3. Ravan parametara $\tau_1 O \tau_2$

koordinatni početak, i određena je uglom α. Kretanje po pravoj opisujemo parametrom θ, tako da važi

$$\begin{aligned} \tau_1 &= \theta \cos(\alpha) \\ \tau_2 &= \theta \sin(\alpha) \end{aligned} \tag{16}$$

Radi preglednijeg zapisa, koristićemo oznake $a_1 = \cos(\alpha)$ i $a_2 = \sin(\alpha)$. Karakteristični polinom sistema izražen u zavisnosti od parametra θ je

$$f(s, \theta) = s^2 + 2se^{-sa_1\theta} + e^{-sa_2\theta}. \tag{17}$$

Za $\theta_0 = 0$, lako je utvrditi da je sistem stabilan. Za primenu algoritma neophodno je dokazati da nejednakost (4) važi na luku konture $\mathcal{C}$. U našem primer, nejednakost (4) postaje

$$\begin{aligned} &\left| 2s\left(e^{-sa_1\theta} - e^{-sa_1\theta_0}\right) + \left(e^{-sa_2\theta} - e^{-sa_2\theta_0}\right) \right| \\ &\qquad < \left| s^2 + 2se^{-sa_1\theta_0} + e^{-sa_2\theta_0} \right|. \end{aligned} \tag{18}$$

Smenom $s = \rho e^{j\varphi}$ za $\rho \to \infty, \varphi \in [-\frac{\pi}{2}, \frac{\pi}{2}]$, nejednakost (18) je očigledno ispunjena za bilo koju vrednost parametara θ i α zbog većeg stepena promenljive s u desnoj strani nejednakosti u odnosu na levu.

Ovde ukazujemo na razlog za korišćenje fundamentalne teoreme analize. Dokazivanje nejednakosti (18) na imaginarnoj osi za $s = j\omega$ nije jednostavno. Primenom fundamentalne teoreme analize dobijamo nejednakosti sa kojima je jednostavnije raditi. U skladu sa tim, dozvoljeni korak $\Delta\theta$ određujemo iz nejednakosti (14) umesto direktno iz (18).

Naravno, veličina koraka u opštem slučaju zavisi od parametara θ i α. Prvo ćemo da posmatramo imenilac iz nejednakosti (14) koji možemo označiti sa

$$c = \frac{1}{\max_{\theta_0 \leq \beta \leq \theta} \left| \frac{\partial f}{\partial \theta}(j\omega, \beta) \right|}. \tag{19}$$

Diferenciranjem funkcije $f(s, \theta)$ i smenom $s = j\omega$ dobijamo

$$\frac{\partial f}{\partial \theta} = 2\omega^2 a_1 e^{-j\omega a_1\theta} - j\omega a_2 e^{-j\omega a_2\theta}. \tag{20}$$

Apsolutna vrednost ovog izraza je

$$\left| \frac{\partial f}{\partial \theta} \right| = \sqrt{A^2 + B^2}, \tag{21}$$

gde je

$$\begin{aligned} A &= 2\omega^2 a_1 \cos(\omega a_1 \theta) - \omega a_2 \sin(\omega a_2 \theta) \\ B &= -2\omega^2 a_1 \sin(\omega a_1 \theta) - \omega a_2 \cos(\omega a_2 \theta) \end{aligned} \tag{22}$$

Dalje računamo vrednost c određenu izrazom (19). Pošto vrednost c figuriše u izrazu čiju minimalnu vrednost tražimo, možemo umesto tačne vrednosti promenljive c koristiti njenu konzervativniju granicu:

$$\begin{aligned} c &= \frac{1}{\max_{\theta_0 \le \beta \le \theta} \left| \frac{\partial f}{\partial \theta}(j\omega, \beta) \right|} \\ &\ge \frac{1}{\max_{\beta \in \mathbb{R}} \left| \frac{\partial f}{\partial \theta}(j\omega, \beta) \right|} \\ &\ge \frac{1}{\left| 2\omega^2 a_1 + \omega a_2 + j(2\omega^2 a_1 + \omega a_2) \right|} \\ &= \frac{1}{\sqrt{2}(2a_1\omega^2 + a_2\omega)} \end{aligned} \tag{23}$$

Sada nejednakost (14) postaje

$$\Delta\theta < \frac{1}{\sqrt{2}} \min_{\omega} \frac{|f(j\omega, \theta_0)|}{2a_1\omega^2 + a_2\omega}. \tag{24}$$

Računanjem apsolutne vrednosti u brojiocu, nejednakost postaje

$$\Delta\theta < \frac{1}{\sqrt{2}} \min_{\omega} \frac{(\omega^4 - 4\omega^3 s_1 + \omega^2(4 - 2c_2) + 4\omega s_\alpha + 1)^{\frac{1}{2}}}{2a_1\omega^2 + a_2\omega}. \tag{25}$$

gde je $s_1 = \sin(a_1\omega\theta_0)$, $c_2 = \cos(a_2\omega\theta_0)$ i $s_\alpha = \sin(\sqrt{2}\omega\theta_0 \cos(\alpha + \frac{\pi}{4}))$. Traženje minimuma ovog izraza bi u opštem slučaju bilo složeno, te ćemo opet tražiti konzervativniju granicu minimuma. Primetimo da za izraz od kog tražimo minimum važi

$$\begin{aligned} &\frac{(\omega^4 - 4\omega^3 s_1 + \omega^2(4 - 2c_2) + 4\omega s_\alpha + 1)^{\frac{1}{2}}}{2a_1\omega^2 + a_2\omega} \\ &\ge \frac{(\omega^4 - 4\omega^3 + 2\omega^2 - 4\omega + 1)^{\frac{1}{2}}}{2\omega^2 + \omega} = g(\omega). \end{aligned} \tag{26}$$

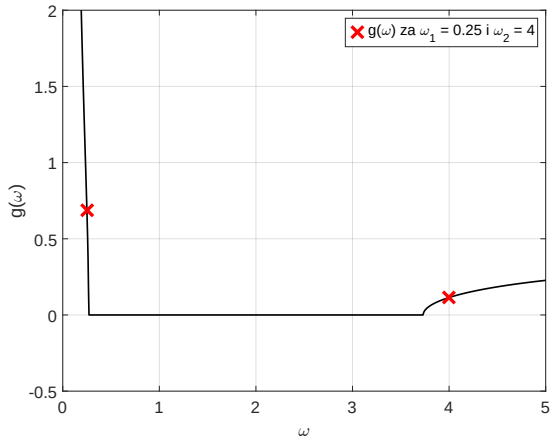
Nejednakost je dobijena uzimajući u obzir činjenicu da sinusne i kosinusne funkcije uzimaju vrednosti u intervalu $[-1, 1]$ (podsećamo da su a_1 i a_2 takođe prostoperiodične funkcije).

Lako je utvrditi da za $\omega \in (0, \frac{1}{4}] \cup [4, \infty)$ važi

$$\frac{(\omega^4 - 4\omega^3 + \omega^2 - 4\omega + 1)^{\frac{1}{2}}}{2\omega^2 + \omega} > 0.1. \tag{27}$$

Ovo tvrđenje je ilustrovano slikom 4. Skrećemo pažnju na činjenicu da je izraz pod korenom brojioca u (25) sigurno nenegativan. Uzimanjem najmanjih mogućih vrednosti trigonometrijskih funkcija, u korenu brojioca funkcije $g(\omega)$ dobijen je polinom koji je negativan za neke vrednosti promenljive ω. Zbog toga vrednost funkcije $g(\omega)$ može biti kompleksna. Prilikom crtanja ove funkcije, na mestima gde funkcija poprima kompleksne vrednosti nacrtana je vrednost 0.

Na osnovu (27) zaključujemo da je i vrednost izraza pod minimumom u nejednakosti (25) veća od 0.1 za $\omega \in (0, \frac{1}{4}] \cup [4, \infty)$. Dalje ćemo minimum tražiti na intervalu $\omega \in [\frac{1}{4}, 4]$. Zbog ograničenosti intervala, možemo uvesti gornje ograničenje na vrednost imenioca i tvrditi



Slika 4. Prikaz funkcije $g(\omega)$

$$\begin{aligned} &\frac{1}{\sqrt{2}} \min_{\omega \in [\frac{1}{4}, 4]} \frac{(\omega^4 - 4\omega^3 s_1 + \omega^2(4 - 2c_2) + 4\omega s_\alpha + 1)^{\frac{1}{2}}}{2a_1\omega^2 + a_2\omega} \ge \\ &\frac{1}{36\sqrt{2}} \Big(\min_{\omega \in [\frac{1}{4}, 4]} (\omega^4 - 4\omega^3 s_1 + \omega^2(4 - 2c_2) + 4\omega s_\alpha + 1) \Big)^{\frac{1}{2}}. \end{aligned} \tag{28}$$

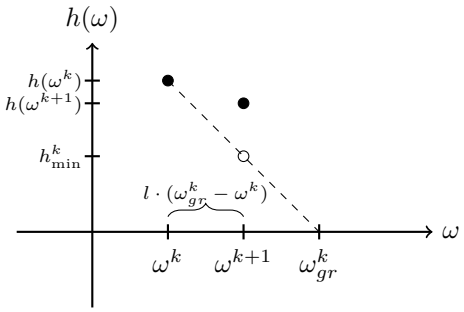
Problem se sada svodi na određivanje minimuma izraza $h(\omega) = \omega^4 - 4\omega^3 s_1 + \omega^2(4 - 2c_2) + 4\omega s_\alpha + 1$ na intervalu $\omega \in [\frac{1}{4}, 4]$. Zbog ograničenosti intervala, moguće je pronaći gornju granicu za vrednost izvoda funkcije h na posmatranom intervalu:

$$\max_{\omega \in [\frac{1}{4}, 4]} \left| \frac{dh}{d\omega} \right| \le 500 + |\theta_0|(256a_1 + 32a_2 + 16\sqrt{2}|a_3|) \tag{29}$$

Izvedenu gornju granicu izvoda funkcije h ćemo označiti simbolom γ. Koristeći ovu činjenicu, moguće je numerički pronaći donju granicu minimuma funkcije $h(\omega)$ na intervalu $\omega \in [\frac{1}{4}, 4]$. Počnimo od tačke $\omega = \frac{1}{4}$. Računanjem vrednosti funkcije u toj tački (koja je sigurno pozitivna) i poznavajući gornju granicu izvoda funkcije na datom intervalu, zaključujemo da je funkcija $h(\omega)$ sigurno veća ili jednaka nuli na intervalu $\omega \in [\frac{1}{4}, \frac{h(\frac{1}{4})}{\gamma}]$. Najveću vrednost ω za koju možemo da tvrdimo da važi $h(\omega) \ge 0$ označićemo sa $\omega_{gr} = \frac{h(\frac{1}{4})}{\gamma}$. Možemo uzeti proizvoljan broj iz intervala $[\frac{1}{4}, \omega_{gr})$ kao novu vrednost promenljive ω, i ponoviti postupak. Na ovaj način iterativno prelazimo interval $[\frac{1}{4}, 4]$, uvek birajući novu vrednost promenljive ω po pravilu

$$\omega^{k+1} = \omega^k + l(\omega_{gr}^k - \omega^k), \tag{30}$$

gde promenljiva $l \in (0, 1)$ određuje na koji način biramo novu vrednost ω^{k+1} iz dozvoljenog intervala $[\omega^k, \omega_{gr}^k]$.



Slika 5. Grafički prikaz numeričkog traženja minimuma

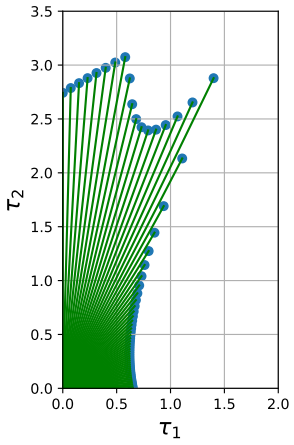
Donju granicu minimuma prilikom svakog koraka ažuriramo po pravilu

$$h^k_{\min} = \min\left(h^{k-1}_{\min},\ (1-l)h(\omega^k)\right). \tag{31}$$

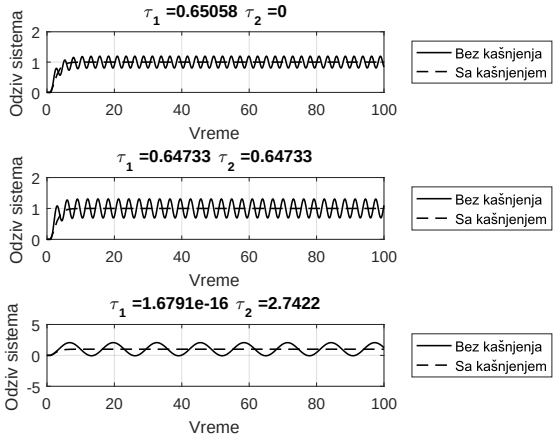
Jedna iteracija ovog postupka grafički je prikazana na slici 5. Zanimljivo je proučiti uticaj parametra l na efikasnost algoritma. Što je l veće, broj iteracija neophodnih da se "pređe" interval $[\frac{1}{4}, 4]$ biće manji, ali će dobijena donja granica minimuma biti konzervativnija. Tokom eksperimenata smo za vrednost ovog parametra koristili $l = \frac{1}{2}$.

Koristeći navedene postupke, moguće je implementirati Algoritam 1 za primer koji smo analizirali u ovom radu. Dati algoritam pronalazi granicu stabilnosti duž jedne prave u ravni parametara $\tau_1 O \tau_2$. Na slici 6 nalazi se grafički prikaz rezultata algoritma primenjenog na više prava u ravni parametara, koristeći kriterijum zaustavljanja $\varepsilon = 10^{-5}$. Računarski kod koji reprodukuje dobijene rezultate javno je dostupan na repozitorijumu https://github.com/Moon-Raven/stability_analysis. Numeričkom simulacijom moguće je proveriti stabilnost krajnje tačke sa svake prave, kako bi se utvrdio ispravan rad algoritma. Rezultati simulacije za tri takve tačke prikazani su na slici 7. Za svaku tačku dat je uporedan prikaz odziva sistema bez kašnjenja i sa kašnjenjima odgovarajućim za tu tačku. U svim slučajevima sistem je pobuđen step funkcijom $u(t) = h(t)$.

U ovom relativno jednostavnom primeru, stabilnost je moguće ispitati i iterativnom primenom Nikvistovog kriterijuma. Diskutovaćemo dobijene rezultate i na ovaj način, u cilju dodatne verifikacije dobijene oblasti stabilnosti. Posmatrani sistem sadrži unutrašnju i spoljašnju petlju, kao što se vidi na slici 1. Zbog toga je neophodno Nikvistov kriterijum primeniti prvo na unutrašnju petlju sistema, potom na spoljašnju. Na slici 8 prikazani su Nikvistovi dijagrami za tri odabrane tačke iz ravni parametara. Sve tri tačke pripadaju pravoj određenoj uglom $\alpha = \frac{\pi}{4}$ i za svaku tačku su prikazani dijagrami unutrašnje i spoljašnje petlje. Na osnovu skicirane oblasti stabilnosti sa slike 6, očekivano je da se prva tačka nalazi unutar oblasti stabilnosti, da je druga tačka bliska granici stabilnosti i da je treća tačka nestabilna. Prikazani Nikvistovi dijagrami potvrđuju očekivane rezultate; unutrašnja petlja sistema je uvek stabilna, ali je spoljašnja petlja granično stabilna za drugu i nestabilna za treću tačku.



Slika 6. Grafički prikaz rezultata

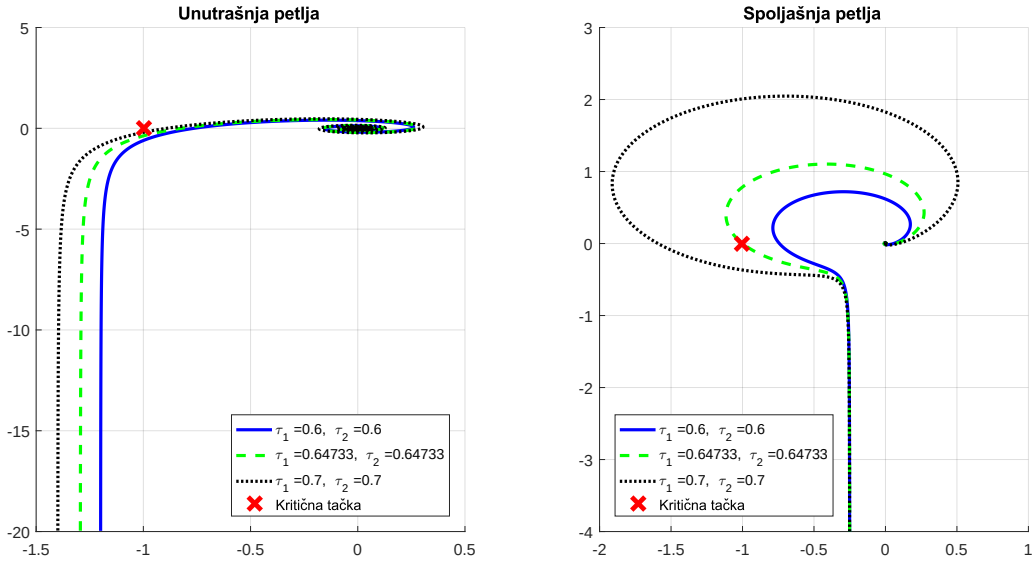


Slika 7. Numerička simulacija krajnjih tačaka intervala

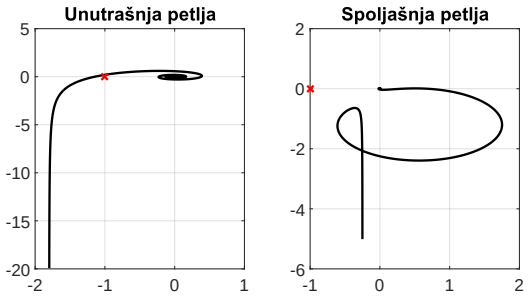
Ovde ukazujemo i na činjenicu da je za diskusiju Nikvistovog dijagrama spoljašnje petlje potrebno poznavati broj nestabilnih polova unutrašnje petlje. Drugim rečima, neophodno je porediti broj obuhvata oko kritične tačke za spoljašnju i unutrašnju petlju. Ovo ćemo ilustrovati Nikvistovim dijagramima posmatranog sistema za kašnjenja $\tau_1 = \tau_2 = 0.8$. Odgovarajući dijagrami su prikazani na slici 9. Činjenica da Nikvistova kriva spoljašnje petlje ne obuhvata kritičnu tačku $-1 + 0j$ nam ukazuje na to da je sistem nestabilan, zbog toga što Nikvistova kriva unutrašnje petlje obuhvata kritičnu tačku jednom.

IV. Zaključak

U ovom radu smo prikazali algoritam za ispitivanje stabilnosti dinamičkih sistema sa višestrukim vremenskim kašnjenjima. Iako je postupak ilustrovan na primeru sistema sa dva vremenska kašnjenja, jasno je da se na direktan način može proširiti i na sisteme sa većim brojem kašnjenja. Prikazanim postupkom, stabilnost se ispituje duž fiksiranih



Slika 8. Nikvistovi dijagrami spoljašnje i unutrašnje petlje za tri odabrane tačke iz ravni parametara



Slika 9. Nikvistovi dijagrami sistema za $\tau_1 = \tau_2 = 0.8$

pravaca u prostoru kašnjenja. Međutim, kao što je pokazano u radu, uzastopnom primenom na niz ovakvih pravaca moguće je skicirati oblik oblasti stabilnosti.

U daljem radu, bavićemo se proširenjem prikazanog algoritma. Konkretno, bavićemo se stabilnošću šire klase sistema sa beskonačno stepeni slobode, uključujući i sisteme sa distribuiranim parametrima. Takođe, po ugledu na [7] i [8], bavićemo se proširenjem algoritma kojom se stabilnosti direktno garantuje u oblasti, a ne samo duž pravca.

Zahvalnica

M.R.R. se zahvaljuje na podršci Ministarstvu Prosvete, nauke i tehnološkog razvoja R. Srbije, projektima TR32012 i TR33018.

Literatura

[1] Eric A. Faulkner. *Introduction to the Theory of Linear Systems*. Springer, 1969.

[2] Su Juing-Huei, Fong I-Kong, and Tseng Chwan-Lu. Stability analysis of linear systems with time delay. *IEEE Transactions on Automatic Control*, 39(6):1341–1344, June 1994.

[3] Frédéric Gouaisbaut and Dimitri Peaucelle. Delay-dependent stability analysis of linear time delay systems. *IFAC Proceedings Volumes*, 39(10):54 – 59, 2006. 6th IFAC Workshop on Time Delay Systems.

[4] Grienggrai Rajchakit. Stability analysis of linear systems with time delays. *Journal of Sound and Vibration - J SOUND VIB*, 51, 01 2012.

[5] Abdelaziz Hmamed, Hicham El Aiss, and Ahmed El hajjaji. Stability analysis of linear systems with time varying delay: An input output approach. In *2015 IEEE 54th Annual Conference on Decision and Control (CDC)*, pages 1756–1761, 12 2015.

[6] Tomislav B. Šekara and Milan R. Rapaić. A revision of root locus method with applications. *Journal of Process Control*, 34:26 – 34, 2015.

[7] *Stability of fractional incommensurate systems*, July 2016.

[8] Rachid Malti and Milan Rapaić. Sufficient stability conditions of fractional systems with perturbed differentiation orders. *IFAC-PapersOnLine*, 50(1):14557 – 14562, 2017. 20th IFAC World Congress.

[9] C. Berg. *Complex Analysis*. Matematisk Afdeling, Københavns Universitet, 2008.

[10] S.G. Krantz and S.G. Krantz. *Handbook of Complex Variables*. Birkhäuser Boston, 1999.

Abstract

This paper focuses on the problem of obtaining a stability region of linear, time-invariant systems with multiple time delays. The presented method is iterative and is based on the application of the Rouché's theorem and the fundamental theorem of calculus. The method is illustrated with an example.

Stability of linear dynamical sistems with time delays
V. Turkulov, M. R. Rapaić, R. Malti

Podešavanja dinamike kliznih režima višeg reda kod linearnih sistema sa jednim ulazom

Boban Veselić, *Member, IEEE,* Čedomir Milosavljević, Branislava Draženović, *Senior Member, IEEE*, i Senad Huseinbegović, *Member, IEEE*

***Apstrakt*—U radu se razmatra problem podešavanja dinamike u kliznom režimu višeg reda kod linearnih sistema sa jednim ulazom. Predložena je metoda izbora klizne površi koja istovremeno obezbeđuje neophodni relativni red klizne promenljive za željeni klizni režim višeg reda, kao i željenu dinamiku po uspostavljanju datog kliznog režima. Pokazano je da je rešenje ovog problema jedinstveno i dat jednostavan način za njegovo nalaženje. Teorijski dobijeni rezultati su potvrđeni na numeričkim primerima i ilustrovani simulacionim rezultatima.**

***Ključne reči*—Klizni režimi višeg reda; Projektovanje klizne površi; Dinamika sistema u kliznom režimu; Podešavanje polova.**

I. Uvod

Sistemi upravljanja promenljive strukture (SUPS) sa kliznim radnim režimom (KR) [1,2] su jedna od značajnijih tehnika robusnog upravljanja zbog teorijske invarijantnosti u KR na poremećaje koji deluju u prostoru vektora upravljanja [3]. Ova osobina se u praktičnim realizacijama svodi na veliku robusnost sistema na parametarske i spoljne poremećaje. Osnovna prepreka širokoj primeni ove upravljačke tehnike je pojava tzv. četeringa usled neidelanosti prekidačkih elemenata u sistemu i nemodelovane dinamike. Ovaj neželjeni efekat se ispoljava kroz pojavu visokofrekvencijskih oscilacija, nedopustivih u nekim primenama.

U sklopu pokušaja pronalaženja načina za redukciju četeringa nastao je koncept KR višeg reda [4]. KR višeg reda su najpre razmatrani u sistemima sa jednim ulazom, a potom u sistemima sa više ulaza kao i u sistemima sa vremenski diskretnom obradom informacija [5-8]. Najznačajni rezultati su postignuti kod KR drugog reda [9,10].

Prvi korak u projektovanju SUPS je izbor klizne površi, tj. klizne promenljive. Ovim izborom se definiše željena dinamika sistema u KR duž te površi. Kod konvencionalnih KR (KR prvog reda) relativni red klizne promenljive mora biti jednak jedinici. Predloženo je nekoliko metoda izbora klizne promenljive koja obezbeđuje željenu dinamiku u KR prvog reda. Jedan pristup je transformacija sistema u tzv. regularnu formu [2] u kojoj je redukovana dinamika kliznog režima jasno uočljiva. U slučaju sistema sa jednim ulazom, moguće je projektovati kliznu površ bez transformacije sistema na način koji se bazira na primeni Akermanove formule [11]. Međutim, razvijena je i sveobuhvatna metoda kojom se takođe bez tranformacije sistema jodnostavno projektuje klizna površ za sisteme sa jednim ili više ulaza [12,13].

U slučaju KR višeg reda, relativni red klizne promenljive mora biti jednak redu KR. Dakle, u projektovanju klizne površi treba ostvariti dvojaki zadatak. Obezbediti željenu dinamiku redukovanog reda u KR i istovremeno zadovoljiti preduslov potrebnog relativnog reda klizne promenljive. Mali je broj radova koji tretiraju ovu problematiku, i do sada su posmatrani sistemi sa jednim ulazom. Generalizacija formule Akerman-Utkin [11] za projektovanje klizne površi proizvoljnog relativnog reda sa ciljem podešavanja dinamike sistema izvršena je u [14,15].

U ovom radu je predložen još jedan način projektovanja klizne površi KR višeg reda u sistemima sa jednim ulazom. Ova metoda se bazira na pristupu [12,13] koji se zasniva na analogiji sa projektovanjem standardne povratne sprege po stanju. Pored zahteva da se KR r-tog reda ostvari predefinisana redukovana dinamika $(n-r)$-tog reda zadavanjem željenog spektra polova, izbor klizne promenljive mora da zadovolji i preduslov da je njen relativni red r. U radu je pokazano da je kod sistema n-tog reda sa jednim ulazom moguće istovremeno zadovoljiti oba zahteva. Štaviše dokazano je da se formirani sistem jednačina sastoji od n linearno nezavisnih jednačina sa n nepoznatih, tj. da postoji jedinstveno rešenje za izbor klizne promenljive. Ovo omogućava da se do tog rešenja dođe na jednostavan način primenom pseudoinverzije matrica, koja će u ovom slučaju dovesti do ispravnog rešenja. Validnost predloženog rešenja je potvrđena na numeričkim primerima i ilustrovana simulacionim rezultatima.

II. Postavka problema

Posmatrajmo linearni sistem upravljanja sa skalarnim upravljanjem, čiji je model u prostoru stanja dat sa

$$\dot{x} = Ax + b(u + d). \qquad (1)$$

$x \in \mathbb{R}^n$ je raspoloživi vektor stanja, a $u, d \in \mathbb{R}$ su upravljanje i nepoznati ograničeni poremećaj, respektivno. A i b su konstantne matrice odgovarajućih dimenzija i sistem je kontrolabilan, tj. matrica kontrolabilnosti $Q_c = [b \;\; Ab \;\; \cdots \;\; A^{n-1}b]$ ima puni rang $(\mathrm{rank}Q_c = n)$. Očigledno je da poremećaj zadovoljava uslove poklapanja, tj.

Boban Veselić – Elektronski fakultet, Univerzitet u Nišu, Aleksandra Medvedeva 14, 18000 Niš, Srbija (e-mail: boban.veselic@elfak.ni.ac.rs).

Čedomir Milosavljević – Elektrotehnički fakultet, Univerzitet u Istočnom Sarajevu, Istočno Sarajevo, Bosna i Hercegovina, (e-mail: cedomir.milosavljevic@elfak.ni.ac.rs).

Branislava Draženović i Senad Huseinbegović – Elektrotehnički fakultet, Univerzitet u Sarajevu, Sarajevo, Bosna i Hercegovina (e-mails: brana_p@hotmail.com, shuseinbegovic@etf.unsa.ba).

Rezultati prikazani u radu su deo istraživanja u okviru projekta III44004 Ministarstva prosvete, nauke i tehnološkog razvoja Republike Srbije.

deluje na sistem kroz upravljački kanal. Treba primetiti da su sve promenljive u (1) funkcije vremena, ali da su vremenski argumenti radi kraće notacije tu i nadalje izostavljeni. Zadatak je organizovati KR r-tog reda u sitemu (1) pri čemu je potrebno ostvariti željenu dinamiku u kliznom režimu.

Neka je klizna promenljiva $g \in \mathbb{R}$ definisana kao

$$g = cx, c \in \mathbb{R}^{1\times n}. \quad (2)$$

Kretanje sistema (1), (2) u potprostoru definisanog sa

$$g = \dot{g} = \ddot{g} = \cdots = g^{(r-1)} = 0 \quad (3)$$

se naziva KR r-tog reda, [4]. Upravljanje koje obezbeđuje da se u sistemu (1), (2) za konačno vreme ostvari uslov (3) mora biti diskontinualno, barem na skupu (3), [6], uz uslov da klizna promenljiva bude relativnog reda r u odnosu na upravljanje. To znači da se upravljanje pojavljuje tek u r-tom izvodu g, tj. u $g^{(r)}$. Za r-ti izvod klizne promenljive se dobija

$$g^{(r)} = cA^r x + \sum_{j=0}^{r-1} cA^{r-1-j} b(u^{(i)} + d^{(i)}). \quad (4)$$

Uslov da klizna promenljiva ima zahtevani relativni red se može iskazati kao

$$c \cdot [b \quad Ab \quad \cdots \quad A^{r-2}b] = 0_{1\times(r-1)}, \quad (5)$$

$$cA^{r-1}b \neq 0. \quad (6)$$

Tada r-ti izvod klizne promenljive (4) postaje

$$g^{(r)} = cA^r x + cA^{r-1}b(u + d). \quad (7)$$

Dinamika sistema u KR r-tog reda je $(n-r)$-tog reda [7]. Ona se može naći korišćenjem ekvivalentnog upravljanja, koje se može odrediti iz uslova

$$g^{(r)}\big|_{u=u_{eq}} = 0, \quad (8)$$

što daje

$$u_{eq} = -(cA^{r-1}b)^{-1}cA^r x - d. \quad (9)$$

Zamenom ekvivalentnog upravljanja u (1) dobija se opis sistema u KR r-tog reda

$$\dot{x} = Ax + b(u + d)|_{u=u_{eq}} = [I - b(cA^{r-1}b)^{-1}cA^{r-1}]Ax = PAx = A_{eq}x. \quad (10)$$

Iz dinamike sistema u (idealnom) KR r-tog reda (10) se vidi da nema uticaja poremećaja d, te je sistem invarijantan na dejstvo poremećaja koji zadovoljavaju uslove poklapanja. Naravno, ekvivalentno upravljanje (9) je nemoguće fizički ostvariti, budući da zahteva poznavanje poremećaja. Takođe, lako se može uočiti da je matrica P idempotentna matrica, tj. da važi $P^2 = P$. To znači da P predstavlja projektor. Na osnovu osobina projektorskih matrica [16], nalazi se da je $\mathrm{rank}(P) = n - 1$. Onda sledi da je $\mathrm{rank}(A_{eq}) = \mathrm{rank}(PA) < n$, što znači da je $\det(A_{eq}) = 0$ te A_{eq} predstavlja singularnu matricu. Očigledno je da sopstvene vrednosti matrice A_{eq} definišu dinamiku posmatranog sistema u KR, koja treba biti stabilna.

Dakle, potrebno je odrediti vektor c koji će obezbediti željenu dinamiku sistema (10) u KR r-tog reda, kao i zahtevani relativni red r klizne promenljive (2). Jednostavan način izbora c predložen je u narednoj sekciji.

III. Projektovanje klizne površi

Kako je upravljanje (9) linearno, može se sagledati kao tradicionalna povratna sprega po stanju. Naime, model (10) koji opisuje dinamiku u KR se može predstaviti kao

$$\dot{x} = A_{eq}x = (A - b(cA^{r-1}b)^{-1}cA^r)x = (A - bk)x, \quad (11)$$

gde je $k \in \mathbb{R}^{1\times n}$ vektor pojačanja povratne sprege po stanju $u = -kx$. Kako je dinamika sistema u KR redukovanog reda, u slucaju KR r-tog reda dinamiku odlikuje r nultih sopstvenih vrednosti i $n - r$ sopstvenih vrednosti različitih od nule. Nenulte sopstvene vrednosti $\lambda_1, \lambda_2, \ldots, \lambda_{n-r}$ treba izabrati da budu stabilne i da definišu željenu dinamiku sistema u KR, koja se može opisati sledećim karakterističnim polinomom

$$\phi(s) = s^r(s - \lambda_1)(s - \lambda_2)\cdots(s - \lambda_{n-r}) = s^n + \beta_1 s^{n-1} + \cdots + \beta_{n-r}s^r. \quad (12)$$

Iz teorije je poznato da se može naći jedinstveni vektor pojačanja povratne sprege k u kontrolabilnom sistemu (11) koji će obezbediti željenu dinamiku (12). Jedan od načina nalaženja k je primena Akermanove formule

$$k = e_1 Q_c^{-1}\phi(A), e_1 = [0_{1\times(n-1)} \quad 1]. \quad (13)$$

Tako određeno k se može iskoristiti za nalaženje nepoznatog vektora c. Na osnovu (11) se može uspostaviti veza između c i k koja je data relacijom $(cA^{r-1}b)^{-1}cA^r = k$, koja se može posle sređivanja iskazati jednačinom $cA^{r-1}(A - bk) = 0_{1\times n}$. Vektor c koji zadovoljava ovu jednačinu obezbediće željenu dinamiku (12) sistema u KR r-tog reda. Međutim, pored ovog zahteva, c mora da obezbedi i traženi relativni red klizne promenljive u odnosu na upravljanje (5) i (6). Često se uslov (6) bira u obliku $cA^{r-1}b = 1$ radi daljeg uprošćavanja. Dakle, c mora da zadovolji sledeće jednačine

$$cA^{r-1}(A - bk) = 0_{1\times n}, \quad (14)$$

$$c \cdot [b \quad Ab \quad \cdots \quad A^{r-2}b] = 0_{1\times(r-1)}, \quad (15)$$

$$cA^{r-1}b = 1. \quad (16)$$

Sistem jednačina koji se formira na osnovu (14)-(16) se sastoji od $n + r$ skalarnih jednačina sa n nepoznatih elemenata vektora c. Ovaj sistem jednačina se može u matričnom obliku predstaviti kao

$$c[A^{r-1}(A - bk) \quad b \quad Ab \quad \cdots \quad A^{r-2}b \quad A^{r-1}b] = [0_{1\times n} \quad 0_{1\times(r-1)} \quad 1] \quad (17)$$

Sledećom teoremom je dato jedinstveno rešenje datog sistema jednačina.

Teorema 1: U kontolabilnom sistemu (1), vektor c koji predstavlja jedinstveno rešenje sistema jednačina (14)-(16) se može naći kao

$$c = [k \quad 0_{1\times(r-1)} \quad 1] \cdot [A^r \quad b \quad Ab \quad \cdots \quad A^{r-1}b]^{\dagger}, \quad (18)$$

gde simbol † označava pseudoinverziju.

Dokaz: Kako je sistem (1) kontrolabilan, postoji nesingularna matrica transformacije T $(x = T\hat{x})$ koja transformiše sistem u kontrolabilnu kanoničku formu, koja je opisana matricama

$$\hat{A} = T^{-1}AT = \begin{bmatrix} 0_{(n-1)\times 1} & I_{n-1} \\ -a_n & -a \end{bmatrix}, \hat{b} = T^{-1}b = \begin{bmatrix} 0_{(n-1)\times 1} \\ 1 \end{bmatrix},$$
$$a = [a_{n-1} \quad a_{n-2} \quad \cdots \quad a_1],$$
$$\det(sI - A) = s^n + a_1 s^{n-1} + \cdots + a_{n-1}s + a_n. \quad (19)$$

Uzimajući u obzir relacije

$$A = T\hat{A}T^{-1}, b = T\hat{b}, c = \hat{c}T^{-1}, k = \hat{k}T^{-1}, \quad (20)$$

lako se može pokazati da je sistem jednačina (14)-(16) ekvivalentan sledećem sistemu jednačina

$$\hat{c}\hat{A}^{r-1}(\hat{A} - \hat{b}\hat{k}) = 0_{1\times n}, \quad (21)$$

$$\hat{c} \cdot [\hat{b} \quad \hat{A}\hat{b} \quad \cdots \quad \hat{A}^{r-2}\hat{b}] = 0_{1\times(r-1)}, \quad (22)$$

$$\hat{c}\hat{A}^{r-1}\hat{b} = 1. \quad (23)$$

Vektor pojačanja povratne sprege $\hat{k}$ transformisanog sistema koji će obezbediti željenu dinamiku (12) se može naći primenom Akermanove formule

$$\hat{k} = e_1\hat{Q}_c^{-1}\phi(\hat{A}), \quad (24)$$

gde je $\hat{Q}_c = [\hat{b} \quad \hat{A}\hat{b} \quad \cdots \quad \hat{A}^{n-1}\hat{b}]$. Poznato je iz [17] da važi

$$e_1\hat{Q}_c^{-1} = e = [1 \quad 0_{1\times(n-1)}],$$
$$e\hat{A}^i = [0_{1\times i} \quad 1 \quad 0_{1\times(n-1-i)}], r \le i < n, \quad (25)$$
$$e\hat{A}^n = [-a_n \quad -a].$$

Korišćenjem (24), (19) i (12) uz osobinu (25), za $\hat{k}$ se dobija

$$\hat{k} = [-a_n \quad \cdots \quad -a_{n-r+1} \quad -(a_{n-r} - \beta_{n-r}) \quad \cdots \quad -(a_1 - \beta_1)]. \quad (26)$$

Ekvivalentna matrica spregnutog sistema $\hat{A}_{eq} = \hat{A} - \hat{b}\hat{k}$ za ovako dobijeno $\hat{k}$ se dobija kao

$$\hat{A}_{eq} = \hat{A} - \hat{b}\hat{k} = \begin{bmatrix} 0_{(n-1)\times 1} & I_{n-1} \\ 0 & \gamma \end{bmatrix}, \quad (27)$$
$$\gamma = [0_{1\times(r-1)} \quad -\beta], \ \beta = [\beta_{n-r} \quad \cdots \quad \beta_1].$$

Po ugledu na (17), sistem jednačina (21)-(23) se može zapisati u obliku

$$\hat{c}\hat{L} = [0_{1\times n} \quad 0_{1\times(r-1)} \quad 1],$$
$$\hat{L} = \left[\hat{A}^{r-1}(\hat{A} - \hat{b}\hat{k}) \quad \hat{b} \quad \hat{A}\hat{b} \quad \cdots \quad \hat{A}^{r-2}\hat{b} \quad \hat{A}^{r-1}\hat{b}\right]. \quad (28)$$

Množenjem (27) sa $\hat{A}^{r-1}$ sa leve strane može se pokazati da prva submatrica $\hat{A}^{r-1}(\hat{A} - \hat{b}\hat{k})$ matrice $\hat{L}$ ima stukturu

$$\hat{A}^{r-1}(\hat{A} - \hat{b}\hat{k}) = \begin{bmatrix} 0_{(n-r)\times 1} & 0_{(n-r)\times(r-1)} & I_{n-r} \\ 0 & 0_{1\times(r-1)} & -\beta \\ 0_{(r-1)\times 1} & M(a_n, a)_{(r-1)\times(r-1)} & N(a,\beta)_{(r-1)\times(n-r)} \end{bmatrix} \quad (29)$$

Takođe, može se pokazati da ostali elementi (kolone) matrice $\hat{L}$ imaju strukturu opisanu sa

$$\hat{A}^{\mathrm{i}}\hat{b} = \begin{bmatrix} 0_{(n-i-1)\times 1} \\ 1 \\ H(a)_{i\times 1} \end{bmatrix}, i = 0,1,\dots,r-1. \quad (30)$$

Da bi se moglo suditi o rešivosti sistema jednačina (21)-(23), tj. (28), potrebno je analizirati rang matrice $\hat{L}$. Najpre se polazi od submatrica $\hat{A}^{r-1}(\hat{A} - \hat{b}\hat{k})$. Na osnovu strukture (29) može se uočiti da rang ove matrice određuju njeni elementi I_{n-r} i $M(a_n, a)$, tj.

$$\mathrm{rank}\left(\hat{A}^{r-1}(\hat{A} - \hat{b}\hat{k})\right) = \mathrm{rank}(I_{n-r}) + \mathrm{rank}(M(a_n,a)) =$$
$$n - r + \mathrm{rank}(M(a_n,a)). \quad (31)$$

U slučaju kada su svi koeficijenti polinoma $\det(sI - A)$ jednaki nuli ($a_n = 0, \quad a = 0_{1\times(n-1)}$) tada je $\mathrm{rank}\left(M(0, 0_{1\times(n-1)})\right) = 0$, pa je $\mathrm{rank}\left(\hat{A}^{r-1}(\hat{A} - \hat{b}\hat{k})\right) = n - r$. Takođe, maksimalni mogući rang $M(a_n, a)$ je $\mathrm{rank}(M(a_n,a)) = r - 1$. To znači da se rang matrice (29) na osnovu (31) nalazi u opsegu

$$n - r \le \mathrm{rank}\left(\hat{A}^{r-1}(\hat{A} - \hat{b}\hat{k})\right) \le n - 1. \quad (32)$$

U slučaju kada matrica $M(a_n, a)$ ima puni rang $r - 1$, na osnovu (30) se može zaključiti da nijedna od $r - 1$ kolona $\hat{A}^{\mathrm{i}}\hat{b}, i = 0,1,\dots,r-2$ matrice $\hat{L}$ ne može doprineti povećanju ranga matrice $\hat{L}$. Sa druge strane kada je $\mathrm{rank}(M(a_n,a)) = r - 1 - j$, $j = 1,2,\dots,r-1$, tada svaka od j kolona $\hat{A}^{\mathrm{i}}\hat{b}, i = 0,1,\dots,j-1$ će povećati rang matrice $\hat{L}$ za jedan. To znači da će deficit ranga matrice $M(a_n, a)$ nadoknaditi određeni broj kolona $\hat{A}^{\mathrm{i}}\hat{b}, i = 0,1,\dots,r-2$ koje slede iza $\hat{A}^{r-1}(\hat{A} - \hat{b}\hat{k})$ u matrici $\hat{L}$. Dakle, može se pisati da je

$$\mathrm{rank}\left(\left[\hat{A}^{r-1}(\hat{A} - \hat{b}\hat{k}) \quad \hat{b} \quad \hat{A}\hat{b} \quad \cdots \quad \hat{A}^{r-2}\hat{b}\right]\right) = n - 1 \quad (33)$$

Poslednja kolona matrice $\hat{L}$ je $\hat{A}^{r-1}\hat{b}$, koja je na osnovu (30) oblika

$$\hat{A}^{\mathrm{r}-1}\hat{b} = \begin{bmatrix} 0_{(n-r)\times 1} \\ 1 \\ H(a)_{(r-1)\times 1} \end{bmatrix}, i = 0,1,\dots,r-1. \quad (34)$$

Može se videti na osnovu strukture (29) matrice $\hat{A}^{r-1}(\hat{A} - \hat{b}\hat{k})$ da će kolona $\hat{A}^{\mathrm{r}-1}\hat{b}$ povećati rang ove matrice za jedan, bez obzira na rang matrice $M(a_n, a)$. Uzimajući u obzir (33), dolazi se do zaključka da je

$$\mathrm{rank}(\hat{L}) = n, \quad (35)$$

te sistem jednačina (28), odnosno (21)-(23), sa n nepoznatih ima jedinstveno rešenje. Samim tim i ekvivalentni sistem jednačina (14)-(16) ima jedinstveno rešenje, tj. postoji jedinstvni vektor c koji zadovoljava sve tri jednačine.

Ako jednačinu (16) uvrstimo u (14), jednačina (14) se tada modifikuje u prostiji oblik

$$cA^r = k. \quad (36)$$

Onda se sistem koga čine jednačine (36), (15) i (16) može zapisati kao

$$c \cdot [A^r \quad b \quad Ab \quad \cdots \quad A^{r-1}b] = [k \quad 0_{1\times(r-1)} \quad 1]. \quad (37)$$

Budući da je ovaj sistem punog ranga, tj. ima n linearno nezavisnih jednačina sa n nepoznatih, do ispravnog jedinstvenog rešenja se može doći korišćenjem pseudoinverzije matrica, na način (18). Ovim je kompletiran dokaz.■

IV. Izbor algoritma upravljanja

Pokazano je u [6] da upravljanje u sistemu (1), (2) koje obezbeđuje da se za konačno vreme ostvari uslov (3) mora biti diskontinualno, barem na skupu (3). Ukoliko je adekvatnim izborom c ova dinamika stabilna, trajektorija sistema asimptotski konvergira duž (3) u koordinatni početak ($x \to 0$ za $t \to \infty$), uprkos dejstvu poremećaja koji zadovoljava uslove poklapanja.

Kako je projektovanjem c obezbeđeno da važi (16), r-ti izvod klizne promenljive (7) postaje

$$g^{(r)} = cA^r x + u + d. \quad (38)$$

Ako se upravljanje u sistemu (1), (2) formira kao u [7]

$$u = -cA^r x - \gamma(\xi), \ \xi = \left(g, \dot{g}, \dots, g^{(r-1)}\right), \quad (39)$$

r-ti izvod po vremenu od g postaje

$$g^{(r)} = -\gamma(\xi) + d. \quad (40)$$

Levant je predložio u [6] skup kvazi-kontinualnih funkcija $\gamma(\xi)$ za uspostavljanje kliznog režima r-tog reda za nelinearne sisteme sa skalarnim upravljanjem. Kompleksnost ovih funkcija raste sa porastom reda r, dok se četering redukuje. Ove funkcije se lako mogu primeniti za linearni slučaj (40).

Funkcije $\gamma(\xi)$ koje garantuju klizni režim višeg reda u konačnom vremenu se mogu izabrati kao nelinearne funkcije date u [5,6]. Na primer, funkcije $\gamma(\xi)$ za $r = 1,2,3$ su

respektivno date kao

$$\gamma(\xi) = \alpha \frac{g}{|g|}, \tag{41}$$

$$\gamma(\xi) = \alpha \frac{\dot{g} + \beta_1 |g|^{\frac{1}{2}} \text{sign}(g)}{|\dot{g}| + \beta_1 |g|^{\frac{1}{2}}}, \tag{42}$$

$$\gamma(\xi) = \alpha \frac{\ddot{g} + \beta_2 \left(|\dot{g}| + \beta_1 |g|^{\frac{2}{3}} \right)^{-\frac{1}{2}} \left(\dot{g} + \beta_1 |g|^{\frac{2}{3}} \text{sign}(g) \right)}{|\ddot{g}| + \beta_2 \left(|\dot{g}| + \beta_1 |g|^{\frac{2}{3}} \right)^{\frac{1}{2}}}, \tag{43}$$

gde su $\alpha, \beta_1, \beta_2 > 0$ izabrani da budu dovoljno veliki.

Treba istaći da je za formiranje upravljanja potrebno poznavanje sukcesivnih izvoda klizne promenljive, tj. $g^{(i)}$, $i = 0,1,\dots,r-1$. Međutim, u sistemima na koje deluju poremećaji koji zadovoljavaju uslove poklapanja, ovi izvodi se na osnovu (4) i (5) mogu dobiti kao

$$g^{(i)} = cA^i x,\ i = 0,1,\dots,r-1. \tag{44}$$

V. Ilustrativni primeri i simulacioni rezultati

Ispravnost predložene metode projektovanja klizne površi u slučaju KR višeg reda je proverena na numeričkim primerima i potkrepljena je simulacionim rezultatima. Posmatra se potpuno kontrolabilni sistem (1) petog reda čije su matrice

$$A = \begin{bmatrix} -1.129 & -2.262 & 2.21 & 0.465 & 1.663 \\ 1.829 & -0.743 & -1.344 & -3.802 & -6.199 \\ -1.06 & 1.87 & -1.375 & -2.324 & 0.181 \\ 0.399 & 4.190 & 1.258 & -1.062 & 0.908 \\ -2.89 & 5.645 & 1.334 & 0.24 & -1.45 \end{bmatrix},$$

$b^{\mathrm{T}} = [0.4136 \quad 0 \quad 0.144 \quad 0 \quad -0.7601]$.

Na sistem deluje složeni poremećaj $d(t) = 2\sin(4\pi t) + 2h(t-3)$. Zadatak upravljanja je da se sistem uprkos dejstvu poremećaja dovede iz početnog stanja $x(0) = [10 \quad -5 \quad 0 \quad -10 \quad 5]^{\mathrm{T}}$ u koordinatni početak organizovanjem kliznog režima drugog reda. Kako je u ovom slučaju $n = 5$ a $r = 2$, dinamika sistema u kliznom režimu je $n - r = 3$ reda. Neka je željena dinamika sistema definisana spektrom polova $p = [0 \quad 0 \quad -3 \quad -3 \quad -3]$. Korišćenjem (13) za pojačanja povratne sprege k se dobija

$k = [4.2365 \quad -12.9816 \quad -3.4063 \quad -0.8998 \quad -2.6053]$

Ovde treba uočiti da je matrica A punog ranga, te je matrica $A^{r-1}(A - bk)$ maksimalnog ranga $n - 1 = 4$. Primenom formule (18) dolazi se do tražene vrednosti vektora c

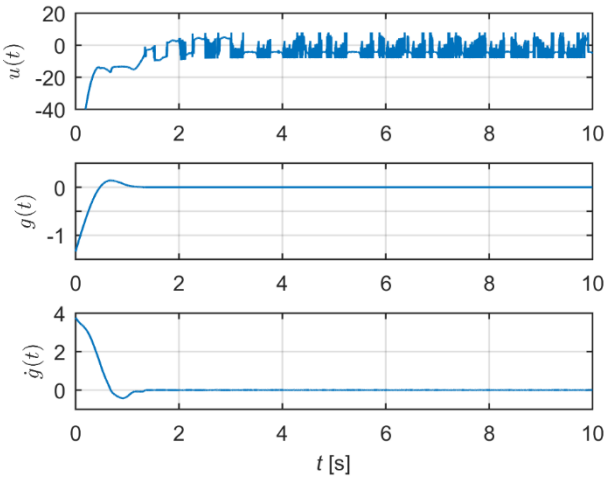
$$c = [0.052 \quad 0.2238 \quad 0.101 \quad 0.0976 \quad 0.0474].$$

Proverom je potvrđeno da dobijeno c zadovoljava jednačine (14)-(16) i da matrica $(A - b(cA^{r-1}b)^{-1}cA^r)$ iz (11) ima spektar sopstvenih vrednosti p.

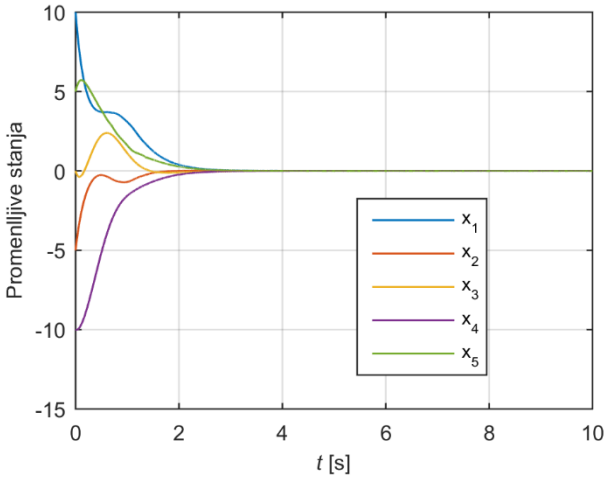
Navedena izračunavanja se lako mogu sprovesti u okviru MATLAB-a izvršavanjem sledećeg seta naredbi:

```
p=[0 0 -3 -3 -3]
k=acker(A,B,p)
c=[k 0 1]*pinv([A^2 B A*B])
```

Upravljanje koje u posmatranom sistemu obezbeđuje nastanak KR drugog reda je dato sa (39) i (42), pri čemu se izvod klizne promenljive $\dot{d}$ dobija korišćenjem (44). Za parametre regulatora je izabrano $\alpha = 8$ i $\beta_1 = 1$. Na Sl. 1 su prikazani upravljački signal $u(t)$, klizna promenljiva $g(t)$ i njen izvod $\dot{g}(t)$. Vidi se da upravljanje obezbeđuje $g = \dot{g} = 0$, tj. Nastanak KR drugog reda za konačno vreme uprkos dejstvu spoljnog poremećaja. Sl. 2 pokazuje da promenljive stanja asimptotski konvergiraju ka koordinatnom početku nakon nastanka KR, sa dinamikom definisanom spektrom polova p.



Sl. 1. Upravljački signal, klizna promenljiva i njen izvod



Sl. 2. Promenljive stanja

Na sledećem primeru se testira slučaj kada je matrica A singularna. Kao objekat upravljanja se posmatra serijska veza pet integratora, čiji model (1) karakterišu matrice

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},\ b = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 20 \end{bmatrix}.$$

Model je u kontrolabilnoj kanoničkoj formi, pri čemu su koeficijenti polinoma $\det(sI - A)$ dati sa $a_1 = a_2 = a_3 = a_4 = a_5 = 0$. U slučaju KR drugog reda, rang matrice $A^{r-1}(A - bk)$ je tada minimalan, tj. $n - r = 3$. Neka kao u prethodnom primeru, na sistem deluje isti spoljni poremećaj i ima istu željenu dinamiku u KR. Za pojačanje povratne sprege se sada na osnovu (13) dobija

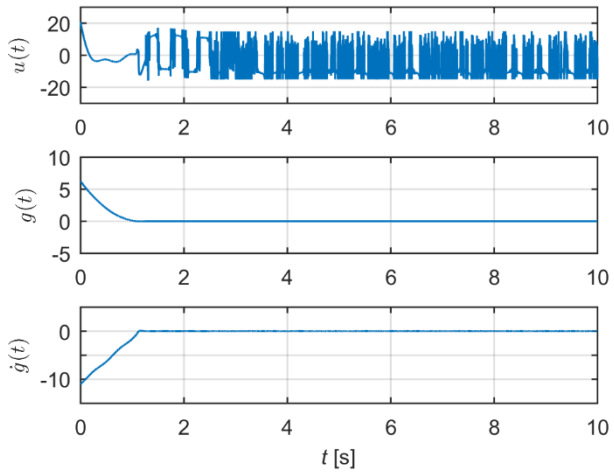
$$k = [0 \quad 0 \quad 1.35 \quad 1.35 \quad 0.45].$$

Korišćenjem (18) za c se dobija

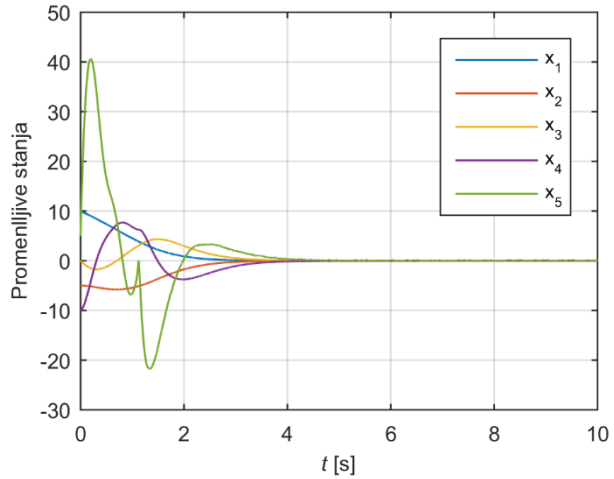
$$c = [1.35 \quad 1.35 \quad 0.45 \quad 0.05 \quad 0],$$

koje zadovoljava jednačine (14)-(16) i obezbeđuje da matrica $(A - b(cA^{r-1}b)^{-1}cA^r)$ ima spektar sopstvenih vrednosti p. Ovo pokazuje da i kada matrica $A^{r-1}(A - bk)$ ima najveći

mogući deficit ranga, formula (18) daje ispravno rešenje.



Sl. 3. Upravljački signal, klizna promenljiva i njen izvod



Sl. 4. Promenljive stanja

I ovde je primenjen isti algoritam upravljanja (39), (42) i (44) gde je $\alpha = 15$ i $\beta_1 = 1$. Simulacioni rezultati prikazani na Sl. 3 i Sl. 4 potvrđuju, da i u ovom slučaju dobijeni vektor klizne promenljive uz odgovarajući upravljački signal obezbeđuju nastanak KR drugog reda sa željenom dinamikom.

VI. Zaključak

U radu je pokazano da se princip projektovanja konvencionalne povratne sprege po stanju može iskoristiti prilikom projektovanja dinamike KR višeg reda u slučaju sistema sa jednim ulazom. Predložena metoda projektovanja klizne površi za slučaj KR višeg reda je jednostavna i ne zahteva velika izračunavanja. Dokazano je da isprojektovani vektor c obezbeđuje željenu dinamiku sistema u KR, kao i neophodni relativni red klizne promenljive koji je preduslov za uspostavljanje KR višeg reda. Validnost analitički dobijenog rešenja je potvrđena kroz numeričke primere i simulacione rezultate.

Dati prilaz projektovanja klizne površi se može u potpunosti primeniti i na diskretne KR r-tog reda dok se problem dosezanja i robusnosti mora rešavati na drugi način, što će biti predmet narednog istraživanja.

Literatura

[1] V. I. Utkin, "Variable structure systems with sliding mode," *IEEE Transactions on Automatic Control*, vol. 22, no. 2, pp. 212–222, 1977.

[2] V. I. Utkin, Sliding modes in control and optimization, Berlin, Germany: Springer-Verlag, 1992.

[3] B. Draženović, "The invariance conditions in variable structure systems," *Automatica*, vol. 5, no. 3, pp. 287-295, 1969.

[4] A. Levant, "Sliding order and sliding accuracy in sliding mode control," *Int. J. Contr.*, vol. 58, no. 6, pp. 1247-1163, 1993.

[5] A. Levant, "Homogeneity approach to higher-order sliding mode design" *Automatica*, vol. 41, pp. 823-830, 2005.

[6] A. Levant, "Quasi-continuous high-order sliding-mode controllers," *IEEE Trans. on Automatic Control*, vol. 50, no. 11, pp. 1812-1816. 2005.

[7] V.I. Utkin, "Discussion aspects of higher-order sliding modes," *IEEE Trans. on Automatic Control*, vol. 61, no. 3, pp. 829-833, 2016.

[8] A. Levant, M. Livne, "Uncertain disturbances' attenuation by homogeneous MIMO sliding mode control and its discretization," *IET Control Theory & Applications*, vol. 9, no. 4, pp. 515-525, 2015.

[9] A. Levant, "MIMO 2-sliding control design," Proc. Europian Control Conference ECC, Cambridge, UK, pp. 916-921, 1-4 Sept., 2003.

[10] G. Bartolini, A. Ferrara, E. Usai, V. I. Utkin, "On multi input chattering free second order sliding mode control," *IEEE Transactions on automatic Control*, vol. 45, no. 9, pp. 1711-1717, Sept., 2000.

[11] J. Ackermann, V. Utkin, "Sliding mode control design based on Ackermann's formula," *IEEE Trans. on Automatic Control*, vol. 43, no. 2, pp. 234-237, 1998.

[12] B. Peruničić, Č. Milosavljević, B. Veselić, V. Gligić, "Comprehensive approach to sliding subspace design in linear time invariant systems," Proc. of IEEE 12th Int. Workshop on Variable Structure Systems (VSS 2012), pp. 473-478, Mumbai, India, 2012.

[13] B. Draženović, Č. Milosavljević, B. Veselić, "Comprehensive Approach to Sliding Mode Design and Analysis in Linear Systems", in B. Bandyopadhyay, S. Janardhanan and S.K. Spurgeon (Eds.), *Advances in Sliding Mode Control: Concept, Theory and Implementation*, Lecture Notes in Control and Information Sciences, Vol. 440, Ch. 1, pp 1-19, Springer Berlin Heidelberg, 2013.

[14] D. Hernandez, F. Castanos, L. Fridman, "Pole-placement in higher-order sliding-mode control," Proc. 19th IFAC Word Congres, pp. 1386-1391, 2014.

[15] I. Castillo, F. Castaños, L Fridman, "Sliding Surface Design for Higher-Order Sliding Modes," in L. Fridman, J.P. Barbot, F. Plestan (eds.), *Recent Trends in Sliding Mode Control*, IET, 2016

[16] O.M.E. El-Ghezawi, A.S.I. Zinober, S.A. Billings, "Analysis and design of variable structure systems using a geometric approach," *Int. J. Control*, Vol. 38, No. 3, pp. 657–671, 1983.

[17] R.L. Williams, D.A. Lawrence, *Linear state space control systems*, John Wiley & Sons, New Jersey, 2007.

Abstract

The paper considers higher-order sliding mode dynamics design in single-input linear systems. The proposed method of sliding manifold selection simultaneously provides necessary relative degree of the sliding variable for a specific sliding mode order and desired system dynamics after establishing that sliding mode. It is shown that the solution of this problem is unique and a simple way of finding it is suggested. Theoretically obtained results are validated through numerical examples and illustrated by digital simulations.

Higher-Order Sliding Mode Dynamics Design in Single-Input Linear Systems

B. Veselić, Č. Milosavljević, B. Draženović, S. Huseinbegović

Prepoznavanje govora iz ograničenog rečnika primenom neuralne mreže

Emilija Kisić, Slobodan Drašković i Vera Petrović

Apstrakt— **Cilj ovog rada je da se pokaže kako se jednom efikasnom i kompleksnom obradom govornih signala i pravilnim izborom arhitekture neuralne mreže može napraviti sistem za prepoznavanje govora iz ograničenog rečnika koji će raditi sa velikom tačnošću. U ovom radu prikazan je kompletan postupak pravljenja sistema za prepoznavanje reči iz ograničenog rečnika i pokazano je da su neuralne mreže u kombinaciji sa poznatim metodama za modelovanje i obradu govornog signala jedan veoma moćan alat za prepoznavanje govora.**

***Ključne reči*—neuralne mreže; prepoznavanje govora; obrada signala.**

I. UVOD

Već decenijama predmet istraživanja brojnih naučnika je automatsko prepoznavanje govora (eng. *Automatic Speech Recognition*). Ova oblast doživljava puni razvoj tek prelaskom sa analognih na digitalne sisteme, a poslednjih godina, obeleženih opštim tehnološkim razvojem, dobija primenu u velikom broju aplikacija koje se mogu sresti u svakodnevnom životu [1], [2].

U ovom radu dat je detaljan opis pravljenja sistema za prepoznavanje govora primenom neuralnih mreža [3], [4]. Pravljenje sistema sastojalo se iz više koraka. Najpre je izvršeno snimanje govornih signala kako bi se oformila baza na osnovu koje će se obučavati neuralna mreža, zatim je usledila predobrada snimljenih signala koja se sastoji od *pre-emphazis* filtriranja, eliminisanja jednosmerne komponente iz signala i filtriranja na osnovni opseg govorne učestanosti [5]. Potom je napravljen efikasan algoritam koji određuje početak i kraj reči, kako bi se eliminisali početni i krajnji deo snimljene sekvence koji su „prazni“, odnosno u njima se nalazi samo termički šum, a zadržao se samo onaj deo signala koji nosi informaciju o reči. Nakon što je cela baza isečena, prelazi se na formiranje vektora karakterističnih obeležja kako bi se dala predstava posmatranih reči u višedimenzionalnom prostoru karakterističnih obeležja. Prvo je izračunat spektrogram govornog signala [8], a nakon toga je formirana pravougaona mreža kojom se spektrogram deli na *N*X*M* pravougaonih delova. Na kraju se prelazi na obučavanje neuralne mreže na bazi isečenih govornih signala koji su predstavljeni odgovarajućim vektorima karakterističnih obeležja. Svi algoritmi su implementirani u *Matlab* softverskom paketu. Na Sl. 1 dat je prikaz algoritma koji opisuje sve korake potrebne za pravljenje ovog sistema.

Sl. 1. Algoritam za prepoznavanje govora iz ograničenog rečnika primenom neuralne mreže

II. FORMIRANJE BAZE ZA OBUČAVANJE NEURALNE MREŽE

Prvi korak u razvoju sistema za prepoznavanje govora bio je snimanje baze koja će služiti za obučavanje neuralne mreže. Sistem je obučavan na ograničenom rečniku i za početak su izabrane reči koje će se u njemu nalaziti. Za elemente rečnika odabrane su cifre nula, jedan, dva, tri, četiri, pet, šest, sedam, osam i devet i reči napred, nazad, kreni, stani, levo i desno. Nakon što je odlučeno da dimenzija rečnika bude 16, obavilo se snimanje govornih signala. Za bazu je dalo glas 10 muških i 10 ženskih govornika, pri čemu je svako od njih izgovorio svaku od 16 reči po 10 puta. Na taj način, oformljena je baza od 3200 govornih signala, odnosno, od po 200 glasova za svaku klasu. Naravno, bilo bi bolje da je baza bila veća, jer bi na taj način sistem bio obučen za još više različitih glasova, ali i 3200 govornih signala predstavlja sasvim zadovoljavajući broj da se sistem kvalitetno obuči.

Svi govorni signali snimljeni su sa frekvencijom odabiranja od 8 kHz kako bi se zadovoljila Šenonova teorema [5]. S obzirom da je opseg učestanosti govornog signala u kome je sadržana gotovo sva informacija koja obezbeđuje potpunu razumljivost govorne poruke i većinu informacija o identitetu govornika od 250 Hz do 4kHz, za njegovu vernu reprezentaciju potrebno je vršiti odabiranje sa učestanošću od 8 kHz. Svi govorni signali snimljeni su u trajanju od 2s, što je bilo sasvim dovoljno da se zabeleže sve bitne informacije i omogući kvalitetno formiranje baze.

Za vreme snimanja govornih signala vodilo se računa o uslovima pri kojima se vrši snimanje. Naime, u zavisnosti od opreme (mikrofona i podešavanja računara) i buke za vreme snimanja koja potiče od okoline zavisi i kvalitet snimljenog signala. Zbog toga se nastojalo da se snimanje

Emilija Kisić– Visoka škola elektrotehnike i računarstva strukovnih studija, Vojvode Stepe 273, 11000 Beograd, Srbija (e-mail: emilija.kisic@viser.edu.rs).

Slobodan Drašković – Visoka škola elektrotehnike i računarstva strukovnih studija, Vojvode Stepe 273, 11000 Beograd, Srbija (e-mail: slobodan.draskovic@viser.edu.rs).

Vera Petrović – Visoka škola elektrotehnike i računarstva strukovnih studija, Vojvode Stepe 273, 11000 Beograd, Srbija (e-mail: vera.petrovic@viser.edu.rs).

vrši u prostoriji gde nema nijednog izvora buke, a za slučaj da se za vreme snimanja javio neki veći šum ili buka koja potiče spolja, snimanje se ponavljalo, u cilju formiranja što kvalitetnije baze. Takođe, veoma važno je bilo kako se drži mikrofon prilikom snimanja, da bude na dovoljnom odstojanju, da se vodi računa o pravilnom disanju, da se mikrofon ne trese i da se reči izgovaraju što spontanije (da se ne viče, ne priča pretiho), itd.

Svi signali snimljeni su u *Matlab*-u, pomoću naredbe *wavrecord*, a zatim su sačuvani kao *.wav* fajlovi. U slučaju da je došlo do pojave koja se zove clipp-ovanje, odnosno ako su određeni odbirci prešli maksimalnu vrednost koju oprema za snimanje dozvoljava, pa im je dodeljena maksimalna dozvoljena vrednost umesto njihove prave vrednosti, takvi signali su izbačeni i snimanje je ponovljeno.

III. PREDOBRADA SIGNALA

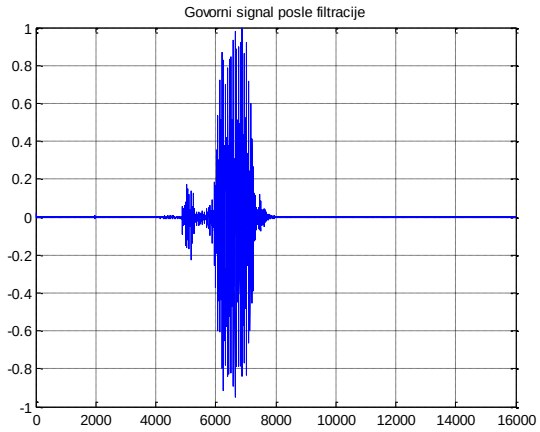
Nakon formiranje baze, sledeći korak je bio predobrada snimljenih signala. Za početak objasnićemo kako su govorni signali bili predstavljeni. Fizički posmatrano govorni signal predstavlja seriju promena vazdušnog pritiska u medijumu između zvučnog izvora i slušaoca. Najčešća i najjednostavnija predstava govornog signala je preko tzv. talasnog oblika. Horizontalna osa predstavlja vremensku osu, dok se na vertikalnoj osi može videti kako se pritisak povećava i smanjuje sa protokom vremena. Svi signali iz baze predstavljeni su u talasnom obliku, a na Sl. 2 nalazi se talasni oblik reči dva.

Svi signali najpre su prošli kroz *pre-emphasis* filtar. Naime, kada posmatramo raspored spektralnih komponenti po frekvencijama možemo uočiti da je njihov intenzitet na učestanostima iznad 1 kHz dosta mali. Međutim ako zanemarimo ove komponente i na primer signal propustimo kroz niskopropusni filtar učinićemo veliku grešku i izgubiti bitne informacije o govornom signalu. Jedna od metoda koja omogućava uspešno modelovanje formanata različitih intenziteta jeste *pre-emphasis*. Ideja je da se energija ulaznog govornog signala poveća za neku promenljivu vrednost u zavisnosti od učestanosti. Na taj način se smanjuje veliki opseg intenziteta u spektru govornog signala i „izravnjava“ spektar, čime se kasnija obrada govornog signala čini robusnijom i manje podložnom na uticaj konačne preciznosti pri računanju. Drugim rečima, *pre-emphasis* omogućava da informacije smeštene na višim učestanostima dođu do izražaja u kasnijoj analizi signala. Ravnanje spektra se realizuje dodavanjem nule u spektar signala, čime se parira padu signala od -12dB/oktava, do koga inače dolazi usled procesa nastanka govornog signala. *Pre-emphasis* filtar se definiše preko vrednosti konstante α. Kada se *pre-emhasis* filtriranje primeni na ulazni govorni signal dobija se:

$$y[n] = s[n] - \alpha s[n-1] \tag{1}$$

Za vrednost konstante α izabrana je vrednost 0.97 za koju se pokazalo da daje dobre rezultate.Nakon propuštanja kroz *pre-emphasis* filtar izvršeno je filtriranje Čebiševljevim VF filtrom šestog reda kako bi se odseklo prvih 250 Hz koji nam nisu potrebni za dalju obradu i kako bi se eliminisali neželjeni efekti nastali prilikom snimanja.

Na kraju je izvršena normalizacija signala tako da maksimalna vrednost apsolutne vrednosti odbirka bude 1 i eliminisana je jednosmerna komponenta koja se javlja na 50 Hz kao signal smetnje uzrokovan gradskom mrežom. Neutralizacija jednosmerne komponente je izvršena tako što je izračunata srednja vrednost svih odbiraka, a zatim je vrednost svakog odbirka umanjena za tu vrednost. Na Sl. 2 prikazan je talasni oblik reči „dva“ nakon celokupne predobrade.



Sl. 2. Talasni oblik reči „dva“ nakon celokupne predobrade

IV. ODREĐIVANJE POČETKA I KRAJA REČI

U procesu snimanja signala rešeno je da svi signali budu snimani u trajanju od 2s. Na taj način pored središnjeg dela snimljene sekvence u kome se nalazi govorna informacija, postoje i početni i krajnji deo koji su „prazni“, odnosno u njima se nalazi samo termički šum. Da bi algoritam mogao da vrši efikasno prepoznavanje reči potrebno je prvo iz snimljenog signala izdvojiti onaj deo signala koji nosi informaciju o reči, od termičkog šuma.

Postoji više načina da se ovaj problem reši. Prva ideja je bila da se pokuša segmentacija signala na osnovu njegove energije. Ovaj pristup izuzev što je jednostavan nema nekih posebnih prednosti, a rezultati koji su dobijeni nisu bili zadovoljavajući. Najveći problem predstavljala je detekcija kraja reči koje sadrže slovo T (pet, šest, devet). Problem sa ovim pristupom je što se na krajevima nekih reči nalaze visokofrekventne komponente u kojima se nalazi informacija o samoj reči, pri čemu je energija sadržana u tom delu govornog signala, dosta manja u odnosu na energiju niskofrekventnih delova signala. Stoga je određivanje granica reči u tim slučajevima jako teško, a često i nemoguće izvršiti samo na osnovu energije signala.

Nešto drugačiji pristup rešavanju ovog problema jeste korišćenje tzv. *Teager* energije za predstavu signala [6].
U [7] je pokazano da korišćenje *Teager* energije daje jako dobre rezultate kada je u pitanju detekcija „slabog“ početka i kraja reči u prisustvu šuma koji ima veću energiju od početka i kraja izgovorene reči. Osnovna ideja upotrebe *Teager* energije je da se istaknu komponente signala na višim frekvencijama. To je ostvareno množenjem odgovarajućih spektralnih komponenti kvadratom frekvencije.

Procedura računanja *Teager* energije se sastoji od nekoliko koraka. Prvo se signal podeli na preklapajuće segmente (okvire) dužine *W*. Za svaki od segmenata računa se *FFT*:

$$X(w) = \sum_{i=-\infty}^{\infty} s_i e^{-jwi} \tag{2}$$

gde je s_i vrednost odbirka signala u posmatranom segmentu.
Amplitude spektralnih komponenti se potom skaliraju vrednošću kvadrata odgovarajuće učestanosti :

$$f_k = w_k^2 X(w_k) \tag{3}$$

gde je sa f_k označena odgovarajuća *Teager* spektralna komponenta, a sa w_k njoj odgovarajuća učestanost. Na kraju se *Teager* energija signala dobija po formuli:

$$T_i = \left(\sum_{k=1}^{K} f_k \right)^{1/2} \tag{4}$$

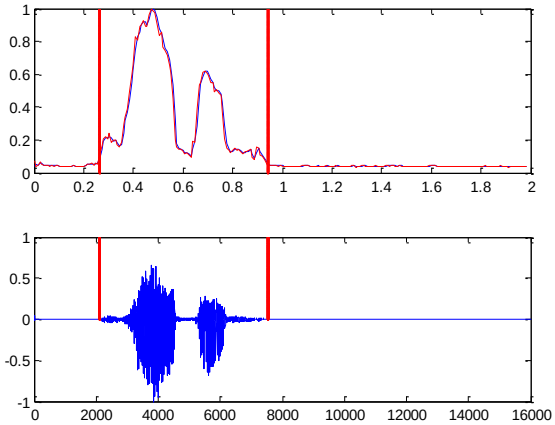
gde je *K* broj tačaka u kojima se računaju spektralne komponente.

Računanje *Teager* energije signala realizovano je je u *Matlab*-u. Prva dva koraka opisanog algoritma predstavljaju računanje spektrograma signala. Za to je korišćena *Matlab* funkcija *spectrogram*. Za širinu prozora uzeto je *W*=128, za veličinu preklapanja izabrana je vrednost *K*=*W*/2, a za broj tačaka u kojima se računa *FFT (Fast Fourier Transform)* uzeto je *M*=512. Izbor vrednosti za navedene parametre može da utiče na rezultat algoritma.

Nakon formiranja spektrograma, vrednost dobijenih spektralnih komponenti su pomnožene sa kvadratom učestanosti. Na taj način dobijena je *Teager* predstava signala, a zatim se *Teager* energija određenog segmenta dobija kao koren sume *Teager* spektralnih komponenti koje pripadaju tom segmentu. Dobijena predstava signala je prilično dinamična kriva sa velikim skokovima, pa je potrebno dodatno je isfiltrirati da bi se smanjila verovatnoća pogrešne detekcije granica reči. To je urađeno usrednjavanjem krive na prozoru dužine *N* (u ovom slučaju za vrednost *N* izabrana je *N*=2).

Sada treba dobijenu predstavu signala iskoristiti za određivanje granice reči. Ideja je da se definišu vrednosti pragova sa obe strane snimljenog signala (sa početka P_L i kraja P_R) i da se odrede uslovi pod kojim se proglašava početak/kraj reči. Algoritam se izvršava poređenjem nivoa *Teager* energije snimljenog signala sa odabranim pragovima. Leva granica se određuje kao trenutak t_L kada signal posmatran od početka ka kraju nadmaši vrednost levog praga. Slično tome desna granica se postavlja u trenutku t_R kad signal posmatran od kraja ka početku nadmaši vrednost desnog praga. Na Sl. 3 dat je prikaz *Teager* energije i talasnog oblika reči jedan sa pragovima t_L i t_R.

Vrednost pragova detekcije je određena na osnovu procenjenog nivoa šuma i usvojena je da bude jednaka *Teager* energiji šuma uvećanoj za 4% od maksimalne vrednosti energije. Ovakav izbor vrednosti praga je utvrđen eksperimentalno i predstavlja neku vrstu kompromisa. Smanjenjem vrednosti praga svakako bi se smanjio procenat nedetektovanih delova problematičnih reči, ali sa druge strane sistem bi bio manje robustan i jako osetljiv na smetnje koje se mogu pojaviti prilikom snimanja signala.



Sl. 3. *Teager* energija i talasni oblik reči "jedan" sa eksperimentalno određenim pragovima t_L i t_R.

Procenjeni nivo šuma je karakterisan *Teager* energijom signala na početku i na kraju snimljene sekvence. Podrazumevano je da se korisni signal nalazi u centralnom delu snimljenog niza, odnosno da se na krajevima niza nalazi samo termički šum.

Problem sa opisanim načinom detekcije granica reči je u tome što je jako teško izabrati veličinu praga koja bi odgovarala svim glasovima iz baze. Kao moguće rešenje za ovaj problem za svaku stranu detekcije postavljena su po dva praga umesto jednog. Vrednost većeg praga (P_{LH} i P_{RH}) je postavljena dovoljno visoko, tako da u svim slučajevima kada je nadmašena, sa sigurnošću možemo reći da je to usled postojanja govornog signala, a ne kao posledica varijacija termičkog šuma. Analogno tome vrednost manjeg praga (P_{LL} i P_{RL}) se postavlja da bude u stanju da detektuje najmanje varijacije u *Teager* energiji govornog signala. Trenuci u kojima se detektuje da je po prvi put nadmašena vrednost praga označeni su sa t_{LH} i t_{LL} za levu stranu i sa t_{RH} i t_{RL} za desnu stranu. Postavljen je i dodatni uslov, a to je da vrednost rastojanja između trenutaka kada su nadmašeni viši i niži prag mora biti manja od neke unapred definisane maksimalne vrednosti ($t_{L\max}$ i $t_{R\max}$). Odnosno taj uslov se može napisati kao:

$$t_{LH} - t_{LL} < t_{L\max} \tag{5}$$

$$t_{RH} - t_{RL} < t_{R\max} \tag{6}$$

Algoritam kojim se određuje početak i kraj reči se izvršava tako što se prvo odredi tačka u kojoj *Teager* energija signala postaje veća od višeg praga (t_{LH}), a potom se pretraga vrati za $t_{L\max}$ trenutaka unazad i počev od te pozicije treba naći momenat u kome je nadmašena vrednost nižeg praga (t_{LL}). Trenutak t_{LL} je proglašen za početak reči $t_L = t_{LL}$. Ukoliko postoji potreba moguće je još preciznije određivanje početka i kraja reči. U tom slučaju obično se računa i neka druga predstava signala i definiše se dodatna vrednost praga P_D u odnosu na tu predstavu signala. Potom se na intervalu $t_{LL} - t_{LH}$ traži trenutak kada je vrednost tog praga nadmašena. Na sličan način se određuje i kraj reči, sa razlikom što u tom slučaju pretraga počinje od kraja snimljene sekvence.

Sve vrednosti pragova su određene eksperimentalno, tako da daju dobre rezultate za postojeću bazu signala. Pri tome se naročito vodilo računa o osobenostima elemenata baze, pa tako na primer $t_{R\max}$ mora biti dovoljno dugačko da se ne bi odseklo slovo T iz pojedinih reči (devet, pet,...), odnosno prag t_{LL} mora biti dovoljno nizak da detektuje početak reči koje počinju bezvučnim samoglasnikom (sedam, šest,...).

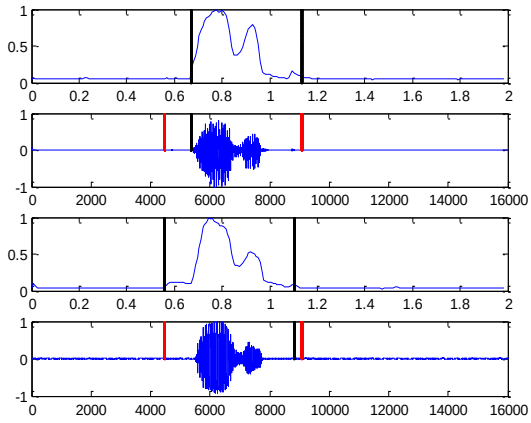
Prethodnim algoritmom ispravljena je većina nedostataka raznih drugih metoda koja se odnosila na lošu detekciju visokofrekfentnih komponenti i dobijeni su dosta dobri rezultati. Međutim, ostao je problem sa kasnom detekcijom početka nekih reči (dva, devet,...). Rešenje koje je primenjeno u ovom radu je da se opisani algoritam primeni dva puta. Prvi put, kao što je to i do sada bio slučaj, algoritam se primenjuje na govorni signal u osnovnom opsegu učestanosti (250Hz-4kHz), dok se u drugom slučaju algoritam primenjuje na signal koji je prethodno propušten kroz filtar propusnik opega učestanosti (250Hz-550Hz). Na taj način dobijaju se dva para granica reči (t_{L1}, t_{L2}) i (t_{R1}, t_{R2}) na osnovu kojih se donosi konačna odluka. Primenjen je sledeći princip odlučivanja:

$$t_L = \min\{t_{L1}, t_{L2}\} \tag{7}$$

$$t_R = \min\{t_{R1}, t_{R2}\} \tag{8}$$

gde su t_L i t_R granice reči.

Na Sl. 4 prikazano je ponašanje algoritma u slučaju izgovorene reči "devet". Reč se posmatra jer ilustruje dva bitna problema: detekciju kraja reči koje se završavaju na T i detekciju početka reči koje počinju na D. U gornjoj polovini Sl. 4 nalazi se grafik *Teager* energije signala i njegov talasni oblik, dok se na donjoj polovini nalazi prikaz *Teager* energije signala filtriranog na opseg 250-550 Hz i odgovarajući talasni oblik.

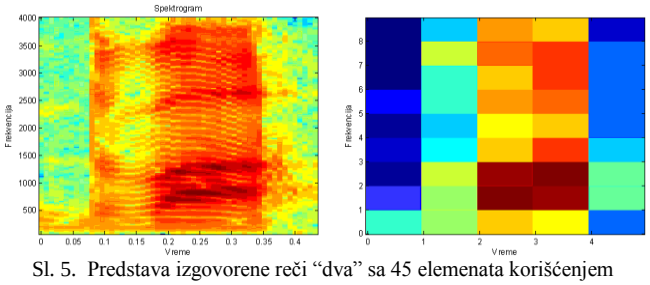


Sl. 4. Ilustracija *Teager* algoritma za izgovorenu reč "devet"

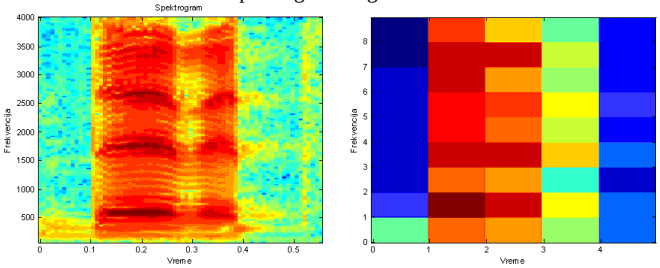
Na graficima *Teager* energije crnim vertikalnim linijama označene su granice reči onako kako to vidi algoritam za pojedinačne slučajeve signala. Kombinacijom ta dva rezultata opisan u (5) i (6) dobijaju se granice reči koje se mogu videti na slikama talasnih oblika gde su označene crvenim vertikalnim linijama. Kao što se može videti korekcija je neophodna bez obzira koji signal posmatrali. Algoritam opisan u ovom poglavlju primenjen je na sve signale iz baze i na taj način je dobijena baza skraćenih signala, na kojoj će se vršiti obučavanje neuralne mreže.

V. IZBOR KARAKTERISTIČNIH OBELEŽJA

Vektor karakterističnih obeležja daje predstavu posmatrane reči u višedimenzionalnom prostoru karakterističnih obeležja. Dimezije tog prostora i izbor samih obeležja zavise od mnogih faktora, a tačnost algoritma za prepoznavanje reči je u velikoj meri uslovljena dobrim izborom obleležja. Vektor karakterističnih obležja formiran je korišćenjem spektrograma signala. Nakon toga se formira pravougaona mreža kojom se spektrogram deli na *N*x*M* pravougaonih delova. Dimenzije mreže, odnosno izbor vremenske i frekvencijske rezolucije podele spektrograma, može značajno da utiče na tačnost sistema za prepoznavanje. Kada se odrede vrednosti za *N* i *M*, za svaki od pravougaonih segmenata treba izračunati srednju vrednost logaritama spektralnih komponenti unutar njega. Na ovaj način dobija se karakterističan vektor čija je dimenzija *N*x*M* elemenata. Kao što će kasnije biti pokazano, najbolji rezultati dobijeni su za vrednosti *N*=9 i *M*=5. Do ovog rezultata došlo se eksperimentalnim putem, tako što se neuralna mreža obučavala na bazi isečenih govornih signala koji su predstavljeni odgovarajućim vektorom karakterističnih obeležja za različite vrednosti *N* i *M*, pa je uzeta ona vrednost za koju je neuralna mreža dala najmanju grešku. Osim srednje vrednosti logaritama spektralnih komponenti, računat je i maksimum spektralnih komponenti unutar pravougaonih segmenata, kako bi se videlo u kom slučaju se dobijaju bolji rezultati. Eksperimentalnim putem utvrđeno je da se mnogo bolji rezultati dobijaju kada se računa srednja vrednost, tako da je za optimalnu podelu spektrograma izabrano 9X5 elemenata, pri čemu je vrednost svakog od ovih elemenata zapravo srednja vrednost logaritama spektralnih komponenti unutar odgovarajućeg pravougaonog segmenta. Na Sl. 5 data je predstava izgovorene reči "dva" pomoću spektrograma (levo) i vektora obeležja sa 45 elemenata (desno). Na Sl. 6 data je predstava izgovorene reči "devet" pomoću spektrograma (levo) i vektora obeležja sa 45 elemenata (desno).



Sl. 5. Predstava izgovorene reči "dva" sa 45 elemenata korišćenjem spektrograma signala.



Sl. 6. Predstava izgovorene reči "devet"sa 45 elemenata korišćenjem spektrograma signala.

Sa slika se vidi da se spektrogrami za ove dve reči veoma jasno razlikuju. Drugim rečima, predstave govornih signala na ovaj način biće veoma slične za govorne signale iz iste klase, a razlikovaće se za govorne signale iz različitih klasa. Kada svi govorni signali budu bili predstavljeni na ovaj način dobiće se baza na kojoj će se neuralna mreža obučiti.

VI. DIZAJNIRANJE NEURALNE MREŽE I DOBIJENI REZULTATI

Kao što je ranije rečeno, veoma je bitno koje su vrednosti uzete za vremensku i frekvencijsku rezoluciju, odnosno na koliko pravougaonih segmenata će spektrogram biti podeljen. Kako za *N* i *M* ne treba uzeti ni previše male, a ni previše velike vrednosti, odlučeno je da se izvrši analiza za sledeća tri slučaja: *N*=8 i *M*=6, *N*=8 i *M*=7 i *N*=9 i *M*=5. Neuralna mreža koja je korišćena za problem prepoznavanja izgovorenih reči iz ograničenog rečnika je višeslojna *feedforward* mreža. Ovakav tip mreže izabran je zbog višedimenzionalnog problema koji jednoslojna mreža ne bi mogla adekvatno da reši [3,4]. Ako bi se izabrala prva podela, mreža bi imala 48 ulaza, za drugu podelu bi imala 56 ulaza, a za treću podelu bi imala 45 ulaza. Svaki od ulaza ima po 3200 vrednosti, po 200 za svaku od 16 reči, a svaka vrednost predstavlja odgovarajući pravougani segment spektrograma. Za obučavanje neuralne mreže primenjeno je obučavanje sa nadzorom, jer je broj klasa unapred definisan, odnosno u svakom trenutku mreža zna šta bi trebalo da bude na izlazu (jedna od 16 izgovorenih reči). Kod obučavanja sa nadgledanjem (supervizorskog obučavanja), u svakom diskretnom trenutku vremena, kada je određeni signal na ulazu, odgovarajući željeni izlaz je dat. Tada neuralna mreža tačno zna šta treba da bude na izlazu [9]. Algoritam koji je korišćen za obučavanje mreže jeste algoritam propagacije greške unazad (eng. *backpropagation)* [3]. Na ulaz neuralne mreže dovodi se obrađena baza na način koji je opisan u prethodnim poglavljima. Svaka izgovorena reč predstavljena je vektorom karakterističnih obeležja, na način koji je opisan u prethodnom poglavlju.

Kako bi se došlo do zaključka koja struktura mreže daje najbolje rezultate, mreža je obučavana tako što se pravilo krosvalidaciono okruženje [9]. To znači da se obučavajući skup (raspoloživa baza izgovorenih reči) deli u određenom procentu na obučavajući i testirajući (validacioni) skup (u ovom radu uzeto je 50% za obučavanje, 50% za testiranje). Obavlja se 2 iteracije algoritma, tako što se u prvoj iteraciji uzima polovina podataka za potrebe validacije, a druga polovina se koristi za učenje, a u drugoj iteraciji obrnuto. U obe iteracije algoritma uzima se ista struktura mreže. Pri svakoj iteraciji računa se greška, odnosno koliko je reči pogrešno klasifikovano. Na kraju se računa prosečna uspešnost (greška) na nivou obe iteracije, jer tako izračunate mere uspešnosti daju bolju sliku o performansama algoritma. Mera uspešnosti, odnosno izračunata greška daje informaciju o tome koliko je reči pogrešno klasifikovano, odnosno koliko uspešno je algoritam klasifikovao reči iz baze. Kada se izabere optimalna struktura, neuralna mreža se obučava na celom skupu.

Da bi se izabrala optimalna struktura mreže koja dovodi do najmanje greške, najpre se uzela siromašna struktura mreže (jedan sakriveni sloj sa pet čvorova), broj epoha bio je 100, konstanta obučavanja 0.1, a algoritam propagacije unazad *traingd* (*Gradient descent backpropagation*) i dobila se greška od 90.19%. U svim strukturama koristile su se *purelin* i *tansig* aktivacione funkcije [3]. Zatim se broj skrivenih slojeva i čvorova povećavao (najviše se uzimalo četiri skrivena sloja do 50 čvorova). Menjao se broj epoha obučavanja (od 100 do 1500). Menjala se i vrednost konstante obučavanja (od 0.1 do 10). Sa povećanjem broja epoha, skrivenih slojeva i čvorova kao i konstante obučavanja greška se smanjivala, dok se nije došlo do najmanje greške. Takođe, korišćeni su različiti algoritmi propagacije unazad koji su davali različite rezultate. Za strukturu od 4 skrivena sloja sa po 48,30,48,30 čvorova, 1500 epoha, algoritam propagacije unazad *traincgp* (*Conjugate gradient backpropagation with Polak-Ribiere updates*), primenjena je regularizacija [3] i konstanta obučavanja je 2, greška je 5.63%. Za strukturu od 4 skrivena sloja sa po 45 čvorova, 1500 epoha, algoritam propagacije unazad *trainoss* (*One step secant backpropagation*), primenjena je regularizacija i konstanta obučavanja je 7, greška je 4.5%. Nakon detaljne analize različitih struktura mreže kojih je bilo mnogo, pa se ne mogu sve navesti, došlo se do zaključka da najmanju grešku od 3.41% daje struktura gde je izabrano 4 skrivena sloja sa po 45 čvorova, 1500 epoha, algoritam propagacije unazad *traincgb* (*Conjugate gradient backpropagation with Powell-Beale restarts*) [10], primenjena je regularizacija i konstanta obučavanja je 7. Kada se uporede najmanje greške koje se dobijaju za tri različite podele (4.03%, 5.34% i 3.41%) vidi se da se najbolji rezultat dobija za podelu *N*=9 i *M*=5 i gore izabranu strukturu. Takođe, bolji rezultati (manja greška) dobila se kada se vektor karakterističnih obeležja deli na 45 pravougaonih segmenata koji se računaju kao srednja vrednost logaritamskih vrednosti na odgovarajućem segmentu.

Nakon što je izabrana optimalna struktura, mreža je obučena na celoj bazi. Da bi se videlo kolika je tačnost rada ovog sistema za prepoznavanje govora iz ograničenog rečnika, sistem je najpre testiran na skupu na kojem je obučen. Od dobijenih rezultata pravi se matrica konfuzije koja je data u vidu tabele. Matrica konfuzije daje tačan prikaz koliko je svaka reč tačno klasifikovana, i za slučaj da je došlo do greške, vidi se kako je sistem prepoznao reč, odnosno u koju klasu je pogrešno klasifikovao [9]. Jasno je da je idealna matrica konfuzije dijagonalna, što znači da su sve reči pravilno klasifikovane. Kada je sistem testiran na istom skupu na kojem je i obučen, dobijena je greška od 0% (odnosno dijagonalna matrica konfuzije) što je očekivan i veoma zadovoljavajuć rezultat.

Zatim, sistem je testiran na 3 govornika koji se nalaze u bazi na kojoj je sistem obučen. Govornici su po 5 puta ponovo rekli svaku reč iz ograničenog rečnika, i dobijeni rezultati dati su u Tabeli 1. Na kraju, u Tabeli 2 dati su rezultati kada je sistem testiran na govornicima koji nisu bili u bazi.

Klase u matricama konfuzije obeležene su na sledeći način: Klasa 1-Nula, Klasa 2- Jedan, Klasa 3-Dva, Klasa 4-Tri, Klasa 5-Četiri, Klasa 6-Pet, Klasa 7-Šest, Klasa 8-Sedam, Klasa 9-Osam, Klasa 10-Devet, Klasa 11-Napred, Klasa 12-Nazad, Klasa 13-Kreni, Klasa 14-Stani, Klasa 15-Levo, Klasa 16-Desno. Matrica konfuzije u Tabeli 1 nije dijagonalna. Greška klasifikacije od 2.5% nije velika greška, tako da je dobijeni rezultat sasvim dobar. Sistem je od 240 reči 6 reči pogrešno klasifikovao. Dva puta je prepoznao reč „dva“ kao „nula“, jednom reč „sedam“ kao „jedan“, jednom reč „napred“ kao „levo“, jednom reč „stani“ kao „devet“ i

jednom reč „stani“ kao „nazad“. U svim ostalim slučajevima reči su tačno prepoznate

TABELA I
MATRICA KONFUZIJE KADA JE SISTEM TESTIRAN NA TRI GOVORNIKA IZ BAZE NA KOJOJ JE OBUČAVAN

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	2	0	13	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	15	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	15	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	15	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	15	0	0	0	0	0	0	0	0	0
8	0	1	0	0	0	0	0	14	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	15	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	15	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	14	0	0	0	1	0
12	0	0	0	0	0	0	0	0	0	0	0	15	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	15	0	0	0
14	0	0	0	0	0	0	0	0	0	1	0	1	0	13	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	15	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	15

.

TABELA II
MATRICA KONFUZIJE KADA JE SISTEM TESTIRAN NA GOVORNICIMA KOJI NISU BILI U BAZI NA KOJOJ JE SISTEM OBUČEN

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	14	0	0	0	0	0	1	0	0	0	0	0	0	0	0
3	0	0	23	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	17	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	17	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	12	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	22	0	0	0	3	0	0
11	0	0	0	0	0	0	0	0	0	0	22	0	0	0	0	0
12	0	0	0	1	0	0	0	1	0	0	0	25	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	28	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	13	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	3	0	23	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	27

Matrica konfuzije u Tabeli 2 je najvažnija matrica, jer prikazuje sa kolikom tačnošću sistem prepoznaje reči koje su izgovorili govornici van baze. Od 322 izgovorene reči sistem je pogrešno klasifikovao 9 reči, a sve ostale je tačno prepoznao. Dobijena je greška od 2,8% što je veoma zadovoljavajuć rezultat koji pokazuje da je sistem dobro obučen da prepoznaje reči koje izgovaraju govornici koji se ne nalaze u bazi na kojoj je sistem obučen.

VII. ZAKLJUČAK

Dobijeni rezultati u ovom radu su sasvim zadovoljavajući i pokazuju da se efikasnim algoritmom za predobradu govornih signala i pravilnim izborom arhitekture neuralne mreže može napraviti sistem za prepoznavanje reči iz ograničenog rečnika koji će raditi sa velikom tačnošću, uz sva ograničenja koja se na sistem nameću i umanjuju tačnost klasifikacije.

Prvi problem koji se javlja jesu uslovi pod kojima se reči snimaju. U slučaju da postoji neki izvor buke koji potiče ili od bliže okoline, ili da dolazi spolja, ili od same opreme pomoću koje se vrši snimanje, sistem će ga registrovati kao šum ili će ga protumačiti kao deo reči, a svakako će umanjiti razumljivost same reči.

Drugi problem koji se javlja jeste način na koji govornici izgovaraju reči. S obzirom da je ovo sistem koji je nezavistan od govornika, odnosno nije pravljen da prepoznaje jednog govornika, nego je pravljen da prepoznaje razne govornike, način na koji se izgovaraju reči je veoma bitan. Iz ovih razloga veoma je teško napraviti jedinstven algoritam koji klasifikuje reči nezavisno od toga koji je govornik izgovorio reč. Sa povećanjem broja govornika u obučavajućem skupu svakako da bi se poboljšao kvalitet dobijenih rezultata.

I pored gore navedenih ograničenja i problema na koje se nailazilo prilikom projektovanja jednog ovakvog sistema, dizajniran sistem za prepoznavanje govora iz ograničenog rečnika daje veoma zadovoljavajuće rezultate. Neuralne mreže su zbog svojih osobina i danas veoma atraktivne za naučnike i istraživače kada je u pitanju prepoznavanje govora [11].

LITERATURA

[1] L. R. Rabiner, R. W. Schafer, *Digital processing of speech signals*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey 07632, 1978.

[2] L. Rabiner, B.-H. Juang, B. Yegnarayana,, *Fundamentals of speech recognition*, Pearson, India, 2010.

[3] C.-T. Lin, C. S. G. Lee, *Neural Fuzzy systems: a neuro-fuzzy synergism to intelligent systems*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1996.

[4] J. Tebelskis, “Speech Recognition using Neural Networks”, Ph.D. dissertation, Carnegie Mellon University Pittsburgh, 1995.

[5] Z. Dobrosavljević, Lj. Milić, *Uvod u Digitalnu Obradu Signala*, Elektrotehnički fakultet Univerziteta u Beogradu, 1999.

[6] G.S.Ying, C.D. Mitchell, L.H. Jamieson, “Endpoint Detection of Isolated Utterances Based on A Modified Teager Energy Measurement”, In Proc. IEEE ICASSP-92, Minneapolis, Minnesota, USA, pp.732-pp.735, 1992.

[7] L. Gu, S. A. Zahorian, "A New Robust Algorithm for Isolated Endpoint Detection", *Proc. IEEE ICASSP*, vol. 0, pp. 4161-4164, 2002.

[8] C. Hory, N. Martin and A. Chehikian, “Spectrogram Segmentation by Means of Statistical Features for Non-Stationary Signal Interpretation”, IEEE Transactions on Signal Processing, Vol.50, No.12, pp. 2915-2925, Dec 2002.

[9] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, Boston, 1980.

[10] M.J.D Powell, “Restart procedures for the conjugate gradient method,” *Mathematical Programming*, Vol. 12, 1977, pp. 241–254

[11] A. Graves, A. Mohamed, A., G.E. Hinton, “Speech recognition with deep recurrent neural networks” *IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, Canada 6645-6649, 2013.

ABSTRACT

The goal of this paper is to show how effective and complex speech signals processing with the correct choice of neural network architecture can make a speech recognition system from a limited vocabulary that will work with great accuracy. This paper presents the complete procedure for creating a word recognition system from a limited vocabulary and it has been shown that neural networks combined with well-known methods for modeling and speech signal processing are a very powerful speech recognition tool.

Speech recognition from a limited vocabulary using a neural network

Emilija Kisić, Slobodan Drašković and Vera Petrović

TECHNOLOGY-SUPPORTED THERAPEUTIC APPROACHES FOR STROKE REHABILITATION: FROM DESIGN TO CLINICAL TRANSLATION (INVITED PAPER)

Emilia Ambrosini, NearLab, Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy

ABSTRACT

Stroke is the third most common cause of death and the main cause of acquired adult disability in high-income countries. Hemiparesis, which is motor impairment of one side of the body, affects about 80% of stroke survivors. Restoration of gait and gait-related activities is one of the main goals of stroke rehabilitation, while the recovery of arm functions is crucial for the capability to perform activities of daily life (ADL), which increases independence and quality of life.

Neuroplasticity is the basic mechanism for functional recovery after stroke. Rehabilitative interventions should make effective use of neuroplasticity, proposing high-intensity, repetitive, task-specific, interactive and individualized training. Technology-supported therapeutic approaches are emerging as a solution to support therapists in providing such training for a long duration, allowing the participants to progress in task difficulty, so as to increase their motivation.

In the last thirty years, several robotic devices for both lower and upper-limb rehabilitation have become commercially available. Robot-assisted gait training, supporting (partially or totally) the body weight and the movement of patients, allow intensive and highly repetitive training of complex gait cycles, with a reduced effort for the therapists. A recent Cochrane review showed that robotic gait training in combination with physiotherapy might improve recovery of independent walking in stroke survivors. Leg cycling training may represent a low-cost and safe alternative to robot-assisted gait training. Indeed, cycling shares a similar locomotor pattern with walking and can be performed safely with a sufficient intensity soon after stroke since it does not require standing balance.

Robotic devices have been strongly proposed also to support upper limb stroke rehabilitation and recent systematic reviews demonstrated that stroke patients who receive electromechanical and robot-assisted arm and hand training might improve ADL, arm and hand function and strength.

To increase the therapeutic benefits of robotic rehabilitation, other approaches have been combined with robots, such as Functional Electrical Stimulation (FES) and Virtual Reality (VR); these combined approaches can make robotic based-interventions more functional and engaging. FES has been strongly used to enhance functional recovery of the paretic arm or leg. The combination of FES with robotic devices helps overcome one of the main limits of FES, e.g., the early onset of muscle fatigue. VR is instead used to provide an interactive and individualized training modality, which can provide sensorimotor training in enriched environments, so as to maximize patient's engagement.

This talk will provide an overview of the main technology-supported therapeutic approaches for upper and lower limb stroke rehabilitation. The main training elements, which technology should guarantee in order to maximize cortical plasticity and consequently motor re-learning, will be summarized. Interventions for both upper and lower limb recovery will be mentioned. Special focus will be given to active-assistive control modalities of robotic devices which implement "assisted-as-needed" rehabilitation therapy aimed at maximizing patient's involvement and self-esteem. Finally, the importance of the design of randomized controlled trials (RCT) to evaluate the efficacy of technology-supported therapeutic approaches with respect to usual care will be highlighted. The examples of two RCTs recently conducted by our group will be also provided. Specifically, one RCT recruited a population of 68 subacute stroke survivors and evaluated the effects of a biofeedback training involving FES-augmented cycling training and balance exercises on motor recovery and walking ability. The second RCT was a multi-center clinical trial conducted within the European project RETRAINER and evaluated the efficacy of training with the support of a hybrid robotic system, consisting of an anti-gravity arm exoskeleton combined with EMG-triggered FES, on a sample of 68 post-acute stroke patients.

Rules for Estimation of Gait Phases from Data Acquired by the Gait Teacher Insoles

Vladimir Džepina, Aleksandar Gogić, Dejan B. Popović, *Member, IEEE*

***Abstract*—The task of this study was to develop a user-friendly method for estimation of gait events from the sensor's data integrated into the wearable system called Gait Teacher. Ten pressure sensors and two six-axis inertial measurement sensors built into a pair of insoles wirelessly send signals to the host computer sampled at 100 Hz. We collected and processed data by a custom-designed software developed in LabVIEW. We heuristically segmented the gait cycle to the phases typically used for the gait analysis (*Toes off, Terminal swing, Heel contact, Foot flat and Heel off,*) and then applied the automatic segmentation of data. The performance of the method was tested by comparing the heuristic and automated segmentation. The results suggest that the averaged accuracy of the segmentation reached almost 100% for all four healthy subjects who participated in the study. These results indicate that Gait Teacher could be used for evaluation of the gait performance in sports, recreation, and rehabilitation.**

***Index Terms*—Gait; Insole; Inertial measurement unit (IMU); Ground Reaction Force (GRF); Gait segmentation.**

I. Introduction

The description of the gait is of interest to clinicians for assessing the patient's course of recovery during the rehabilitation [1]. Therefore, it is essential to provide clinicians with a system that would allow gait data acquisition and processing. Two different types of gait data recording systems are: 1) camera-based laboratory system with force plates [2, 3] and 2) wearable sensors with instrumented insoles [4, 5]. The complexity of use and the price make wearable systems much more attractive for clinical applications.

Wearable sensors need to acquire the ground reaction forces and kinematics of bodily segments. The kinematics can be assessed by inertial measurement units (IMU) which integrate accelerometers, gyroscopes, and magnetometers. Today, these sensors are miniature, have low power consumption and generate a reproducible digital signal that characterizes the motion of the bodily segments on which they are mounted. These sensors need a power supply and need to be linked to a wireless communication chip, to send data to a computer. There are systems at the market which satisfy most of the requirements [e.g., 6, 7, 8, 9]. The ground reaction force is delicate to measure information, and no wearable system can generate comparable precision to the one from the force platform. The sensory systems available on the market use piezoelectric or piezoresistive material on distributed over shoe insoles [e.g., 10, 11, 12, 13]. One of the systems that we single out because it is by the intention of an application similar to the Gait Teacher is the Gait Tutor [14]. The Gait Tutor uses separate IMU positioned at the body segments for biofeedback and monitoring of the gait performance.

The new instrument (Gait Teacher) used in this study comprises in each insole one IMU unit (3D gyroscope and 3D accelerometer) and five ground reaction pressure transducers and all 2 x 11 raw signals are synchronously sent via a specific protocol to the host computer [15]. The system is the successor of the Walky system [16] that has been used in combination with IMU in several clinical studies. The construction of the Walky system is very similar to the Gait Tutor. The target of the investigation was to establish a method for sequencing the gait cycle into phases determined by specific gait events. To clarify the objective, we show a sketch of the gait cycle (Fig. 1). The movements of the left and right legs are time-shifted for 50% of the gait cycle.

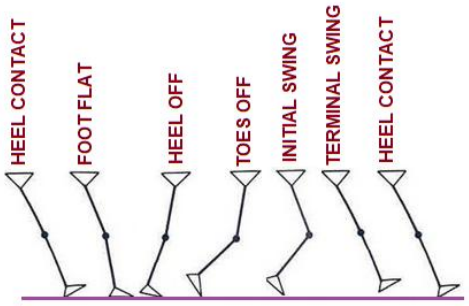


Fig. 1: The events segmenting the gait cycle to phases

II. The Methods

A. Instrumentation

We have used a custom designed system Gait Teacher [15]. The system comprises two insoles instrumented with five industrial pressure sensors and one inertial measurement unit (IMU) per insole. Position of pressure sensors and the orientation of the IMU axis are shown in Fig. 2. Z-axis is positioned towards the ground.

Vladimir Džepina is a Ph.D. student at the School of Electrical Engineering, University of Belgrade, Bulevar kralja Aleksandra 73, 11000 Belgrade, Serbia, (e-mail: dzepina.vladimir@gmail.com).

Aleksandar Gogić is a Ph.D. student at the School of Electrical Engineering, University of Belgrade, Bulevar kralja Aleksandra 73, 11000 Belgrade, Serbia, (e-mail: gogicaleksandar@yahoo.com).

Dejan B. Popović is with the Serbian Academy of Sciences and Arts, Knez Mihailova 35, 11000 Belgrade, Serbia, and Aalborg University, Aalborg, Denmark (e-mail: dbp@etf.rs).

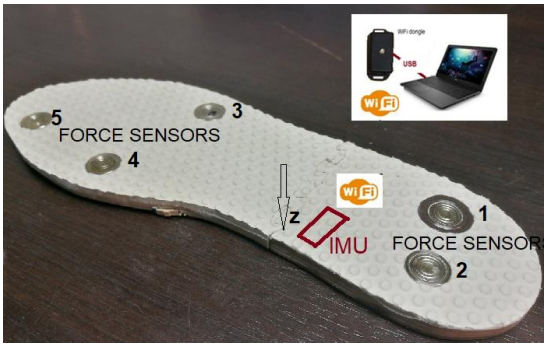


Fig. 2: Gait Teacher insole comprising five pressure sensors, a six-axis digital IMU, Bluetooth communication chip and the battery [15]

Signals from insoles are sent wirelessly to the host PC and read with custom acquisition software. This application was designed in LabVIEW [17] software. It has options to show, calibrate, and log signals acquired from insoles. There is no preprocessing of the raw signals.

B. Subjects and procedure

We recorded data for four healthy subjects with no known history of neuromuscular disorders. All subjects signed Informed Consent approved by the Local Ethics Committee. All subjects were females, with a shoe size of 39 (same size of insoles for all subjects).

Insoles were placed in both shoes of subjects. Subjects were asked to lift feet from the ground one by one, to set the zero for pressure sensors (calibration).

Subjects were then asked to walk along the straight 10m long path with the self-selected comfortable speed. The test was repeated in total four times for each subject.

C. Data processing

Signals were processed offline in LabVIEW 2015 software. We reduced the number of signals for this analysis. The heuristics were based on the evidence gained from the observation that some sensors provide redundant information.

We decided to segment the gait cycle into five consecutive phases. The phases were determined with the gait events as defined below:

- *Toes off* – event when a subject lifts the foot from the ground and starts the swing
- *Terminal swing* – event when the foot is approaching the ground
- *Heel contact* – event when the subject contacts the ground with the heel at the beginning of the stance
- *Foot flat* – a phase where both the heel and toes are in contact with the ground
- *Heel off* – event when the subject lifts the heel, and the support remains only on the tarsus and toes.

After a thorough analysis of recorded data, we have heuristically selected rules for three models that we used for segmentation of the gait cycle events. The signals used in the models are in Table 1.

TABLE 1
LIST OF SIGNALS WITH POSITION AND ORIENTATION OF SENSORS

SIGNAL	DESCRIPTION
P1 HEEL	PRESSURE SENSOR ON THE HEEL (1 IN FIG. 2)
P2 HEEL	PRESSURE SENSOR ON THE HEEL (2 IN FIG. 2)
P1 META	PRESSURE SENSOR ON METATARSAL (3 IN FIG. 2)
P2 META	PRESSURE SENSOR ON METATARSAL (4 IN FIG. 2)
P TOE	PRESSURE SENSOR ON THE TOE (5 IN FIG. 2)
AX	ACCELERATION OF IMU ON THE X-AXIS
AY	ACCELERATION OF IMU ON THE Y-AXIS
AZ	ACCELERATION OF IMU ON THE Z-AXIS
WX	ANGULAR VELOCITY OF IMU ON THE X-AXIS (ROLL)
WY	ANGULAR VELOCITY OF IMU ON THE Y-AXIS (PITCH)
WZ	ANGULAR VELOCITY OF IMU ON THE Z-AXIS (YAW)

The terms used in the tables to follow are the following: dwy/dt is the first derivative of wy, Treshold1 and Treshold2 are the minimal offsets of the pressure signals used to eliminate error caused by noise, Max and Min are maximum and minimum of values inside the brackets, respectively. OR and AND are Boolean operators.

TABLE 2
RULES FOR MODEL 1

GAIT PHASE	RULES
TOES OFF	DWY/DT<0 AND AX>0 AND AZ≤1
TERMINAL SWING	DWY/DT>0 AND AX<0 AND AZ>1
HEEL CONTACT	P1 HEEL>TRESHOLD1 OR P2 HEEL>TRESHOLD1
FOOT FLAT	MAX(P1 META,P2 META)>MIN(P1 HEEL,P2 HEEL)
HEEL OFF	MAX(P1 HEEL, P2 HEEL)<THRESHOLD1 AND P TOE>THRESOLD2

TABLE
RULES FOR MODEL 2

GAIT PHASE	RULES
TOES OFF	DWY/DT<0
TERMINAL SWING	DWY/DT>0
HEEL CONTACT	P1 HEEL>TRESHOLD1 OR P2 HEEL>TRESHOLD1
FOOT FLAT	MAX(P1 META,P2 META)>MIN(P1 HEEL,P2 HEEL)
HEEL OFF	MAX(P1 HEEL, P2 HEEL)<THRESHOLD1 AND P TOE>THRESOLD2

TABLE 4
RULES FOR MODEL 3

GAIT PHASE	RULES
TOES OFF	DWY/DT<0 AND MAX(P1 HEEL, P2 HEEL, P1 META, P2 META,P TOE)<TRESHOLD1
TERMINAL SWING	DWY/DT [K]>0 AND DWY/DT [K-1]≤0
HEEL CONTACT	P1 HEEL>TRESHOLD1 OR P2 HEEL>TRESHOLD1
FOOT FLAT	MAX(P1 META,P2 META)>MIN(P1 HEEL,P2 HEEL)
HEEL OFF	MAX(P1HEEL, P2HEEL)<THRESHOLD1 AND PTOE>THRESHOLD2

Rules for models 1, 2 and 3 are shown in Tables 2, 3 and 4, respectively. In the first two models, phases can switch only in sequential order, the same as the order in the table. In the third model, skip of heel-off is made possible.

III. Results

All signals recorded from one foot are in Fig. 3. The stance phase is denoted by a yellow background, while the swing phase by white for more natural visual perception. Fig. 4 shows one gait cycle extracted from the signal in Fig 3.

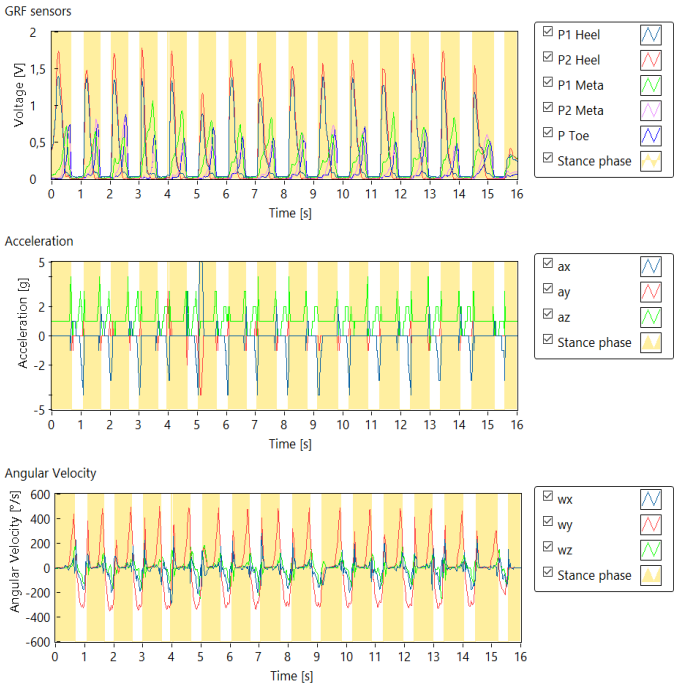


Fig. 3: Signals of one foot during one 10m walk

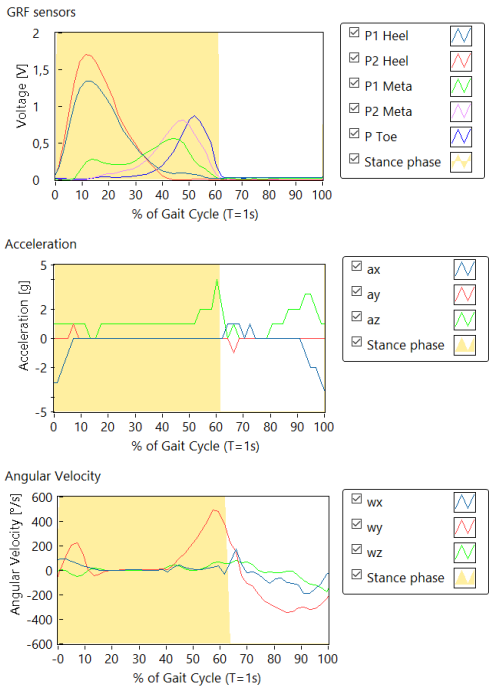


Fig. 4: A typical gait cycle extracted from Fig. 3

The results from the four clinical tests are presented in tables 5, 6 and 7.

TABLE 5
THIS NUMBER OF GAIT PHASES IS ESTIMATED BY HEURISTICS AND FROM MODEL 1 AUTOMATICALLY.

	Toes off	Terminal swing	Heel contact	Foot flat	Heel off
# of phases Heuristics	300	300	300	300	300
# of phases Automatic	289	289	289	289	289

TABLE 6
THIS NUMBER OF GAIT PHASES IS ESTIMATED BY HEURISTICS AND FROM MODEL 2 AUTOMATICALLY.

	Toes off	Terminal swing	Heel contact	Foot flat	Heel off
# of phases Heuristics	300	300	300	300	300
# of phases Automatic	286	300	300	300	300

TABLE 7
THIS NUMBER OF GAIT PHASES IS ESTIMATED BY HEURISTICS AND FROM MODEL 1 AUTOMATICALLY.

	Toes off	Terminal swing	Heel contact	Foot flat	Heel off
# of phases Heuristics	300	300	300	300	300
# of phases Automatic	300	300	300	300	300

Column chart (Fig. 5) is showing the percentage of successful individual gait phase recognitions. Every color of a bar represents one model that was used for classification.

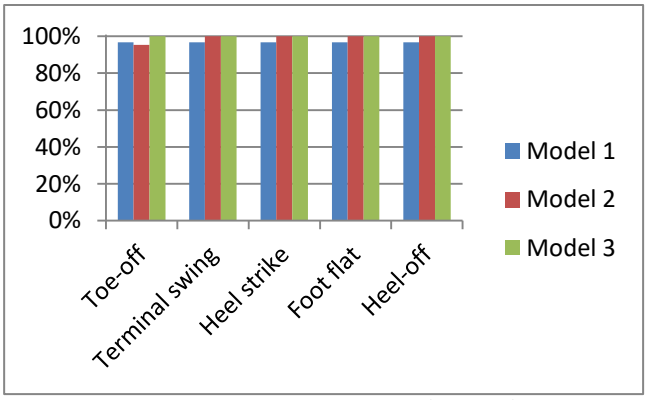


Fig. 5: Bar diagram showing the percent of successful automatic gait event/phase recognitions for the three models tested

IV. Discussion

Model 1 gives acceptable results (Fig. 5). Almost every phase is recognized. Problem is in the duration of phase *Toes off*. Also, there is a problem of not "recognizing" the steps within some of the tests. The problem occurred most likely due to sometimes unreliable acceleration data (noise) that has

been acquired from IMU. This measurement was not reproducible at the level necessary. Therefore, rules that implement accelerometer data can “miss” the gait event that would be seen by heuristics.

Classification model 2 gave better results compared to the results from model 1 for all phases except the phase *Toes off* (Fig. 5). There are two reasons for the decrease in accuracy. One possible problem for missing the *Toes off* event in the initial step comes from the actual modality of movement of a subject (lifting the foot in the air and moving it forward compared to leaning forward and pushing the contralateral leg). The error might also occur since some subjects tend to slightly swing their foot in front at the beginning of *Toes off*. Therefore, the system could falsely register *Terminal swing* phase. Heuristics do not miss all of this, but it is lost by automatic segmentation. Model 3 gives perfect segmentation when compared to heuristics for all phases (Fig. 5).

V. Conclusion

Gait Teacher provides accurate real-time ground reaction force information from only one per leg wireless gadget. The donning and doffing of the system is trivial since the self-contained insole with the small LiI battery allowing eight hours of operation between recharging is identical to many insoles available and uses only wireless communication. The data can be presented to the user online and off-line, or sent to a remote evaluator allowing the feedback based gait training. An example of a possible application is the detection of freezing of gait (FOG) in patients with Parkinsonian disease [18].

The analysis of the system performance suggested that it is appropriate to estimate the length and variability of each gait phase (not shown in this paper). Analyzing the signals used for segmentation defined by model 3, we found approaches for reducing the number of signals that need to be acquired. This conclusion comes from the analysis of level walking. Future studies need to increase the population size and consider perturbations and change in the gait modality (level walking, slope, steps, side walking, backward stepping, etc.).

For future improvement of hardware and classification methods, we are planning to design a smartphone application (Android, iOS) to provide online feedback in sports, recreation, and rehabilitation.

Acknowledgment

Authors would like to thank all volunteers for participation in this study. The work was partly supported by the Serbian Academy of Sciences and Arts, project F-137

Authors would like to thank Vladimir Kojić and the company Rehabshop doo [15] for developing and providing the Gait Teacher system for the tests.

References

[1] Hashimoto K, Higuchi K, Nakayama Y, Abo M. Ability for basic movement as an early predictor of functioning related to activities of daily living in stroke patients. Neurorehabil. Neural Repair 2007, 21, 353–357.

[2] Davis RB, Ounpuu S, Tyburski D, Gage JR. A gait analysis data collection and reduction technique. Human movement science. 10(5), 575-587, 1991.

[3] Aminian K, Trevisan C, Najafi B, Dejnabadi H, Frigo C, Pavan E, Telonio A, Cerati F, Marinoni EC, Robert P, Leyvraz PF. Evaluation of a mobile system for gait analysis in hip osteoarthritis and after total hip replacement. Gait & posture. 20, no. 1: 102-107, 2004.

[4] Muro-de-la-Herran A, García-Zapirain B, Méndez-Zorrilla A. Gait analysis methods: An overview of wearable and non-wearable systems, highlighting clinical applications. Sensors 2014, 14, 3362–3394.

[5] Tao W, Liu T, Zheng R, Feng H. Gait analysis using wearable sensors. Sensors 2012, 12, 2255–2283.

[6] Giggins OM, Persson UM, Caulfield B. Biofeedback in rehabilitation. J Neuroeng Rehabil 2013, 10:60.

[7] Fong DT, Chan YY. The use of wearable inertial motion sensors in human lower limb biomechanics studies: a systematic review. Sensors (Basel) 2010, 10(12):11556–11565.

[8] Cuesta-Vargas AI, Galán-Mercant A, Williams JM. The use of inertial sensors system for human motion analysis. Phys Ther Rev 2010, 15(6):462–473.

[9] https://www.xsens.com (accessed in February 2019)

[10] http://novel.de/novelcontent/pedar (accessed in February 2019)

[11] http://www.feetme.fr/en/index.php (accessed in February 2019)

[12] https://www.tekscan.com/applications/force-sensitive-insole (accessed in February 2019)

[13] https://www.tekscan.com/product-group/medical/in-shoe (accessed in February 2019)

[14] http://mhealthtechnologies.it/products-wearable-sensors-smartphone-apps/rehabilitation-gait-tutor/ (accessed in May, 2019)

[15] https://rehabshop.rs/PROJECTS.php (accessed in May 2019)

[16] Topalović I, Popović DB. Estimation of gait parameters based on data from inertial measurement units. Proc of 4th Intern Conf on Electrical, Electronics and Computing Engineering, IcETRAN 2017, Kladovo, Serbia, June 05-08, ISBN 978-86-7466-693-7, pp. BTI2.4.1-5.

[17] http://www.ni.com/en-rs.html (accessed in February 2019)

[18] Popovic, M.B., Djuric-Jovicic, M., Radovanovic, S., Petrovic, I. and Kostic, V., 2010. A simple method to assess freezing of gait in Parkinson's disease patients. *Brazilian Journal of Medical and Biological Research*, *43*(9), pp.883-889.

Multi-sensor acquisition system for noninvasive detection of heart failure

Aleksandar Lazović, Lana Popović-Maneski and Ljupčo Hadžievski

***Abstract*—To research the possibility of noninvasive detection of heart failure we developed an acquisition system with multiple sensors. The system synchronously measures cardiovascular pulsations, heart sounds and ECG using different types of sensors positioned only on the patient's body. The system has a modular structure with five modules: 1. Module for controlling the light source (MWLS) 2. Module for data acquisition from fiber optical sensors (FBGA) with the compact optical spectral analyzer 3. Module for the acquisition of hearth sounds (PCG) with four ports for microphones; 4. Module for the acquisition of standard ECG signals; 5. Module for data acquisition from three accelerometers and three photoplethysmography sensors (ACC/PPG).**

***Keywords*— multi-sensor device, heart failure.**

I. Introduction

Heart failure is an abnormality of cardiac structure and function leading to the incapability of the heart to deliver oxygen to maintain natural metabolic balance in the organism. The most common method for the detection of heart failure is echocardiogram in which some functional and structural changes of the heart related to heart failure can be detected [1].

One of the main reasons for the rare detection of heart failure in the early phase of the disease is non-existence of symptoms. When symptoms (weakness, fatigue, breathlessness) become noticeable, the disease has already progressed to the stage when treatment is difficult, and a chance of mortality is very high compared to the patient with early detected heart failure. Also, due to the expensive equipment and skilled personnel, the echocardiogram is typically not part of primary medical care. So, it would be very useful to find an alternative method, which would be cheaper and easier for use in primary care and available to more patients. In order to achieve that, we took the first step of the research and developed a measurement system that will eventually be transformed in such a device that preventive primary care requires. Noninvasive system for early detection of heart failure, which can be cheap and easy for use in primary care, would provide a high impact on the health care system. [2,3].

To investigate the possibility of noninvasive detection of heart failure we developed a new acquisition system for simultaneous measurement of several biomedical signals. By using this system, we will simultaneously measure different types of cardiovascular pulsations, heart sounds, and ECG signals. To measure cardiovascular pulsation, we use fiber optical sensors, accelerometers, compact semiconductor, lasers, and for measurement of heart sounds we use microphones.

In this work, we present the multi-sensor system which will be used for data acquisition of heart-related biomedical signals and data analyses of their properties for early detection of heart failure.

Aleksandar Lazović – PhD student at University of Belgrade, Studentski trg 1, 11000 Belgrade, Serbia (email: aca.lazovic@gmail.com)

Lana Popović-Maneski - The Institute of Technical Sciences, Knez Mihajlova 35, 11000 Belgrade, Serbia (e-mail: lanapm13@gmail.com).

Ljupčo Hadžievski - Vinča Institute of Nuclear Sciences, University of Belgrade, 11001 Belgrade, Serbia (ljupcoh@vinca.rs)

II. System description

Multi-sensor acquisition system, that we developed is composed of multi-sensor acquisition device and PC acquisition software. The multi-sensor acquisition device has a modular structure (Figure 1) with five different modules:

1. A module that controls a light source (MWLS) with bandwidth 1510 nm-1590 nm
2. Module for data acquisition from fiber optical sensors (FBGA) with a spectrum analyzer in the same bandwidth as the light source (1510 nm-1590 nm)
3. Module for the acquisition of heart sounds (PCG module) from up to four different microphones
4. Module for the acquisition of ECG signals with 12 channels (ECG module)
5. Module for the acquisition from three accelerometer sensors and three photoplethysmography (PPG) sensors (ACC/PPG module).

A. MWLS and FGBA module

The primary purpose of the MWLS and FBGA module is a real-time measurement of the transmitted spectrum passed through some fiber optical sensor (FBG or LPG sensor). MWLS module emits light through a fiber optical sensor in a wide bandwidth range. As optical sensors are fabricated with a grating, transmitted light spectrum on the opposite side of the sensor has resonant deeps. If the sensor is stretched or curved with a small radius, resonant deep is moving towards lower or higher frequencies linearly. This characteristic of the sensor can be used for the measurement of respiratory and cardiovascular pulsations [4,5].

Both MWLS and FBGA modules are an official product of company BaySpec. FBGA module is optical spectrum analyzer which can acquire spectrum at a maximum sampling rate of 5 KHz. Also, each spectrum sample is a group of 512

points in the spectrum range 1510 nm-1590 nm, which means that spectrum resolution is 0.15625 nm. By analyzing the spectrum of measured signals, we concluded that the sampling frequency of 500Hz is good enough. Because a large amount of data needs to be sent in real time, the system uses USB full speed communication for data transmission to PC. USB is also used for control of the MWLS module. FBGA module has a triggering ability which is crucial for synchronization between all modules.

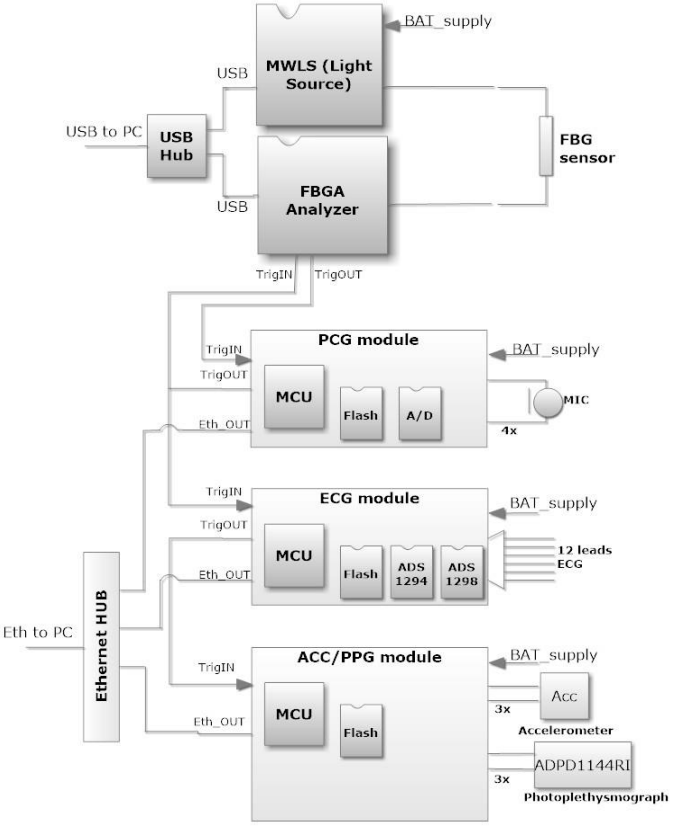


Figure 1. Schematic description of multi sensor device

B. PCG module

PCG module is a module for the acquisition of data from up to 4 different microphone sensors. The module is based on STM32F407 microcontroller which acquires data from microphones. The acquisition sample rate is 1 kHz with 16-bit A/D conversion. Sensors are MEMS microphones placed in an analog stethoscope bell. All sensors are intended to be used for the measurement of heart sounds (S1, S2, S3, S4) [6]. As significant frequency components of heart sounds are below 200 Hz, sample frequency was set to 1kHz. Also, MEMS microphones are selected to have the high sensibility and wide frequency bandwidth. PCG module is a master module for control of synchronization triggering pulses. This module transmits data to PC over Ethernet.

C. ECG module

ECG module is based on Texas Instruments' chip ADS1298. This chip has a lot of advanced functionalities necessary for easy acquisition of ECG signals. It consists of 8 channels with 8 differential amplifiers and 24bit A/D converter with a maximum sampling frequency of 32 KHz. For the regular acquisition of ECG signals we used 500 Hz sampling frequency. Some additional features of this chip are embedded antialiasing digital filter, RLD driver and lead-off detection. ECG chip is controlled by a microcontroller (STM32F407) over SPI. Data is acquired in real time and transmitted to PC over Ethernet. Electrodes are DC coupled to the inputs of ECG chip.

D. ACC/PPG module

ACC/PPG module is one common module for data acquisition from accelerometer sensors [7] and PPG sensors [8]. It consists of one microcontroller STM32F407 which acquires data from sensors over I2C. The accelerometer sampling rate is 500Hz with acceleration range +/-1g. It is possible to use up to 3 sensors from each group of sensors. PPG sensor ADPD174GGI is Analog Devices' chip with integrated all necessary components for reflective photoplethysmography measurement. It has incorporated photodiode, two green, and one infrared LED. Ambient light and offset rejection, photodiode amplifier gain control and LED intensity control are features of the sensor that simplify its usage and processing is not needed when data is acquired. The sampling frequency of the photodiode A/D converter is 100 Hz.

Data is synchronously acquired from both groups of sensors and transmitted to PC over Ethernet.

E. Power supply

The power supply of the system is a lead-acid battery of capacity 7 Ah. The battery can be charged using an external AC/DC power supply. To make the system comply with safety standards, charging is enabled only when the device is turned off. Working capacity of the device is approximately 10 hours.

F. Synchronization and communication with PC

MWLS and FBGA modules communicate with PC over USB. Other three modules communicate with PC over Ethernet. As standard PC doesn't have three Ethernet ports, an Ethernet hub is integrated inside the device. When synchronization measurement is initiated, the PCG master module transmits triggering pulses to synchronize acquisition. Data is transferred to PC independently from each module in real time in the form of packets.

G. Acquisition software

Acquisition software is realized in Visual Studio C#. When the device is connected to PC synchronous acquisition can be started and received data from each module is recorded in an independent file. When the measurement is finished, acquisition software has to merge synchronized data from separate files into one file. In real time, up to 6 different graphs from different types of sensors can be shown on the main screen giving examiner possibility to see if some electrode or sensor was not placed correctly.

III. RESULTS

On the following three pictures (Figure 2, Figure 3 and Figure 4) are shown realized acquisition device, sensors and electrodes used in measurement and example of sensor positioning on the patient.

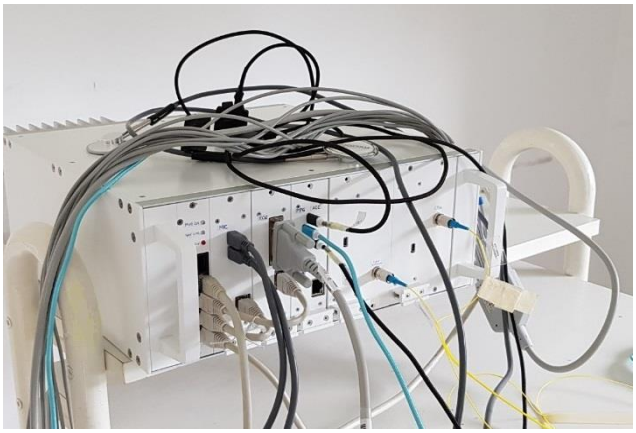


Figure 2:Realized multi-sensor acquisition device

Figure 3:Different sensors used in measurement - fiber optical sensor (up left), ECG electrode cable (up right), photoplethysmography sensor (bottom left), accelerometers (bottom center), digital stethoscope (bottom right)

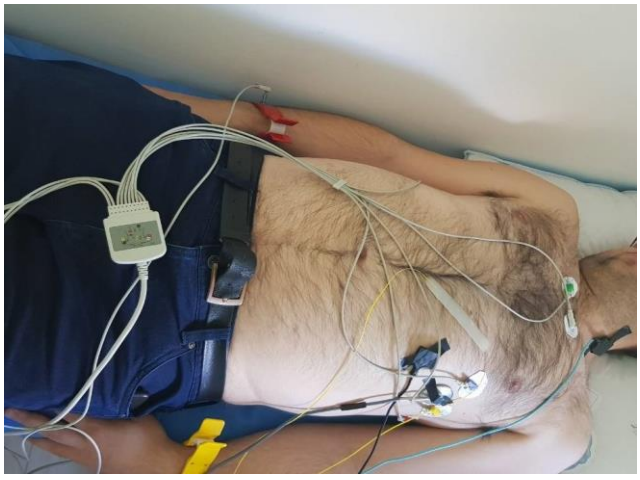


Figure 4: Electrodes and sensors placed on the patient

On the following two pictures are shown signals that were measured independently from each sensor. Figure 5 is shown signal from the microphone placed on the chest of the healthy person. It can be noticed that heart sounds S1 and S2 are visually detectable which means that the position of the microphone in stethoscope bell was adequate.

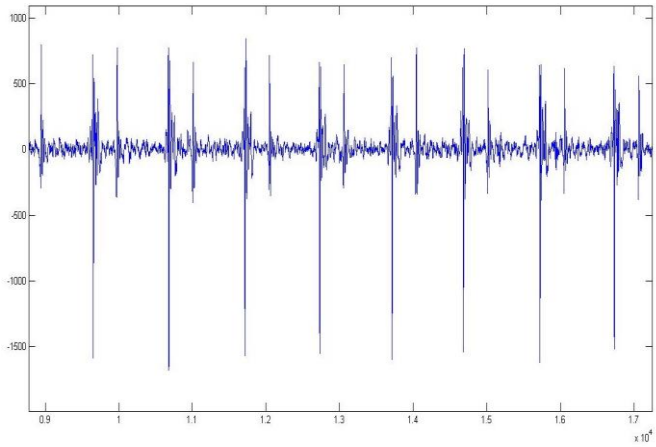


Figure 5. Signal measured from PCG sensor (digital stethoscope)

Figure 6 shows the signal measured from an accelerometer positioned on the chest of the patient below the heart. Also, Figure 6 are visible clear heartbeats with a periodic signal pattern which indicate that the measurement is successful.

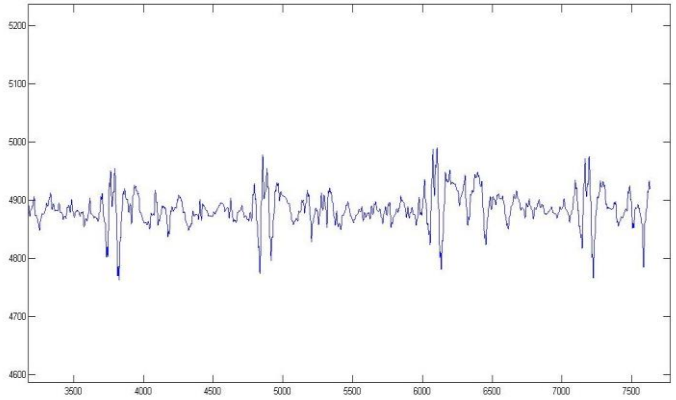


Figure 6. Signals measured from an accelerometer positioned on the chest below the heart

Figure 7 shows multiple signals from multiple different sensors measured simultaneously. This kind of measurement shows that all of the signals are well synchronized.

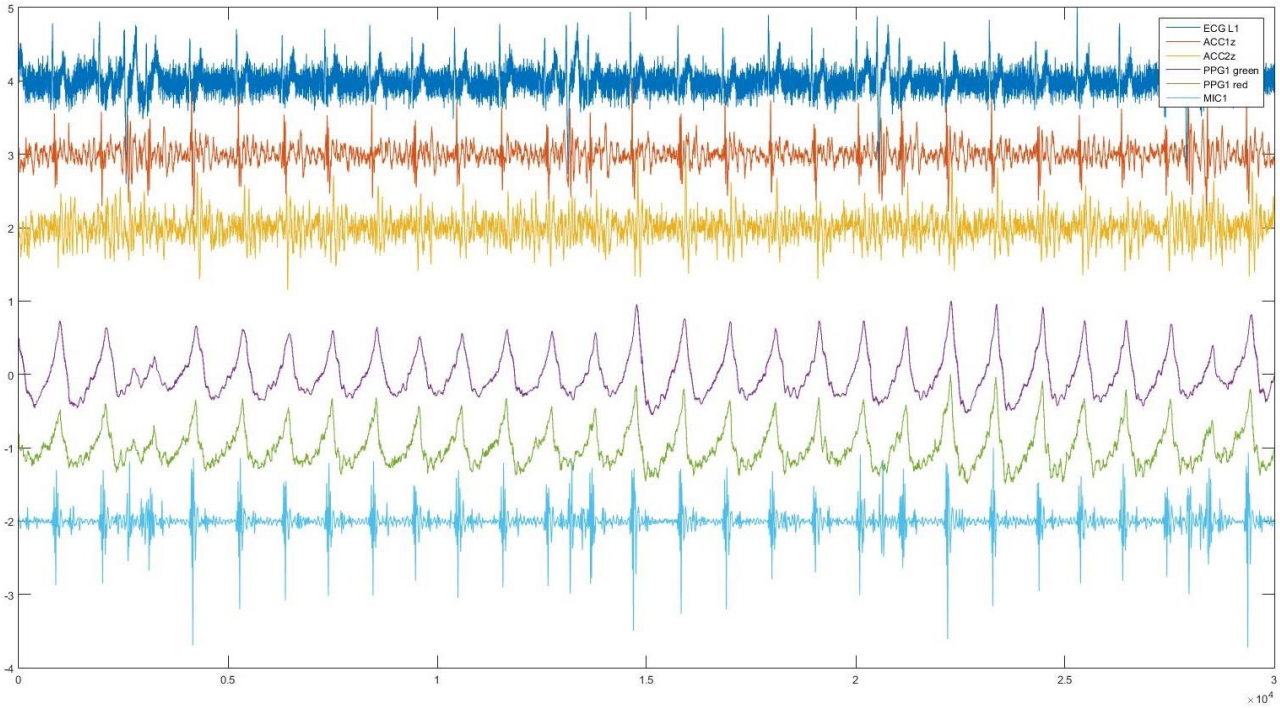


Figure 7. Simultaneous measurements of different types of signals. Shown signals are (from up to down): ECG signal channel I, two accelerometer signals from lower and upper chests, two signals from PPG sensors positioned on the carotid artery and signals measured from PCG sensors.

IV. Conclusion

Presented acquisition device will be used to record multiple heart-related mechanical, electrical and sound signals for further analyses of their properties and space and time correlations to investigate the possibility for early detection the heart failure. The first set of measurements was used to optimize the configuration and positioning of the sensors on the human body to get good quality raw signals with increased signal to noise ratio. The results show that the device is delivering well-synchronized signals from all sensors with good signal to noise ratios appropriate for further analyses. The next step is an adaptation of the multisensory device for use in a clinical study which will include patients with detected heart failure in early and progressed phases. For that purpose, we are developing user-friendly software and preprocessing algorithms to enable easy handling of the device by the medical personnel in clinical environment. Due to its modular structure and possibility to use multiple sensors, the device can be easily reconfigured for some other biomedical researches.

Acknowledgement

We acknowledge support from the Ministry of Education, Science and Technological Development of the Republic of Serbia, Grants III45010 and III44008.

References

[1] McMurray, J. J., Adamopoulos, S., Anker, S. D., Auricchio, A., Böhm, M., ... & Gomez-Sanchez, M. A. (2012). ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure 2012: The Task Force for the Diagnosis and Treatment of Acute and Chronic Heart Failure 2012 of the European Society of Cardiology. Developed in collaboration with the Heart Failure Association (HFA) of the ESC. *European journal of heart failure*, *14*(8), 803-869

[2] S. Stewart, K. MacIntyre, D. J. Hole, S. Capewell, and J. J. V McMurray, "More 'malignant' than cancer? Five-year survival following a first admission for heart failure," Eur. J. Heart Fail., vol. 3, no. 3, pp. 315–322, 2001.

[3] C. Berry, D. R. Murdoch, and J. J. McMurray, "Economics of chronic heart failure.," *Eur. J. Hear. Fail. J. Work. Gr. Hear. Fail. Eur. Soc. Cardiol.*, vol. 3, no. 3, pp. 283–291, 2001.

[4] M. D. Petrović, J. Petrovic, A. Daničić, M. Vukčević, B. Bojović, L. Hadžievski, T. Allsop, G. Lloyd, and D. J. Webb, "Non-invasive respiratory monitoring using long-period fiber grating sensors," *Biomed. Opt. Express*, vol. 5, no. 4, p. 1136, 2014.

[5] T. Allsop, G. Lloyd, R. S. Bhamber, L. Hadzievski, M. Halliday, D. J. Webb, and I. Bennion, "Cardiac-induced localized thoracic motion detected by a fiber optic sensing scheme," *J. Biomed. Opt.*, vol. 19, no. 11, p. 117006, 2014.

[6] P. J. Arnott, G. W. Pfeiffer, and M. E. Tavel, "Spectral analysis of heart sounds: Relationships between some physical characteristics and frequency spectra of first and second heart sounds in normals and hypertensives," J. Biomed. Eng., vol. 6, no. 2, pp. 121–128, Apr. 1984.

[7] I. Starr and H. A. Schroeder, "BALLISTOCARDIOGRAM. II. NORMAL STANDARDS, ABNORMALITIES COMMONLY FOUND IN DISEASES OF THE HEART AND CIRCULATION, AND THEIR SIGNIFICANCE.," J. Clin. Invest., vol. 19, no. 3, pp. 437–50, May 1940.

[8] Allen, John. "Photoplethysmography and its application in clinical physiological measurement." Physiological measurement 28.3 (2007): R1.

Drowsiness detection using machine learning approaches based on cardiopulmonary signals

Anita Lupšić, Predrag Tadić, Veljko Mihailović and Milica Janković, *Member, IEEE*

***Abstract*—Drowsiness detection systems can be behavioral-based (e.g. tracking face/eyes expressions, vehicle-based measuring) or physiological-based (monitoring features of physiological signals such as electroencephalography, electrooculography, respiration rate, electrocardiography). The aim of this paper is the development of an algorithm for the detection of drowsiness based on the variability of heart rate and breathing rate features. A group of ten healthy adults participated in the experiment in which they were exposed to multimedia content. The measurement of electrocardiography (ECG) and the respiration curve was performed using Smartex Wearable Wellness System (Pisa, Italy). The video of the subject's face was recorded as the reference signal for drowsiness. All data were acquired while subjects were awake, sleepy and in the early stage of sleep. Time, frequency and fractal feature extraction of heartrate variability and breathing rate was performed. Machine learning approaches (Support Vector Machines (SVM), k-nearest neighbors (kNN) and ensemble methods) were implemented for the multinomial classification with three output classes: "awake", "drowsy" and "fallen asleep". Unequal risks of different error types were considered because the consequences are significantly higher if the "drowsy" and "fallen asleep" classes are not properly detected, in comparison to the "awake" class. The upper accuracy of 77% and 80% was obtained in the validation and test process, respectively.**

***Index Terms*— drowsiness, ECG, respiration, support vector machine, k-nearest neighbors, ensemble method.**

I. Introduction

DROWSINESS is defined as a transitional state between wakefulness and sleep associated with a desire or inclination to sleep (sleepiness) in which the "sleep onset process" has already begun [1],[2]. Assessment tools for sleepiness include behavior measures (e.g. head movements, facial expressions, eye closing etc.), subjective rating scales (e.g. Stanford Sleepiness Scale - SSS, Karolinska Sleepiness Scale - KSS etc.) and electrophysiological measurements (Multiple Sleep Latency Test - MSLT, polysomnography, cerebral evoked potentials etc.) [3]. Measurement of sleepiness is significant from the point of view of clinical evaluation [4], [5] as well as in conditions when healthy population intends to state awake (e.g. while driving or at work).

According to the data collected by the American Automobile Association, drowsiness is one of the main causes of traffic accidents (about 20% of all) [6]. From an economical point of view, around 30 billion dollars is spent on traffic accidents caused by drivers' drowsiness. Those facts caused expansion in the development of commercial systems for drowsiness detection based on arteficial intelligance approaches performed on features extracted from vehicle behavior signals (e.g. deviation from the central lane position, steering wheel disturbance etc.), camera recordings (e.g. position of the head, yawning, degree of eye closure etc.) or physiological signals (e.g. electrocardiogram (ECG), electrooculogram (EOG), electroencephalogram (EEG), respiration curve) [7], [8], [9], [10]. The most reliable systems for sleepiness assessment are based on physiological signals, but they usually require uncomfortable contact sensors for measuring physiological state of the body. Contactless systems (e.g. radar-based systems) for precise vital signs monitoring (heartrate and breathing rate) can overcome such problems and enable using objective and reliable measurement of sleepiness based on physiological features [11], [12].

The aim of this paper is the development of a drowsiness detection algorithm based on the variability of heart rate and breathing rate features using different machine learning approaches. This type of algorithm could be further implemented for use in non-contact drowsiness detection systems which rely on the measurement of heart rate and breathing rate. In Section II we have presented the experiment setup, signal preprocessing methods, feature extraction procedure and classification methods. The classification results are reported in Section III. Finally, the conclusion and plans for future work are given in Section IV.

II. The Method

A. Experiment description

A group of ten healthy volunteers participated in the experiment (four males and six females, 26±5 years). The written informed consent form was signed. The experiment is carried out in accordance with the ethical standards of the Declaration of Helsinki. They were in a semi-seated position and exposed to multimedia content from a laptop (movie,

Anita Lupšić is with the NovelIC L.L.C., Veljka Dugoševića 54/A3, 11000 Belgrade, Serbia and with the School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail: anita.lupsic@novelic.com).

Predrag Tadić is with the University of Belgrade - School of Electrical Engineering, Bulevar kralja Aleksandra 73, 11120 Belgrade, Serbia (e-mail: ptadic@etf.rs).

Veljko Mihajlović is with the NovelIC L.L.C., Veljka Dugoševića 54/A3, 11000 Belgrade, Serbia (e-mail: veljko.mihajlovic@novelic.com).

Milica Janković is with the University of Belgrade - School of Electrical Engineering, Bulevar kralja Aleksandra 73, 11120 Belgrade, Serbia (e-mail: piperski@etf.rs)..

music or noise, depending on the subject's wishes). Laptop was positioned in front of the subject so that the camera could capture the face and eyes of the subject, Fig. 1. Electrocardiogram (ECG) and respiration curve were acquired using the reliable textile-based Wearable Wellness System (WWS, Italy, Pisa) that was positioned as a belt around the subject's torso [13]. WWS sends ECG and respiration data via bluetooth to the laptop where data logging is performed. A screen shot of the acquisition software for ECG and breathing monitoring is presented in Fig. 2.

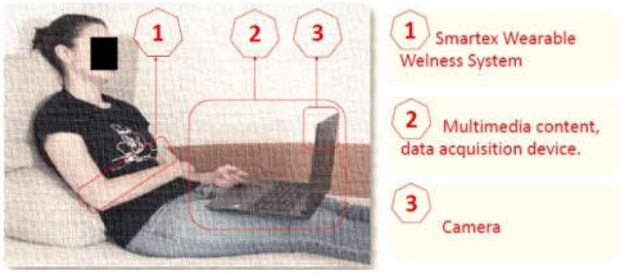


Fig. 1. Experiment setup

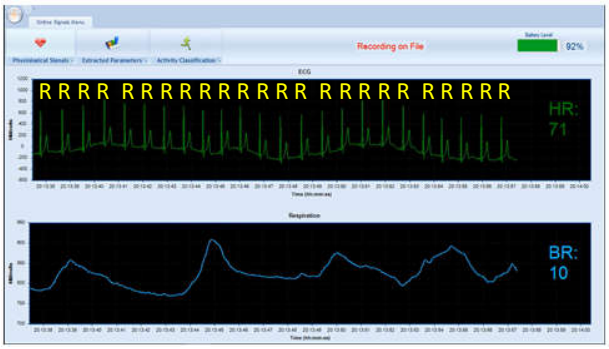


Fig. 2. A screen shot of WWS data acquisition software

B. *Signal preprocessing and feature extraction*

Sampling rates for ECG and breathing were 250Hz and 25 Hz, respectively. Beat-to-beat intervals (position of R waves in ECG) were extracted using the Pan Tompkins algorithm [14]. Time, frequency and fractal feature extraction for heart rate variability (HRV) and breathing rate parameters was performed. Table I and Table II show HRV and breathing rate parameters, respectively, selected as features for classification, as in [15].

C. *Classification*

Data labeling was performed manually, based on video recordings in three output classes: 1-"awake", 2-"drowsy" and 3-"fallen asleep". 80% of data was used for training and validation, and 20% of data was used for testing. K-fold cross-validation method (k=10) was used for evaluation of machine learning approaches [16]. Classification procedure was performed using: Support Vector Machine (SVM) with Gaussian kernel [17], k-nearest neighbors (kNN) and ensemble methods (boosted trees, bagged trees, random subspace method) [18]. Trained model needs to have two main characteristics: high classification accuracy and low risk of false alarm. After determination the optimal model, data are ones again run through the selected model for finding the optimal hyper-parameters. The final metrics, reported in Section III, were obtained by running the held-out test data through the tuned models.

TABLE I
HEART RATE VARIABILITY FEATURES

Parameters	Units	Definition
Time domain		
MeanRR	ms	Mean value of the RR intervals in the fixed time window.
STDRR	ms	Standard deviation of RR intervals in the fixed time window.
RMSSD	ms	Root mean square of RR intervals in the fixed time window.
NN50	count	Number of neighbour RR intervals with maximum differences of 50 ms.
Frequency domain		
VLF,LF,HF	ms	Absolute power of frequency ranges.
LF,HF	n.u.	Normalised power of different frequency ranges.
Non-linear parameters		
SD1	ms	Standard deviation of Poincare plot perpendicular to the line-of-identity (45° line to the normal axis).
SD2	ms	The standard deviation of Poincare plot along the line-of-identity (45° line to the normal axis)
ApEn	-	Entropy for short-time signals.

TABLE II
BREATHING RATE FEATURES

Parameters	Units	Definition
Time domain		
MeanBB	ms	Mean value of BB intervals in the fixed width time window.
STDBB	ms	Standard deviation of BB intervals in the fixed width time window.
RMSSD_BB	ms	Root mean square of BB intervals in the fixed width time window.

III. RESULTS

Confusion matrix for SVM with Gaussian kernel is presented in Fig.3. The optimization of hyper parameters was done for each model separately. A one-versus-all approach was employed – three binary classifiers were built, each of which distinguishes between one of the classes and the remaining two. Fig. 3 shows that 33% instances of the class "drowsy" is misclassified as "awake". The overall accuracy was 73%.

Fig. 3. SVC confusion matrix for Gaussian kernel

kNN was trained for different distance functions, as well as for different numbers of parameter k. The optimization was done for eleven different distance functions [19] (Spearman, Seuclidean, Mincowski, Mehalanobis, Jaccard, Hamming, Euclidean, Cosine, Correlation, Chebyshev, City block Distance) and for each function validation errors were calculated for k=1 to k=100 neighbors. Fig. 4 presents the objective function model through iterations for different distance functions and different numbers of neighbors.

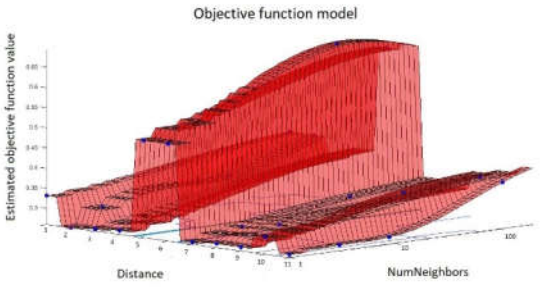


Fig. 4. kNN objective function model for eleven distance functions and k=1,100. (1-Spearman,2- Seuclidean,3- Mincowski, 4-Mehalanobis, 5-Jaccard, 6-Hamming, 7-Euclidean, 8-Cosine, 9-Correlation,10- Chebyshev, 11-City block Distance)

Optimization for these two parameters (distance function and number of neighbors k) resulted in the model with the confusion matrix presented in Fig. 5. The probability of misclassification for the "drowsy" class is 29%, 18% of instances from this class were labeled as "awake", and 10% as "fallen asleep".

New optimization was done by the weight which should be assigned to every sample. The model which gave the best results was obtained for k=10 neighbors with Euclidean distance, with reciprocal weights. 24% of instances from the "drowsy" class was misclassified: 14% was mislabeled as "awake", and 11% as "fallen asleep", Fig. 6. The overall accuracy was 75.7%.

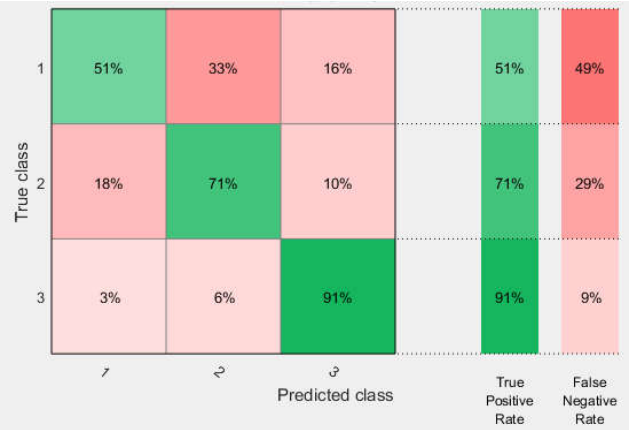


Fig. 5. kNN confusion matrix (optimized for two parameters)

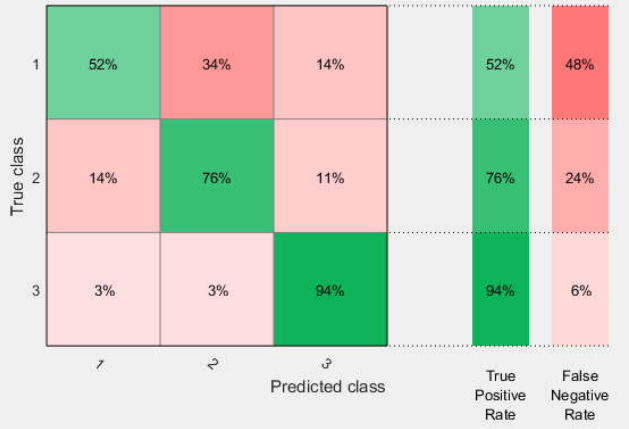


Fig. 6. kNN confusion matrix (optimized for four parameters)

Confusion matrix for different types of ensemble models (boosted tree, bagged tree and random subspace model with kNN classifiers as week learners) is presented in Fig. 7.

Boosted tree - We have optimized boosted tree model over the following hyper parameters: learning rate, number of trees and the maximum depth of each tree. Based on this the best result is achieved with model with *number of trees*=52, *maximal number of branching*=16, *learning constant*=0.1. Figure 7A shows the confusion matrix for boosting algorithm. The "drowsy" class was labeled correctly with probability 78%, while 17% of instances were labeled as "awake". The overall accuracy was 71%.

Bagged tree - Every tree is trained on a different, bootstrapped dataset, and the final decision is made by a majority vote, [20]. Figure 7B shows the confusion matrix for the bagged tree model with the learning rate of 0.1 and with 100 week learners. The overall accuracy was 74%.

Random subspace – This method was used for improving

the performance of the kNN model. kNN was used to train the weak learners, [21]. After a certain point the overall error gets bigger when the number of nearest neighbors is increased. The minimum is achieved for ten nearest neighbors, with the overall accuracy converges to 77%. The confusion matrix for this choice of parameters is represented in Fig. 7C.

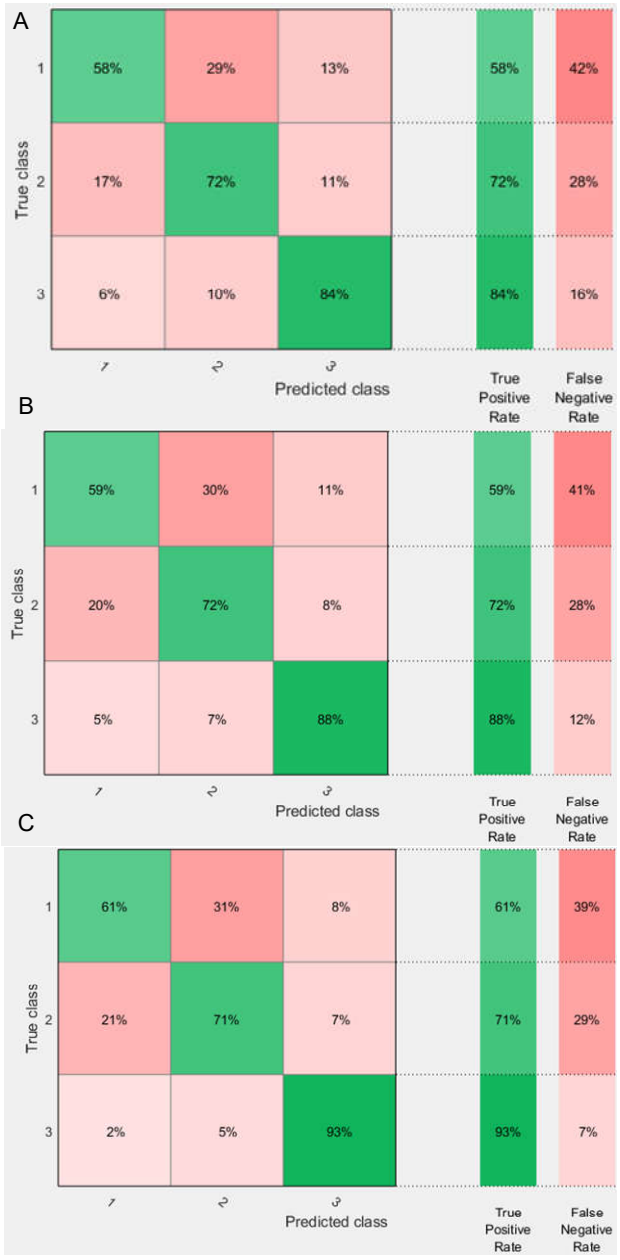


Fig. 7. Ensemble methods confusion matrix for: A) boosted trees method, B) bagged trees method, C) random subspace method

The model with the best performance was the ensemble method with kNNs as weak learners. The main problem is slightly higher error when classifying the "drowsy" class. Model which gave a slightly higher estimation of validation error, but better performance on the samples from the "drowsy" class is the kNN model (optimized for four parameters). Fig. 8 and 9 show confusion matrices on the test set for these two models.

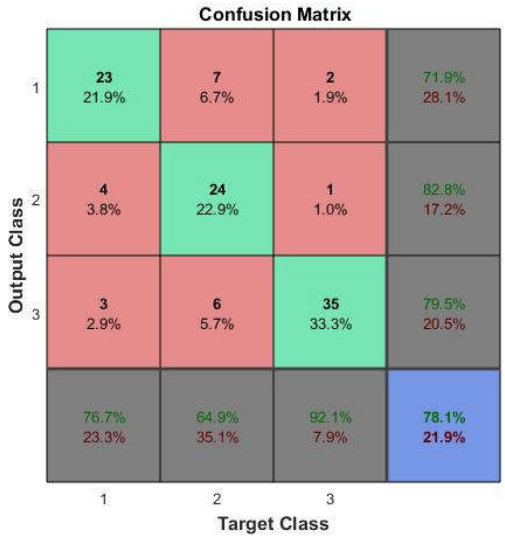


Fig. 8. Confusion matrix for random subspace model on test data

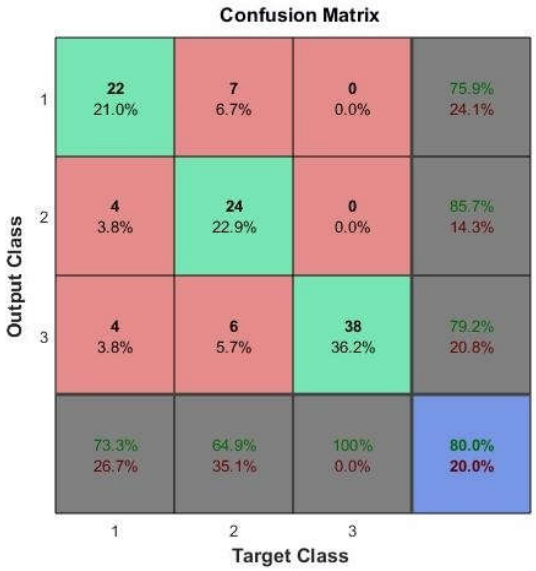


Fig. 9. Confusion matrix for kNN model on test data

IV. Conclusion

The study presented in this paper compared results of several machine learning approaches for detection the degree of alertness, based on features of cardiopulmonary signals. The upper accuracy of 77% and 80% was obtained in the validation and test process, respectively. Results could be improved by a more precise and automatic data labeling. Video reference could be replaced by electroencephalographic (EEG) recording signal that offers reliable detection of

sleeping phases from the EEG signal [22]. This approach would not only lead to the better precision, but also to the possibility of distinguishing more number of classes [23]. Also, more realistic environment for inducing subjects into sleeping phases is desirable (e.g. driving simulator in virtual reality). This would give us a much more realistic picture of the transition between the sleeping stages.

ACKNOWLEDGMENT

This research was partly supported by the Innovation fund of Serbia (no.ID50053) and the Ministry for Education, Science and Technology Development of Serbia, Belgrade, Serbia (no. OS175016).

REFERENCES

[1] J. Shen, J. Barbera and C. M. Shapiro, "Distinguishing sleepiness and fatigue: focus on definition and measurement," *Sleep medicine reviews,* vol. 10, pp. 63-76, 2006.

[2] J. D. Slater, "A definition of drowsiness: One purpose for sleep?," *Medical hypotheses,* vol. 71, pp. 641-644, 2008.

[3] R. Cluydts, E. De Valck, E. Verstraeten and P. Theys, "Daytime sleepiness and its evaluation," *Sleep medicine reviews,* vol. 6, pp. 83-96, 2002.

[4] D. Neu, P. Linkowski and O. Le Bon, "Clinical complaints of daytime sleepiness and fatigue: how to distinguish and treat them, especially when they become'excessive'or'chronic'?," *Acta neurologica belgica*, vol. 110, p. 15, 2010.

[5] G. Fallone, J. A. Owens and J. Deane, "Sleepiness in children and adolescents: clinical implications," *Sleep medicine reviews,* vol. 6, pp. 287-306, 2002.

[6] B. C. Tefft, "Asleep at the wheel: The prevalence and impact of drowsy driving," 2010.

[7] A. Sahayadhas, K. Sundaraj, M. Murugappan, "Detecting driver drowsiness based on sensors: a review", *Sensors*, vol. 12, no. 12, pp. 16937-16953, 2012.

[8] S. Ftouni, T. L. Sletten, M. Howard, C. Anderson, M. G. Lenné, S. W. Lockley, S. M. Rajaratnam, "Objective and subjective measures of sleepiness, and their associations with on-road driving events in shift workers", *Journal of sleep research*, vol. 22, no. 1, pp. 58-69, 2013.

[9] R. R. Johnson, D. P. Popovic, R. E. Olmstead, M. Stikic, D. J. Levendowski, C. Berka, "Drowsiness/alertness algorithm development and validation using synchronized EEG and cognitive performance to individualize a generalized model", *Biological psychology*, vol. 87, no. 2, pp. 241-250, 2011.

[10] B.-G. L. a. W.-Y. Chung, "Driver alertness monitoring using fusion of facial features and bio-signals," *IEEE Sensors Journal,* vol. 12, no. 7, pp. 2416-2422, 2012.

[11] X. Gu, L. Zhang, Y. Xiao, H. Zhang, H. Hong, X. Zhu, "Non-contact Fatigue Driving Detection Using CW Doppler Radar", Proc. IEEE MTT-S International Wireless Symposium (IWS), 6-10 May, Chengdu, China, pp. 1-3, 2018.

[12] K. Staszek, K. Wincza, S. Gruszczynski, "Driver's drowsiness monitoring system utilizing microwave Doppler sensor," Proc. 19th International Conference on Microwaves, Radar & Wireless Communications, 21-23 May, Warsaw, Poland, vol. 2, pp. 623-626, 2012.

[13] M. D. Rienzo, V. Racca, F. Rizzo, B. Bordoni, G. Parati, P. Castiglioni, P. Meriggi, M. Ferratini, "Evaluation of a textile-based wearable system for the electrocardiogram monitoring in cardiac patients", *Europace*, vol. 15, no. 4, pp. 607-612, 2013.

[14] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm," *IEEE Trans. Biomed. Eng,* vol. 32, pp. 230-236, 1985.

[15] M. Mahachandra, I. Z. Sutalaksana, K. Suryadi and others, "Sensitivity of heart rate variability as indicator of driver sleepiness," in *Network of Ergonomics Societies Conference (SEANES), 2012 Southeast Asian*, 2012.

[16] R. Fletcher, Practical methods of optimization, John Wiley & Sons, 2013.

[17] N. Cristianini, J. Shawe-Taylor and others, An introduction to support vector machines and other kernel-based learning methods, Cambridge university press, 2000.

[18] Z.-H. Zhou, Ensemble methods: foundations and algorithms, Chapman and Hall/CRC, 2012.

[19] V. B. Prasath, H. A. A. Alfeilat, O. Lasassmeh and A. Hassanat, "Distance and Similarity Measures Effect on the Performance of K-Nearest Neighbor Classifier-A Review," *arXiv preprint arXiv:1708.04321,* 2017.

[20] L. Breiman, "Bagging predictors," *Machine learning,* vol. 24, no. 2, pp. 123--140, 1996.

[21] I. Barandiaran, "The random subspace method for constructing decision forests," *IEEE transactions on pattern analysis and machine intelligence,* vol. 20, no. 8, 1998.

[22] A. Vuckovic, V. Radivojevic, D. A.C.N. Chen, Popović, "Automatic recognition of alertness and drowsiness from EEG by an artificial neural network," *Medical engineering & physics,* vol. 24, no. 5, pp. 349--360, 2002.

[23] M. W. Johns, "A new method for measuring daytime sleepiness: the Epworth sleepiness scale," *sleep,* vol. 14, no. 6, pp. 540--545, 1991.

Analysis of PVC microfluidic system for antibacterial solutions delivery in dentistry

Anđela Stojanović, Jovana Jevremov, Bojan Petrović, Sanja Kojić, Jovana Lazarević and Goran Stojanović, *Member, IEEE*

Abstract- Microfluidic systems can be used for oral diagnostics but they can also provide a new approach for effective drug delivery. The aim of this investigation was to evaluate the applicability of PVC microfluidic setup for the purposes of controlled release of two antibacterial mouthwashes commonly used in dentistry and to evaluate basic physical properties of liquids within the channel of microchips. Eight PVC chips were fabricated, they were Y-channel chips without any obstacles, with the input channels set at an angle of 60 °, and the width of 500-700 μm. The analyzed parameters included the passage, speed and necessary pressure for the laminar flow of solutions. The liquid diffusion was observed with a USB camera. All chips were crossable for both tested solutions. A laminar flow for both liquids was achieved with a pressure of 40 mbar. The minimum pressure on which flow was possible was 1 mbar for Eludril and 5 mbar for Curasept. The obtained data indicate that controlled drug delivery for routine use in dental clinical practice utilizing microfluidic setups require additional preclinical confirmation, calibration of all relevant parameters and the improvement of merge of existing medical and engineering technologies.

Keywords. Mouthwash solutions; drug delivery; microfluidics; PVC chips

I. Introduction

Microfluidics offers numerous possibilities for various clinical applications since it is still young and rather underdeveloped area of engineering that allows us to control liquids on the micro and nano scale, factors enabling basic medical science to extend its diagnostic, research and innovation properties. Microfluidic setups generally consist of fluid channels through which liquids are passed, pumping units and electronics for data collection and control. During the last couple of decades, exploration of the possibilities fordesigning less invasive and more accurate diagnostic tests gained increased attention, together with the raising concern for targeted and individualized drug delivery which has intensified research into the opportunity of salivary use as a diagnostic fluid, and intraoral drug delivery as a favorable drug delivery route. It has been clearly shown that among non-invasive organic fluids, saliva is one of the most preferred and most practical samples for monitoring general and oral health because it is always easily accessible, effortlessly gathered and stored [1,2]. When analyzing the drug delivery patterns, the intraoral route is without doubt the most preferred from patients' perspective. In addition, numerous factors render the oral cavity tissues, mucosa and teeth, appealing and practical sites for systemic and local drug delivery. The mucosa of oral cavity is rather permeable with a good vascularization and innervation, and exhibits rapid healing period after trauma or damage [3,4]. However, intraoral administration of drugs has its disadvantages for both oral and general diseases, such as unpredictable patients' compliance, unstable effective concentrations, clearance within the oral cavity, mixing properties with saliva and penetration to targeted tissues and degradation within the gastrointestinal tract. All abovementioned limits encouraged the consideration of other similar tissues and mucosae as probable drug release routes and sites for drug absorption [5]. Inside the oral cavity, drug delivery is categorized into three groups: 1) sublingual delivery, 2) buccal delivery, 3) local delivery.

Chlorhexidine, especially in a form of digluconate (CHX) is an extensively assessed and commonly used antibacterial agent in dentistry, since it exhibits the potential for efficient disinfection in the oral cavity against numerous bacterial strains and also has antifungal properties. The application indications for CHX are numerous, ranging from prophylactic and preventive, preoperative and postoperative disinfection in surgical procedures, and finally, to therapeutic use for the therapy of gingivitis and periodontal disease [6]. CHX is used in various forms and concentrations for both at home and professional applications, and nowadays it is used in concentrations ranging from 0.1% up to 20%.

Studies on the controlled and slow release of antimicrobial agents, including CHX, from some polymer based systems, and the assessment of their clinical performances have been described [7]. And at the moment, various polymer systems such as fibers, micro beads, films and strips are used for oral drug delivery systems, using chemical or physical release regulation with the aim of adjusting the release level, without the possibility of automation and target delivery. When it comes to CHX, targeted tissues are condition dependent, varying from subgingival periodontal pocket, gingival surface, to exposed mucosal wound, issues that can be easily targeted with using microfluidic systems. However, at the same time it has been clearly pointed out to the fact that transitioning from the research laboratory microfluidic setups into the routine clinical applications faces some important challenges. With

Anđela Stojanović is with the Faculty of Medicine, University of Novi Sad, Hajduk Veljkova 3, 21000 Novi Sad, Serbia (e-mail: stojanovic.andjela96@yahoo.com)

Jovana Jevremov is with the Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia (e-mail: jevremov.jovana@gmail.com)

Bojan Petrović is with the Faculty of Medicine, University of Novi Sad, Hajduk Veljkova 3, 21000 Novi Sad, Serbia (e-mail: bojan.petrovic@mf.uns.ac.rs)

Sanja Kojić is with the Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia (e-mail: sanjakojic@uns.ac.rs)

Jovana Lazarević is with the Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia (e-mail: jlazarevic.ftn@gmail.com)

Goran Stojanović is with the Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia (e-mail: sgoran@uns.ac.rs)

increasing number of microfluidic devices and applications emerging for clinical use, it is mandatory to assess the selection of substrate and fabrication method in order for the product to meet biocompatibility, clinical safety and regulatory criteria, and finally to contribute to answering the question: "*when does a microfluidic device become a medical device, as opposed to a research product*" [8]. Bearing in mind that the area of microfluidic setups rapidly transfers from research laboratories to everyday clinical practice, the careful choice of materials, systems, setups that comply with medical prerequisites turn out to be very important.

The aim of this investigation was to evaluate the applicability of PVC microfluidic setup for the purposes of controlled release of two antibacterial mouthwashes commonly used in dentistry and to evaluate basic physical properties of liquids within the channel of microchips.

II. Material and methods

Microfluidic properties of two different CHX based mouthwash solutions were analyzed, and the chemical composition of the investigated liquids is presented in Table I). Eight PVC chips were fabricated, they were Y-channel chips without any obstacles inside of the channel, with the input channels set at an angle of 60 ° (to aviod turbulance) , and the width of 500-700 μm and observation field (Figure 1). The analyzed parameters included the passage, speed and necessary pressure for the laminar flow of solutions. The liquid diffusion was observed with a USB camera (Figure 2).

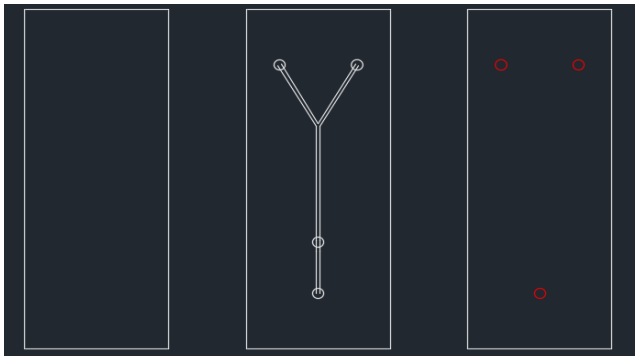


Figure 1. The scheme of the PVC chip design.

TABLE I
COMPOSITIONS OF INVESTIGATED MOUTHWASHES

Eludril	CURASEPT
Glycerin	Water
Chlorhexidine digluconate (0.1%)	Chlorhexidine digluconate (0.2%)
Water	Xylitol
Ethanol	Propylene glycol
Chlorbutanol	Hydrogenated castor oil

For the visualization purposes, tested fluids were painted with gentian violet (Figure 3). Plotter/cutter was used to cut the foils (125 microns thick) in order to manufacture the microfluidic chips. The parameters used for different layer elements are given in Table II, the same experimental and setup parameters were used in our previous reports [9,10], where for the realization of the microfluidic chips, the xurographic technique was used. In order to examine what pressure in the channel is needed for laminar flow of liquid, the fluid pressure at the inlet was changed.

TABLE II
PARAMETERS OF XUROGRAPHIC METHOD FOR CHIPS FABRICATION

	Channel Cutting Parameters	Reference values	Edge cutting parameters	Reference values
Speed	5	1-8	10	1-8
Acceleration	1	1-64	1	1-64
Force	28	1-38	28	1-38

III. Results

All fabricated microfluidic chips were passable for both tested solutions. A laminar flow for both liquids was achieved with a pressure of 40 mbar. The minimum pressure on which flow was possible was 1 mbar for Eludril and 5 mbar for Curasept. Experimental setup enabled satisfactory control, visualization of the flow and recording of the trial. The formation of the bubbles was recorded in both tested solutions but it did not significantly affect the procedure.

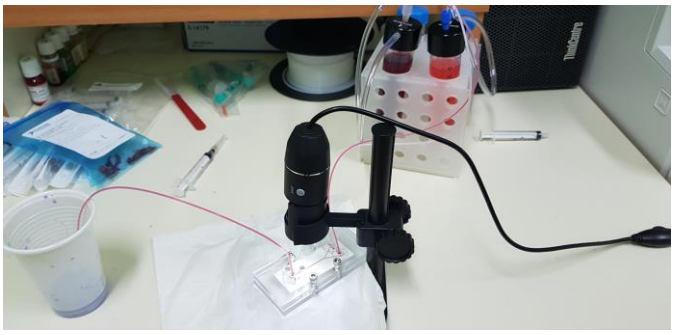


Figure 2. Experimental setup

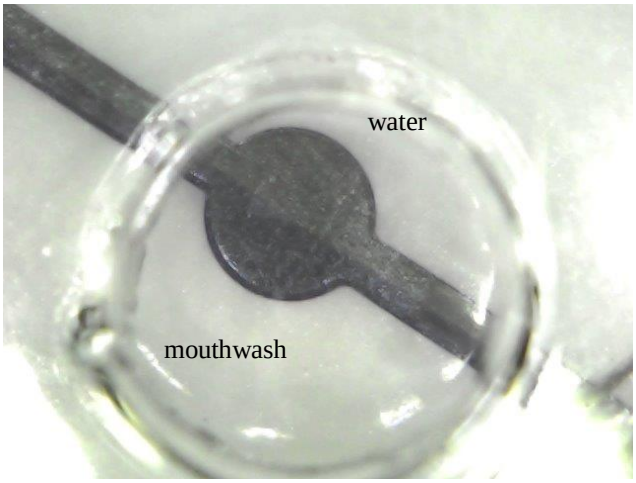


Figure 3. Observational field visualizing laminar liquid flow.

IV. Discussion

This study was conducted in order to assess the possibilities of integration of salivary diagnostics within the microfluidic systems. The chips, experimental setups and all employed technologies were for the first time, according to the authors' best knowledge, used for the purposes of evaluating the microfluidic behavior of the antibacterial solutions within PVC based microfluidic systems. This paper, together with previously published reports from our group [9,10], investigates the possibilities of designing optimal microfluidic system for salivary theranostics based on economic and easily fabricated PVC technology, following the criteria set for contemporary medical devices, and defined as ASSURED by WHO [11], that suggested that appliances used for routine procedures in general health monitoring need to be affordable, sensitive, specific, user-friendly, rapid and robust, without the use of complex equipment.

For more than forty years CHX have been used in preventive, prophylactic and therapeutical procedures in dentistry, and its efficiency has been attributed to a high potency, moderate side effects and the lack of bacterial resistance formation [12]. In the present investigation, two commercially available CHX solutions, Eludril and Curasept were chosen for integration into microfluidic system. Controlled release devices, for a prolonged treatment have already been developed and clinically tested within polymer based systems and with different effects [12], and mainly for the use in periodontal disease treatment. However, the use of CHX based solutions is not limited to the periodontal pockets treatment, since the local delivery of antiseptic agents to hard dental tissues and oral mucosa has a lot of indications, comprising dental plaque accumulation prevention, both bacterial and fungal infections, oral ulcers and various types of stomatitis. In addition, during several past decades, the researchers and clinicians search for a therapeutic option employing the device to be placed in the oral cavity, with the possibility of adherence to hard dental tissues, intraoral appliance or oral mucosa, with the aim of treating surgical wounds, as well as wounds that are caused by mechanical, physical or chemical trauma, and other types of oral ulcerations [13], which is for the most part important in the therapy of children experiencing oral and dental trauma, or persons with disabilities, whose non-acceptance behavior makes problematic to put on dressings or local medications.

The expansion of microfluidic setups for salivary diagnostics, controlled drug delivery, possibility for incorporation into intraoral appliances offers significant advantages such as minimal risk of infectivity with no emotional and physical pain together with mechanization, combination, ability for multiple and simultaneous biomarker analysis. This approach offers opportunity for rapid testing using small samples amounts and requires minimal training for final user. Despite the fact that the significant academic interest in microfluidics has been observed, together with the increased interest in salivary diagnostics and intraoral route of drug delivery, the commercial applications have not progressed at a similar rate.

The majority of the materials currently used in microfluidic research and experimental setups, such as polydimethylsiloxane (PDMS) has not translated over to production well due to matters with manufacturability and scaling. In contrast to that, the lack of research data with respect to other, alternative materials that can be used for microfluidic chip fabrication, such as glass and thermoplastic polymers has been observed, and this is recognized as the major obstacle for rapid transfer of these technologies from experimental laboratories to industry and finally, to everyday clinical practice. On top of that, salivary theranostics carries its own limits, criteria and prerequisites that must be met before actual clinical application. That is the reason why the PVC chips with simple design were fabricated and tested against the commonly used antibacterial agent in the present investigation. There is evident size match between microfluidics and salivary constituents, investigated metabolites and diagnostic analites. Active drug formulation particles sizes are also in the micrometer scale range. This overlap in dimensions will certainly facilitate application of microfluidic setups in salivary theranostics, and justify the need for incorporation of these appliances within intraoral devices.The present study defined the criteria for several basic experimental parameters when it comes to behavior of two similar, but not identical solutions within the microfluidic systems, revealing the fact that every different specimen requires different experimental parameters. This suggests that future applications require specific chip designs, controlled canals dimension, adequate pressure and sophisticated analytical and control mechanism.

V. CONCLUSIONS

The obtained data indicate that controlled drug delivery for routine use in dental clinical practice utilizing microfluidic setups require additional preclinical confirmation, calibration of all relevant parameters and the improvement of merge of existing medical and engineering technologies.

Acknowledgment

This paper received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 690876.

References

[1] F. Sun, EJ. Reichenberger, "Saliva as a source of genomic DNA for genetic studies: review of current methods and applications," *Oral Health Dent Manag*., vol. 13, pp. 217–222, 2014.

[2] L. Nunes, R. Brenzikofer, D. V. Macedo, "Reference intervals for saliva analytes collected by a standardized method in a physically active population," *Clin Biochem.*, vol. 44, pp. 1440–1444, 2011.

[3] V. Agarwal, B. Mishra, "Design, development, and biopharmaceutical properties of buccoadhesive compacts of pentazocine," *Drug Development and Industrial Pharmacy*, vol. 25, no. 6, pp. 701–709, 1999.

[4] A. Ahuja, R. K. Khar, J. Ali, "Mucoadhesive drug delivery systems," *Drug Development and Industrial Pharmacy*, vol. 23, no. 5, pp. 489–515, 1997.

[5] A. H. Shojaei, "Buccal mucosa as a route for systemic drug delivery: a review," *Journal of Pharmacy and Pharmaceutical Sciences*, vol. 1, no. 1, pp. 15–30, 1998.

[6] M. Scholz, T. Reske, F. Boehmer, A. Hornung, N. Grabow, H. Lang, "In vitro chlorhexidine release from alginate based microbeads forperiodontal therapy," *PLoS ONE* vol. 12, no. 10, e0185562, 2017, https://doi.org/10.1371/journal.pone.0185562.

[7] M. Bruschi, O. Freitas, “Oral Bioadhesive Drug Delivery Systems,” *Drug Development and Industrial Pharmacy*, vol. 31, pp. 293–310, 2005.
[8] J. S. Kuo, D. T. Chiu, “Disposable microfluidic substrates: Transitioning from the research laboratory into the clinic,” *Lab Chip*, vol. 11, pp. 2656–2665, 2011.
[9] A. Stojanović, J. Jevremov, “Mogućnosti primene mikrofluidnih PVC čipova u dijagnostici rizika za nastanak oralnih oboljenja,” *Kongres studenata Medicinskog fakulteta, Univerziteta u Novom Sadu*, 2019.
[10] S. Kojić, A. Stojanović, J. Jevremov, J. Lazarević, B. Petrović, G. Stojanović, “Design of microfluidic PVC chip based systems for salivary diagnostics,” *Int. scientific conference in dentistry*, 2019.
[11] Bulletin of the World Health Organization, vol. 95, pp. 639–645, 2017, doi: http://dx.doi.org/10.2471/BLT.16.187468.
[12] T. K. Schuck, “Intelligent intraoral drug delivery microsystem,” *Proc. IMechE*, vol. 220, Part C: J. Mechanical Engineering Science.
[13] M. S. Silva, N. L. Neto, “Biophysical and biological characterization of intraoral multilayer membranes as potential carriers: A new drug delivery system for dentistry,” *Materials Science and Engineering C* vol. 71, pp. 498–503, 2017.

Performances of Microfluidic Mixing Regulated using Active Pressure Controller

Jovana Jevremov, Ivana Podunavac, Jovana Lazarević, Sanja Kojić, *Student Member, IEEE*, Vasa Radonić, *Member, IEEE,* and Goran Stojanović*, Member, IEEE*

***Abstract*— Microfluidics studies show how fluid dynamic changes at the microscale level. Interest in microfluidic technologies has been driven by associated developments in bio-related fields such as cell biology, genomics, drug delivery, high-throughput screening and diagnostics, as well as a recognized need to perform fast and efficient experiments on small-sample volumes. Fluid behaviour in the microfluidic channel depends on channel geometry, fluids inside the channels, used material for chip fabrication, chip complexity, presence of external force (passive or active) and many other factors. Rapid and uniform mixing are fundamental principles on which effective design and development of micro-mixer relies on. In this paper COMSOL Multiphysics simulation software was used to investigate the flow characteristics within Y shaped microfluidic channel model for different pressure signals (sin, ramp, step), their periods (1 and 1.5 s) and amplitudes (10 mbar, 50 mbar and 100 mbar). Results were compared with experimentally gained data obtained using commercial flow control system. The results confirm that the best mixing performance was achieved with step signal shape, on shorter periods, and with the higher pressure.**

***Index Terms*—microfluidics, fluids, active mixer, biomedicine, COMSOL Multiphysics.**

I. Introduction

Microfluidics represents cutting-edge combination of science and technology, which roots date back to the 1950s, when the microfluidic was used for the realization of inject printer head. It is based on control and manipulation of fluids at a microliter scale. Thanks to related advantages such as reduced sample volume, scalability, laminar flow (therefore predictable fluid behaviour), short time analysis and low-cost fabrication, microfluidics is becoming one of the fastest growing area of science [1]. These major advances made possible for microfluidics to be incorporated and applied in other fields such as tissue engineering, biosensors, medical diagnostics, ecology monitoring, etc. The "organ-on-a-chip" technology, which represents cellularized constructs integrated in microfluidics platform, faithfully imitates physiological, and pathological conditions of complex tissues, thus revolutionizing existing approaches to drug screening and toxicology studies [2]. More than often, this remarkable studies are followed with extensive and thorough computerized simulations in order to verify and test in vitro models [3]. The computational analysis and simulations are also performed alongside microfluidic experiment, in order to get more reliable insight in chip performance [4].

Jovana Jevremov, Jovana Lazarević, Sanja Kojić and Goran Stojanović are with the Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia (e-mails: sanjakojic@uns.ac.rs, sgoran@uns.ac.rs).

Ivana Podunavac and Vasa Radonić are with the BioSense Institute, University of Novi Sad, Dr Zorana Đinđića 1a, 21000 Novi Sad, Serbia (e-mail: vasarad@biosense.rs).

A microfluidic chip consists of set of micro-channels etched or moulded into different material (glass, silicon, ceramics, or polymers) and fabricated using different fabrication technologies such as photolithography, soft-lithography, PDMS (polydimethylsiloxane), LTCC (Low Temperature Co-fired Ceramics), laser micromachining or xurography [5-7]. The selection of the appropriate materials and technologies depend on the concrete application, chip complexity, applied detection principle, operating temperature, biocompatibility and many other factors.

The new generation of microfluidic chips are composed of network of micro-channels, chambers, and reagent storages connected together in order to achieve the desired features (mix, pump, sort, or somehow otherwise process the fluid) and can be integrated with other components such as micro-pumps, valves, electronics or optics. This system of micro-channels and components realized inside the microfluidic chip can be connected to the outside by inputs and outputs pierced through the chip that serves as an interface between the macro- and micro-world [8]. In that manner, different fluids can be injected and removed from the microfluidic chip through tubing, syringe adapters or even simple holes in the chip with external active systems (pressure controller, syringe or peristaltic pump) or with passive hydrostatic pressure.

New microfluidic platforms appear every day as a toolbox for the development of new testing kits and solutions, most commonly in combination with other technology such as PCB (Printed Circuit Board), LTCC, cellulose paper, glass, etc. The complexity of the chips and features of the systems depend on the application-specific requirements that can vary from very simple devices to complex lab-on-chip platforms [1]. The special attention in paid on the development of novel materials for chip fabrication and development of low-cost, disposable microfluidic chip for rapid in-field testing.

The aim of this study was to investigate mixing performance of low-cost microfluidic chip controlled with external active system. The standard Y geometry shaped microfluidic channel was used in order to test effects of different pressures and their nature (signal shapes and period) on percentage of the fluids mixing. Results obtained in COMSOL Multiphysics modelling software are compared with experimentally obtained results. In this paper, we use a novel hybrid technology concept that can be easily used for the rapid fabrication of low-cost microfluidic chips [5]. The proposed fabrication process combines Polyvinyl Chloride (PVC) foils and green tapes, and relies on the cost-effective xurography technique and laser micromachining process. The experimental results confirm that the best mixing performance was achieved with step signal. Potential biomedical applications in area of drug delivery and personal treatment are further discussed.

II. Materials and Methods

A. Materials

For the fabrication of the proposed microfluidic chips PVC foil—A4 hot lamination foil (MBL® 80MIC, Serbia) with the thickness of 80 µm was used. The middle chip layer has been realized using Ceram Tape GC (CERAMTEC GmbH®, Germany) and Heraeus CT800 (Heraeus Electronics LTCC Materials, Germany) green tapes.

Distillate water was the selected fluid. Water and water based food dye colorant (Aroma 1990®, Belgrade, Serbia) were mixed with ratio 5:1. All experiments were recorded with Digital Microscope.

B. Equipment

Plotter cutter (CE6000-60 PLUS®, Graphtec America, Inc., Irvine, CA, USA) with the 45° cutting blade (CB09U) and the cutting mat (12" Silhouette Cameo Cutting Mat, Sacramento, USA) were used for carving inlets, outlets and edges of PVC layers for microfluidic chips. Ceram Tapes GC and Heraeus CT800 tapes were cut out with laser (Rofin-Sinar Power Line D-100, Germany). Bondage between PVC and green tapes are performed through lamination with A4 card laminator (FG320, Minoan Binding Laminating, Serbia).

Profiler Huwitz Panasis with bioimaging software for 3D profile of microfluidic channels was used for profiler analysis and measurement of the channel width.

Microfluidic flow control system ElveflowOB1 [9] was used to set different pressures, periods of relaxation and shapes of pressure signal. In this experiment, three signal shapes were used: sine, ramp and step. Pressures were set to pressure amplitudes of 10 mbar, 50 mbar, and 100 mbar, while periods of pumping were set to 1 s and 1.5 s.

C. Methods

COMSOL Multiphysics® software was used for initial simulations of microfluidic active mixers and testing of their performances. Experimental testing has been performed on the fabricated chips using microfluidic set-up that consists of ElveflowOB1 microfluidic flow control system, PTFE tubing, fittings, holder, connections and digital microscope. Mixing efficiency in the channel was detected optically using digital camera, and mixing rates were determined using image processing algorithm developed in Matlab.

III. Chip Fabrication

The proposed microfluidic chip consists of three layers, as shown in Fig. 1. Top and bottom layers were realized using PVC foils, while the green tape was used for the middle layer. Fabrication of the chip was realized through of several steps. In the first step, laser cutting of the middle chip layer in the green tape, as for standard preparation of layers for LTCC technology was performed. Exact laser parameters used for microfluidic chip fabrication were: current 28 mA, frequency 10 kHz, and cutting speed of 15 mm/s. Plotter cutting of PVC layers, as for standard xurography technique, was used for cutting the inlets and outlet of the microfluidic channels. In the final step, lamination of the cut layers as for standard xurography technique (laminated first: Layer 1 and 2, and then Layer 3) has been accomplished at the temperature of 130 °C. Fig. 1b shows photographs of the fabricated Y-mixer microfluidic chip. In all designed chips, the width of the microfluidic channels has been set to 200

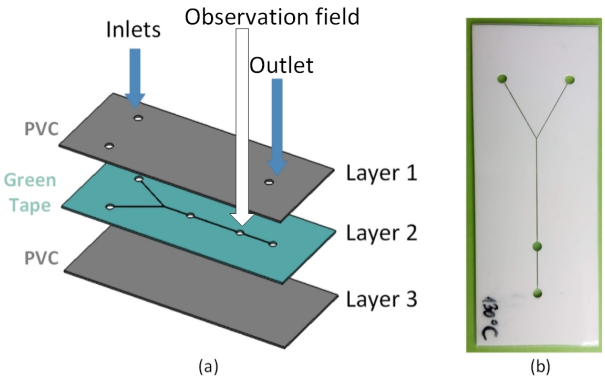


Fig. 1. Microfluidic chips fabricated using proposed hybrid technology (a) 3D model of the microfluidic chip, (b) Fabricated chip with Y-mixer.

µm, while the inlet and outlet holes, and observation field have been manufactured with the diameter of 2 mm. The total length of the microfluidics Y-mixer is chip is 50 mm.

In the process of the fabrication, dimensions slightly changes due to imperfection of laser cutting and lamination processes. Therefore, we fabricated 10 microfluidic chips, five using Ceram Tape GS and five using Hereaus CT800 as a middle layer. Profiler Huwitz Panasis was used for the characterization of the microfluidic channel widths. The width of the channel was measured at eight different points. Fig. 2 shows the measured channel and standard deviation of the channel width. It can be seen that the measured width of channels realized in Ceram Tape GC is smaller than predefined value, while relatively good agreement is obtained for Hereaus CT800 tape. The variation of the channel width in the worst case was below 15%. For experimental testing we used the chip with width of 150 µm and with a minimal deviation of 5% (chip marked as #1GC).

IV. COMSOL Multiphysics Simulations

COMSOL Multiphysics® software was used for simulation and testing of the proposed microfluidic mixer. COMSOL Multiphysics® software uses finite element method for numerical solving of different equations dictated by physical laws. In microfluidic simulations, fundamental equation used for the description of fluid motion in micro channels is Navier-Stokes equation [10]. Multiphysics simulation software was used to investigate the flow characteristics within Y shaped channel for different input pressure signals (sin, ramp, step), their periods (1 and 1.5 s) and amplitudes (10 mbar, 50 mbar and 100 mbar). Three different signal shapes were used for inlet pressure in simulations: sine, step and ramp. In Fig. 3 used signal shapes with period of 1 s are shown.

For visualization of the mixing process, different colours of fluids were used, representing the concentration of fluids. The concentration of one liquid was set to 500 mol/m^3 and concentration of the other one was set to zero. Fig. 4 represents the part of the micro channel in two specific cases.

Fig. 4a shows simulation results for constant pressure flows of 50 mbar at both inlets, while Fig. 4b shows simulation results for sine shaped pressure signals at both inlets. It can be seen that liquids mix only in their contact region, which is the property of laminar flow. Fig. 4b represents mixing of liquids with sine shaped pressure signal on both inlets with amplitude of 50 mbar. Mixing in this case covers almost the whole channel and it can be seen that mixing of liquids is significantly better.

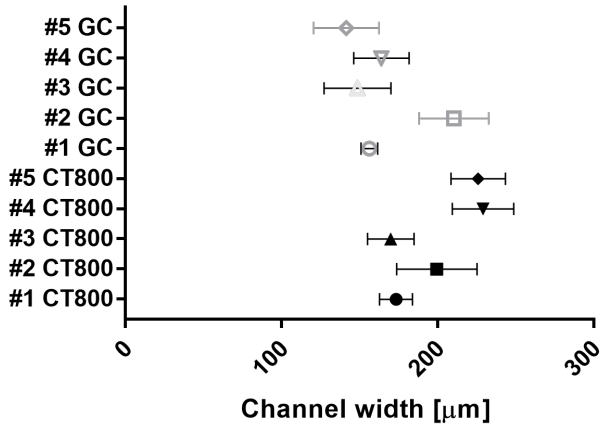


Fig. 2 The measured channel width and its standard deviation for 10 fabricated chips.

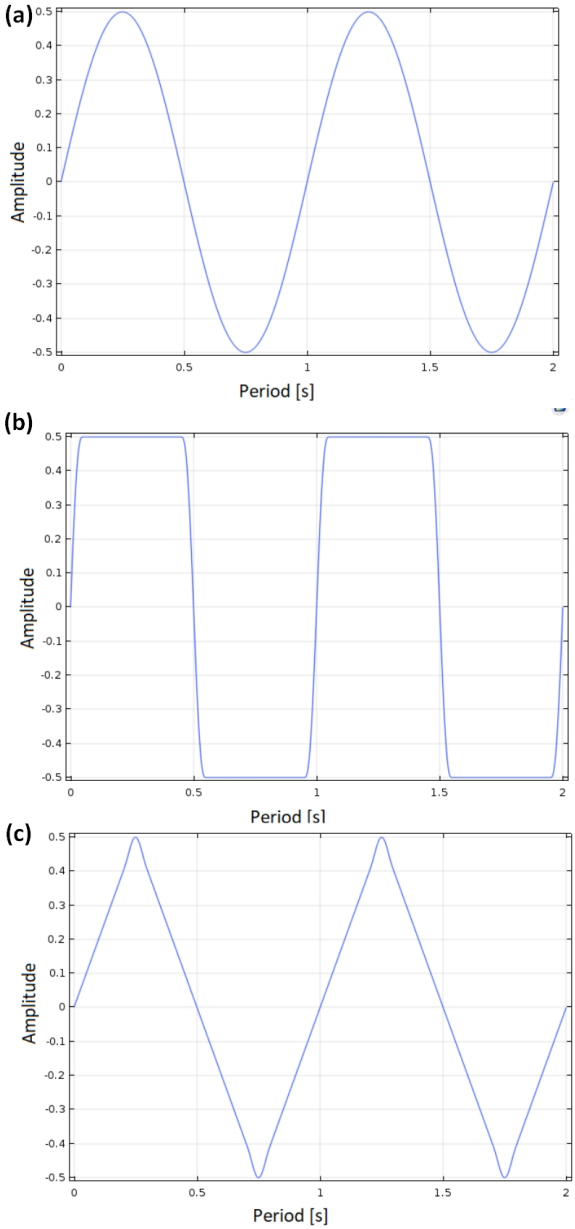


Fig. 3. Pressure signal at inlet of the chip: (a) Sine, (b) step, and (c) ramp.

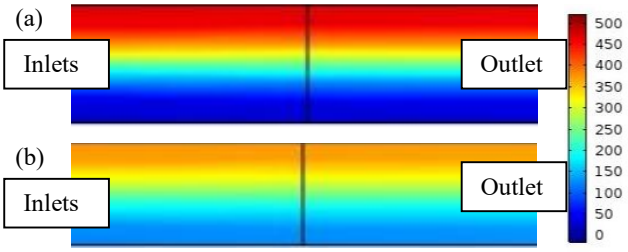


Fig. 4. Mixing of liquid for: (a) Constant pressure flows at inlets of mixer, and (b) Sine shaped pressure signal at inlets.

Fig. 5 shows the average value of concentration along the observation zone for the different signal shapes and pressures during 10 s. Results presented in Fig. 5 are simulation results with 1 s signal period. As it was expected, the better mixing of two fluids can be achieved for higher pressure values. Therefore, for 100 mbar pressure, mixing is achieved after 2 s, for 50 mbar after 4 s and for 10 mbar 10 s is not enough to achieve proper mixing of two fluids. Fig. 5 also shows that step signal gives the best mixing results in term of response time and mixing concentration.

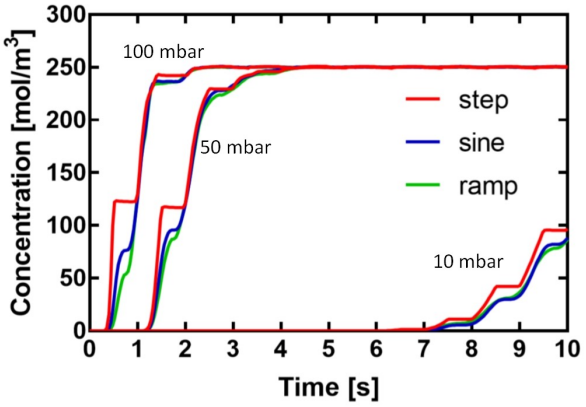


Fig. 5. Average value of concentration along the observation zone. Simulation results for step, sine and ramp signal shapes at pressures of 10, 50 and 100 mbar.

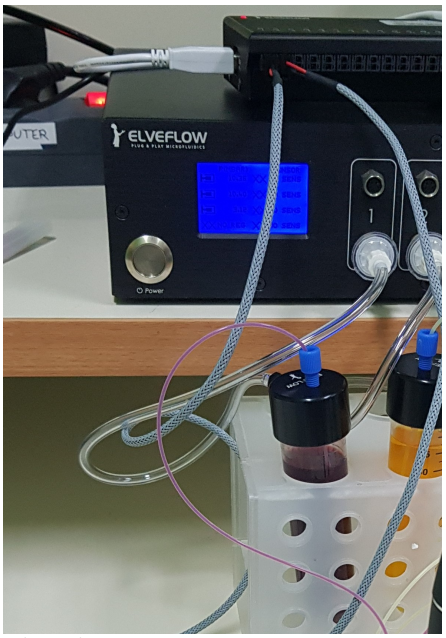


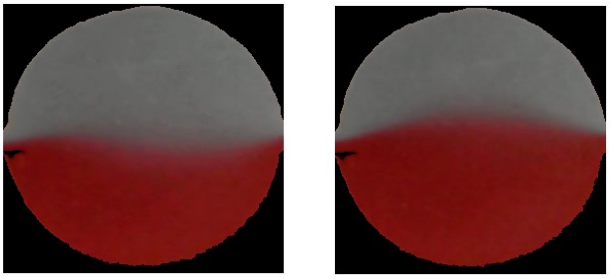
Fig. 6. Experimental set-up

V. Experimental Results

In order to verify simulation results, experimental testing has been performed on the fabricated chips. Set-up for conducting the experiment, shown in Fig. 6, consists of microfluidic flow control system ElveflowOB1, PTFE tubing, fittings, holder, connections and digital microscope.

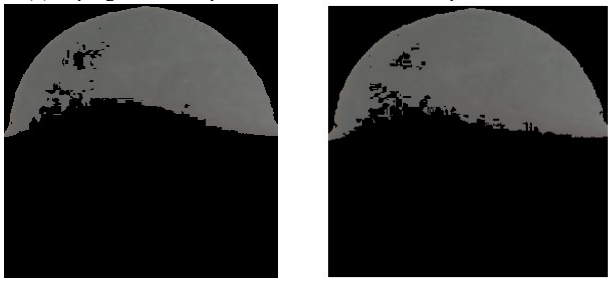
Mixing in the channel was detected visually using digital camera and mixing rates were determined in Matlab. For every flow pressure signal, amplitude, and period, 12 independent photographs during one period of signal were recorded. Photographs had high resolution of 3648 x 2736 dpi. Two examples of the recorded photos are shown in Fig 7. All pictures are further processed in Matlab in order to determine mixing efficiency.

Matlab software was used to isolate the observation field, and isolate the pure red and green color in the image, Fig. 8. Because of the equal flows at both inlets we assume that there is the equal amount of non-mixed red and yellow liquid. In the following step R, G and B components were determined in order to isolate pixels in the observation field, Fig. 9. Based on that for non-isolating, i.e. non-mixed liquids, pixels were classified and counted. The code is written to count the total number of pixels of the observation field, to take away the number of pure yellow/red, and to calculate the mixing efficiency. These results are presented as heat maps for the different amplitude and periods, after first period of signal. In Fig 9. heatmap of results is shown in percentage for different pressure, amplitude and period recorded after period of first impulse. As it was expected, better mixing is performed on shorter periods of pulsation, and higher pressures.



(a) (b)

Fig. 7. Photographs of the observation field after stimulation with: (a) sin, and (b) step signal with amplitude of 50 mbar within the period of 1 s.



(a) (b)

Fig.8. Isolated pixels of unmixed fluid based on 3 colour channels.

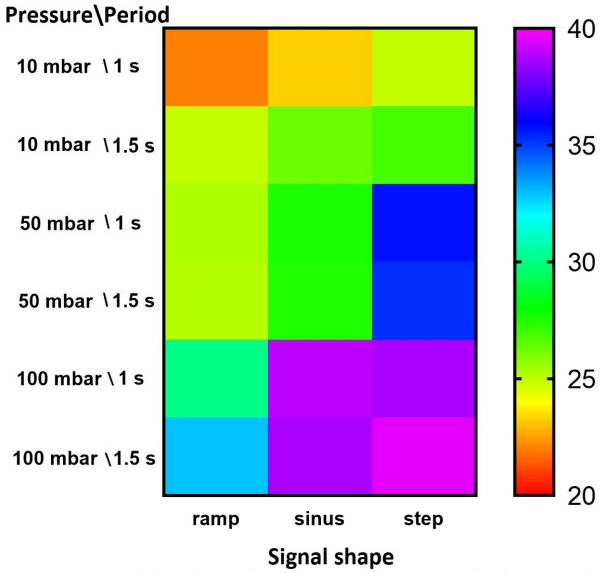


Fig. 9. Results of the mixing for different pressures, amplitude and period recorded after period of the first impulse.

The best mixing occurs for the stimulation with the step signal shape, because of the nature of signal itself, although mixing with sine and ramp signals give almost the same results. With the change of the shape of the signals and their period at the inlets of the microfludic chips the efficiency of the mixing can be increased for more than 30%.

VI. Conclusion

In this paper, we use novel technology concept for rapid fabrication of robust microfluidic mixer. The proposed fabrication process combines PVC foils and green tapes, and relies on the cost-effective xurographic technique and laser micromachining process. The comprehensive study of mixing performances on Y-shape microfluidic channel have been performed. Characteristics of mixing performances of Y shaped channel for different pressure signals (sin, ramp, step), their periods and amplitudes have been investigated using simulation and experimentally verified. Obtained results shows that the shape of signal at the inlets of active mixer, their amplitude and period can influence the mixing performances and can improve mixing efficiency.

The impact of the signal shape is demonstrated on a Y-mixer, but the same effect can also be applied to complex fluid mixing systems. Therefore, the proposed concept can be implemented in complex microfluidic mixing systems for applications in all types of lab-on-chip analysis where rapid mixing of two or more liquids is required.

Acknowledgment

This work is funded in the framework of the project III44006 and 142-451-2459/2018 as well as this paper received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 690876, MEDLEM.

References

[1] A. R. Perestrelo, A. C. Aguas, A. Rainer, G. Forte, "Microfluidic Organ/Body-on-a-Chip Devices at the Convergence of Biology and Microengineering," *Sensors (Basel),* vol. 15, no. 12, pp. 31142-31170, Dec. 2015.

[2] M. G. Whitesides, “The origins and the future of microfluidics,” *Nature*, vol. 442, no. 7101, pp. 368-373, Jul. 2006.

[3] S. Naher, D. Orpen, D. Brabazon, C. R. Poulsen, M. M. Morshed, “Effect of micro-channel geometry on fluid flow and mixing,” *Simulation Modelling Practice and Theory*, vol. 19, no. 4, pp. 1088-1095, Apr. 2011.

[4] S. S. Das, B. Patawari, P. K. Patowari, S. Halder, “Computational Analysis for Mixing of Fluids Flowing through Micro-Channels of Different Geometries,” Conference: AIMTDR-2014, ITT Guwahati, India, vol. 236, Dec. 2014.

[5] S. P. Kojic, G. M. Stojanovic, V. Radonic, “Novel Cost-Effective Microfluidic Chip Based on Hybrid Fabrication and Its Comprehensive Characterization,” *Sensors,* vol. 19, no. 7, 1719, Apr. 2019.

[6] M. D. Ville, P. Coquet, P. Brunet, R. Boukherroub, “Simple and low-cost fabrication of PDMS microfluidic round channels by surface-wetting parameters optimization,” *Microfluid. Nanofluidi,* vol. 12, pp. 953–961, Dec. 2011.

[7] N. Jankovic, V. Radonic, “A Microwave Microfluidic Sensor Based on a Dual-Mode Resonator for Dual-Sensing Applications,” *Sensors,* vol. 17, no. 12, 2713, Dec. 2017.

[8] R. M. Chanmanwar, R. Balasubramaniam., L. N. Wankhade, “Application and Manufacturing of Microfluidic Devices: Review,” *IJMER*, vol. 3, no. 2, pp. 849-856, Apr. 2013.

[9] https://www.elveflow.com/

[10] “Microfluidics User's Guide”, pp. 46-47. COMSOL Multiphysics® v. 4.3, 2012.

Increase of the Energy Efficiency of an Urban Type Wind Turbine in a Smart Energy Building

Christos Mademlis, *Senior Member, IEEE*

***Abstract*—This paper presents an efficiency increase control strategy for an urban type wind turbine in a nearly zero energy building (nZEB). The efficiency increase is attained by employing the flux-weakening control technique for the electrical generator and the maximum power point tracking control for the wind turbine. Thus, maximum power harvesting from the whole wind energy conversion system (WECS) is achieved and additionally expansion of the exploitable wind speed region towards the lower-speed range is accomplished.**

Specifically, the developed control technique has been based on previous research work of the author and it is properly adjusted so as, it can be utilized as generic method for any urban type wind turbine for smart energy buildings. Therefore, the developed control method can narrow the advantage of the permanent magnet synchronous generator with respect to the energy efficiency and therefore, a low cost and high efficiency wind generation system can be provided. The usability of the developed wind system is highly important, since the urban type wind turbines have not been widely spread as a renewable energy source in nZEBs, due to the higher cost compared to the photovoltaic systems. It should be noted that, although the proposed control system has been developed with a squirrel cage induction generator (SCIG), due to the advantage of the low cost; it can be applied to a permanent magnet synchronous generator, as well.

In the developed control method, a Minimum Electric Loss (MEL) controller is introduced in order to minimize the generator electric loss and a Maximum Power Point Tracking (MPPT) controller is used in order to maximize the wind turbine output power. Common input to the two optimal controllers is only the generator speed, while the measurement of the wind speed is not required. The two controllers determine the optimal *d*- and *q*-axis stator current components of the SCIG through optimal conditions and therefore, fast dynamic response of the WECS is accomplished. An experimental procedure is proposed to determine the MEL and MPPT controller parameters. Therefore, neither the knowledge of SCIG loss model nor the characteristic curves of the wind turbine are required. The effectiveness and the operational improvements of the suggested optimal control scheme have been verified experimentally.

***Index Terms*— Nearly zero energy building, wind power generation, optimal control, power generation, generators, variable speed drives.**

I. Introduction

NOWADAYS, the wind power industry is growing very fast and the modern wind turbines are efficient, reliable and produce power at reasonable cost [1]. Utilizing a squirrel cage induction generator (SCIG) as a mean to convert the mechanical energy captured by the wind turbine to electrical energy has a series of advantages, making it an attractive option for small and medium size WECS [2]. A SCIG has low manufacturing cost with robust construction compared to a permanent magnet synchronous generator. Also, it requires little maintenance compared to a wound rotor synchronous generator and a doubly fed induction generator.

On the other hand, buildings are the largest energy consuming sector in the world [3]. They are accountable for the one-third of the global energy consumption and consequently highly responsible for the increase of the carbon emission. Thus, several technologies have been developed to increase the efficiency and sustainability of the buildings and also, various policy measures have been adopted to support investments in RES and cultivate residents' consciousness in energy saving [4].

According to the Energy Performance of Buildings Directive (EPBD, 2010/31/EC), the definition of the nZEB is a building with very high energy performance where the required energy is covered by renewable energy sources (RES) produced onsite or nearby [5]. The RES in a nZEB can be any type, although more common are the photovoltaics and the wind turbines [6]. The research efforts for cost reduction of transforming an existing building to nZEB are towards the directions of reducing the construction cost of the renewable energy sources (RES) and on seeking cost-effective solutions for the constructional interventions [7], [8]. Key elements for the above techniques are the improvement of the quality and the cost reduction of the materials and devices, as well as the onsite study of the building needs in order to provide the best in case solutions for constructional interventions that can reduce the energy losses of a building.

However, considerable reduction of the cost for transforming a conventional building to nZEB can be attained by reducing the power capacity of the required RES, increasing the exploitation of the ESS and minimizing the complementary electric energy that is to be absorbed by the grid in order to fully cover the energy needs of a nZEB. These can be realized by increasing the efficiency of the RES, improving the energy management of the building microgrid by properly cooperating and highly exploiting the ESS and also, increasing the efficiency of the building's devices. Therefore, although the energy needs of a building in the level of use and comfort of the residents remain the same, the power and energy capacity requirements of the RES and ESS can be reduced (that result to the CAPEX reduction), as well as the energy consumption in the building can be confined (that results to lower OPEX). Moreover, constructional conversions of high cost and low energy saving result can be avoided, since the same level of

C. Mademlis is Professor with the Faculty of Electrical and Computer Engineering, Aristotle University of Thessaloniki, GR-54 124, Thessaloniki, Greece (e-mail: mademlis@auth.gr).

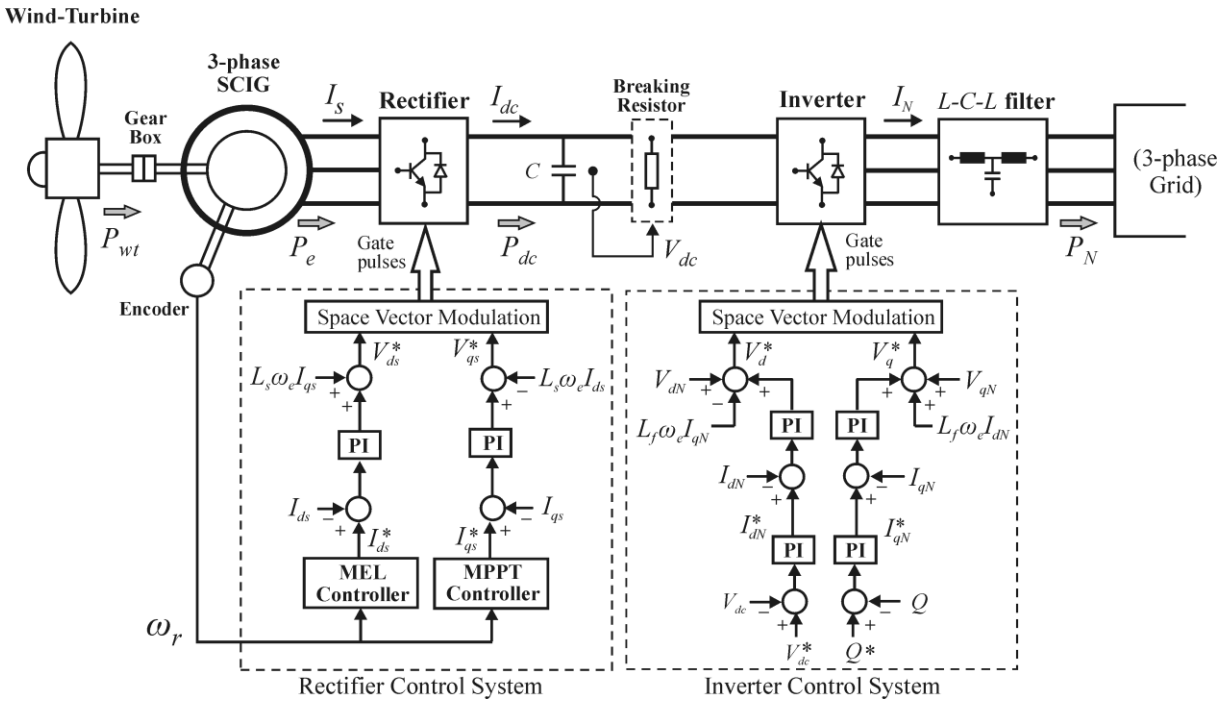


Fig. 1. Structure of the optimally controlled WECS with a 3-phase squirrel cage induction generator (SCIG).

energy autonomy and reliability in a nZEB can be achieved with lower installation cost.

The RES that is mainly utilized in a nZEB consists of photovoltaic and small wind turbines of urban type. In this paper, the efficiency increase in a RES is referred to the wind system, since this exhibits more margins for efficiency improvement.

Efficiency of a WECS is of great importance to provide maximum power harvesting from the incident wind. Variable speed capability of a wind turbine enables operation at its maximum aerodynamic efficiency and also, provides minimum torque perturbation in the drive train. Apart from the tip speed ratio control that requires an anemometer to measure the wind speed, several MPPT control schemes have been reported in the technical literature which are mainly based on optimal torque control [9], search control [10], fuzzy-logic control [11] and neural networks control methods [12]. Also, a fast tracking control algorithm has been presented in [13], an adaptive fuzzy-logic based scheme has been proposed in [14] and an MPPT controller with adaptive compensation control has been proposed in [15]. Finally, an MPPT algorithm for WECS with doubly fed induction generator, which takes advantage of the rotor inertia power has been presented in [16].

Additionally to MPPT control, the efficiency of the whole WECS can be improved by increasing the efficiency of the electric generator. This can be achieved through the appropriate control of the generator flux-linkage by regulating the d-axis stator current. A fuzzy-logic control method has been presented in [17] and search control techniques have been proposed in [18] and [19]; however, the WECS response is very slow and cannot follow the fast changes of the wind. A WECS control scheme of combined SCIG minimum ohmic loss controller with either a search or a fuzzy-logic MPPT controller has been presented in [20]. Model based optimal efficiency control methods for SCIG have been proposed in [21] and [22]; however, in [21] the variation of iron loss with frequency was disregarded and in [22] the accurate wind speed measurement is required. An optimized capacitance design for the dc-link of a back-to-back converter for wind turbine power generation was presented in [23]. A loss minimization method (considering copper and iron losses) for an induction generator that can be used for a stand-alone system for battery charging was presented in [24]. Finally, a model based optimally controlled method for interior PM synchronous generator in combination with an MPPT algorithm has been presented in [25].

From the above, it is concluded that there is a need for a cost-effective optimal efficiency control strategy for WECS with SCIG that has fast dynamic response and can be easily implemented. This control system can be applied to high as well as low power wind turbines, since it can be easily implemented and does not require any hardware modifications, but only replacement of the existing firmware of the WECS. The latter advantage is very important, since by not affecting the cost of the energy conversion system, the developed control technique can be adopted by any low power wind turbine. Therefore, the developed control method can contribute to narrowing the advantage of the counterpart permanent magnet synchronous generator against the SCIG, with respect to the energy efficiency of the wind system and thus, a low cost and high efficiency wind system can be provided that can be used for any smart energy building.

The developed control system for WECS with SCIG has been based on the methodology proposed by the author in the

[26] and [27], and it is properly adapted in order to by applied, as a generic control method, for any urban type wind turbine for smart energy buildings. The developed control method, minimizes the electric loss of the SCIG and maximizes the mechanical power extraction by the wind turbine and therefore, provides maximum efficiency of the whole WECS.

It should be noted that all electric loss of the SCIG has been considered. Specifically, two controllers have been introduced for the *d*- and *q*-axis stator current components, and the optimal efficiency is attained by properly controlling the above current components by means of optimal conditions. Therefore, both increase of the efficiency as well as fast dynamic performance of the WECS can be achieved. In particular, the MEL controller regulates the *d*-axis stator current in order to minimize the SCIG's loss and the MPPT controller adjusts the q-axis stator current in order to maximize the aerodynamic power in the wind turbine. Common input to the two optimal controllers is only the generator rotor speed.

An experimental procedure is proposed in order to determine the controller parameters and therefore, the proposed control scheme can be easily implemented because neither the loss model of the SCIG nor the characteristics of the wind turbine are required. Moreover, the wind speed information is not required and hence, the implementation of the suggested control system is cost-effective.

The effectiveness of the developed control technique WECS has been validated experimentally in a laboratory emulator of a wind system. Selective experimental results are provided in order to validate the theoretical considerations and demonstrate the operational improvements of the proposed control scheme.

II. WECS CONFIGURATION

Fig. 1 illustrates the block diagram of the optimally controlled WECS. The wind turbine is coupled to the shaft of a SCIG through a gear box that is inserted in order to adapt the low rotational speed of the wind turbine to the high speed of the generator. The SCIG is connected to the power grid through two back-to-back converters. For the rectifier, the field oriented control technique with two control loops is utilized. The inner PI control loop regulates the *d* and *q* axis current components for providing the respective *d* and *q* axis voltages through space vector modulation. The outer control loop contains the MEL controller that determines the optimal I_{ds} current so as the electric loss of the generator is minimized and the MPPT controller that determines the optimal I_{qs} current so as the mechanical power, that is provided from the wind turbine to the electric generator, is maximized. Common input to the MEL and MPPT controllers is the generator rotor speed ω_r.

For the line-side inverter, the voltage oriented control is employed by applying a double loop control system, as well. An inner PI control loop that regulates the *d* and *q* axis current components for providing the respective *d* and *q* axis voltages through space vector modulation and an outer PI control loop for the decoupled control of the active and reactive power

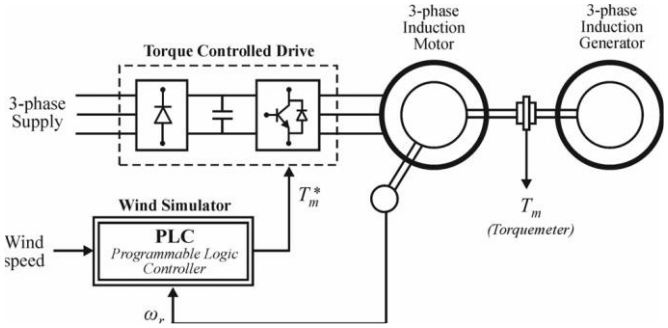


Fig. 2. Structure of the WECS emulator.

provided to the grid by controlling the I_{dN} and I_{qN} grid current components, respectively. In particular, the one PI controller of the outer loop is utilized to keep the *dc*-link voltage constant and thus, to control the active power. The other PI controller is used to regulate the line-side power factor and thus, to provide the demanded reactive power to the grid.

III. WIND TURBINE CHARACTERISTICS AND SCIG LOSS MODEL

The power captured by a wind turbine is given by

$$P_{wt} = \frac{1}{2}\rho\pi R^2 C_p u^3 \tag{1}$$

where ρ is the air density, R is the radius of the blades, C_p is the wind-turbine power coefficient and u is the wind speed. For each blade pitch angle β, the value of the tip-speed ratio λ_{opt} is constant for all MPPs and the optimum wind turbine shaft speed $\omega_{wt_{opt}}$ at a wind speed u is calculated by

$$\omega_{wt_{opt}} = \frac{\lambda_{opt} u}{R} \tag{2}$$

In this paper, a wind turbine emulator is used for the laboratory tests, implemented by an inverter fed induction motor, as per [28], [34] and the test WECS is illustrated in Fig. 2. A programmable logic controller (PLC) obtains wind speed values and, by using turbine characteristics and induction motor speed, calculates the torque value of the wind turbine. This is the reference value to the torque-controlled drive that forces the three-phase induction motor to act like a real wind turbine to the energy conversion system.

The relation of C_p versus λ of a three-blade horizontal axis wind turbine for various blade pitch angles β is illustrated in Fig. 3. The curves have been obtained by using the following equation that is commonly used in wind turbine simulators [12], [36]

$$C_p(\lambda,\beta) = 0.5176\left(\frac{116}{\lambda_i} - 0.4\beta - 5\right)e^{-\frac{5}{\lambda_i}} + 0.0068\lambda \tag{3}$$

with

$$\frac{1}{\lambda_i} = \frac{1}{\lambda + 0.08\beta} - \frac{0.035}{\beta^3 + 1} \tag{4}$$

A three-blade horizontal axis wind turbine with radius of 3.5 m is emulated at the system. Fig. 3 illustrates the steady state power-speed characteristics (solid curves) for a blade pitch angle of 0 degrees ($\beta = 0^0$). The maximum power point curve (dashed line) is attained at each wind speed with $C_p = 0.46$. Since a gear box is used, the wind turbine speed ω_{wt} and torque $T_{wt} = P_{wt} / \omega_{wt}$ are converted to the generator level on the basis of the gear ratio n as follows: $\omega_r = n\omega_{wt}$ and $T_r = T_{wt}/n$.

In field-oriented control on SCIG, decoupled control of d- and q-axis stator current components is provided by aligning the rotor flux linkage ψ_r to the d-axis ($\psi_{dr} = \psi_r$ while $\psi_{qr} = 0$). Therefore, the d- and q-axis rotor current components are given by [29]

$$I_{dr} = \frac{\psi_r}{L_r} - \frac{L_m}{L_r} I_{ds} \tag{5}$$

and

$$I_{qr} = -\frac{L_m}{L_r} I_{qs} \tag{6}$$

In Laplace transformation, the rotor flux linkage is given by

$$\psi_r = \frac{L_m}{(L_r / R_r)s + 1} I_{ds} \tag{7}$$

From (7), it is concluded that at steady state the rotor flux-linkage is given by

$$\psi_r = L_m I_{ds} \tag{8}$$

and consequently, from (5) results

$$I_{dr} = 0 \tag{9}$$

The d- and q-axis air-gap flux-linkage components are given as follows

$$\psi_{dm} = L_m (I_{ds} + I_{dr}) \tag{10}$$

$$\psi_{qm} = L_m (I_{qs} + I_{qr}) \tag{11}$$

and using (6) and (9) yields

$$\psi_{dm} = L_m I_{ds} \tag{12}$$

$$\psi_{qm} = I_{qs} (L_m - \frac{L_m^2}{L_r}) \tag{13}$$

The SCIG electric power loss is given in terms of d- and q-axis current components as follows [29]-[31]

$$P_{loss} = 3R_s (I_{ds}^2 + I_{qs}^2) + 3R_r (I_{dr}^2 + I_{qr}^2) + c_{Fe}\omega_e^2 \psi_m^2 + c_{str}\omega_e^2 (I_{dr}^2 + I_{qr}^2) \tag{14}$$

where c_{Fe} and c_{str} are the iron and stray loss coefficients, respectively. The mechanical and harmonic losses are not included in the above SCIG loss equation (14) because they are not controlled by the proposed method. However, harmonic loss is indirectly controlled and it is reduced through flux weakening. From (6), (9), (12) and (13) yields

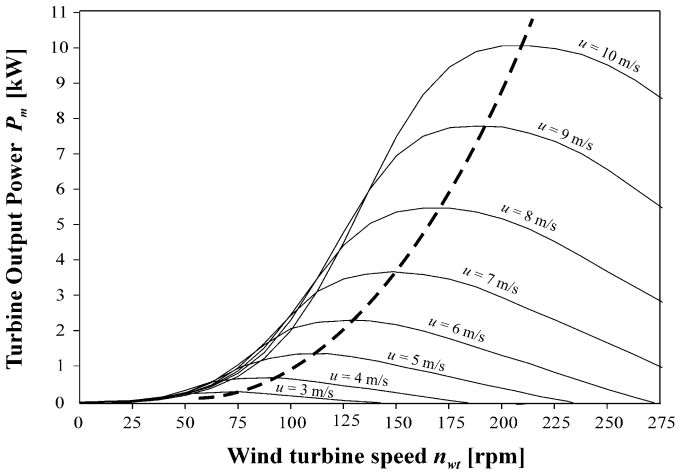


Fig. 3. Power-speed characteristics of a three-blade horizontal axis wind turbine, for various wind speeds and blade pitch angle at 0 degrees ($\beta = 0^0$).

$$P_{loss} = aI_{ds}^2 + bI_{qs}^2 \tag{15}$$

where

$$a = 3R_s + c_{Fe}\omega_e^2 L_m^2 \tag{16}$$

and

$$b = \left[3R_s + 3R_r \frac{L_m^2}{L_r^2} \right] + \left[c_{Fe}(L_m - \frac{L_m^2}{L_r})^2 + c_{str} \frac{L_m^2}{L_r^2} \right] \omega_e^2 \tag{17}$$

Due to decoupled control of the d- and q-axis stator currents, the electromagnetic torque expression is given by

$$T_e = \frac{3}{2} p \frac{L_m^2}{L_r} I_{qs} I_{ds} \tag{18}$$

IV. WECS Optimal Efficiency Conditions

The control objectives of a WECS include the following: a) electric loss minimization of the induction generator, b) maximum power extraction of the wind turbine, c) fast dynamic performance and d) cost-effective implementation of the control scheme.

The MEL controller regulates the flux-linkage through the I_{ds} current in order to minimize the electric loss of the SCIG. The MPPT controller adjusts the rotor speed of the SCIG through the I_{qs} current in order to maximize the wind turbine mechanical power. Both controllers operate simultaneously, so as optimal efficiency of the whole WECS is achieved.

Fast dynamic performance is achieved because the optimal I_{ds} and I_{qs} currents are determined through optimal conditions. Also, cost-effective implementation is provided because wind

speed measurement is not required and only the SCIG rotor speed is needed.

WECS operation is studied in two separate control modes. Specifically, SCIG loss minimization is investigated for constant electromagnetic torque T_e and speed ω_e operation while MPPT operation is examined for constant I_{ds} current operation.

A. MEL Controller

The electric loss minimization condition at steady state (T_e and ω_e constant), with respect to I_{ds} is given by

$$\left.\frac{\partial P_{loss}}{\partial I_{ds}}\right|_{\omega_e=\text{const.}} = 0 \tag{19}$$

Using (15), condition (19) is satisfied when

$$aI_{ds} + bI_{qs}\frac{\partial I_{qs}}{\partial I_{ds}} = 0 \tag{20}$$

Since the electromagnetic torque is constant with respect to I_{ds} current

$$\left.\frac{\partial T_e}{\partial I_{ds}}\right|_{\omega_e=\text{const.}} = 0 \tag{21}$$

and using (18) results

$$\frac{\partial I_{qs}}{\partial I_{ds}} = -\frac{I_{qs}}{I_{ds}} \tag{22}$$

Substituting (22) into (20), and using (16) and (17), the MEL condition is given by

$$I_{ds_{opt}} = \left|I_{qs}\right| G_d \sqrt{\frac{1+T_a^2\omega_e^2}{1+T_b^2\omega_e^2}} \tag{23}$$

where

$$G_d = \sqrt{1+\frac{R_r L_m^2}{R_s L_r^2}} \tag{24}$$

$$T_a = L_m \sqrt{\frac{c_{Fe}(L_r - L_m)^2 + c_{str}}{3(R_s L_r^2 + R_r L_m^2)}} \tag{25}$$

and

$$T_b = L_m \sqrt{\frac{c_{Fe}}{3R_s}} \tag{26}$$

In (23), the absolute value of I_{qs} current is used because the I_{qs} current is negative in the generator operation.

Condition (23) means that I_{ds} current should be supplied according to a torque demand (i.e. according to I_{qs} current). As speed increases, the iron loss increases. Therefore, iron loss can be reduced by reducing the field current through the factor $T_b\omega_e^2$, at the denominator of (23). If speed increases beyond a certain value, further reduction of field current would increase stray loss and iron loss due to leakage inductance. Such an effect can be avoided by keeping the field current I_{ds} constant through the factor $T_a\omega_e^2$ at the nominator of (23).

B. MPPT Controller

When maximum power harvesting from the incident wind is accomplished by the wind turbine, the maximum power coefficient $C_{p_{opt}}$ is achieved. Therefore, from (1) and (2), the maximum wind turbine torque for a given wind speed is given by

$$T_{wt_{opt}} = \frac{\rho\pi R^5}{2n^3\lambda_{opt}^3} C_{p_{opt}} \omega_{r_{opt}}^2 \tag{27}$$

The net mechanical torque, that results from $T_{wt_{opt}}$ after the subtraction of the mechanical loss, is equal to the SCIG electromagnetic torque and thus,

$$T_{e_{opt}} = T_{wt_{opt}} - T_{ml} \tag{28}$$

where T_{ml} is the mechanical loss torque at the speed-up side of the wind turbine due to gear box, bearings etc. Since the mechanical loss power is proportional to the third power of rotational speed [33], from (27), (28) and by using (2) yields

$$T_{e_{opt}} = \left[\frac{\rho\pi R^5}{2n^3\lambda_{opt}^3} C_{p_{opt}} - c_m\right] \omega_{r_{opt}}^2 \tag{29}$$

where c_m is the mechanical loss coefficient. Therefore, from (18) and (29), it is concluded that the MPP condition is given by

$$I_{qs_{opt}} = G_q \frac{\omega_{r_{opt}}^2}{I_{ds}} \tag{30}$$

where

$$G_q = L_r \frac{\rho\pi R^5 C_{p_{opt}} - 2c_m n^3 \lambda_{opt}^3}{3pL_m^2 n^3 \lambda_{opt}^3} \tag{31}$$

C. WECS Optimal Control

In order to achieve maximum efficiency of the whole WECS, both MEL and MPP conditions of (23) and (30), respectively, should be active simultaneously. Therefore, by combining (23) and (30), the optimal conditions that determine the *d*- and *q*-axis components of the SCIG stator current are given by

$$I_{ds_{opt}} = \omega_r \sqrt{G_d G_q} \left[\frac{1+T_a^2\omega_e^2}{1+T_b^2\omega_e^2}\right]^{1/4} \tag{32}$$

and

$$\left|I_{qs_{opt}}\right| = \omega_r \sqrt{\frac{G_q}{G_d}} \left[\frac{1+T_b^2\omega_e^2}{1+T_a^2\omega_e^2}\right]^{1/4} \quad (33)$$

V. Implementation of the Optimal Efficiency Control Scheme

The parameters of MEL and MPP conditions of (32) and (33), respectively can be determined experimentally by the following procedure.

Specifically, the MEL controller parameters can be adjusted with off-line laboratory experiments, as follows:

1) A wind turbine emulator is employed that it provides steady mechanical torque to the SCIG equal to 15% of its nominal value. The WESC output power to the grid is measured.
2) The SCIG rotates at low speed (about 10% - 15% of its nominal value) and in this case condition (23) becomes

$$I_{ds_{opt}} \approx \left|I_{qs}\right| G_d \quad (34)$$

The gain G_d is adjusted so that the maximum power to the grid is achieved.

3) The SCIG speed is increased to 40%-60% of its nominal value and under this case, condition (23) becomes

$$I_{ds_{opt}} \approx \left|I_{qs}\right| G_d \sqrt{\frac{1}{1+T_b^2\omega_e^2}} \quad (35)$$

The parameter T_b is adjusted so that the maximum power to the grid is attained.

4) The speed is increased to its nominal value and the parameter T_a is adjusted so that the maximum power to the grid is attained.
5) Steps 2 up to 4 are repeated until the desired accuracy is obtained.

An adaptive neuro-fuzzy logic-based control method can be adopted for the experimental identification of the MPPT controller parameter G_q at the real wind turbine, as that presented in [32]. Specifically, the SCIG is operated with the MEL controller as determined from condition (23) and for any instant wind speed, the MPPT is achieved by adjusting the I_{qs} current through the adaptive neuro-fuzzy logic control method. Due to degradation that may occur owing to aging of the mechanical system of the wind turbine, the above experimental procedure should be periodically repeated.

The L_m and L_r vary due to saturation while R_s and R_r vary due to temperature; hence, they may affect the WECS optimal controller gains and parameters. Specifically, gain G_q is proportional to the term (L_r/L_m^2), the value of which increases with saturation. As SCIG speed ω_r increases, saturation increases due to I_{ds} increase and consequently, gain G_q should be an increasing function of ω_r speed, as given by

$$G_q = G_{q1} + G_{q2}\omega_r \quad (36)$$

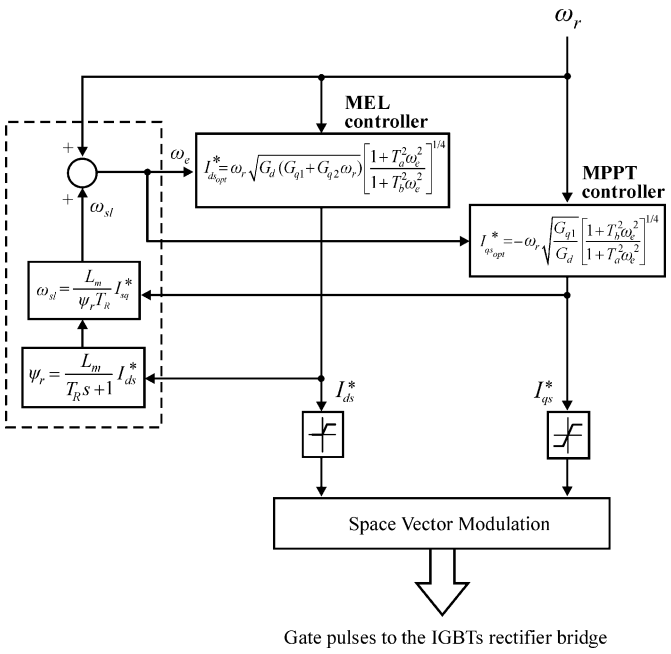


Fig. 4. Control algorithm for the rectifier converter of the optimal efficiency WECS.

By applying the aforementioned adaptive neuro-fuzzy logic based technique and recording the G_q and ω_r values at various wind conditions, the coefficients G_{q1} and G_{q2} of expression (36) are determined.

The gain G_d remains unaffected by saturation and temperature variations because a resistance and a magnetic inductance are both in nominator and denominator of (24) and therefore, their variation is neutralized. The parameters T_a and T_b are affected by temperature variation. However, since they vary in both nominator and denominator of (32) and (33), the variation of I_{ds} and I_{qs} with temperature is narrow. For similar reasons, T_a is unaffected by saturation. Contrarily, T_b varies with saturation and specifically it reduces as SCIG speed increase, because of I_{ds} increase. However, the variation of T_b with saturation is partially compensated by the increase of G_q, as given in (36).

From the above analysis and specifically from (32) and (36), it is concluded that the MEL condition is given by

$$I_{ds_{opt}} = \omega_r \sqrt{G_d (G_{q1} + G_{q2}\omega_r)} \left[\frac{1+T_a^2\omega_e^2}{1+T_b^2\omega_e^2}\right]^{1/4} \quad (37)$$

In condition (33), G_q increases with saturation whereas T_b reduces with saturation. Therefore, the variation of I_{qs} with saturation is narrow and consequently the MPP condition (33) is given as follows

$$\left|I_{qs_{opt}}\right| = \omega_r \sqrt{\frac{G_{q1}}{G_d}} \left[\frac{1+T_b^2\omega_e^2}{1+T_a^2\omega_e^2}\right]^{1/4} \quad (38)$$

Fig. 4 illustrates the block diagram of the proposed control algorithm for the rectifier of the optimal efficiency WECS.

TABLE I
3-PHASE, 10-kW, INDUCTION MACHINE AND OPTIMAL CONTROLLER PARAMETERS

V_s= 380 V (rms)	I_s= 21 A (rms)	
f_e= 50 Hz	$2p$= 4 (number of poles)	
R_s= 0.7 Ω	R_r= 1 Ω	
L_m= 0.2 H	L_{ls}= 0.01 H	L_{lr}= 0.01 H
G_d= 1.558	G_{q1}= 2.23·10^{-2}	G_{q2}= 2.9·10^{-4}
T_a= 2.51·10^{-3}	T_b= 1.82·10^{-2}	

The MEL controller determines the I_{ds} current through condition (37) and the MPPT controller determines the I_{qs} current through condition (38). Common input to the two controllers is only the generator rotor speed, while the measurement of the wind speed is not required. The stator frequency ω_e is obtained from the calculation of the slip frequency ω_{sl}. The MEL controller parameters are determined with off-line laboratory experiments and the MPPT controller parameters are determined with on-line experiments during the WECS operation.

It should be noted that the adaptive neuro-fuzzy logic technique is used only for the experimental identification of the MPPT controller parameters, because the MPPT controller is implemented through the condition (38).

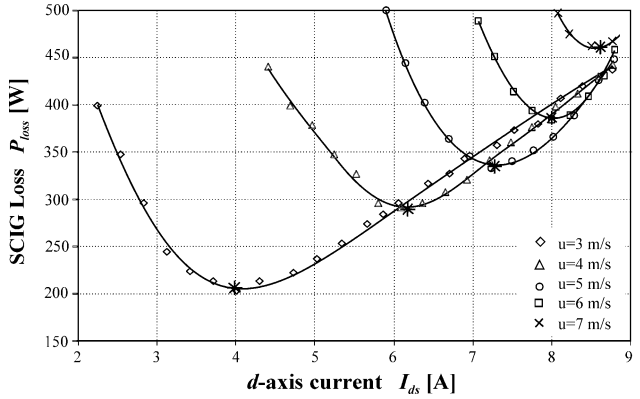
VI. PROPER ADAPTATION OF THE DEVELOPED OPTIMAL EFFICIENCY CONTROL METHOD TO URBAN TYPE WIND TURBINES

In order to properly adjust the developed optimal efficiency control method to the specific needs of an urban type WECS, the following procedure should be followed:

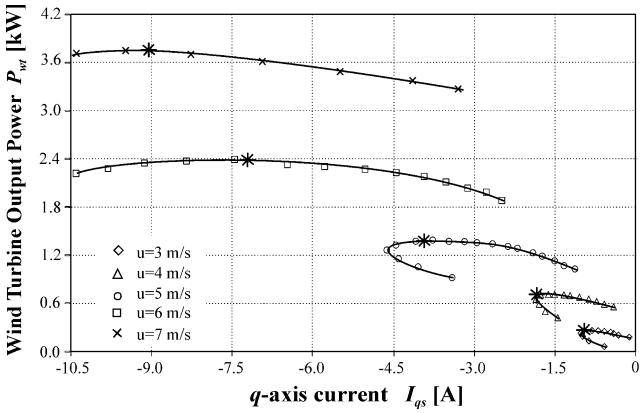
a) The experimental procedure of the Section V (steps 1-5) should be carried out, to determine the parameters of the two optimal controllers MEL and MPPT for the loss minimization of the SCIG and the maximum energy extraction by the wind turbine.
b) The adaptive neuro-fuzzy logic control procedure should be carried out to experimentally identify the MPPT controller parameter G_q.
c) The calibration procedure for the G_q parameter should be conducted, as determined by condition of (36).
d) Then, the optimal controllers MEL and MPPT are established by the conditions (37) and (38), respectively, and the optimal values of the I_d and I_q stator current components are online determined.

It is worth noting that in case of a vertical type wind turbine, there is need for yaw control. Contrarily, in the case of a horizontal axis wind turbine, the nacelle of the turbine should be properly controlled in order to correctly oriented to the direction of the wind. For the latter case, the improved active yaw control technique can be employed, in which the wind turbine is aligned to the wind direction through the error between the rotor speed determined by the optimal tip speed ratio and the real wind speed commanded by the maximum power point tracking control, as per [35].

As in the optimal efficiency control method, the aforementioned yaw control is cost effective since only changes in the yaw control software are required and the highly expensive remote sensing instruments based on laser and hypersonic technologies can be avoided. Therefore, the yaw control method of [35] is affordable for low power wind systems. Moreover, it is highly accurate, since the yaw misalignment is



(a)



(b)

Fig. 5. Variation of: (a) SCIG loss versus I_{ds} stator current for various wind speeds (at each wind speed, the generator rotor speed corresponds to maximum power of the wind turbine) and (b) wind turbine output power versus I_{qs} stator current of the SCIG for various wind speeds (I_{ds} stator current is equal to the nominal value).

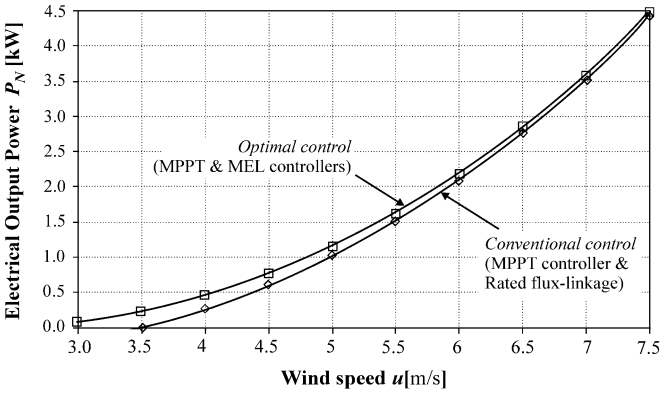
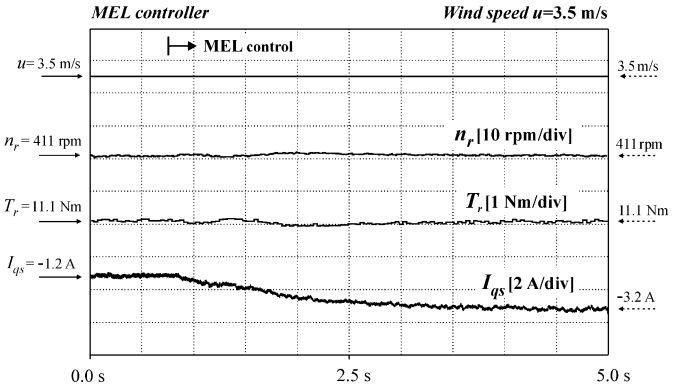
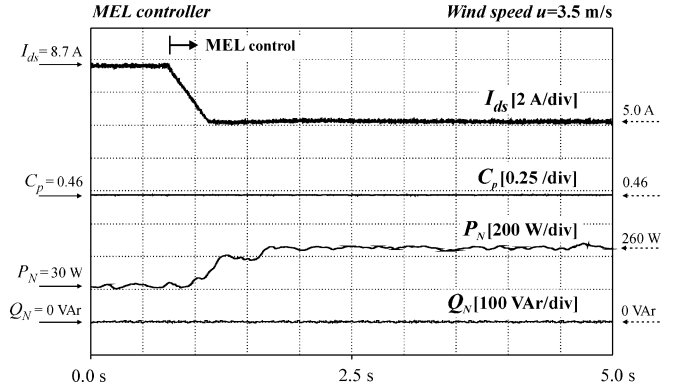


Fig. 6. Comparison between the WECS improved control and conventional control in electrical output power.

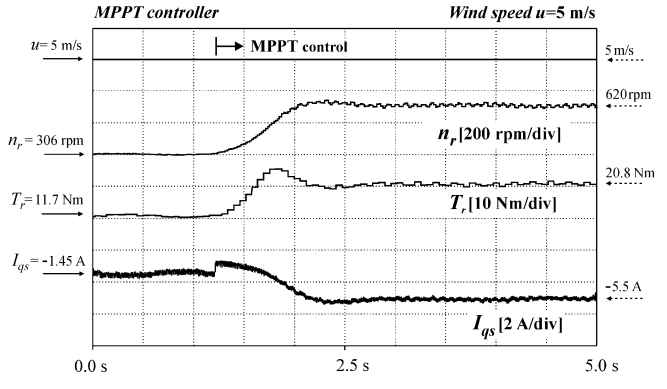


(a)

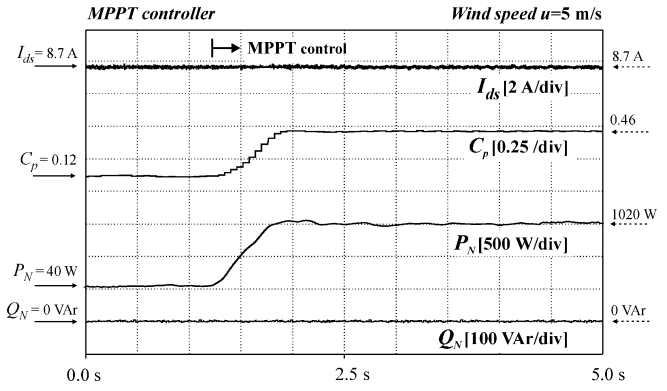


(b)

Fig. 7. WECS performance when only the MEL controller is active. The wind speed is constant at 3.5 m/s and the reference generator speed is $n_r^* = 411$ rpm that corresponds to optimal speed of the MPPT control.



(a)



(b)

Fig. 8. WECS performance when only the MPPT controller is active. The wind speed is constant at 5 m/s and the I_{ds} current is equal to the nominal value.

indirectly determined by the rotor speed error of the turbine and thus, it does not suffer from inaccuracies that are caused by the vortices in the wake of the flow downstream of the wind turbine, for the case that the wind direction is directly measured by a sensor installed at the rear end of the nacelle.

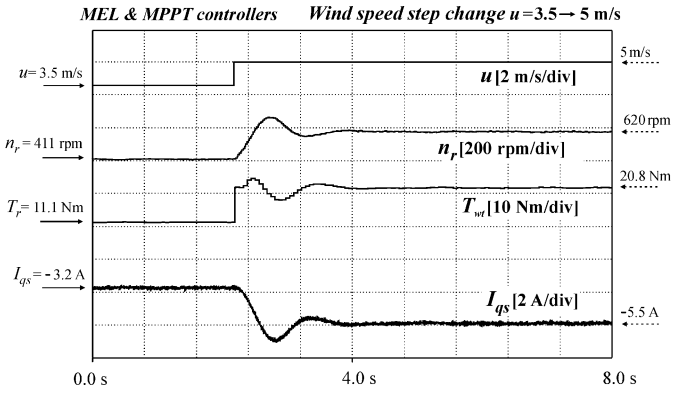
VII. Experimental Results

The parameters of the SCIG and the WECS optimal controllers are given in Table I. The gear ratio of the gear box is n=5.2. A gear box of ratio 1/5.2 is used in the wind-turbine emulator system. A capacitor of 2mF and a braking resistor of 85Ohm are used at the *dc*-link. The *LCL* filter between the inverter and the grid is composed of two series inductances of 1.8mH each and a capacitor of 15μF.

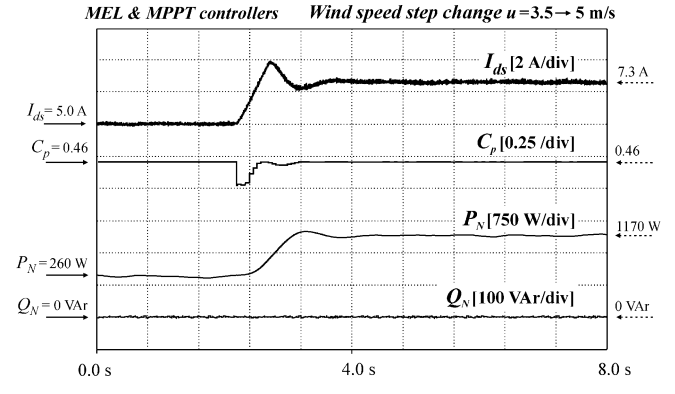
Two dSPACE DS1104 controller boards are used for the implementation of the rectifier and the inverter control schemes. The rectifier controller board houses the MPPT and the MEL controllers. The sampling periods of both MEL and MPPT controllers are 5ms. The inverter controller board houses a PI controller for regulating the I_{qN} current that maintains constant the *dc*-link voltage at 600V and a PI controller for regulating the I_{dN} current that controls the reactive power injected to the grid. Both PI controllers on the inverter side run at sampling periods of 5ms. For the experiments, a power factor of unity is considered and therefore, the inverter only injects active power into the grid. Thus, the inverter output power factor is slightly capacitive to compensate the filter reactive power.

The existence of exact values of I_{ds} and I_{qs} stator currents that provide SCIG minimum electric loss and maximum power of the wind turbine, for each wind speed, are experimentally verified in Fig. 5. Specifically, Fig. 5 (a) illustrates the variation of SCIG loss versus I_{ds} stator current, for various wind speeds. In this case, the I_{ds} current is manually regulated, and the generator speed corresponds to maximum power of the wind turbine (MPPT controller is active). As can be seen, for each wind speed, there is one specific value of I_{ds} current at which SCIG loss is minimized (all points noted by asterisk). Fig. 5 (b) illustrates the variation of wind turbine output power P_{wt} versus I_{qs} stator current of the SCIG, for various wind speeds. In this case, the I_{ds} current is equal to the nominal value. As can be seen, for each wind speed, there is one specific value of I_{qs} current at which wind turbine output power is maximized (all points noted by asterisk).

Fig. 6 validates the improvement of the optimal controlled WECS in the produced electrical power versus wind speed, for power factor equal to unity at the point of common cou-

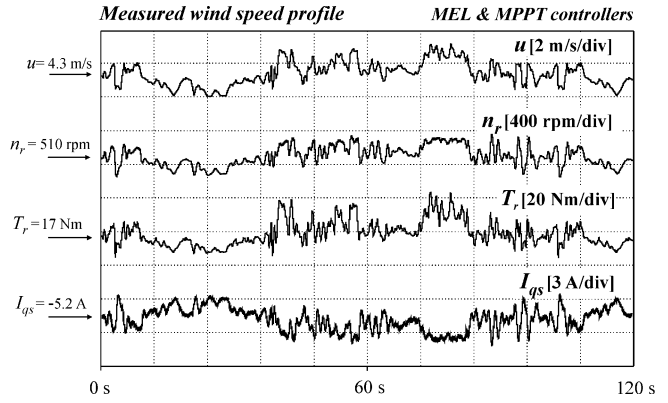


(a)

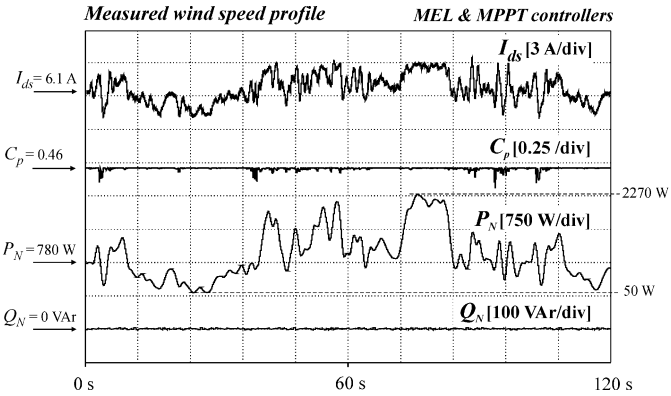


(b)

Fig. 9. Response of the optimal efficiency WECS to a step change of the wind speed (both MPPT and MEL controllers are in operation).



(a)



(b)

Fig. 10. Response of the optimal efficiency WECS to a real wind speed profile obtained by measurements (both MPPT and MEL controllers are in operation).

pling to the grid. The diagrams extended up to wind speed of 7.5 m/s because this region covers the most probable mean and average wind speeds [33]. Specifically, Fig. 6 compares the produced electrical power to the grid accomplished when both MEL and MPPT controllers operate, against that produced when only MPPT controller is active and the generator operates with nominal flux-linkage.

It can be seen in Fig. 6 that, the increase of the produced electrical power is higher at low wind speeds, since at this region the load power is low and therefore considerable reduction of generator loss can be accomplished by reducing the generator flux-linkage. Additionally, the exploitable wind speed region is expanded towards the lower range. Specifically, with the optimal control (both MEL and MPPT controllers in operation) the WECS starts to provide electrical power to the grid from a wind speed of 3.0 m/s, whereas with the conventional control (only MPPT in operation and generator under nominal flux-linkage) electrical power production is attained only above 3.5 m/s wind speed.

Figs. 7 and 8 examine the separate response of the MEL and the MPPT controllers. Specifically, Fig. 7 illustrates the WECS performance for wind speed 3.5 m/s and $\beta = 0^0$, when only the MEL controller is active. The generator reference speed is $n_r^* = 411\,\text{rpm}$ (or 411/5.2=79 rpm of the wind turbine) that corresponds to the optimal value of the MPPT controller. As can be seen, the MEL controller is very fast and finds the minimum SCIG loss operating point in less than 0.5 s. Due to the MEL control, additional electric power of 230 W can be provided to the grid and thus, the efficiency of the WECS is increased.

Fig. 8 illustrates the WECS performance for wind speed 5 m/s and $\beta = 0^0$, when only the MPPT controller is active. The I_{ds} current is equal to the nominal value. The generator initially rotates at a random speed value of 306 rpm (or 306/5.2=58.8 rpm of the wind turbine) and when the MPPT controller is activated, the generator reaches very fast the optimal speed of 620 rpm (or 620/5.2=119.2 rpm of the wind turbine) that corresponds to the maximum C_p coefficient of 0.46.

Fig. 9 illustrates the WECS response to an abrupt change of the wind speed from 3.5 to 5 m/s, when both MEL and MPPT controllers are in operation and $\beta = 0^0$. The wind speed values have been chosen on purpose, in order to verify the effectiveness of the proposed optimal controlled WECS in wind speed step changes. It can be seen that the proposed MEL and MPPT controllers can find very fast the optimal operating point at which minimum SCIG loss and maximum

power of the wind turbine are accomplished and thus, maximum efficiency of the whole WECS is achieved.

Fig. 10 verify the effectiveness of the proposed optimal efficiency WECS in real wind speed conditions. The wind speed ranges from 3m/s up to 6.6 m/s and therefore the WECS output power ranges from 50W up to 2270W. This large range can be justified by the large range in the wind speed variation of 3.3 m/s during the measured period of 120s (see the wind turbine power-wind speed characteristics of Fig. 3). Note that, Fig. 3 illustrates wind turbine power curves while Fig. 10 gives the WECS output power P_N (the active power P_N is the mechanical output power of the wind turbine subtracting the losses on the turbine, the generator and the converters).

The system operates with power factor equal to unity at the point of common coupling to the grid. It can be seen that, the two optimal controllers (MEL and MPPT) can follow the wind speed variations and they well cooperate in providing both minimum electric losses of the induction generator (through MEL controller) and maximum power harvesting of the wind turbine (through MPPT controller). This is achieved because the simultaneously MPPT and MEL operation is accomplished through conditions that can define the optimal operating point for each given wind speed.

VIII. Conclusions

In this paper, an optimal efficiency control strategy for WECS with SCIGs was presented. The developed control system provides minimum electric loss of the SCIG and maximum power harvesting of the wind turbine. Additionally, expansion of the exploitable wind speed region towards the lower speeds is achieved. The generator is connected to the power grid by means of two back-to-back PWM converters, both employing the space vector control technique.

The developed control technique has been based on previous research work of the author and it is properly adjusted so as, it can be utilized as generic method for any urban type wind tur-bine for smart energy buildings. Two optimal controllers have been introduced. The MEL controller adjusts the SCIG *d*-axis stator current according to torque conditions and accomplishes minimum electric loss of the SCIG. The MPPT controller regulates the SCIG *q*-axis stator current according to wind speed conditions and maximizes the wind turbine output power. The optimal *d*- and *q*-axis stator currents are online determined through optimal conditions. Therefore, fast dynamic response of the WECS is accomplished.

An experimental procedure is adopted for determining the parameters of the WECS optimal controllers. Also, the implementation of the suggested control scheme is cost-effective because common input to the two optimal controllers is only the SCIG rotor speed, while the wind speed measurement is not required. Selective experimental results on a WECS with emulated wind turbine have been presented in order to validate the resulting improvements of the proposed control scheme.

Acknowledgment

This work has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE (project code: T1EDK-00399).

References

[1] E. Hau, '*Wind Turbines, Fundamentals, Technologies, Application, Economics*', Berlin: Springer-Verlag, 2006.

[2] S. J. Chapman, '*Electric Machinery Fundamentals*', New York: McGraw-Hill, 1991.

[3] "Climate Finance for Cities and Buildings," Programme, United Nations Environment, UNEP"2014, (http://www.eldis.org/go/home&id=69822&type=Document#.WPzPjaIYm98).

[4] S. Sharma, B. K. Panigrahi and A. Verma, "A Smarter Method for Self-Sustainable Buildings: Using Multiagent Systems as an Effective Alternative for Managing Energy Operations," *IEEE Consumer Electronics Magazine*, vol. 7, no. 2, pp. 32-41, 2018.

[5] "Towards nearly zero-energy buildings Definition of common principles," European Commission, (Available at: https://ec.europa.eu/energy/sites/ener/files/documents/nzeb_full_report.pdf, 2013).

[6] "Synthesis Report on the National Plans for Nearly Zero Energy Buildings (NZEBs)," European Commission, (Available at: http://publications.jrc.ec.europa.eu/repository/bitstream/JRC97408/reqno_jrc97408_online%20nzeb%20report%281%29.pdf, 2016).

[7] B. Tofield and M. Ingham, "An EU Strategy for Energy Efficiency and Climate Action Led by Building Refurbishment", BUILDwithCaRE project financed by EU (Avaliable at: www.buildwithcare.eu).

[8] "Transition to sustainable building, Strategies and Opportunities to 2050", International Energy Agency, IEA, 2013 (Available at: https://www.iea.org/publications/freepublications/publication/Building2013_free.pdf).

[9] A. Mirecki, X. Roboam, and F. Richardeau, 'Architecture complexity and energy efficiency of small wind turbines', *IEEE Trans. Ind. Electron.*, vol. 54, no. 1, pp. 660-670, Feb. 2007.

[10] R. Datta and V. T. Ranganathan, 'A method of tracking the peak power points for a variable speed wind energy conversion system', *IEEE Trans. Energy Convers.*, vol. 18, no. 1, pp. 163-168, March 2003.

[11] R.M. Hilloowala and A.M. Sharaf, 'A rule-based fuzzy logic controller for a PWM inverter in a stand-alone wind energy conversion scheme', *IEEE Trans. Ind. Appl.*, vol. 32, no. 1, pp. 57-65, Jan./Feb. 1996.

[12] M. Pucci and M. Cirrincione, 'Neural MPPT control of generators with induction machines without speed sensors', *IEEE Trans. Ind. Electron.*, vol. 58, no. 1, pp. 37-47, Jan. 2011.

[13] V. Agarwal, R. K. Aggarwal, P. Patidar, and C. Patki, 'A novel scheme for rapid tracking of maximum power point in wind energy generation systems', *IEEE Trans. Energy Convers.*, vol. 25, no. 1, pp. 228-236, March 2010.

[14] V. Galdi, A. Piccolo, and P. Siano, 'Designing an adaptive fuzzy controller for maximum wind energy extraction', *IEEE Trans. Energy Convers.*, vol. 23, no. 2, pp. 559-569, June 2008.

[15] C.T. Pan and Y.L. Juan, 'A novel sensorless MPPT controller for a high-efficiency microscale wind power generation system', *IEEE Trans. Energy Convers.*, vol. 25, no. 1, pp. 207-216, March 2010.

[16] K. H. Kim, T. L. Van, D. C. Lee, S. H. Song, and E. H. Kim, 'Maximum output power tracking control in variable-speed wind turbine systems considering rotor inertial power', *IEEE Trans. Ind. Electron.*, vol. 60, no. 8, pp. 3207–3217, Aug. 2013.

[17] M. G. Simões, B. K. Bose, and R. J. Spiegel, 'Design and performance evaluation of a fuzzy-logic-based variable-speed wind generation system', *IEEE Trans. Ind. Appl.*, vo. 33, no. 4, pp. 956-965, July/Aug. 1997.

[18] A. Mesemanolis, C. Mademlis, and I. Kioskeridis, 'Maximum efficiency of a wind energy conversion system with a PM synchronous generator', in *Proc. MedPower 2010 Int. Conf.*, pp. 1-9.

[19] A. Mesemanolis, C. Mademlis, and I. Kioskeridis, 'Maximum Electrical Energy Production of a Variable Speed Wind Energy Conversion System, in *Proc. IEEE ISIE'2012*, pp. 1029-1034.

[20] A. Mesemanolis, C. Mademlis, and I. Kioskeridis, 'High-efficiency control for a wind energy conversion system with induction generator', *IEEE Trans. Energy Convers.*, vol. 27, no. 4, pp. 958-967, Dec. 2012.

[21] A. G. Abo-Khalil, H. G. Kim, D. C. Lee, and J. K. Seok, 'Maximum output power control of wind generation system considering los minimization machines', in *Proc. IEEE Int. Conf. IECON 2004*, pp. 1676-1681.

[22] A. G. Abo-Khalil, 'Model-based optimal efficiency control of induction generator for wind power systems', in *Proc. Conf. Rec. ICIT-2011*, pp. 191-197.

[23] J. Espí and J. Castelló, 'Wind turbine generation system with optimized dc-link design and control', *IEEE Trans. Ind. Electron.*, vol. 60, no. 3, pp. 919-929, March 2013.

[24] R Leidhold, G. Garcia, and M. I. Valla, 'Field-oriented controlled induction generator with loss minimization', *IEEE Trans. Ind. Electronics*, vol. 49, no. 1, pp. 147-156, Feb. 2002.

[25] S. Morimoto, H. Nakayama, M. Sanada, and Y. Takeda, 'Sensorless output maximization control for variable-speed wind generation system using IPMSG', *IEEE Trans. Ind. Appl.*, vol. 41, no. 1, pp. 60-67, Jan./Feb. 2005.

[26] A. Mesemanolis, C. Mademlis, and I. Kioskeridis, 'High-efficiency control for a wind energy conversion system with induction generator', *IEEE Trans. Energy Convers.*, vol. 27, no. 4, pp. 958-967, Dec. 2012.

[27] A. Mesemanolis, C. Mademlis, and I. Kioskeridis, "Optimal Efficiency Control Strategy in Wind En-ergy Conversion System with Induction Generator", *IEEE Trans on Power Electronics, Journal of Emerging and Selected Topics in Power Electronics*, vol. 1, no. 4, pp. 238-246, Dec. 2013.

[28] H. M. Kojabadi, L. Chang, and T. Boutot, 'Development of a novel wind turbine simulator for wind energy conversion systems using an inverter-controlled induction motor', *IEEE Trans. Energy Convers.*, vol. 19, no. 3, pp. 547-552, Sept. 2004.

[29] B. K. Bose, '*Power Electronics and Motor Drives*', Elsevier, Oxford: 2003.

[30] P.G. Cummings, W.D. Bowers, and W.J. Martiny, 'Induction motor efficiency test methods', *IEEE Trans. Ind. Appl.*, vol. IA-17, no. 3, pp. 253-265, May/June 1981.

[31] H. Li and R.S. Curiac, 'Energy conservation, Motor efficiency, efficiency tolerances, and the factors that influence them', *IEEE Industry Applications Magazine*, vol. 18, no. 1, pp. 62-68, Jan./Feb. 2012.

[32] A. Mesemanolis and C. Mademlis, 'A self-tuning maximum power point tracking control for wind generation systems', in *Proc. Conf. ICCEP-2013*, pp. 442-448.

[33] F. D. Bianchi, H. De Battista and R. J. Mantz, '*Wind Turbine Control Systems, Principles, Modelling and Gain Scheduling Design*', Springer-Verlag, London: 2007.

[34] N. Karakasis, A. Mesemanolis, and C. Mademlis, 'Wind Turbine Simulator for Laboratory Testing of a Wind Energy Conversion Drive Train', in *Proc. MedPower 2010 Int. Conf.*, pp. 1-6.

[35] N. Karakasis, A. Mesemanolis, T. Nalmpantis, and C. Mademlis, "Active Yaw Control in a Horizontal Axis Wind System without Requiring Wind Direction Measurement", *IET Renewable Power Generation*, vol. 10, no. 6, pp. 1441-1449, Oct. 2016.

[36] S. Heier, '*Grid Integration of Wind Energy Conversion Systems*', John Wiley & Sons, New York: 1998.

Propagation of Electromechanical Waves in Conventional Power Grids

Ruzica Cvetanovic, Filip Cvejic and Slobodan N. Vukosavic, *Senior member, IEEE*

***Abstract*—The paper reinstates the impact of distributed generation, consumption and accumulation on the dynamic response and transient stability of the electric power system. Modern power systems consist of distributed loads and power generation with both conventional synchronous generators and grid-side inverters. In comparison to the discrete model, the continuum model, discussed here, offers the opportunity to gain insight into the spatial aspects of transient stability and therefore has the potential to play a role of great importance in stabilizing the modern, distributed power system. In the paper, the phenomenon of electromechanical power waves in a string of conventional synchronous generators is considered and studied. Disturbance waves propagate considerably slower than electromagnetic waves and exhibit multiple reflections at the points of inhomogeneity. There are cases where the interference of direct, indirect and reflected waves can drive the system to instability. The paper considers the practical implementation of the simplified form of the previously proposed wave-quenching control law which is based on local measurements. The requirements point out that the practical implementation of the wave-quenching control law strongly relies on power electronic devices.**

***Index Terms*—Distributed power generation, Impedance matching, Power system dynamics, Power system stability.**

I. Introduction

The wave nature of the electromechanical disturbances in power systems was first considered almost 50 years ago [1], but it is only in the past decade that the research in this field has gained considerable attention. Previous work in this area comprises several different aspects: derivation of the continuum model [2], verification of the analytical model by the observations in actual power systems [3], control strategies for suppressing disturbances [4], [5], practical aspects for the implementation of the proposed control strategies [6]. Most recently, the continuum model was for the first time applied to networked inverters where basic modeling and control techniques of the electronic power waves were presented [7].

The modern power system involves distributed power generation as well as distributed loads, both controlled by power electronic converters. The discrete model described in [8], which assumes centralized representation of the loads and power generation and relies on the nodal description of the components and interconnections, is of no use anymore.

The main shortcoming of this model is its inability to take into account spatial aspects of the disturbance propagation. The correct modeling of the modern power system includes representation of both loads and sources as spatially distributed quantities. Moreover, the dynamics of the grid-side converters is completely different from the one of the traditional synchronous generators, as power electronic devices exhibit considerably faster transient phenomena [9]. Thus, in order to describe and work towards stabilizing the transient behavior of the modern inverter-rich power system, an alternative approach must be applied.

This paper provides recapitulation of the previous research in the study of electromechanical waves using systematic, methodical approach. Our main tool is the derived model for a one dimensional power system consisting only of conventional synchronous generators, which treats system to be continuum. According to the swing equation of the synchronous machine, this approach results in the power system being described by the second order differential equation. This equation corresponds in the form to the wave equation, thereby alluding to the wave nature of the disturbance propagation. In order to suppress multiple reflections at points of inhomogeneity, control strategies based on the impedance matching principle are proposed as in [5]. Practical implementation of the aforementioned wave-quenching compensation law implies digital control of the power at certain points of the grid and relies only on local, readily available measurements. Conclusions drawn from the continuum model are verified by the means of computer simulations of the discrete system. It is important to note that similar concepts can be applied to the string of inverters [9], as well as to the more realistic case of the two dimensional grid containing both grid-side inverters and traditional synchronous generators.

II. Wave Propagation on a String of Conventional Synchronous Generators

In order to gain better understanding of the general concept, we will introduce some approximations so that the use of complex mathematical tools is avoided. It can be shown that the principles derived for the simple model also hold true in the more realistic case [2]. The developments brought up in this section are largely based on those published in [1-5] and [7-9] and represent their summary.

Ružica Cvetanovic is with the School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail:ruzica996@ gmail.com).

Filip Cvejic is with the School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail:fcvejic@ gmail.com).

Slobodan N. Vukosavić is with the School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail:boban@ etf.bg.ac.rs).

A. Derivation of one dimensional wave equation

The wave equation will be derived following the procedure provided in [2], for a one dimensional power system comprising only traditional synchronous generators (Fig. 1) which are considered to be equidistant and interconnected with transmission lines of the same parameters (G [p.u] and B [p.u]) and lengths (Δ [m]). In each node there is a generator, with the inertia constant H [s] and damping D [s], providing variable power (P_m [p.u]) at the constant voltage (E [p.u]), thereby assuming uniform voltage magnitude along the string. The load is represented as the constant power sink (P_l [p.u]).

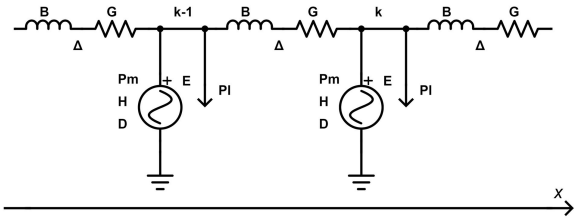


Fig. 1. A string of conventional synchronous generators.

Net power at each node is denoted as difference between generation and load ($P_g = P_m - P_1$). For the purpose of the following analysis we consider fluctuations of the frequency around the nominal value ω_s [rad/s] to be small enough, so that the small-signal stability analysis could be applied as in [4]. According to the swing equation and power balance equation for the k-th node, deviation of the k-th generator's mechanical angle δ_k [rad] is defined by:

$$\begin{aligned}\frac{2H}{\omega_s}\frac{\partial^2\delta_k}{\partial t^2} &= P_g - D\frac{\partial\delta_k}{\partial t}\\ &-E^2\left(B\left(\sin\left(\delta_k-\delta_{k+1}\right)-\sin\left(\delta_{k-1}-\delta_k\right)\right)\right)\\ &+E^2\left(G\left(\cos\left(\delta_k-\delta_{k+1}\right)+\cos\left(\delta_{k-1}-\delta_k\right)-2\right)\right)\end{aligned} \quad (1)$$

If we, as in [4], introduce per unit length parameters of the generators ($h=H/\Delta$ [s/m], $d=D/\Delta$ [s/m], $p_g=P_g/\Delta$ [p.u/m]) and transmission lines ($b=B\cdot\Delta$ [p.u·m], g=[p.u·m]), (1) results in:

$$\begin{aligned}\frac{2h}{\omega_s}\frac{\partial^2\delta_k}{\partial t^2} &= p_g - d\frac{\partial\delta_k}{\partial t}\\ &-\frac{E^2}{\Delta^2}\left(b\left(\sin\left(\delta_k-\delta_{k+1}\right)-\sin\left(\delta_{k-1}-\delta_k\right)\right)\right)\\ &+\frac{E^2}{\Delta^2}\left(g\left(\cos\left(\delta_k-\delta_{k+1}\right)+\cos\left(\delta_{k-1}-\delta_k\right)-2\right)\right)\end{aligned} \quad (2)$$

Focusing on the case with d=0 and g=0, for the small angular variations, equation (2) is well approximated with:

$$\frac{2h}{\omega_s}\frac{\partial^2\delta_k}{\partial t^2} = p_g - \frac{E^2\cdot b}{\Delta^2}\left(2\delta_k-\delta_{k-1}-\delta_{k+1}\right) \quad (3)$$

By taking the limit $\Delta\to 0$ of (3) and considering the definition of the spatial derivative of the rotor angle:

$$\lim_{\Delta\to 0}\frac{\delta_{k+1}-2\delta_k+\delta_{k-1}}{\Delta^2} = \frac{\partial^2\delta}{\partial x^2} \quad (4)$$

We arrive at the second order linear differential equation known as the wave equation [4]:

$$\frac{2\cdot h}{\omega_s}\cdot\frac{\partial^2\delta}{\partial t^2} = p_g + E^2\cdot b\cdot\frac{\partial^2\delta}{\partial x^2} \quad (5)$$

Notice that in the more realistic case of the two dimensional grid with distributed generator ($h(x,y)$ and $d(x,y)$) and transmission line parameters ($b(x,y)$ and $g(x,y)$), with the large angular disturbances where the approximation $sin(x)=x$ is not valid, an analogous derivation can be made. In this case derivation results in a second-order nonlinear partial differential equation [2].

Let $P(x,t)$ and $\omega(x,t)$ denote the deviation (from their nominal values) of the power flow and angular frequency in the direction of increasing x at the point x and the instant t. By taking the limit $\Delta\to 0$ of the power flow between discrete nodes, power flow in the continuum model is determined by:

$$P(x,t) = -E^2\cdot b\cdot\frac{\partial\delta}{\partial x} \quad (6)$$

Differentiation of (6) and substitution of the appropriate variables into (5) results in the following set of equations [5]:

$$\frac{\partial P}{\partial t} = -E^2\cdot b\cdot\frac{\partial\omega}{\partial x} \quad (7)$$

$$\frac{2\cdot h}{\omega_s}\cdot\frac{\partial\omega}{\partial t} = p_g - \frac{\partial P}{\partial x} \quad (8)$$

It is important to note that the acquired set of equations resembles the *Telegraph equations* which describe electromagnetic transient behavior on the transmission lines.

B. Traveling waves and their properties

The general solution of the derived wave equation (5) assumes the following form [2]:

(2)

$$\delta(x,t) = \delta\left(x-\frac{t}{v}\right)+\delta\left(x+\frac{t}{v}\right) \quad (9)$$

Where v is the velocity of the electromechanical waves defined by:

$$v = \sqrt{\frac{\omega_s\cdot b\cdot E^2}{2h}} \quad (10)$$

Hence, when the sudden disturbance occurs at some point of the considered string (due to the random fluctuations in the loads and generation levels), it propagates from that point along the string in the form of a traveling wave of power, frequency and angle. Note that the velocity of these waves depends on the per unit length values of the generator´s inertia constant h and transmission line susceptance b. The functional dependency is such that in the case when the generators are more densely packed (Δ decreases) or they are more capable of storing kinetic energy (H increases), the electromechanical waves propagate more slowly. Analogously, when the transmission lines are more prone to oppose power flow (b decreases), the speed of electromechanical waves decreases.

In order to acquire deeper understanding of the physics of the problem, it is useful to make a rough estimation of the electromechanical wave's propagation speed. Consider the previously analyzed string of generators assuming nominal line voltage to be 220kV (50Hz), with the base power of 100MVA (resulting in the base impedance of 484 Ω) and transmission line reactance of 0.4 Ω/km. Let voltage be equal to its nominal value (E=1p.u). The distance between each two generator is considered to be $\Delta = 120km$ and each generators' inertia constant H=10s (with respect to the given base power). The substitution of all of the above in (10) results in the velocity of approximately 1500km/s. Authors in [5] and [6] state that the values of the velocities observed in the real power systems by the use of synchrophasor measurements and Wide Area Measurement System ranges from 500km/s to 1000km/s. The deviation of the estimated value from the observed range is due to the fact that the real power system is a two dimensional grid with inhomogeneous distribution of inertia constants and transmission line parameters. In addition to this, real power systems contain centralized (as well as distributed) generation and load, which have an impact on the properties of the electromechanical waves [6]. However, the order of magnitude of our rough estimation is in accordance with the observed range.

As the wave propagates along the string, the ratio between the forward travelling waves of frequency $\omega^+$and of power flow P^+ stays constant [4]. This enables us to define a characteristic impedance through the procedure given in [4]:

$$Z_0 = \frac{\omega^+}{P^+} = \sqrt{\frac{\omega_s}{2h}\frac{1}{E^2 \cdot b}} \tag{11}$$

Note that, taking into account that (7) and (8) are analogus to the *Telegraph equations*, Z_0 [rad/sW] is analogus to the ohmic impedance defined as a ratio between voltage and current waves in electromagnetic waves. It will be demonstrated in the subsequent sections that the characteristic impedance plays a crucial role in determining a compensation law used to enhance power system transient stability [5].

III. Multiple Reflections at Points of Inhomogeneity

When the forward travelling (incident) wave reaches the point of inhomogeneity of the grid, it splits into two components: the reflected and transmitted waves (absorption is neglected here). The ratio between those two components depends on the boundary condition at the point of inhomogeneity that needs to be satisfied. Unless we act in a proper manner that would diminish the negative effects of the reflections or preferably completely prevent them from happening, multiple reflections of the electromechanical waves can cause the severe problems and endanger the stability of the power system. Without the lack of generality, in the proceeding analysis we will examine the end of the string as an example of the aforementioned inhomogeneity. Waves exhibit similar behavior in all of the other cases (the change in the transmission lines' or generators' parameters, DC/AC interconnections, etc). Developments brought up in this section are largely based on those published in [1-5] and [7-9] and represent their summary.

A. Problem description

In order to quantify and mathematically describe reflections at the point of inhomogeneity, the reflection coefficient of the frequency wave R is introduced, defined as the ratio between magnitude of the reflected (ω^-) and incident (ω^+) frequency wave [4]. Note that an analogous definition can be made for the power wave. If we assume that it is possible to terminate a string in such a way that the ratio between the frequency variation at the end of the string $(\hat{\omega}_{end})$ and the power flow variation at the end of the string $(\hat{P}_{end})$ is held in constant proportion, we can express the impedance that terminates the string as:

$$Z = \frac{\hat{\omega}_{end}}{\hat{P}_{end}} = \frac{\omega^+ + \omega^-}{P^+ + P^-} \tag{12}$$

Where P^+ and P^- respectively denotes incident and reflected power flow waves [4]. According to (11), (12), and the fact that $\frac{\omega^+}{P^+} = -\frac{\omega^-}{P^-}$, the definition of the reflection coefficient introduced above results in:

$$R = \frac{\omega^-}{\omega^\mp} = \frac{Z - Z_0}{Z + Z_0} \tag{13}$$

We will discuss two extremes, as far as string endings go: an infinite bus ($Z = 0$) and an open end ($Z \to \infty$). In the first case, the frequency variation at the end of the string is equal to zero, meaning that the incident and reflected waves of the frequency must cancel out at the end of the string ($R = -1$). In the case of an open end, the power flow at the end of the string is equal to zero, meaning that incident and reflected waves of the power flow at the end of the string must cancel out ($R = 1$). In this case the reflected wave resembles the incident wave. Thus, after the reflection the wave that travels along the string in the opposite direction may have the magnitude that is twice the magnitude of the incident wave (total reflection occurs). As the frequency wave represents variations of the rotors' angular velocities from their nominal values, this leads to the unwanted acceleration of certain

generators and can potentially trigger the loss of synchronism. Therefore, from the transient stability aspect, the case of an open end is considered to be the most important one to examine.

B. Problem solution

Conventional frequency control strategies consist of the primary (Governor/Droop control), secondary (Automatic Generation Control) and tertiary (Economic dispatch) frequency control. This concept stems from the discrete model which was traditionally used to examine power system stability of the centralized power system comprising only classical synchronous generators and AC transmission lines. Conventional control strategies preserve grid from losing its synchronism with a certain stability margin. However, when this margin is reached the system becomes unstable.

The preceding analysis points to the underlying physics behind the stability issue. Namely, disturbances travel as electromechanical waves and reflect at nodes of inhomogeneity. These reflections could accumulate to the point where the stability margin is exceeded, even though conventional control algorithms are designed such that this is a rare occurrence. The chances of this happening are considerably increasing with the escalating distribution of generation and load which is present in modern power system. In addition to this, the modern power system is mixed, meaning that it contains both conventional synchronous generators and renewable energy sources connected to the grid through power electronic devices. In such an inverter-rich power system electronic power waves also exist [7]. They are considerably faster than the electromechanical waves, thus making the stabilization process based on the conventional control strategies even more difficult. Moreover, modern power systems comprise both DC and AC transmission lines which increases the number of points of inhomogeneity. Thus, in order to ensure the conditions for the modern power system to function normally, it makes sense to develop an alternative way of stabilizing the system which would consider wave nature of the disturbances.

In the subsequent analysis an open ended string is considered, as this case was previously shown to be the worst one in terms of transient stability. Following (13) one can infer that if the string was terminated with the impedance that is equal to its characteristic impedance:

$$Z = \frac{\hat{\omega}_{end}}{\hat{P}_{end}} = Z_0 \tag{14}$$

The reflection coefficient of the frequency wave would be equal to zero [4]. This means that the reflections would be eliminated provided that the ratio between frequency and power flow deviations at the end of the string is held constant. Thus, (14) is the control objective we strive towards.

In order to determine the appropriate control action that would enable the implementation of the desired compensation law, let us focus on the swing equation for the last node [4]:

$$\frac{2 \cdot H}{\omega_s} \cdot \frac{\partial \hat{\omega}_{end}}{\partial t} = P_g + P_{end} \tag{15}$$

Where P_{end} represents the total power flow of the last node and is equal to the sum of the steady state power flow $\bar{P}_{end}$ and its deviation due to transients ($P_{end} = \bar{P}_{end} + \hat{P}_{end}$). Analogously, P_g represents the total net power of the last node and is equal to the sum of the steady state net power $\bar{P}_g$ and its variation due to transients ($P_g = \bar{P}_g + \hat{P}_g$). The steady state power balance for the last node results in $\bar{P}_{end} = -\bar{P}_g$.

The following analysis considers $\hat{P}_g$ as the control variable [5]. If we, according to (14), express $\hat{P}_{end}$ as a function of $\hat{\omega}_{end}$ and Z_0, and along with the statements given above, substitute it in (15), we arrive at the required control action:

$$\hat{P}_g = \frac{2 \cdot H \cdot Z_0}{\omega_s} \cdot \frac{\partial \hat{P}_{end}}{\partial t} - \hat{P}_{end} \tag{16}$$

However, this control action requires having information of the first time derivative of the variable $\hat{P}_{end}$, which turns out to be a problem because this quantity cannot be easily measured. Alternatively, this information can be obtained from (8) but it would require non-local measurements which we want to avoid, due to potential delays in communication and difficulties in information exchange. Our aim is to derive wave-quenching control action that would rely only on local measurements which are already available from the Phasor Measurement Units. With regard to the considerations explained above, the following compensation law is proposed:

$$\hat{P}_g = -\frac{\hat{\omega}_{end}}{Z_0} \tag{17}$$

Which provides an approximate tracking of the desired control objective and is a simplified version of the controller proposed in [5]. Even though it does not enable us to instantaneously achieve the impedance matching objective, it successfully suppresses reflections as it will be demonstrated in the next section by the means of computer simulations. The proposed wave-quenching compensator acts in the following way: if the frequency deviation at the end of the string is positive, compensator takes power from the grid that is proportional to this deviation; if the frequency deviation is negative it injects power in the grid that is proportional to this deviation. In both situations, the proportionality constant is equal to the inverse of the string's characteristic impedance.

The proposed compensation law implies digital control of the power at the end of the grid, or in the general case, wherever there exists a risk of reflections occurring. It is important to note that it relies only on local, readily available measurements.

IV. Simulation Results

In order to examine the validity of the derived continuum model, a string comprising a total of 64 synchronous generators is modeled and simulated in MATLAB. The model is implemented as a linked list of blocks of generators [2]. The number of the elements in the list was chosen such that it is large enough to illustrate the phenomena of interest, but small enough to keep simulation demands at a reasonable level. Each of the 64 blocks is formed with respect to (1), by introducing per unit time according to the following base time constant:

$$T_b = \sqrt{\frac{2 \cdot H}{\omega_s}} \tag{18}$$

This yields the equation that describes generator dynamics in our Simulink model:

$$\begin{aligned} \frac{\partial^2 \delta_k}{\partial T^2} &= P_g - D\frac{\partial \delta_k}{\partial T} \\ &-E^2\left(B\left(\sin\left(\delta_k - \delta_{k+1}\right) - \sin\left(\delta_{k-1} - \delta_k\right)\right)\right) \\ &+E^2\left(G\left(\cos\left(\delta_k - \delta_{k+1}\right) + \cos\left(\delta_{k-1} - \delta_k\right) - 2\right)\right) \end{aligned} \tag{19}$$

Where T is per unit time ($T=t/T_b$). Note that the swing equation for the last node (15), used for derivation of the desired compensation law, could be derived from (19) simply by neglecting D, substituting $T=t/T_b$ into (19) and expressing the sum of the second and third term on the right side of (19) as P_{end}.

In comparison to the model of the generator we used for analytical derivation of the wave equation, for the purpose of simulation, there is no need to neglect transmission line conductance and generator damping, nor to linearize the swing equation in terms of applying the approximation $sin(x)=x$. Note that in simulation tests only the generator's electromechanical dynamics is considered (exciter and turbine dynamics are neglected).

The disturbance is considered to be a sudden increase in the 7th generator's net power which is modeled as a power pulse with both magnitude and duration of 1p.u. The parameters of the generators and transmission lines used in the simulations correspond to the ones used to estimate the velocity of electromechanical waves: $\Delta = 120km,\ B = 10p.u,\ G = 1p.u,\ E = 1\,p.u,\ D = 0.01\,p.u,\ \omega_s = 2 \cdot \pi \cdot 50\ rad/s,\ H = 10s$. The base time is therefore 225ms. It is also assumed that the net power of each node is equal to zero ($P_g = 0$).

Simulation results shown in Fig. 2 represent the change of the rotor mechanical angle at the connection points of corresponding generators in the case of an open ended string of generators [5]. The waveforms are intentionally scale-shifted in the vertical direction, so that the properties of the travelling electromechanical wave could be discussed.

It takes roughly $22T_b$=4.95s for the wave to reach the end of the string. With the string length of 120·64=7680km, the speed of the electromechanical waves is found to be 1552km/s. Note that such a long string is unrealistic for the Serbian power system, but can be apllied to North American power system. Speed estimation given above is in agreement with the estimation based on (10). Thus, the validity of the derived continuum model is verified. After the wave reaches the end of the string, it gets reflected and travels back thereby threating the system to lose stability. The reflection process is repeated whenever a wave reaches an open end. In Fig. 2, multiple reflections are clearly noticeable.

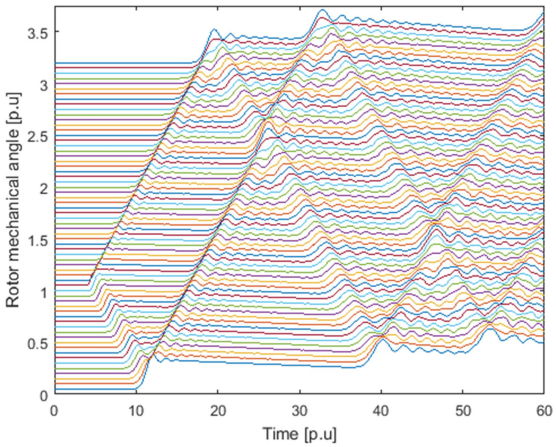


Fig. 2. Disturbance propagation in an open-ended string of synchronous generators without wave-quenching compensator.

The simulation results given in Fig. 3 are obtained by adding the wave-quenching compensator. The compensation law based on (17) is applied at the right end of the string. The reflection at the right end of the string is successfully suppressed even though an approximate compensation law is implemented [4]. It is important to note that the compensation law works extremely well although it was tested on the nonlinear model with both generator damping and transmission line conductance assuming non-zero values. Thus, results given in Fig. 3 attest to the robust nature of the proposed compensation law [5].

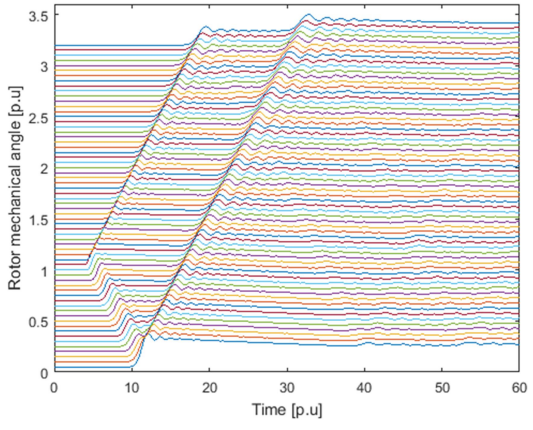


Fig. 3. Disturbance propagation in an open-ended string of synchronous generators with wave-quenching compensation.

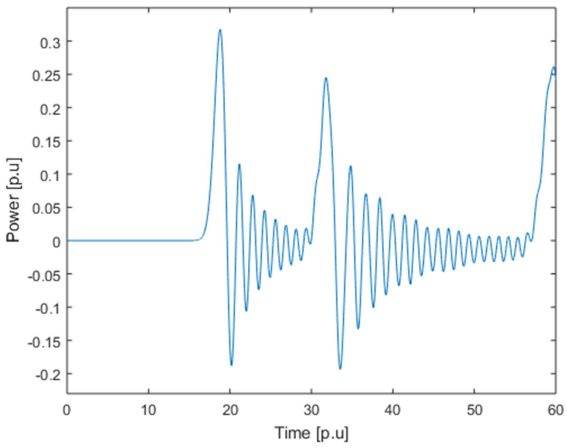


Fig. 4. The power drawn from the wave-quenching compensator at the end of the string, obtained simultaneously with the results plotted in Fig. 3.

The wave-quenching compensator has to supply the transient power pulses illustrated in Fig. 4. Note that the rate of change of the power during this transient is considerably larger than what a conventional synchronous generator is capable of. The limiting factor is the physical nature of exciter and turbine regulation. Thus, a practical implementation of this wave-quenching compensator would require a power electronics device and a local form of energy storage. Batteries are not fast enough for this application. Capacitors do not have enough capacity to store the required amount of energy. Therefore supercapacitors are found to be an adequate solution for the local form of energy storage. The above mentioned power electronics devices are mostly designed as modular multilevel converter topologies (MMC) because they have to operate at high voltages. Instead of a single switch, several H-Bridge cells are connected in series to form a chain [9]. The converter consists of six of these chains in a three-phase inverter configuration (Fig. 5).

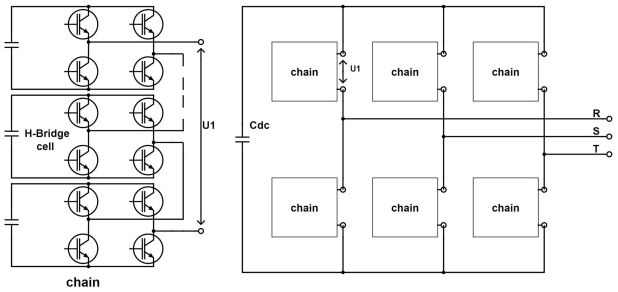


Fig. 5. Modular multilevel converter topology.

V. Conclusions

The main issues addressed in this paper are techniques for modelling and stabilizing power systems with spatially distributed sources and loads. A one dimensional string of conventional synchronous generators is considered, and it is determined that the electromechanical power waves exist. In order to examine validity of the derived continuum model and verify the conclusions drawn from such a model, computer simulations of the discrete system are implemented. It is noticed that, at the points of inhomogeneity, multiple reflections take place. They endanger power system stability, and therefore must be suppressed. A simplified version of the previously proposed wave-quenching compensation law, which is based only on local, readily available measurements is discussed and tested by the means of computer simulations.

Due to the fast transient response of power and the large amount of energy that the wave-quenching compensator has to supply during these transients, the practical implementation of the proposed control law implies the use of power electronics devices for the actuators and supercapacitors for the energy storage. Taking into account the voltage level at which those actuators have to operate, it is most likely that they would be realized as modular multilevel converters.

Further studies on the wave nature of the disturbance propagation in the power system include: a more detailed description of the conventional power system, simulations of distributed feedback control strategies, the implementation of the continuum modelling techniques on networked inverters as well as on a mixed grid, containing both traditional synchronous generators and grid-side inverters.

References

[1] A.Semlyen, "Analysis of disturbance propagation in power systems based on a homogenous dynamic model," *IEEE Trans. Power App. Syst*, 1974.

[2] J. S. Thorp, C. E. Seyler, and A. G. Phadke, "Electromechanical wave propagation in large electric power systems," *IEEE Trans. Circuits Syst.I, Fundam.Theory Appl.*, vol. 45, no. 6, pp. 614-622, June 1988.

[3] R. L. Cresap and J. F. Haue, "Emergence of a new swing mode in the western power system," *IEEE Trans.Power App. Syst.*, 1981.

[4] B. C. Lesieutre, E. Scholtz, and G. C. Verghese, "A zero-reflection controller for electromechanical disturbances in power networks," in *Proc. 14th Power Syst. Comput. Conf*, Sevilla,Spain, June, 2002, pp. 1-7.

[5] B. C. Lesieutre, E. Scholtz, and G.C. Verghese, "Impedance Matching Controllers to Extinguish Electromechanical Waves in Power Networks," in *IEEE International Conference on Control Applications*, Glasgow, Scotland,U.K, September, 2002.

[6] G. Ledwich T. Li, Y.Mishra, and and A. Vahidnia J. Chow, "“Wave Aspect of Power System Transient Stability-Part I:Finite Approximation," *IEEE Trans. Power Systems*, 2017.

[7] S.N. Vukosavic and A.M. Stankovic, "Electronic Power Waves in Network of Invereters," in *2018 North American Power Symposium (NAPS)*, Fargo, ND, USA, 9-11 Sept. 2018, pp. 1-6, DOI: 10.1109/NAPS.2018.8600614.

[8] Peter W. Sauer, M. A. Pai, and Joe H. Chow, *Power System Dynamics and Stability*.: Wiley-IEEE, 2018.

[9] S.N. Vukosavic, *Grid Side Converters-Design and Control*.: Springer, 2018.

A Fault-Tolerant DC UPS System Based on a Battery Charger with an Automatic Load Transfer Function

Vladimir Dj. Vukić

***Abstract*—A direct current uninterruptible power supply (DC UPS) system, designed for operation in a hydro power plant, is presented in this paper. Standard DC power supply configurations for large power plants consist of two AC/DC converters with accompanying storage batteries. To improve the system's reliability, an active fault tolerant control system (FCTS) was utilised. Aside from the two primary AC/DC converters, a third device was implemented with the same nominal characteristics, albeit with some additional advanced functions. This secondary AC/DC converter has two high-power contactors on its output, enabling it to operate (either manually or automatically) in parallel with one of the primary rectifiers. In order to achieve the high reliability of this facility, a serial communication was not established between battery chargers. Instead, a procedure for the turnout detection of a remote battery charger was utilised, based only on data from the auxiliary contacts of its main switch and the input contactor. Based on these logic conditions, an algorithm for the automatic load transfer was devised for implementation in cases when some of the inoperative primary battery chargers would automatically be replaced by a secondary AC/DC converter operating in a "hot reserve".**

***Index Terms*—AC/DC converter, Fault-tolerant control system (FTCS), Uninterruptible power supply (UPS), Algorithm.**

I. Introduction

IN previous years, the area of static uninterruptible power supply systems (UPS) demonstrated intensive technical progress [1]-[2], especially due to a reduction in power converters' weight and size [3]. Furthermore, there were increased demands for highly reliable power converters [4]-[5]. One way to improve the reliability of power supply is to implement fault-tolerant control systems (FTCS) capable of continually operating after a failure has occurred. Some authors divide these systems into passive (PFTCS) and active fault-tolerant control systems (AFTCS) [6]. Usually, passive FTCS (PFTCS) means ordinary redundant systems, such as modular power converters in N+1 configurations (with at least one spare module), enabling the power converter to continue its operations with nominal load even in the case of module failures. Therefore, modular construction with reserve modules is a frequent topology for modern power converters in industrial UPS systems. Nonetheless, often PFTCS are not even considered to be true fault-tolerant systems, since it is not possible to isolate the failure or continue device operations with modified characteristics [6]. Therefore, if a customer has high demands for the secure operation of their power plant, among the first requests would be highly reliable UPS devices, particularly in the case of DC voltage. Accordingly, the implementation of redundant batteries and active fault-tolerant AC/DC converters would be strongly justified.

Another important issue is to realise a simple and reliable state diagnostic of a primary ("master") battery charger. If such a diagnostic relies on a serial communication, it would depend on the regular operation of a battery charger control unit. Nevertheless, in the case of a battery charger control unit fault, or the loss of its power supply voltage, a serial communication would simultaneously fail. Consequently, it would be preferable to develop some mode of diagnostics that could solely rely on a secondary ("slave") battery charger, not depending on the correctness of a primary charger operation. Nonetheless, such a method, with few available signals, should be capable of reliably detecting the state of the monitored device, avoiding both false alarms and oversight of real faults.

In this paper, an active fault-tolerant DC UPS system was presented, implemented in the "Djerdap 2" hydro power plant. A detailed description was given of a method for fault diagnostics in primary battery chargers, as well as a utilised automatic transfer function.

II. Battery Chargers

A. DRI-PT Series of Thyristor AC/DC Converters

Owing to the implemented design, based on a programmable logic controller (PLC) and a touch-sensitive human machine interface (HMI), this specific kind of AC/DC converter distinguishes itself among other thyristor rectifiers, made in the second half of the previous decade. The devices of the DRI-PT series are rectifiers developed in the "Nikola Tesla" Electrical Engineering Institute, designed to be implemented as battery chargers in UPS systems [7]-[9]. Seven thyristor rectifiers were produced in three lots (based on mutually significantly different designs) and commissioned in three hydro power plants (HPP) in Serbia: "Djerdap 2" [7], "Bajina Bašta" [8] and "Djerdap 1" [9].

All of these three types of AC/DC converters had different circuit topologies to achieve fault-tolerant operations. The original devices (DRI 220-70PT) were configured to operate

Vladimir Dj. Vukić is with University of Belgrade, Electrical Engineering Institute "Nikola Tesla", Koste Glavinića 8a, PO Box 139, 11000 Belgrade, Serbia (e-mail: vvukic@ieent.org).

as two battery chargers with accompanying storage batteries, keeping a third battery charger, without its own battery, in a "hot reserve" [7].

The second system, comprised of two rectifiers DRI 48-50PT, was commissioned in the "Bajina Bašta" HPP [8]. This was preferably PFTCS, with a single battery charger and redundant DC load power supply. The thyristor rectifier was expanded with 100% redundant DC/DC converters in its output stage, supplying critical loads with a nominal voltage of 48 V for numerous possible variations in battery voltage (between 36 V and 75 V) [8].

The third system was the most complex of all, forming a conjoint facility of two AC/DC converters (DRI 220-160PTD) and a DC distribution board with two mutually independent busbar systems, designed for "Djerdap 1" HPP [9]. The DRI 220-160PTD line-frequency phase-controlled rectifiers have a two-channel control system, a primary from a PLC microprocessor unit, and a secondary from an analogue PI controller, activating in cases of PLC failure [9].

B. Construction of DRI 220-70PT AC/DC Converters

DRI 220-70PT Rectifiers (Fig. 1) are line-frequency self-commutated power converters, with a thyristor half-bridge serving as an actuator, designed for nominal output parameters of 220 V and 70 A. These AC/DC converters were foreseen for use in lead-acid or nickel-cadmium storage battery charging, as well as the power supply of a critical DC load (either with or without a connected storage battery) in an additional facility of the "Djerdap 2" HPP (situated on the border between Serbia and Romania, on the river Danube). There are "master" (DRI 220-70PTM) and "slave" (DRI 220-70PTS) varieties of these AC/DC converters, with the "slave" type being somewhat more complex (see Fig. 2). Three DRI 220-70PT rectifiers, together with two lead-acid storage batteries, represent the UPS system for a load operating with 220 V DC safety voltage in the power plant (presented in Fig. 3).

Fig. 1. DRI 220-70PTM1 (left), DRI 220-70PTS (in the middle) and DRI 220-70PTM2 (right) AC/DC converters in an additional facility of the "Djerdap 2" HPP.

The control unit is primarily based on the "Omron" CJ1-series PLC. Analogue circuits and drivers, specific to the line-frequency self-commutated rectifiers, were separately consolidated on the "DIGISP 06" motherboard. Occupying the role of the human-machine interface (HMI) was the "Omron" NS5 touch panel, replacing all the push-buttons, instruments and switches (except the main switch) on the AC/DC converter's cabinet front door. Above the HMI, the 20-diode LED panel was mounted.

Fig. 2. Interior appearance of the DRI 220-70PTM1 (left) and DRI 220-70PTS (right) AC/DC converters. Note the two output contactors in the upper back field of the "slave" device.

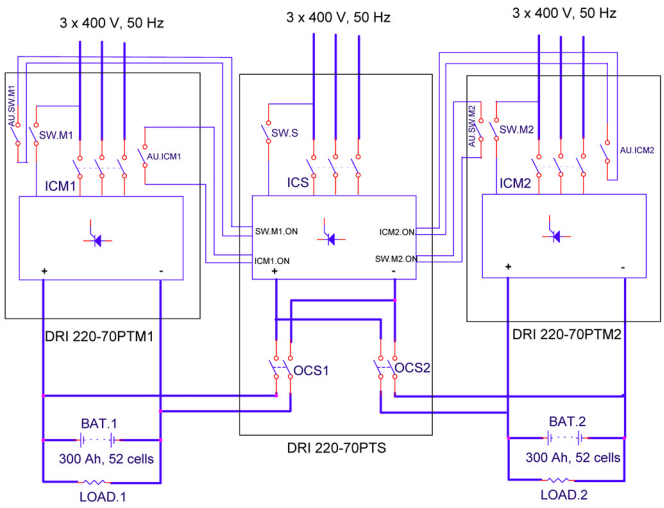


Fig. 3. Simplified schematic circuit diagram of the DC UPS system installed in the "Djerdap 2" HPP. While every battery charger has its input contactor (IC), the output contactors (OCS1 and OCS2) are only part of the DRI 220-70PTS "slave" battery charger. Note the auxiliary contacts of the main switches (AU.SW.M1 and AU.SW.M2) and input contactors (AU.ICM1 and AU.ICM2) of two "master" rectifiers.

When an input contactor is turned off, the power circuit of a battery charger is isolated from the mains grid voltage. As soon as the main switch (on the front door of a battery charger cabinet; see Fig. 1) is turned on, the input contactor activates, connecting a power circuit to the mains grid voltage (3 x 400 V, 50 Hz). Nonetheless, the battery charger is only turned

on after the activation of the virtual "START" taster on an HMI panel.

At an HMI panel, user can make wide variations of the preset parameters, such as the modes of operation, references of output voltage and current, thresholds of overvoltage and undervoltage protections, load transfer options, as well as parameters of PI controllers. A variation of the reference voltage is a particularly important function in the case when one or more battery cells were short-circuited. Rectifier may charge the battery with the reference voltage even if some of battery cells are missing, but this shouldn't be performed for a long time. In such a case, referent voltages in all modes of the battery charging operation (float, boost and equalizing charge) should be reduced, in order to prevent the battery damage due to the operation with excessive voltage. Thus, for all the battery chargers, it is important to operate with the same number of battery cells. The secondary battery charger can have only one set of reference voltages, for the one exact number of the storage battery cells. Therefore, in order to permanently keep active the automatic load transfer function of the secondary battery charger, both storage batteries should have the same number of the short-circuited cells. Nonetheless, it should be noted that the short-circuiting of the storage battery cells is an improvisation that should be avoided if not necessary, and that the spare battery cell should be mounted as soon as possible after the perceived failure of a storage battery string.

In order to enable autonomous operation of all battery chargers of the "DRI-PT" series, they do not have measurement data on the output current of other rectifiers in the same facility. Thus, these rectifiers do not have a possibility for equal current sharing during the parallel operation. Usually, for the same values of the reference voltages, the DRI 220-70PT battery chargers should perform some division of the total load current. Nonetheless, there is no chance for overload of any of the parallel operating rectifiers. In the worst case, if the total output current is too high, one battery charger would reach the current limit, while the other would take the remaining load. All the battery chargers were dimensioned to operate with the nominal load current for the entire operational life, that is, at least, twenty years.

DRI 220-70PT rectifiers have multiple monitoring and test functions implemented, such as a battery presence check and the service mode of operation. There are also several fault detection procedures, either for failure of the complete power devices (rectifiers or storage batteries) or just parts of a battery charger. The former procedures are primarily related to the secondary device DRI 220-70PTS to detect fault (or its cessation) of either of the primary devices. The letter procedures are related to all AC/DC converters, since there is a need to detect either unacceptable operation states or failures that affect the device's operation but can be tolerated (earth fault, low output voltage, etc.).

There are five internal states ("hard faults") that cause immediate the shutdown of the DRI-PT battery chargers: high voltage, overload, mains voltage fault, high output voltage ripple and drive pulse loss (or the total thyristor bridge failure – described in the following section). If a shutdown happens (excluding cases of mains voltage fault), thirty seconds later an automatic restart begins and the rectifier enters the soft-start phase, gradually increasing the output voltage. If the cause of failure ceases to exist, after another 30 seconds the soft-start phase is over and the device reaches the preset nominal output voltage. If the cause of failure is still present, rectifier protection will react and the device would again be turned off. There is a programmed procedure for three automatic restarts: if after the third attempt the rectifier protection reacts intermittently, the automatic restart procedure is discontinued and the input mains contactor is turned off, leaving an inoperative AC/DC converter without mains voltage power supply.

III. Detection of Faults

A. Internal Detection of a Thyristor Bridge Fault

If the failure of some of the actuator's elements occurs (such as particular thyristor or ultrarapid fuse), an AC/DC converter may continue its operation, though with degradation of its characteristics. However, in order to achieve the "graceful degradation" and timely reduction of its output characteristics, fault-tolerant devices need reliable procedures even to detect minor failures.

The most common practice for fuse failure detection means the use of auxiliary contacts to provide an independent logical signal for any one of the blown fuses. On the other hand, this simple method cannot be implemented in cases of thyristor failure, let alone for every power semiconductor in a thyristor bridge.

On DRI 220-70PT rectifiers, auxiliary contacts for the detection of failures in a thyristor bridge were not used. As was first implemented in the "DRI" series of thyristor rectifiers [10], in the "DRI-PT" series the detection of the second mains voltage harmonic was utilised as an indicator of a (partial) thyristor power converter failure, or, alternatively, "a thyristor bridge asymmetry" [10].

If some of the actuator element faults occur (such as particular thyristor or fuse), a battery charger may continue its operation. Nevertheless, it is not impossible for a battery charger to remain in operation with a completely dysfunctional power semiconductor actuator. There may be many reasons for such a failure, from a heavy incident in the surrounding environment, but also seemingly trivial incidents, such as the loss of drive pulses on the power semiconductor switches (even by simply pulling out an appropriate connector from a printed circuit board). In the latter case, a battery charger control unit may continue its operations, and, due to the presence of a storage battery at the output terminals of an AC/DC converter, consequently be unable to detect such a fault for a long time. Thus, a new protection function was developed for the series of DRI-PT rectifiers, designed to detect drive pulse loss on a thyristor bridge.

This protection function operates in the following way: if the output voltage declines for more than 10 V, in comparison with a voltage reference, and if the output current in a battery is lower than 2 A continuously for a period of twenty seconds,

then a battery charger is inoperative and a signalisation of the drive pulses loss activates. Activation of this protective function leads to an automatic battery charger turn-off, with the triple automatic restart procedure simultaneously beginning.

B. Detection of Faults Occurring at External AC/DC Converters

A simplified diagram of the UPS facility, comprised of three battery chargers from the DRI-PT series, is presented in Fig. 3. If a "hard fault" (described in section *II.B*) occurs on some of the primary devices (DRI 220-70PTM1 or DRI 220-70PTM2), it is necessary for the secondary device to detect it. To begin the normal operations of an AC/DC converter, it is, at first, necessary to turn on the main switch. The normal rectifier operation means that when the main switch is turned on, the input mains contactor is also turned on. However, if the automatic triple restart procedure fails and the input contactor subsequently turns off, this is a sign of a hard fault of a battery charger. This simple condition is used for a primary device fault diagnostic from the secondary rectifier. Therefore, the turnout detection of an external battery charger is accomplished in the following way: if the mains contactor is off and the main switch is on, then the conclusion may be drawn that the remote rectifier is inoperative. This simple procedure enables the fault detection of a primary rectifier even if it suffers a central processing unit failure or power supply loss, without the need for the establishment of a serial communication between AC/DC converters. It also disables inadequate automatic load transfer activation, such as in the case when one of the primary battery chargers is turned off, *i.e.,* not operating, whilst without the simultaneous "hard fault" signalisation.

The auxiliary contacts of a main switch and the input contactor of both primary devices (DRI 220-70PTM1 and DRI 220-70PTM2) were wired to the secondary device (DRI 220-70PTS). The local power supply voltage of 24 V DC, from the secondary rectifier DRI 220-70PTS, was used as the command voltage for the aforementioned auxiliary contacts in primary devices, enabling the safe monitoring of these logical conditions by the secondary device. This was also a method of precise fault detection in either of the two primary rectifiers, enabling only the activation of the proper output contactor of the secondary battery charger. Additionally, if the logical condition of the primary device fault ceased to exist, this is a signal to turn off the output contactor and put the secondary rectifier into the idle mode again.

IV. Automatic Load Transfer

A. Description of the Load Transfer Function

The DRI 220-70PTS thyristor rectifier, as opposed to its accompanying DRI 220-70PTM device, has an implemented load transfer function (both manual and automatic). During normal operations, two primary devices, DRI 220-70PTM, charge their storage batteries and the supply load connected to their DC current distribution boards. The load is connected to some of the two 220-V DC-Busbar systems, each supported with a separate battery and AC/DC converter. On the other hand, the secondary DRI 220-70PTS device operates in idle mode, regulating the output voltage and monitoring the operation of primary battery chargers.

In the appropriate menu of an HMI display, a user may select either a manual or automatic load transfer. Before the option of a manual load transfer is activated, a user has to select, by pressing virtual tasters on an HMI panel, the secondary device connection either to the first (DRI 220-70PTM1) or the second (DRI 220-70PTM2) of the primary AC/DC converters. Thus, the activation of a manual load transfer leads to the immediate connection of a secondary battery charger to one of the selected primary devices. Connection is established using one of two output contactors in rectifier DRI 220-70PTS. Rectifier DRI 220-70PTS therefore operates in parallel with one of the selected DRI 220-70PTM devices. If a virtual switch for a manual load transfer is turned off, the secondary DRI 220-70PTS rectifier disconnects from its primary rectifier and returns to the idle mode of operation.

More importantly, there is also the possibility of users choosing the automatic load transfer function. If the automatic load transfer function is active, the secondary device monitors the operation of the two primary battery chargers. If either of the primary devices endures a "hard fault" and the automatic triple restart fails, the mains input connector is automatically turned off. In such a case, the DRI 220-70PTS rectifier automatically connects (through one of its two output contactors) to the output of the failed device after the selected time delay (from one to ten minutes) expires. In the meantime, DC load is supported only by the storage battery.

In the manual mode of the load transfer, the secondary rectifier may operate with the inoperative device's battery and load if necessary, until the manual command is given to an output contactor to turn off. If the secondary battery charger operates with an active option of the automatic load transfer, a failed primary rectifier, after the repair, resumes operations, whilst the secondary device automatically disconnects from a load and a battery, after the expiration of the preset time delay.

B. An Algorithm for the Automatic Load Transfer

An algorithm for the automatic load transfer, implemented in the DRI 220-70PTS battery charger, was realised as a software function, performed by the central processing unit of a PLC. A simplified algorithm of the implemented automatic load transfer function is presented in Fig. 4.

In order to implement the automatic load transfer, several logical conditions have to be fulfilled, with the appropriate values of the corresponding software flags. At first, the "slave" battery charger, DRI 220-70PTS, has to be in a stationary state, when the phases of initialisation (F.Start = 1; Fig. 4) and "soft start" (F.Soft.Start = 0) are over. Further, no detection of any "hard fault" is allowed (F.Hard.Fault = 0), and a flag for the automatic load transfer should be active

(F.Auto.L.Trans = 1; Fig. 4). In the next step, there is a need to detect the main switches of both the remote "master" battery chargers (SW.M1.ON = 1 or SW.M2.ON = 1) powering on. Finally, it is necessary to detect the inactive state of the input contactors of two "master" AC/DC converters (IC.M1.ON = 0 or IC.M2.ON = 0). This way, when all the specified logical conditions are fulfilled, a counting sequence commences, which may last from one to ten minutes (presented in Fig. 4 utilising RC circuits). When the preset time delay is over, the "slave" device turns on the appropriate output contactor, either for connection to the output terminals of the DRI 220-70PTM1 battery charger (using the output contactor OCS.1) or the second rectifier, DRI 220-70PTM2 (utilising the output contactor OCS.2). As may be seen from Fig. 4, both output contactors must not be turned on at the same time. Finally, if any of the necessary logical conditions are no longer being fulfilled, after a new countdown sequence the active output contactor is turned off, and the "slave" battery charger again goes back into idle mode.

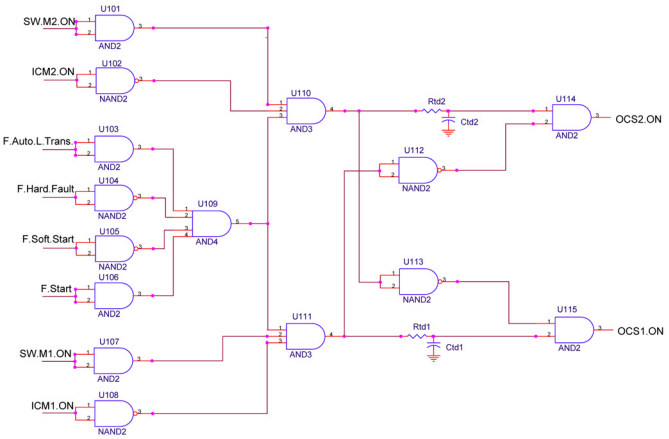


Fig. 4. Logical conditions for the execution of the automatic load transfer, utilised on the "slave" battery charger DRI 220-70PTS. A simplified algorithm was presented in the form of a Boolean (switching) circuit, utilising AND and NAND elements. Time delays were presented with RC circuits.

V. Conclusion

For implementation in the DC UPS system of an HPP, a fault-tolerant facility was developed, comprised of three AC/DC converters. In order to enable the quick replacement of a potentially failed rectifier, a load-transfer procedure (automatic or manual) was utilised in one of these three battery chargers.

In the automatic mode of operation, two primary ("master") AC/DC converters are charging the accompanying storage batteries and DC load, while the secondary ("slave") device operates in idle mode, simultaneously monitoring the state of the two primary devices. If a hard fault of either of the primary devices was detected, the secondary device would, after the preset delay time, connect to the failed rectifier output contacts, supplying the accompanying battery and load. If the primary device comes back into operation, the secondary AC/DC converter would, after the same predefined delay time, disconnect from its output contacts and return to the idle mode of operating.

In order to enable the implementation of the automatic load transfer, it was necessary to utilise a simple procedure for the on-line monitoring of two primary battery chargers without the implementation of a serial communication. Such a procedure was realised utilising only two auxiliary contacts on both of the monitored AC/DC converters, obtaining data on the state of the input switch and input contactor of each battery charger. The fault detection of the monitored battery charger is accomplished in the following way: if a mains contactor is off and a main switch is on, then the conclusion may be drawn that a remote AC/DC converter is inoperative.

In order to enable the fault detection of a primary battery charger even in the case of drive pulse loss on a thyristor bridge, a new protective function was developed. It operates in the following way: if the output voltage significantly decreases, whilst the output current in a battery is simultaneously negligible, then the conclusion may be drawn that the battery charger is inoperative and the signalisation of the drive pulses loss activates. The utilisation of the drive pulse loss protection function, together with other "hard faults" (overvoltage, overload, high output voltage ripple), enables the timely and reliable fault detection of the monitored battery charger.

The operation of the load transfer functions, both automatic and manual, was tested in all phases of the construction of the described facility, *i.e.*, during the phases of development, manufacture and commissioning of the DRI-PT series of battery chargers. Tests were successfully performed with a resistive load in a workshop, as well as with an accompanying DC load and lead-acid batteries in a "Djerdap 2" HPP.

The described technical solution is not restricted only to a field of thyristor power converters but may also be implemented in battery chargers based on high-frequency, switching power converters (PWM rectifiers).

Acknowledgment

This work was supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia under project TR33020, "The power efficiency improvement of the hydro and thermal power plants in the Electric Power Industry of Serbia using the development of power electronics technologies and devices for control and automation".

References

[1] A. C. King and W. Knight, *Uninterruptible power supplies and standby power systems*. New York, U.S.A.: Mc-Graw-Hill Comp. Inc., 2003.

[2] A. Nasiri, "Uninterruptible power supplies," in *Power electronics handbook,* Oxford, U.K.: Butterworth-Heinemann, 2011, ch. *24*, sec. *III*, pp. *627–641*.

[3] E. K. Sato, M. Kinoshita, Y. Yamamoto, T. Amboh, "Redundant high-density high-efficiency double-conversion uninterruptible power system", *IEEE Trans. Ind. Appl.*, vol. 46, no. 4, pp. 1525-1533, Jul. 2010.

[4] S. Yang, D. Xiang, A. Bryant, P. Mawby, L. Ran, and P. Tavner, "Condition monitoring for device reliability in power electronic

converters: A review", *IEEE Trans. Power Electron.*, vol. 25, no. 11, pp. 2734-2752, Nov. 2010.

[5] B. Wang, J. Cai, X. Du, and L. Zhou, "Review of power semiconductor device reliability for power converters", *CPSS Trans. Power Electron. Appl.*, vol. 2, no. 2, pp. 101-117, Jun. 2017.

[6] M. S. Mahmoud and Y. Xia, *Analysis and synthesis of fault-tolerant control systems*. Chichester, U.K.: John Wiley & Sons Ltd, 2014.

[7] V. Vukić, R. Prole, and D. Džepčeski, "Digitally controlled thyristor rectifier based on programmable logic controller with automatic load transfer option," (in Serbian), in *Proc. 28th Simposium JUKO CIGRE,* Vrnjačka Banja, Serbia, September 30 - October 5, 2007, vol. II, pp. 205-212.

[8] V. Vukić, "Thyristor rectifier led by programmable logic controller with modular DC/DC converter in output stage," (in Serbian), *Proceedings, Electrical Engineering Institute "Nikola Tesla"*, vol. 19, pp. 85-92, 2008.-2009.

[9] V. Vukić, R. Prole, and D. Jevtić, "A novel facility with thyristor rectifiers and direct current distribution board for a hydro power plant supply," (in Serbian), *Proceedings, Electrical Engineering Institute "Nikola Tesla"*, vol. 20, pp. 143-156, 2010.

[10] V. Dj. Vukić and R. Dj. Prole, "Diagnostics of the thyristor bridge fault based on the detection of the line-frequency second harmonic in the output current of an AC/DC converter," (in Serbian), *Tehnika*, vol. 67, no. 1, pp. 99-105, Feb. 2018.

Autogenerated Power Distribution Network Model

Lazar Prodanović, Darko Čapko, Aleksandar Erdeljan, *Faculty of Technical Sciences, Novi Sad*

***Abstract*—This research paper covers the development of a smart greedy algorithm, used for autogeneration of a test distribution network model, with predetermined parameters. Autogeneration is done using input data set, which is first clustered in order to reduce its size and thereby simplify the given task. Each cluster provides one representing object that is later used in the optimization algorithm. Proposed optimization algorithm chooses the objects needed to create a network model that is optimal in size and creates a resulting data set. The algorithm is tested with real life data set, extracted from an existing CIM-based distribution network model, and results in a new distribution network model created using the same specification.**

***Index Terms*—Distribution network data model, CIM, clustering, optimization algorithm.**

I. Introduction

Power systems [1] have changed dramatically in recent years, especially those parts that are used for delivering electricity to the customers. Distributed resources are being included (wind farms, photovoltaics, electrical vehicles, batteries, etc.) and the power industry tends to reduce power consumption and losses. By doing that, it challenges power utilities to achieve efficient control and management of the power network. In order to achieve that goal, various software solutions are being used, including sophisticated SCADA systems enhanced with DMS (Distribution Management System), OMS (Outage Management System), DERMS (Distributed Energy Resources Management System), collectively known as ADMS (Advanced Distribution Management System) [2].

These systems are based on models and simulations are being used before making any decisions, be it immediate applied control or planned operations. Size and complexity of these systems lead to big models and require substantial hardware resources for computations. Estimation of needed resources is not an easy task, especially when considering diverse model-dependent ADMS applications. As building a real model represents a long-term process, and resources need to be timely determined, it is desirable to carry out experiments on a similar test model. Therefore, this research proposes a parameterized data model autogenerator, as a fast way of creating those test models. Parameters are selected to represent the model size and they denote total numbers of certain data model elements, like power transformers, power lines, consumers, remotely controlled devices, etc.

Researches on this topic are almost nonexistent. Most test models are created manually, by hand picking model components, one by one. However, there are some simple autogenerator solutions that are based on the multiplication of a smaller model. In this case, the entire model is being multiplied, leading to a bigger model, with the same parameter ratio. As simple as this approach seems, that fixed parameter ratio presents a big problem. If the starting model does not have the desired parameter ratio, that desired ratio is unattainable.

The proposed solution includes setting model parameter values, regardless of the starting model parameter ratio. Multiplying parts/components of the existing model, instead of the entire model, represents the main idea behind this approach and provides greater flexibility when choosing the desired model configuration. Connectivity model is autogenerated, and topological model can be affected using switching devices' status. Aside from this, the attributes of certain components used in multiplication have to be modified, in order to get a valid model, with respect to the power calculations. Some attribute modifications can be done based on simple rules (defined by domain experts), but other require running advanced calculations (power flow, for example). The proposed solution is primarily oriented towards generating models for testing nonfunctional requirements (system attributes such as security, reliability, performance, maintainability and usability), and therefore, does not utilize advanced calculations. After being created, test models are imported into ADMS applications, allowing for all the applications' computations and operations to be run, without affecting consumers. In the background, systems' hardware resources are being monitored, and modifications are made if necessary. For example, hardware resources are added if the application runs slowly, or they can be removed for better cost efficiency if tests show that they are excessive.

Actual model data is usually considered confidential, thereby making its usage in development unworkable or at least limited. The proposed approach goes around these limitations, as it only needs one existing model in order to create countless test models. That existing model can be based on real-life data or it can be manually created, component by

Lazar Prodanović is with the Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia (e-mail: lazar.prodanovic@uns.ac.rs).

Darko Čapko is with the Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia (e-mail: dcapko@uns.ac.rs).

Aleksandar Erdeljan is with the Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia (e-mail: ftn_erdeljan@uns.ac.rs).

component, while making sure it stays valid throughout the process. Standardized data models play a big role in this segment, as they are not origin-dependent. The proposed autogenerator uses models based on Common Information Model (CIM) [3][4], which is an abstract information model used for electrical network modeling and is built on a few basic technologies:

- Unified Modeling Language (UML),
- eXtensible Markup Language (XML) and
- Resource Description Framework (RDF).

RDF/XML [5] files are being used for storing the existing model data and the new test model is exported in the same form, after it is autogenerated.

II. Mathematical Task Description

The aim of the proposed algorithm is to determine a resulting set, consisted of CIM containers – in this case substations and feeders – that comes closest to fulfilling the predetermined distribution network configuration, i.e. total numbers of selected elements in the generated model should be as close as possible to the desired numbers of elements. Therefore, each container has a defining set of parameters, and only some of those parameters are considered when generating a new network model (considered parameters can be added, changed or removed, if needed). Parameters that are used in this research paper are:

1. Number of transformers,
2. Number of signals,
3. Number of SCADA points, and
4. Number of MV consumers (referred to as "consumers").

In other words, the goal of the autogenerator is to analyze the input data set, and pick a new set of substations and feeders out of it, so that the resulting set's total number of transformers, signals, SCADA points and consumers, comes as close as possible to the desired values, given for the new distribution network model.

Mathematical definition of this task can be formulated in the following way. For a given input set of *n* objects, where every object has *m* parameters, determine the set *S*, which consists of those objects, so that each parameter sum becomes as close as possible to the predetermined parameter value. Every object can appear multiple times in the set *S*. All parameter values are whole numbers.

The goal is to determine:

$$\min_{X_i} J = \min_{X_i} \sum_{j=1}^{m} \left(\frac{P_j - \sum_{i=1}^{n} X_i p_{ij}}{P_j} \right)^2 \qquad (1)$$

Where:

- p_{ij} represents *j*-th parameter value for the *i*-th object
- P_j represents predetermined *j*-th parameter value for set S
- X_i represents *i*-th object's number of occurrences in set S

A. Clustering input data

Typical data models consist of hundreds or even thousands of different containers and each container requires additional computational time, if included in the algorithm. Adding more container parameters to this problem increases the number of dimensions and makes it almost impossible to work with the given input data set. In order to improve the speed and the efficiency of the optimization algorithm, the input data set has to be reduced in some way. For this practical reason, k-means clustering [6][7] is performed on the input data. As data clustering represents a process of classifying objects into groups/clusters of similar objects, it is not reducing the input data set itself. However, as each cluster contains similar objects, it can be represented by a single object, without sacrificing diversity. Choosing a representing object is done by calculating every object's distance from its cluster's centroid/center, and choosing the closest one as that cluster's representative. This way, starting data set size is reduced to a desired number of clusters, which can be passed to the algorithm as an input argument.

Moreover, clustering can be used for eliminating the usage of untypical objects/containers in the generated data model. Cluster centroids are positioned as close as possible to all of its belonging containers – total Euclidean distance between the centroid and all of its containers tends to be as low as possible – which makes containers with abnormally high or abnormally low parameter values significantly less likely to be chosen as the cluster representatives.

Although there are clustering algorithms that are better suited for finding the cluster representatives, as well as eliminating the untypical data (k-medoid and c-medoid, for example), the k-means algorithm does almost the same job in considerably smaller time interval, making it ideal for fast optimization algorithms with a lot of input data.

III. System Workflow

The autogenerator consists of:

- Initial data model importer (RDF/XML parser),
- Clustering tool and,
- Optimization algorithm.

System workflow block diagram can be seen in Figure 1. Starting data set is extracted from an existing CIM-based model, given in the form of an RDF/XML file, using the RDF/XML parser, embedded into the autogenerator. Clustering tool then divides that data set into groups, and chooses a representative for each group, thereby creating a reduced data set, used in the optimization algorithm. Optimization algorithm, whose pseudo code is given in the next chapter, creates a new/resulting data set by choosing objects/containers from the reduced data set, while trying to satisfy the predefined network model parameters. When the resulting data set is created, it is exported into a new RDF/XML file, which defines the generated data model of network distribution.

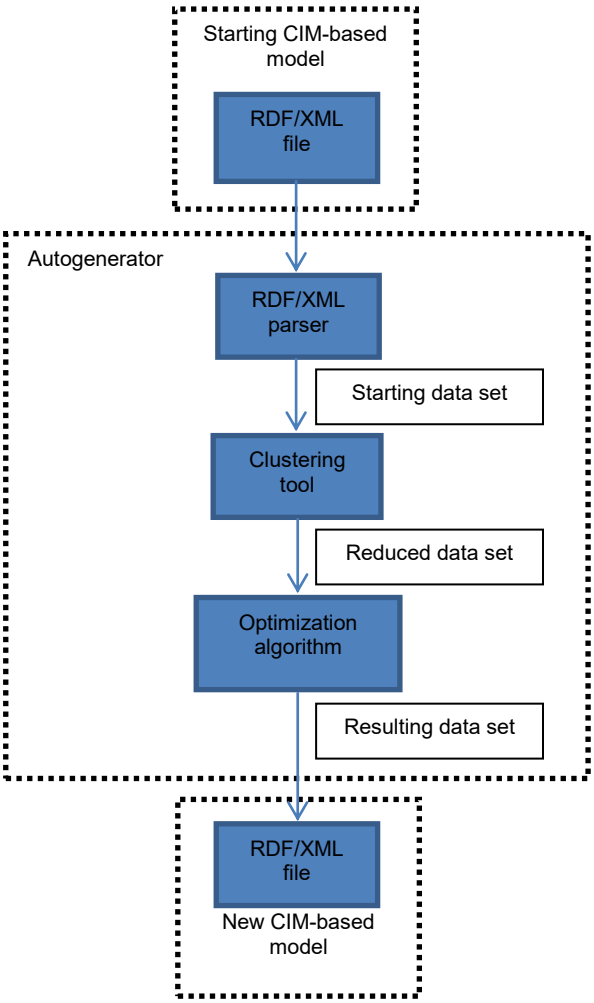


Figure 1 System workflow overview

IV. Implementation

This chapter describes the optimization algorithm, used for determining the optimal configuration of the data model. The algorithm works with a reduced data set, described in the previous chapters. It works as a greedy algorithm, meaning that it is making an optimal choice at each step, while trying to get to the best solution at the end. Greedy algorithms are effective at solving some kinds of problems, but tend to go astray while solving the other kinds, as they always choose the most beneficial element at the moment, instead of the one that makes the final solution better. However, the proposed algorithm is tweaked so it does not go for quantity over quality, and leads to better results in the end.

The other thing that distinguishes this greedy algorithm is the fact that it has multiple iterations. Each iteration produces a different solution, using different reduced data set, thanks to the clustering tool. Different data sets are created using a random number generator that allocates different starting clusters for each iteration. These starting clusters affect cluster centroids, which leads to different representatives at the end of each clusterization process. As there are more iterations and problem solutions at the end, the one that best satisfies the optimization criteria (1) is chosen.

Pseudo code for the proposed smart greedy algorithm:

```
1:  function SmartGreedyAlgorithm(data, needs, numIter, numCluster)
2:      for i = 1:numIter
3:          R = ExtractClusterRepresentatives(data, numCluster)
4:          newJ = CalculateJ(needs)
5:          previousJ = MAX_SE
6:          while newJ <= previousJ
7:              previousJ = newJ
8:              CalculateProfitMatrix(R)
9:              for e = R
10:                 CalculateAverageProfit(e)
11:                 CalculateDeviationForParameters(e)
12:             end for
13:             SelectElementWithMinDeviation(R)
14:             InsertSelectedElementInModel
15:             remainingNeeds = CalculateRemainingNeeds
16:             newJ = CalculateJ(remainingNeeds)
17:         end while
18:         SaveSelectedElements
19:         finalJ = CalculateJ(remainingNeeds)
20:     end for
21:     sol = FindMinFinalJ
22:     ExportSolutionToXML(sol)
23:     end function
```

Input arguments:

- data – starting data set
- needs – predetermined network model parameters
- numIter – assigned number of iterations
- numCluster – assigned number of clusters

Output:

- Resulting data set (line 21), exported into an XML file (line 22), later used by the GDA importer for creating a new distribution network model

A. *Profit matrix calculation*

Main steps of the proposed smart greedy algorithm can be seen in the form of pseudo code. However, profit matrix calculation (line 8) and the selection of the best fitting object (line 13) have to be described with more details.

Object's profit is determined by temporarily inserting that object into the resulting set, and recalculating system's relative errors for each parameter (line 8), in regard to current needs. Object is removed from the resulting set as soon as profit is calculated. In order to make the algorithm smarter, the current fitness of an object is no longer seen as the quantitatively biggest contribution to solving the optimization problem given with the equation (1). Instead, profit mean value is calculated for each object (line 10), thereby enabling the calculation of a deviation for each parameter (line 11). Object that has smallest deviation sum is chosen as the fittest (line 13). This means choosing the object that can even the current relative error levels, as much as possible. By choosing this type of objects, the algorithm tends to keep the predefined parameter/needs ratio until the end, resulting in a better solution, preferably with zero relative errors for all the parameters. After each selection, the fittest object is inserted into the resulting set (line 14), and system needs are

recalculated by subtracting object's parameter values from the current needs (line 15).

B. *Determining the final solution*

Each iteration in the algorithm runs until new *J* value, defined in the equation (1), becomes bigger than the previous *J* value (line 6). After that, the final solution is saved, together with its last *J* value.

When the assigned number of iterations passes, all the collected solutions are compared, and the one that best satisfies the equation (1) is chosen (line 21) and exported to the RDF/XML file (line 22), creating a new distribution network model in the process.

V. RESULTS

The algorithm was implemented and tested for creation of 8 data models each having different set of parameters. For each of those sets, algorithm was run several times, using different numbers of iterations and clusters as input arguments. Iteration numbers used were: 10, 100, 1000 and 2000. Cluster numbers used were: 5, 10, 20, and 50. Those values were chosen randomly (but limited by a size of the starting data set) and are analyzed in the following paragraphs. Table 1 shows results for each test's best combination of the assigned input arguments. Results are given in form ***set value / deviation***. For example, model #1 has 10001 transformers, 19995 signals, 750 SCADA points and 8000 consumers, but its set values were 10000, 20000, 750 and 8000. Each run used the same existing model data for creating the starting data set.

TABLE 1
TEST RESULTS

Set parameter values / final deviations	Number of transformers	Number of signals	Number of SCADA points	Number of consumers
Model #1	10000 / 1	20000 / -5	750 / 0	8000 / 0
Model #2	15000 / 46	20000 / 1	1000 / 1	8000 / 14
Model #3	5000 / 0	15000 / 0	750 / 0	5000 / 0
Model #4	10000 / 1621	10000 / -1169	750 / -10	5000 / -189
Model #5	5000 / -2	7500 / -73	500 / 0	8000 / -13
Model #6	10000 / 15	15000 / -56	1000 / 0	10000 / -16
Model #7	10000 / 0	20000 / -9	500 / 0	5000 / 2
Model #8	1000 / 353	1000 / -10	200 / -32	1000 / 0

Figure **2** shows a graphical view of the results for one of the tests, with all the possible combinations of input arguments.

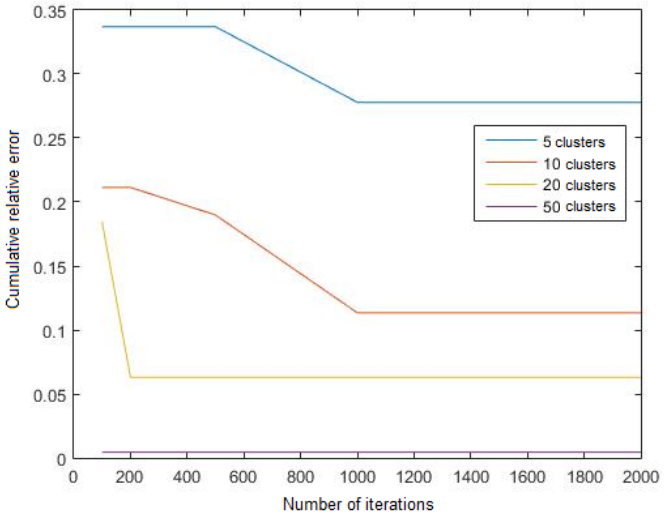


Figure 2 Graphical view of the results

As can be seen from the results provided in Table 1, the proposed algorithm usually provides acceptable results, where relative errors do not exceed 1%. It also points to the unrealistic parameters, as can be seen in tests 4 and 8. This happens when certain parameter values become too high or too low, leaving the clustering algorithm unable to find right representatives. For example, in model #4 test, the desired number of transformers matches the desired number of signals, which leads to big parameter deviations at the end, because there are no suitable containers in the existing model whose standard ratio is 1:2 for those parameters. This means that the starting data set or predefined network model parameters have to be changed, because desired values cannot be achieved using existing objects.

Figure **2** shows the relation between assigned number of iterations, assigned number of clusters and the relative error, that appears at the end. As can be seen, chances of finding an optimal solution decrease significantly if smaller values are taken for any of the input arguments. On the other hand, bigger values lead to a smaller error, which is logical, considering that there are more objects to choose from for a bigger cluster number, and more solutions to choose from at the end for a bigger iteration number. The thing that should be taken into account here is time. Bigger argument values lead to a bigger time consumption and, after some point, have a minor effect on the results. Increasing number of passes leads to a proportional increase in execution time, but increasing number of clusters affects execution time in exponential fashion. For example, changing number of clusters from 5 to 10 leads to a 20-25% increase in execution time, but changing it from 20 to 50 leads to a 20000-25000% increase. With that in mind, input arguments should be chosen carefully. Number of clusters should be 5-20% of the starting data set size, and number of iterations should reflect the desired precision/speed ratio, starting with a few hundred iterations for fast execution, and going all the way to tens of thousands for best precision. These numbers can vary depending on model size and personal precision/speed preferences.

VI. CONCLUSION

This paper describes the algorithm used for autogenerating distribution network test models, with predefined sets of parameters. Autogeneration is done by finding the optimal combination of containers (feeders and substations), extracted from an existing CIM-based data model. These test models are used for testing the nonfunctional requirements in the ADMS software development process.

By looking at the results, a few conclusions can be made:

- For the same number of clusters, the solution is always equal or better when using a bigger iteration number
- The solution is almost always better when the bigger cluster number is assigned
- The algorithm is pretty robust, and provides acceptable results, even for the unrealistic parameter ratio
- For typical network model specifications, relative errors do not exceed 1%

By increasing the iteration and cluster numbers, the execution time rises, so there has to be some compromise when assigning those values.

ACKNOWLEDGMENT

The authors are partially supported by the Serbian Ministry of Education, Science and Technological Development, through grant No. 32018.

REFERENCES

[1] A. Von Meier. "Electric Power Systems: A Conceptual Introduction." A John Wiley & Sons, New Jersey (2006).

[2] M. S. Hossan, B. Chowdhury, J. Schoene, S. Bahramirad," Advanced Distribution Management System: Implementation, Assessment, and Challenges." 2018 IEEE Power & Energy Society General Meeting (PESGM). IEEE, 2018.

[3] L. King, T. Nielsen, S. Neumann, A. Vojdani, P. Parikh, "The Common Information Model for Distribution - An Introduction to the CIM for Integrating Distribution Applications and Systems," EPRI, Palo Alto: CA 1016058 (2008).

[4] A. W. McMorran, "An Introduction to IEC 61970-301 & 61968-11: The Common Information Model," University of Strathclyde 93 (2007): 124.

[5] "RDF 1.1 XML Syntax," W3C Recommendation, W3C (2014), https://www.w3.org/TR/2014/REC-rdf-syntax-grammar-20140225

[6] P. Tan, M. Steinbach, V. Kumar. "Cluster analysis: basic concepts and algorithms." Introduction to data mining 8 (2006): 487-568.

[7] J. McCaffrey, "Machine Learning Using C# Succinctly," Syncfusion Inc. (2014)

Current Sampling Techniques for Digitally Controlled Inverters

Filip Filipović, Milutin Petronijević, Nebojša Mitrović, Bojan Banković and Vojkan Kostić

***Abstract*—The purpose of this paper is to describe common problems that influence accurate current sampling in the case of power inverters. The current sampling is observed for nominal and low load of the inverter. Considered problems include dead time of power switches and small load time constant. Techniques considered for minimization of the influence of these problems are synchronized sampling method, multisampling method and antialiasing filter. The common algorithms for obtaining a single current value in the multisampling method are also considered. Effects of reviewed techniques on current sampling are examined in the case of one power inverter that supplies passive load in isolated grid. The results confirm the necessity of careful current sampling method consideration with respect to the application. Model of the inverter with all considered techniques for accurate current sampling is built using MATLAB/Simulink.**

***Index Terms*—Inverter, Current Sampling, Pulse Width Modulation, Synchronised Sampling, Multisampling, Antialiasing Filter**

I. Introduction

Today, control of inverters is realised using digital logic circuits (DSPs of FPGAs) [1]. Digital control showed several advantages over the analog in terms of smaller susceptibility to temperature variations and component aging, easier algorithm modification and lower bill of material. The downside would be inability to continuously monitor variables of interest, but rather in specific moments in time using the analog to digital converters (ADC). This reduction of continuous time to a discrete time signal is called sampling and some of the problems that can occur during this conversion process that deviate obtained signal from perfect reconstruction may include:

- Aliasing effect,
- Jitter,
- Noise,
- Quantisation and
- Other errors due to non-linear effects [2].

Inverters produce higher order harmonics at the switching frequency and its multiples. In order to minimise the influence of previously mentioned aliasing effect, sampling techniques such as single update pulse width modulation (PWM) [3], double update PWM [4] and multisampling PWM [5-7] are used. These sampling methods can put various technical requirements on the hardware used for sampling. For the single update PWM, only synchronization of PWM carrier signal with the ADC is required. On the other hand, for the multisampling method, more demanding hardware for ADC in terms of speed, along with the appropriate digital circuit for the manipulation of the obtained samples is required.

Digital current controller is the fastest control loop in the system for most inverters. Its accuracy determines the quality of the current waveforms injected into the grid [5]. Transport and sampling delays have a large impact on maximum possible performance of digital current controllers [7]. Any additional delay in sampling process would have a direct effect on current loop performance.

The purpose of this paper is to describe problems related to current sampling and the common ways of overcoming them. The main idea is to design alias-free sampling with as small effect on sampling delay as possible. In the second section, the single sampling and multisampling methods along with the antialiasing filter (AAF) will be considered. The details of test setup, AAF, along with the main results are presented in the third section. Based on the presented results, summary of the results and directions of future work is presented in the conclusion.

II. The Method

Prevention of aliasing is one of the key requirements for the accurate current sampling during inverter operation. The dominant part of inverter's current ripple located in the out-of-band region comes from the PWM. With the ideal sampling, we would be able to perfectly reconstruct all signals below the Nyquist frequency without attenuation and phase lag, while rejecting all signals above it. Commonly used techniques for the PWM ripple suppression in the sampled current include sampling and PWM synchronization, averaging over multiple samples and filtering of analog signal before digital conversion.

Filip Filipović is with the Faculty of Electronic Engineering, University of Niš, 14 Aleksandra Medvedeva, 18106 Niš, Serbia (e-mail: filip.filipovic@elfak.ni.ac.rs).

Milutin Petronijević is with the Faculty of Electronic Engineering, University of Niš, 14 Aleksandra Medvedeva, 18106 Niš, Serbia (e-mail: milutin.petronijevic@elfak.ni.ac.rs).

Nebojša Mitrović is with the Faculty of Electronic Engineering, University of Niš, 14 Aleksandra Medvedeva, 18106 Niš, Serbia (e-mail: nebojsa.mitrovic@elfak.ni.ac.rs).

Bojan Banković is with the Faculty of Electronic Engineering, University of Niš, 14 Aleksandra Medvedeva, 18106 Niš, Serbia (e-mail: bojan.bankovic@elfak.ni.ac.rs).

Vojkan Kostić is with the Faculty of Electronic Engineering, University of Niš, 14 Aleksandra Medvedeva, 18106 Niš, Serbia (e-mail: vojkan.kostic@elfak.ni.ac.rs).

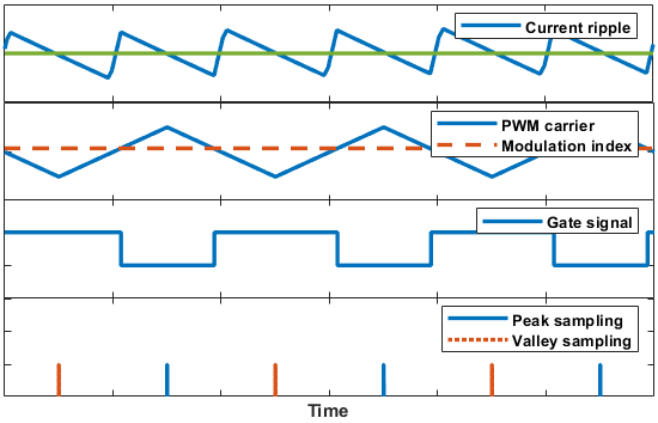


Fig. 1. Synchronized sampling with peak or valley of PWM carrier signal.

A. *Single Update PWM Method*

Since the current control algorithm has a typical closed loop bandwidth of several kHz, its controller cannot be designed without considering delays induced by appropriate low pass filters, if used [9]. Widely used technique today for current ripple suppression is sampling synchronization [10]. It uses the fact that in the ideal case no ripple is present in the sampled current if the sampling is done at the peak or valley of PWM carrier signal. Example of current sampling synchronized with the peak or valley of PWM carrier signal in three phase, three wire, two level voltage source inverter is shown in Fig. 1, for a current measured after L grid filter.

Zero current ripple coincides with peaks and valleys of the PWM carrier if series resistance through which current of interest passes is zero, supply voltage is constant and if there are no other delays in PWM generation. The effects of the PWM dead time and sufficiently large series resistance on the current sampling ripple can be observed in Fig. 2. Also, PWM voltage pulses along with an unneglectable parasitic capacitance can induce oscillations microseconds after the switching. If sampling occurs during that period, considerable sampling error can be induced. Synchronized sampling can be combined with AAF of much higher cutoff frequency than the one that would be needed if sampling without synchronization is done [9].

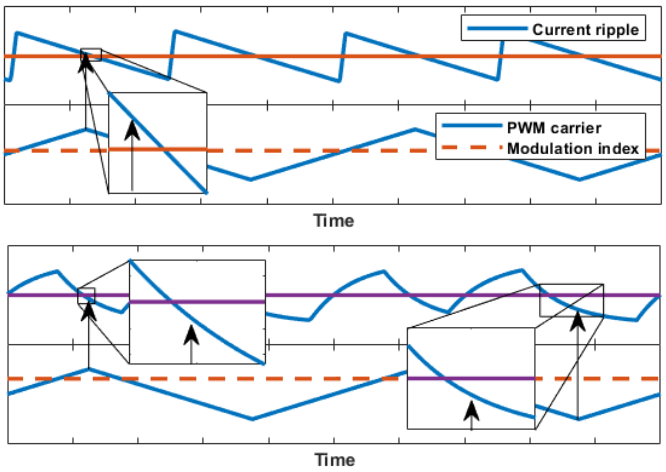


Fig. 2. Effects of dead time (top) and small time constant (bottom) on the sampled current ripple.

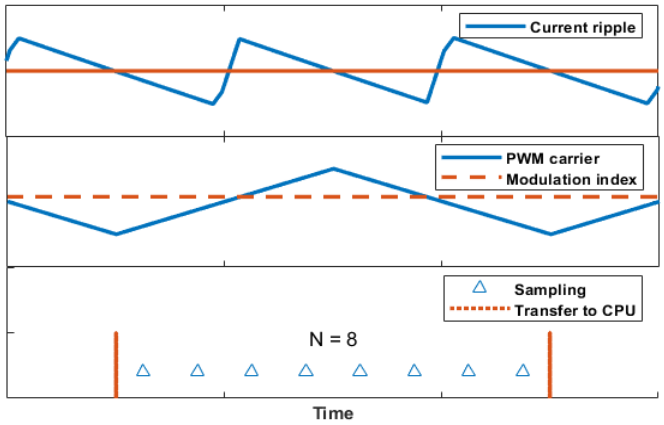


Fig. 3. Synchronized multisampling.

B. *Multisampling PWM Method*

Ever more attention recently is being given to the multisampling methods, due to the fact that it can be implemented in most of the modern hardware used for control. The process of synchronized multisampling can be seen in Fig 3. This method uses N equidistant current measurements between peaks or valleys of PWM carrier signal. N is called multisampling factor and recommendation for its proper selection can be found in [9].

For the triangular PWM carrier signal, single sampling method induces a pure delay phase lag. In the multisampling approach, there is a pure delay, that is a decreasing function of the N. Achieving high multisampling factor can be a hardware constraint. In one way, it can be achieved with timed CPU interrupts for sampling and synchronization with the PWM carrier, although it would result in relatively high CPU workload, even for today's DSPs. Arguably better way in terms of CPU workload is outsourcing of sampling command to direct memory access (DMA) controller. DMA would use timed interrupts to trigger ADC and store result in the dedicated array located in random access memory (RAM). Control algorithm executed on CPU would retrieve this array from RAM on PWM interrupt and based on the extraction algorithm, from N collected samples, it would calculate instance that would be used in control algorithm.

A possible solution for value extraction is usage of low pass filter or moving average filter (MAF). MAF is one of the most widely used filtering technique today, presents a linear phase finite impulse response (FIR) filter and can act as an ideal LPF if certain conditions are met [12]. Adjusting MAFs window length, switching harmonics can be eliminated since their frequency of occurrence is known, but that would induce a delay that is equal to single update PWM's sampling delay. This would eliminate any advantage of multisampling method in terms of phase boost [11]. Another filtering method that allows higher control bandwidth is presented in [13] and will be denoted here as a repetitive ripple estimation filter (RREF). The RREF behaves as an ideal low pass filter only on specific frequencies, while for all other frequencies it has no filtering effect. This implies that it would not be a good solution for sampling noise filtering. For value extraction, reduced version of RREF presented in [11] will be analysed here. Bode gain and phase plots of MAF and RREF are presented in Fig 4.

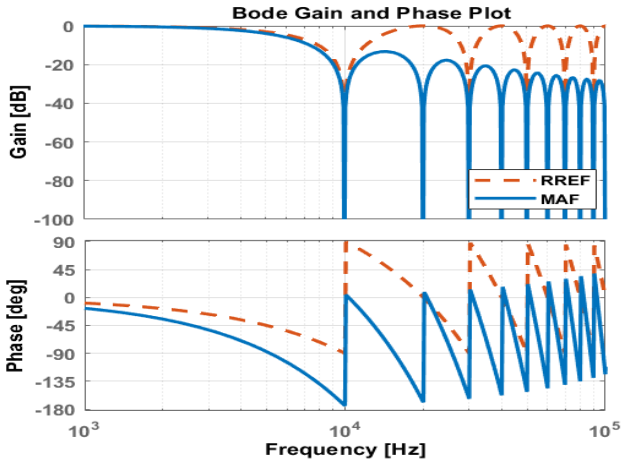


Fig. 4. Bode gain and phase plots of RREF and MAF.

C. Antialiasing Filter

The signal filters can be implemented in analog (prior to ADC) or in digital domain, with a number of differences between those two realizations. An analog filtering is more suitable for sampling speeds above 5 kHz, as it can eliminate high frequency noise before it reaches ADC and, in that way, reduce noise in the out-of-band region and alias signals. Also, with noise peaks removal, analog modulator saturation of the ADC can be avoided. Digital filters can remove noise after the conversion process and digital filters are programmable, thus allowing easier modifications [14].

In data-acquisition systems, analog low pass filters are commonly used for AAF implementation. With all real filters, various trade-offs have to be made, according to the desired application. Popular filter designs from [15] include:

- Butterworth – optimized for maximally flat gain in the pass-band. At the cutoff frequency, gain is -3 dB. Above the cutoff frequency, attenuation is -20 dB/decade/order. Transient response to pulse input shows moderate overshot and ringing.
- Bessel – optimized for maximally flat phase in the pass-band. At the cutoff frequency, gain is -3 dB. Above the cutoff frequency, attenuation is -20 dB/decade/order. Excellent response to pulse input, but slower roll-off.
- Chebyshev – designed with the desired ripple in the pass-band. Cutoff frequency is defined as the frequency at which the response falls below ripple band. Steepest roll-off, but pulse input shows highest overshoot and ringing.

Active filters can be made in a number of different architectures. Some provide better results in the case of stability, susceptibility to noise and the number of elements needed. Two most popular ones are:

- Sallen-Key – has unity gain in the pass band independent of component variation.
- Multiple Feedback (MFB) – has a superior high frequency response.

Transfer function of a second order low pass filter can be written as

$$F_{LPF}(p) = \frac{1}{\left(p\frac{1}{2\pi f_c FSF}\right)^2 + p\frac{1}{2\pi f_c QFSF} + 1} \quad (1)$$

where, p denotes Laplace complex variable, fc is crossover frequency, FSF is frequency scaling factor and Q is quality factor [15].

Low pass filters of the order higher than second can be obtained with a series connection of first and second order filters. Factors Q and FSF are selected following the recommendations from [15] (using provided table) in the way of uniform quality factor increase (lowest Q near the input, highest near the output).

III. Main Results

Testing of all previously mentioned sampling techniques was done on the model of the inverter supplying a passive load. The model is presented in Fig. 5 and it is build using MATLAB/Simulink R2018b using components located in Simulink library Simscape. Details of this setup are displayed in Table I. Algorithm that controls PWM of the inverter work in open loop mode with a 100 µs cycle time. It has a fixed reference frequency of 50 Hz and a fixed line to line reference voltage of 400 V. Dead time of the PWM is 1 µs and there is no dead time compensation in the control algorithm. Current sampling is modelled through 12-bit discretization. AAF is modelled in continuous time using transfer function.

Tested current sampling methods are:

- random sampling – sampling occurs at the speed of the control algorithm at the arbitrarily point on PWM carrier,
- single sampling – sampling occurs at the valley of the carrier,
- multisampling (MAF) – sampling occurs at the frequency of 320 kHz, AAF to be described is present and for the single value extraction MAF is used with a filtering window that corresponds to the frequency of 10 kHz,
- multisampling (RREF) - sampling occurs at the frequency of 320 kHz, AAF to be described is present and for the single value extraction RREF is used with a filtering window that corresponds to the frequency of 10 kHz.

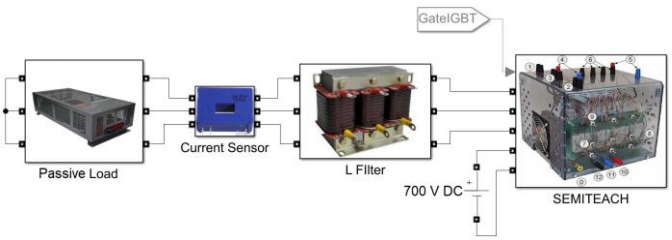


Fig. 5. Test setup used for simulation built in MATLAB/Simulink.

TABLE I
Simulated Setup Details

Inverter type	3 phase, 3 wire, 2-level voltage source inverter
Switching frequency	10 kHz
L filter inductance	6.4 mH
Current sensor	±10 A (±10 V)
ADC converter	12-bit (±10 V)
DC supply voltage	700 V

A. *Design of Antialiasing Filter*

AAF can be used in combination with any of the previously mentioned methods to reduce alias signals, although its implementation may require recalculation of current controller parameters in some cases. The AAF is designed for 12-bit ADC with a 320 kHz sampling frequency. According to the design consideration from [16], to use the full ADC dynamic range, any undesired out-of-band signal components have to be filtered to less than the ADC's Least Significant Bit (LSB) level. This implies that AAF with a gain of -72 dB at the frequency of 160 kHz would be desired. In the real case, there is a tradeoff between filter order, crossover frequency and the gain at the Nyquist frequency. Operational amplifiers with inappropriate characteristics can deteriorate performance of the filter. Using closed loop bandwidth and slew rate as additional design parameters [17], widely accessible LF356 [18] is used for practical filter implementation. For a current measurement system of the inverter, AAF is selected with the parameters provided in the Table II.

TABLE II
ANTIALIASING FILTER DETAILS

Type	Bessel
Architecture	MFB
Order	6^{th}
Crossover frequency	f_c = 30 kHz
First stage	Q = 0.5103 FSF = 1.606
Second stage	Q = 0.6112 FSF = 1.6913
Third stage	Q = 1.0234 FSF = 1.9071

Bode gain and phase plots of the filter are provided in Fig. 6. Schematic of the filter is presented in Fig. 7. In the schematic, the input (usually a buffer circuit and isolation amplifier) and the output (operational amplifier for signal range match and output impedance adjustment) stages are not shown.

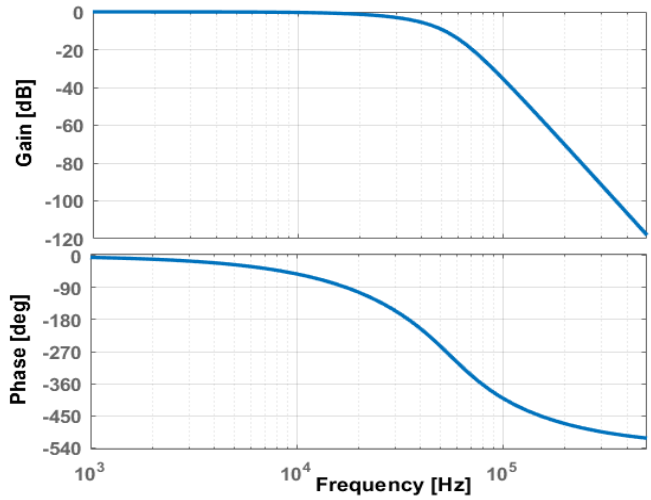


Fig. 6. Bode plots of analog antialiasing filter.

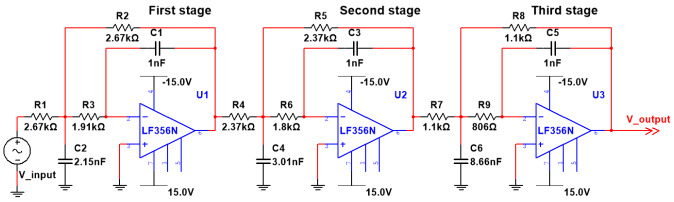


Fig. 7. Schematic of 6^{th} order LPF in MFB architecture with a 30 kHz cut off frequency.

B. *Rated Load Test*

For the first test, comparison of the selected algorithms during nominal load operation (phase resistance of 40 Ω) is chosen. In the Fig. 8. real current waveform is shown, along with the current obtained from sampling with each method. Lower part of the same figure presents Fourier transformation (FFT) of the real signal and each sampling method.

It can be observed that all sampling methods except random sampling during nominal load operation provide similar results. Multisampling (MAF) method provides the most similar result to the original. Random sampling in this case has a sampling phase that does not coincide with the peak or the valley of the PWM carrier and due to that, it has bigger deformation in original signal reconstruction. This can be observed from the FFT decomposition of this signal (larger deviation from the original value on some frequencies). For the multisampling (RREF) there is a problem with the fundamental harmonic estimation. It can be observed from the Fig. 4. that RREF behaves as an ideal low pass filter only for the odd multiples off first filtered frequency. For all other frequencies it has a small or negligible filtering effect. Since the AAF is designed for the crossover frequency of 30 kHz, switching current ripple at the twofold PWM frequency will be present in the sampled current with a negligible attenuation.

C. *Low Load Test*

Although nominal condition are the ones that power inverters are built for, in the case of renewable energy sources, nominal conditions are not so common in practice and nominal load results do not represent authentic situation most of the time. In the case of photovoltaic, typical power output ranges from twenty to fifty percent of the rated power [19]. If the inverter is located in the isolated grid supplying a small number of consumers, load can also vary substantially. Results from the low load operation are presented in Fig. 9. These results are obtained for a load phase resistance of 350 Ω, thus making this test a combination of small load and small load time constant.

From Fig. 9. substantial differences between the sampling methods can be seen even by waveform inspection. The same methods as for the previous test now show significantly different results among themselves. Random sampling method provides arguably worst results. FFT analysis shows substantial presence of even harmonics that results in asymmetry of the waveform. Somewhere better results show synchronized sampling method, although it suffers from the same problem of asymmetry. Synchronisation of the PWM carrier with the current sampling does not provide accurate results in the case of small load time constant.

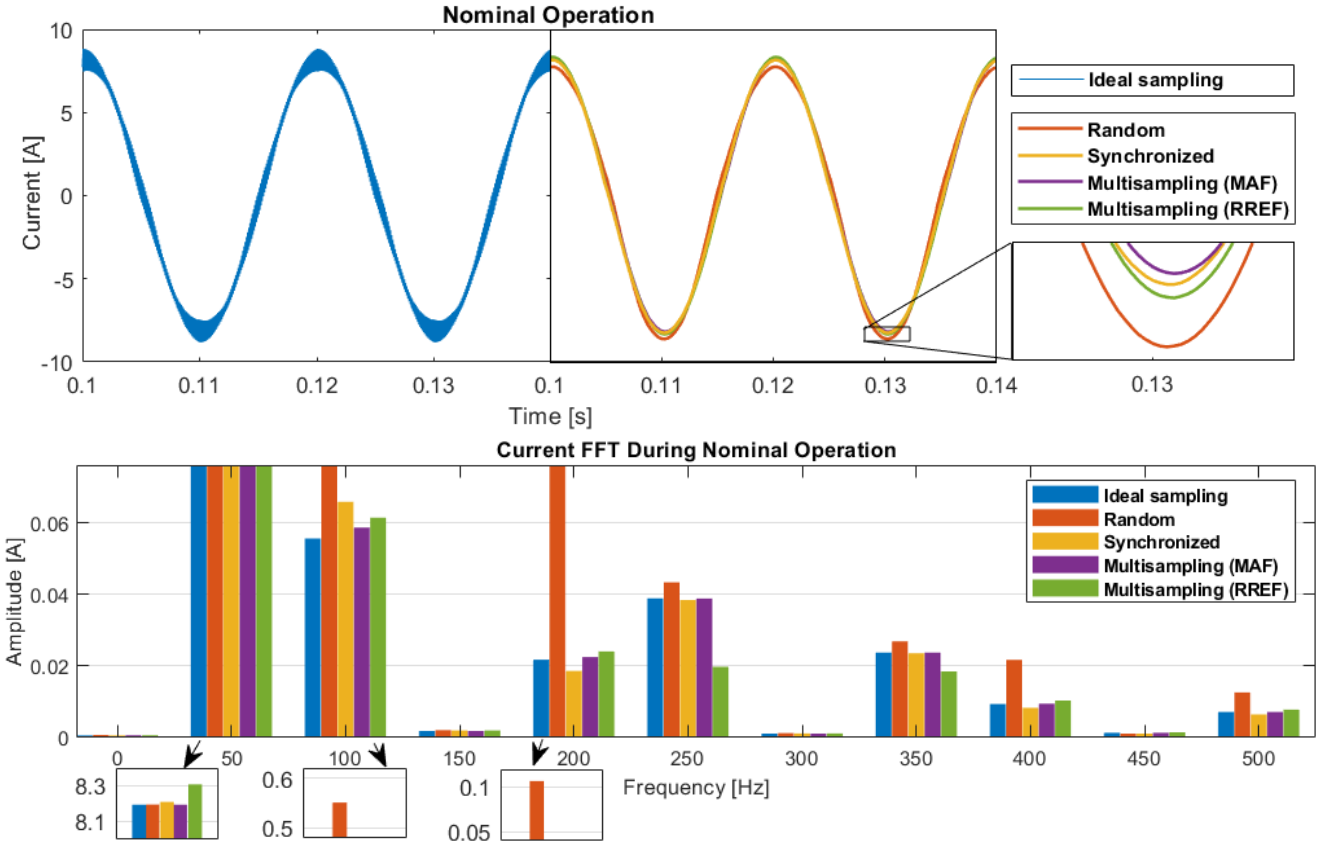


Fig. 8. Sampling of original current (upper) and FFT of it (bottom) for a nominal load operation.

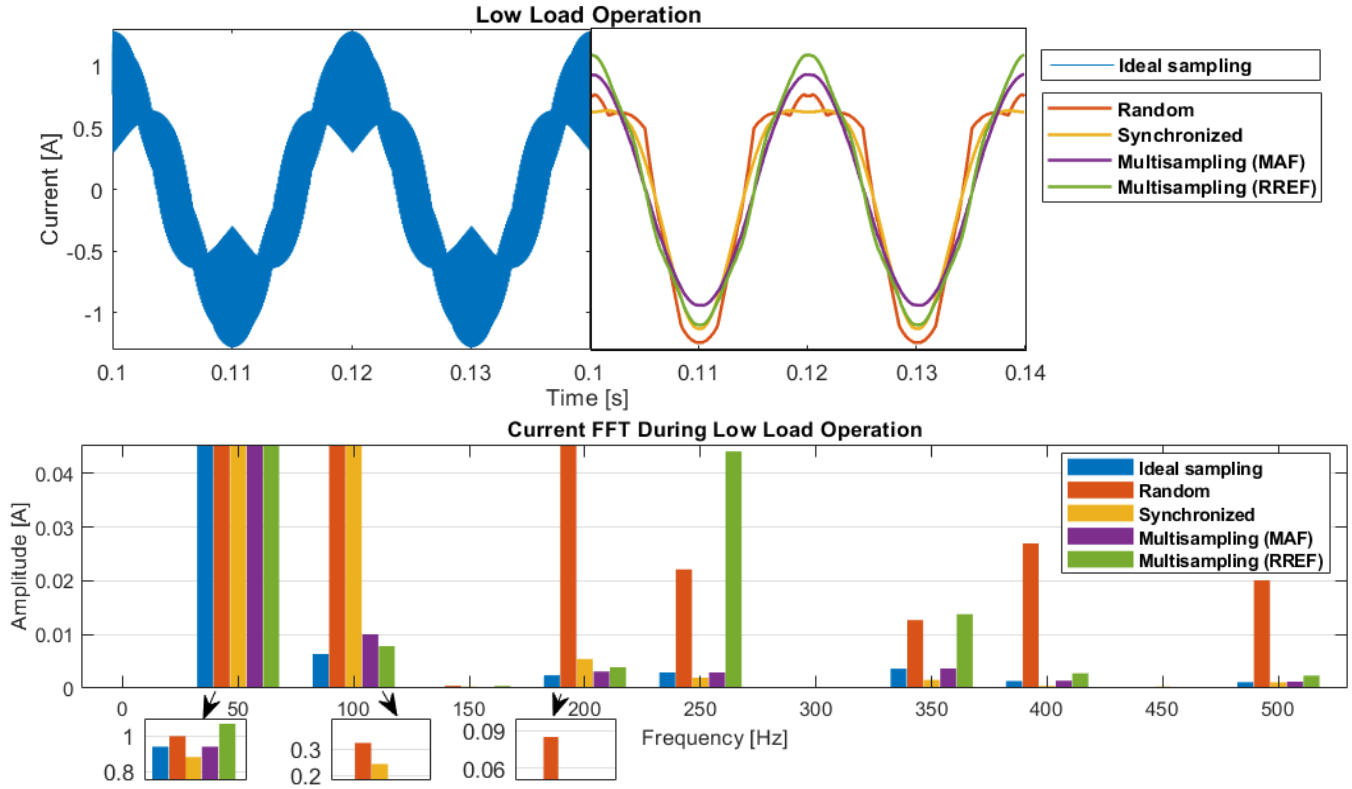


Fig. 9. Sampling of original current (upper) and FFT of it (bottom) for a low load operation.

Using multisampling methods, the influence of current ripple and small time constant on sampled current is reduced significantly. RREF extraction method provides information of higher fundamental harmonic and higher fifth harmonic compared to the real value. MAF extraction provides the most trustworthy results. Current sampled using multisampling technique and single value obtained using averaging method provides most consistent results at the cost of phase lag higher than in other tested method.

IV. Conclusion

The alias signals in sampled currents can induce substantial problems in control algorithms. For a high performance current controllers, selection of proper sampling and filtering method can have large impact on the controller bandwidth or control loop tracking capabilities.

Results comparison shows that random sampling should be avoided in all times if possible. With this method, the amount of out-of-band noise can vary with the variation of the point on the PWM carrier signal at which sampling occur.

Synchronized sampling provides consistent results for a relatively large time constant of the AC power line in respect to the sampling time. With an increase of active resistance present in the load accuracy of this method deteriorates.

Combination of AAF and multisampling with the averaging method provides the best results of all tested methods in terms of credible current waveform. The problem with the averaging method is substantial phase shift that is induced and that presents limiting factor in the case of high performance current controller implementation.

Usage of repetitive ripple estimation in combination with the AAF for a switching frequency harmonic elimination shows moderate results. It only manages to partially suppress occurrence of the PWM ripple in sampled currents. The main reason for the eventual usage of this filtering technique is a small phase lag, providing higher current control loop bandwidth.

AAFs can have limiting impact on the current control loop performance in terms of induced delay. Their design is a tradeoff between phase and gain flatness, filter order and phase delay and better response on higher frequencies or more stable gain with the component variation. For a high performance current controller design, current sampling requires careful planning.

In the future work, techniques for AAF phase lag compensation and more value extraction algorithms for multisampling techniques will be researched. This research will be done as an effort to improve the achievable bandwidth of current controller and the credibility of the sampled current.

Acknowledgment

This work was supported by the Ministry of Science and Technological Development, Republic of Serbia (Project number: III 44004 and III 44006).

References

[1] S. Tahir, J. Wang, M. Baloch, and G. Kaloi, “Digital Control Techniques Based on Voltage Source Inverters in Renewable Energy Applications: A Review,” *Electronics*, vol. 7, no. 2, p. 18, 2018.

[2] “Improving ADC Resolution by Oversampling and Averaging”, Silicon Laboratories, Application note 118, Rev. 1.3(7/13), 2013. [online] (Updated December 19, 2003). Available at https://www.cypress.com/file/236481/download. [Accessed: 03-Apr-2019].

[3] B. Liu, Y. Zha, and T. Zhang, “D-Q frame predictive current control methods for inverter stage of solid state transformer,” *IET Power Electronics*, vol. 10, no. 6, pp. 687–696, 2017.

[4] L. Yang, Y. Chen, A. Luo, K. Huai, L. Zhou, X. Zhou, W. Wu, W. Tan, and Z. Xie, “A Double Update PWM Method to Improve Robustness for the Deadbeat Current Controller in Three-Phase Grid-Connected System,” *Journal of Electrical and Computer Engineering*, vol. 2018, pp. 1–13, 2018.

[5] S. N. Vukosavic, L. S. Peric, and E. Levi, “AC Current Controller with Error-Free Feedback Acquisition System,” *IEEE Transactions on Energy Conversion*, vol. 31, no. 1, pp. 381–391, 2016.

[6] Y. Changzhou, L. Chun, W. Qionglong, Z. Weitang, L. Sicong, and Z. Xing, “Implementation of multi-sampling current control for grid-connected inverters using TI TMS320F28377x,” *2017 32nd Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, 2017.

[7] C.-C. Kuo and Y.-Y. Tzou, “FPGA predictive control for single-phase active NPC grid inverters with multi-sampling technique,” *IECON 2016 - 42nd Annual Conference of the IEEE Industrial Electronics Society*, 2016.

[8] D. G. Holmes, T. A. Lipo, B. P. Mcgrath, and W. Y. Kong, “Optimized Design of Stationary Frame Three Phase AC Current Regulators,” *IEEE Transactions on Power Electronics*, vol. 24, no. 11, pp. 2417–2426, 2009.

[9] S. N. Vukosavic, *Grid-Side Converters Control and Design*. Cham: Springer International Publishing., 2018.

[10] N. Hoffmann, F. W. Fuchs, and J. Dannehl, “Models and effects of different updating and sampling concepts to the control of grid-connected PWM converters—A study based on discrete time domain analysis.” *Proceedings of the 2011 14th European Conference on Power Electronics and Applications*, pp. 1-10, 2011.

[11] L. Corradini, W. Stefanutti, and P. Mattavelli, “Analysis of Multisampled Current Control for Active Filters,” *IEEE Transactions on Industry Applications*, vol. 44, no. 6, pp. 1785–1794, 2008.

[12] S. Golestan, M. Ramezani, J. M. Guerrero, F. D. Freijedo, and M. Monfared, “Moving Average Filter Based Phase-Locked Loops: Performance Analysis and Design Guidelines,” *IEEE Transactions on Power Electronics*, vol. 29, no. 6, pp. 2750–2763, 2014.

[13] E. Tedeschi, P. Mattavelli, D. Trevisan, and L. Corradini, “Repetitive Ripple Estimation in Multi-sampling Digitally Controlled dc-dc Converters,” *IECON 2006 - 32nd Annual Conference on IEEE Industrial Electronics*, 2006.

[14] B. C. Baker, “Anti-aliasing, analog filters for data acquisition systems. *AN699*,” *Microchip Technology Inc*, pp. 1-10, 1999.

[15] J. Karki, “Active low-pass filter design,” *Texas Instruments application report*, pp. 1-24, 2000.

[16] B. M. Ewer, SIGNAL PATH designer®, “Selecting Amplifiers, ADCs, and Clocks for High-Performance Signal Paths,” *Literature Number: SNOA866*, pp. 1-10.

[17] B. C. Baker, “Select the Right Operational Amplifier for your Filtering Circuits - Analog Design Note ADN003,*” Microchip Technology Inc*, pp 1-2, 2003.

[18] “LFx5x JFET Input Operational Amplifiers”, Texas Instruments, Datasheet, Rev. 11/2015, [online] Available at http://www.ti.com/lit/ds/symlink/lf356.pdf. [Accessed: 03-Apr-2019].

[19] M. Ndawula, S. Djokic, and I. Hernando-Gil, “Reliability Enhancement in Power Networks under Uncertainty from Distributed Energy Resources,” *Energies*, vol. 12, no. 3, p. 531, 20

Practical implementation of voltage dip, swell and interruption detection algorithm according to IEC 61000-4-30 standard

Lazar Sladojević, Miodrag Stojanović and Vladeta Milenković

***Abstract*—This paper describes a practical implementation of a part of IEC 61000 standard for Electromagnetic compatibility, part 4-30 which considers Power quality measurement methods for various power quality parameters. The developed algorithm implements real-time digital signal processing techniques for accurate detection and evaluation of the supplying voltage dips, swells and interruptions as key power quality parameters. Necessary prerequisites for this process such as frequency evaluation and hardware configuration are also briefly explained. The signal processing algorithm is written in Python programming language which is open-source and easy to use. Experimental results are evaluated to give some insights on practical problems that should be addressed in the future.**

***Index Terms*—Power Quality; Voltage Dip, Swell, Interruption; Digital Signal Processing; Python**

I. Introduction

Power Quality (PQ) is becoming an increasingly important problem in electrical power systems. The deviation of supplying voltage from the ideal sine wave in many areas is now evident more than ever. This is most obvious in areas with extensive usage of electronic devices, especially big computer centers and factories with large number of electronically controlled drives. These devices generate higher order harmonics which negatively impact the supplying power grid, generating unnecessary losses [1].

However, these are not the only problems in the electrical energy supply chain. Buyers of electrical energy expect suppliers to deliver voltage in a form of a clean sine wave with the rated magnitude and frequency. This is not an easy task, since the buyers are usually the ones that generate harmonics into the power grid [1].

However, there are some aspects that solely depend on the power suppliers. It is their responsibility to ensure reliability of the supplying power and to prevent power interruptions which could severely damage production in industrial facilities, normal functioning of important institutions or even human survivability in hospitals. It has been known for a long time that even the slightest variations in the supply voltage can represent a great deal of failures in various electrical devices [2]. Low reliability also affects residential loads such as common electrical household devices [3].

Besides technical, poor power quality can also produce negative economic effects. It can lead to unplanned interruptions in production lines, increased expenses for electrical energy due to increased losses, premature failure of equipment and therefore unforeseen capital investments etc. [4]. Therefore, it is very important to observe and detect all of these abnormal situations in order to find their root and be able to prevent them in the future.

There is a total of twelve power quality parameters defined in the IEC 61000-4-30 [5] standard that should be addressed, measured and evaluated. The standard defines measurement methods as well as measurement uncertainty and measurement range for each individual parameter, except for measurement of Flickers, Harmonics and Interharmonics, which are described in separate standards. Two classes of measurement equipment are recognized, class A and class S and the main difference between them is the demanded accuracy, with class A being more accurate than class S.

This paper will address the problem of real-time detection and evaluation of supplying voltage dips, swells and interruptions. A data acquisition platform has been developed and used in conjunction with fast ARM processor for the necessary calculations. The goal is to achieve real time detection and evaluation of these events and to present results of practical implementation of the methods proposed in standard.

II. Parameters Description

The IEC 61000-4-30 standard defines voltage dip as a "temporary reduction of the voltage magnitude at a point in the electrical system below a dip threshold" [5]. This threshold is usually a percentage of the rated voltage, but in the medium and high voltage systems a sliding voltage reference can be used too. The sliding voltage reference is one-minute averaged voltage magnitude preceding the dip or swell. In this paper, the threshold was simply a percentage of the rated voltage. This threshold is usually in range of 80 to 90 percent of the rated voltage.

Voltage swell is a "temporary increase of the voltage magnitude at a point in the electrical system above a swell threshold" [5]. This threshold is once again a percentage of the rated voltage or a sliding voltage reference. As in the case

Lazar Sladojević is with Faculty of Electronic Engineering, University of Niš, Aleksandra Medvedeva 14, 18000 Niš, Serbia (e-mail: lazar.sladojevic@elfak.rs).

Miodrag Stojanović is with Faculty of Electronic Engineering, University of Niš, Aleksandra Medvedeva 14, 18000 Niš, Serbia (e-mail: miodrag.stojanovic@elfak.ni.ac.rs).

Vladeta Milenković is with Faculty of Electronic Engineering, University of Niš, Aleksandra Medvedeva 14, 18000 Niš, Serbia (e-mail: vladeta.milenkovic@elfak.ni.ac.rs).

of dip, the first type of threshold was also used for swell detection. The threshold value is normally greater than 110 percent of the rated voltage.

Voltage interruption is "reduction of the voltage at a point in the electrical system below the interruption threshold" [5]. The standard defines that this threshold can only be a percentage of the rated voltage. This threshold is much lower than the dip threshold, and is usually in range of 5 to 10 percent of the rated voltage. Interruption can therefore be observed as a special case of voltage dip, but it should still be separated, at least by post processing.

The above definitions refer to single phase systems, but can also be applied to polyphase systems with some restrictions. For example, in a three-phase system, a voltage dip begins when the voltage magnitude of one phase drops bellow the dip threshold and ends when the magnitudes of all phases increase above the same threshold augmented by the dip hysteresis. Similarly, swell begins when voltage magnitude of one phase raises above the swell threshold and ends when magnitudes of all phases drop below the same threshold reduced by the swell hysteresis value. Unlike these two parameters, voltage interruption begins when magnitude of all three phases drop below the interruption threshold and ends when voltage of any of the phases exceeds interruption threshold plus the interruption hysteresis. This implies that any event can begin on one phase and end on another.

All hysteresis values are introduced to prevent a single event to be counted multiple times which can occur if the magnitude oscillates around the threshold value. Typical value for all hysteresis is 2 percent of the rated voltage [5].

III. Detection Method

The standard defines that the sliding Root Mean Square (RMS) value is used for events detection. RMS is equivalent to magnitude since there is a linear relationship between the two when the signals are simple sine waves. The RMS value is obtained over the window which is one full cycle long and slides half of cycle when updated. This means that the first RMS value will be obtained from the whole first cycle, second one will be calculated from the second half of first cycle and first half of second cycle, third value is calculated from the whole second cycle and so on. First RMS value for each channel should be synchronized with that phase first zero cross. For all three types of events, the event start is timestamped with the start of window whose RMS value initiated that event i.e. the first RMS value that crosses the corresponding threshold if the event is dip or swell or the last one if the event is interruption. Similarly, the event end is timestamped with the end of window whose RMS value ended that event i.e. the last RMS value that returned to normal operating limits including the event hysteresis if the event is dip or swell or the first one whose value is higher than the interruption threshold plus hysteresis if the event is interruption. The accuracy class A of the standard states that maximum error in event detection time can be one full cycle i.e. ±1/2 cycle before event and ±1/2 cycle after event.

The detection method described above implicitly requires three things: RMS value calculation, cycle duration calculation and periodic synchronization with the zero cross for each phase.

A. RMS values calculation

The RMS calculation is very straight-forward, for the sampled voltage signal of N samples, RMS value is simply obtained from (1):

$$U_{RMS} = \sqrt{\frac{\sum_{i=1}^{N} U_i^2}{N}} \tag{1}$$

In (1), U_i is the instantaneous value of sample with index i.

B. Cycle duration estimation

The necessary number of samples N is equal to number of samples in one full cycle. This number can be calculated from the frequency of signal, f and the sampling frequency f_s:

$$N = round\left(\frac{f_s}{f}\right). \tag{2}$$

There are number of frequency estimation methods, some of those are detailed in [6–9], but for this application a simple zero-crossing detection method will be used.

This method utilizes linear regression over the samples around the first and the last zero cross of a signal. Samples involved in the linear regression are the ones with values between the fixed upper and lower threshold. This way, a variable number of samples is used for each linear regression, because the signal always contains some noise. This noise can cause multiple zero crosses and non-equal number of positive and negative samples. Here, linear regression behaves as some sort of a filter, because it constructs the straight line which minimizes the sum of squared distances from all samples involved in regression,

$$\min \sum_{i=1}^{M} \left(U_i - \hat{U}_i\right)^2 = \min \sum_{i=1}^{M} \left(U_i - \left(a \cdot t_i + b\right)\right)^2 . \tag{3}$$

In (3), $\hat{U}_i$ is the projection of sample U_i on the regression line, M is number of samples involved in linear regression, a is the line slope coefficient, b is line intercept coefficient and t_i is time point at which the sample was taken.

The approach is fairly simple and computationally effective, both with processing time and memory requirements, but can give errors when applied to highly distorted signals. In that case, IEC 61000-4-30 standard recommends filtration of a signal before estimation of its frequency. However, primary goal of this paper is developing and evaluation of practical algorithm for detection of voltage dips, swells and interruptions, and for simplicity, that will be done on fairly clean, sine waveform signals. This assumption

is not always fulfilled in practice but processing of highly polluted signal would require a whole other paper, and will be subject of author's future work.

C. Periodic zero cross synchronization

The third requirement is periodic synchronization with each phase zero cross. The standard requires that the processed RMS value be calculated over the window which starts where that phase crosses zero and is one full cycle long. This window is moving half a cycle at each iteration. Therefore, only first zero cross is calculated and other windows' positions for RMS calculation can be estimated. During time however, a small frequency variation takes place in any electrical system and therefore the cycle duration varies slightly. If the synchronizations were not conducted periodically, error in window position would become too large to comply with standard. However, if the voltage interruption occurs, no frequency can be estimated from voltage signal and therefore window position can only be estimated from its last known position.

IV. Experimental Results

For experimental purposes, a data acquisition and processing platform has been developed. The platform contains 16-bit Analog to Digital (A/D) converter with the 25.6 kHz sampling frequency for data acquisition and ARM Cortex-A8 AM335X microprocessor [10] for digital processing of the gathered data. Platform is capable of sampling and processing of all three voltage channels simultaneously.

The algorithm itself was entirely written in Python programming language. Python was chosen because it is simple and open source programming language, yet very powerful when it comes to scientific computing. It runs on Linux operating system which is embedded in the platform.

The algorithm was developed to comply with class A measurement instrumentation from IEC 61000-4-30 standard. Sets of sampled data from each voltage channel arrive in packages of 25600 samples, which means they are sampled for one second before they are processed. This implies that the algorithm has to process the whole previous set of samples in time shorter than one second, before the new samples arrive. Therefore, the algorithm has to be highly optimized. Besides events detection, each second a new signal frequency is estimated and zero cross synchronization is conducted.

For test voltage generation, Omicron CMC 356 test set [11] was used. This test set is primarily intended for relay protection testing but it can as well serve for other purposes, such as this one. The algorithm has been tested on various cases of voltage dips, swells and interruptions. Some of the test cases are shown in Figs. 1 to 5. The rated Phase-to-Neutral voltage in all scenarios was 230 V_{rms}, with frequency of 50 Hz. All test cases contained a repeatable sequence of several signal states and those signals were recorded for 10 seconds. No synchronization between beginning of signal and beginning of recording existed, i.e. examined signal had been present for a random amount of time before the recording started.

Fig. 1 depicts the case of single-phase voltage dip occurring on phase A. The test signal is a repeated sequence of three states: one second rated voltage on all three phases, then 150 ms of rated voltage on phases B and C and 2/5 of rated voltage on phase A, and finally one more second of rated voltage on all phases. One of the several captured dips is shown in the figure, with black vertical lines representing dip start and end times. It is obvious that the dip beginning and end were captured at the right moments, where phase A voltage waveform crosses zero and its RMS values cross corresponding thresholds. Of course, no perfect zero cross can be detected due to quantization error, finite sampling rate and numerical errors in calculation, but the achieved precision was well above class A minimum. Real dip duration can be obtained from the figure by counting number of cycles from dip trigger point up to ending point. Since there are 9 full cycles, and each cycle is 20 ms long, total duration of dip was 180 ms. Calculated duration is obtained as a difference between end time and begin time and is equal to 180.35 ms. The absolute error is therefore 0.35 ms which is well below requested 20 ms.

Fig. 2 represents two phase voltage dip which begins on phase C and ends on phase B. The test signal consists of five states: one second of rated voltage on all phases, 100 ms of voltage dip on phase C, 100 ms of voltage dip on phases C and B, 100 ms of voltage dip only on phase B and finally one second of rated voltage on all phases. It can be seen that once again, dip begin and end times are estimated in the correct moments, when corresponding phases cross dip thresholds. At one point, voltage on phase C which initiated the dip returns to rated value, but there is still voltage dip on phase B, so no dip end is detected there. By counting full cycles and having in mind the phase difference between the voltage signals, real dip duration can be obtained. Since there are 16 plus 2/3 of the full cycle, real duration is 333.3333 ms, while the estimated duration is equal to 333.67 ms. The absolute difference is 0.3367 ms which is once again below the requested 20 ms.

Similar conclusions can be drawn from Figs. 3 and 4, where the only difference was that the event type is voltage swell. It can be observed that both single phase (Fig. 3) and two phase (Fig. 4) voltage swell are detected as precisely as two previous dips. The absolute differences between real and estimated duration of single phase and two-phase voltage swells are 0 and 0.067 ms, respectively. These errors are lower than those obtained during dip detection, but this is only due to fact that swells occurred near the beginning of one second buffer and dips occurred near the end of buffer. The zero error obtained in detection of single phase swell is a consequence of rounding the begin and end times to certain precision. Of course, such precision can never really be achieved in practice. Nevertheless, results are very good and satisfy the class A accuracy demands.

The special case is voltage interruption, presented in Fig. 5.

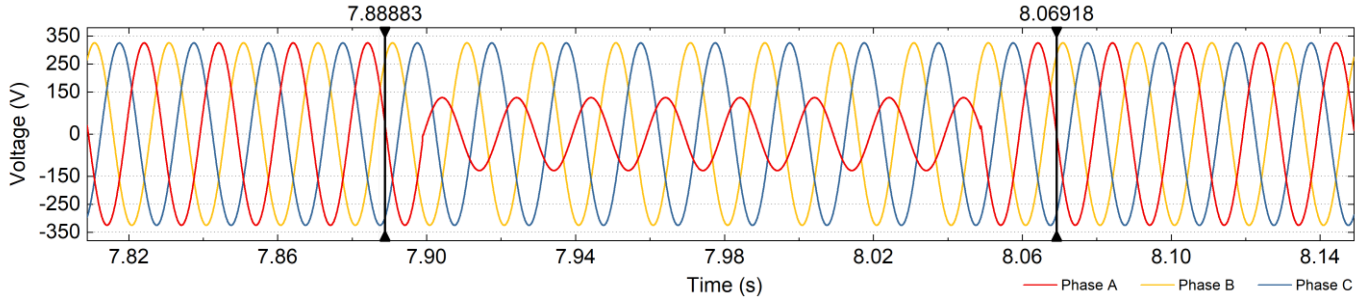


Fig. 1. Phase A voltage dip detection times

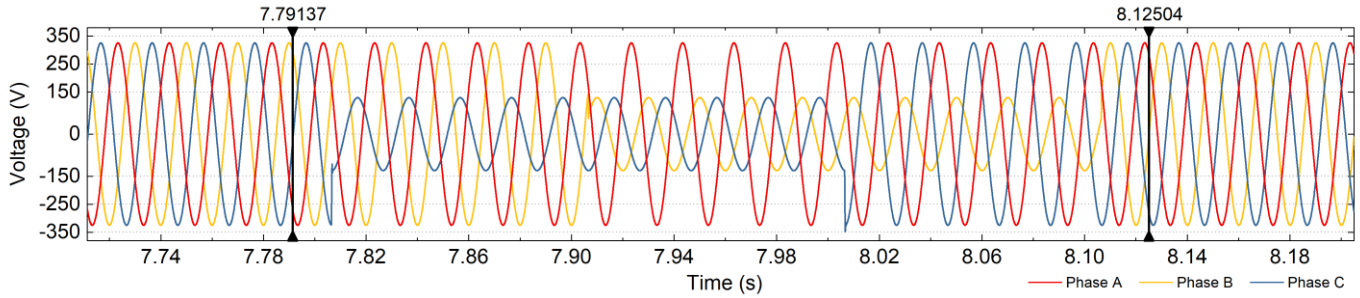


Fig. 2. Two phase voltage dip detection times

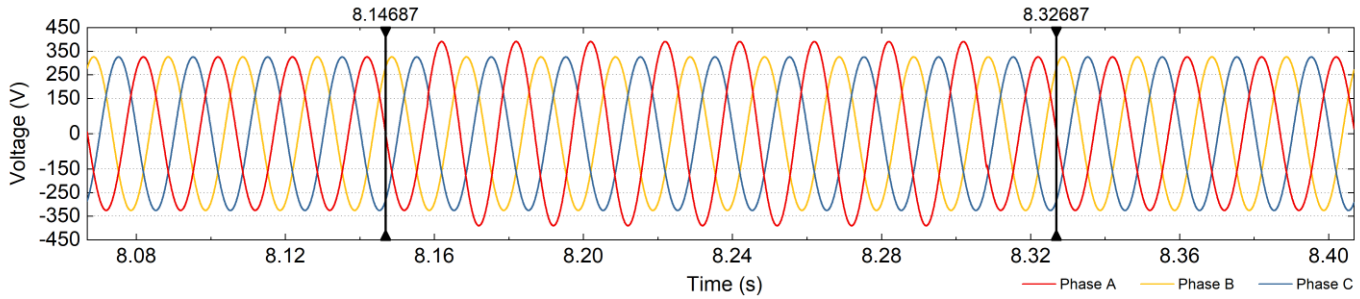


Fig. 3. Phase A voltage swell detection times

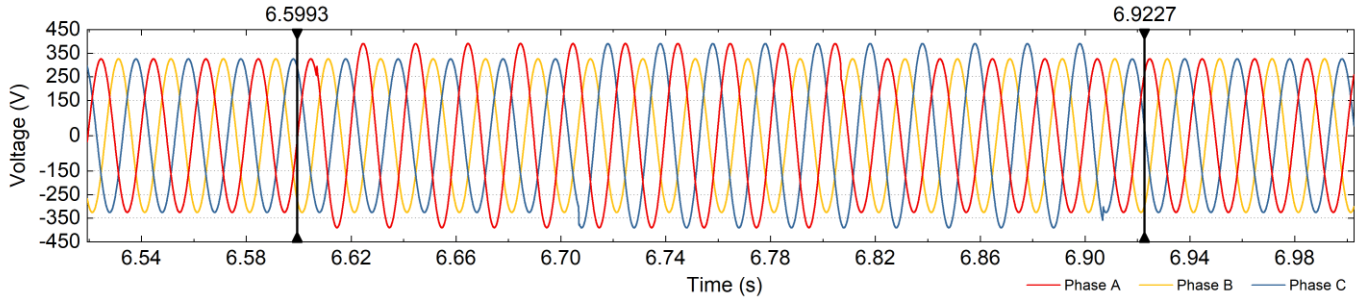


Fig. 4. Two phase voltage swell detection times

Voltage interruption begins when all three phases' voltage RMS values drop below the interruption threshold and ends when at least one raises above the threshold plus hysteresis value. Since all three phase voltages drop to zero at the same time, it is very difficult to estimate the real beginning of interruption. That would require interpolating all three sine waveforms and visually identifying their zero-cross times. However, it can be seen that interruption end was estimated precisely, because it was timestamped where phase B voltage first appears after the interruption. Of course, there have been voltage dips immediately before and after the interruption. These dips have also been detected, but their detection times

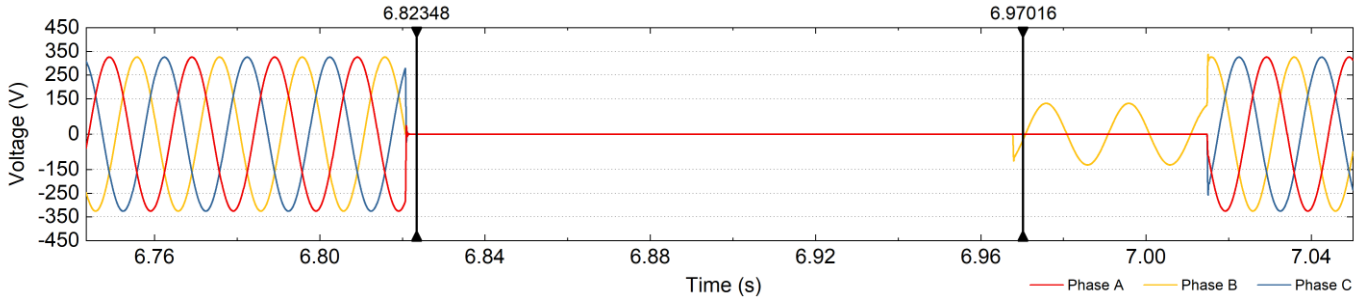


Fig. 5. Interruption detection times

are not shown in Fig. 5 for the sake of clarity. In fact, according to standard's event definitions, there are always a preceding and succeeding dips around interruption event.

The average time needed for processing of one second data buffer was also measured. The measurement results are shown in Table I.

TABLE I
DATA PROCESSING TIMES FOR DIFFERENT EVENT TYPES

Event type	**Average processing time for one buffer of data [s]**
Single phase dip	0.6719
Two phase dip	0.6647
Single phase swell	0.6710
Two phase swell	0.6700
Interruption	0.6605
Total average time	**0.6676**

From Table I it is obvious that all processing times are shorter than one second which is the time available for one buffer of samples to be processed before the new one arrives. This means that the algorithm is capable of conducting real-time events detection and evaluation.

V. CONCLUSION

In this paper a software algorithm for detection and evaluation of supply voltage dips, swells and interruptions was developed. The goal of algorithm was to achieve real time detection of aforementioned irregular events. Measurement methods and accuracy requirements are given in IEC 61000-4-30 standard and were only briefly discussed here. The overall main goal of this paper is not to propose new methods for events detection, but to evaluate practically obtained results with the methods already described in standard. For practical purposes, new data acquisition platform was developed and tested. Five typical cases of voltage dips, swells and interruptions were examined and experimental results are presented. It could be concluded that the proposed algorithm is capable of real-time dip, swell and interruption detection and evaluation with very high precision which complies with class A instrumentation from the standard. However, there are still some drawbacks for this algorithm. The main difficulty is estimation of frequency and zero crossing times for examined signals. In this paper only clean sinusoidal signals were examined, but in reality, these signals can contain a large amount of higher order harmonics which can make calculations either very unprecise or even impossible. Research in this field and improvement of the already developed algorithm will be subject to authors future studies.

ACKNOWLEDGMENT

This research is partly supported by project grant III44006 financed by the Ministry of education, science and technology development of the Republic of Serbia.

REFERENCES

[1] R. Dugan, M. McGranaghan, S. Santoso, H. Beaty, *Electrical Power Systems Quality*, 3rd edition, New York, USA, McGraw-Hill, 2012.
[2] R. P.Bingham, "SAGs and SWELLs", Dranetz-BMI, Feb. 1998.
[3] G. Karady, S. Saksena, B. Shi, N. Senroy, "Effects of Voltage Sags on Loads in a Distribution System", Power Systems Engineering Research Center, Publication 05-63, Oct. 2005.
[4] N. Edomah, "Effects of voltage sags, swell and other disturbances on electrical equipment and their economic implications", Electricity Distribution - Part 1, CIRED, Proceedings of the 20th International Conference and Exhibition, pp. 1-4, Jun. 2009.
[5] IEC 61000 International Standard – Electromagnetic Compatibility, Part 4-30: Testing and Measurement Techniques, - Power Quality Measurement Methods, edition 3.0, 2015.
[6] M. S. Sachdev and M. M. Giray, "A least error squares technique for determining power system frequency," IEEE Trans. Power App. Syst., vol. PAS-104, no. 2, pp. 437–444, Feb. 1985.
[7] H. Karimi, M. Karimi-Ghartemani, and M. R. Iravani, "Estimation of frequency and its rate of change for applications in power systems," IEEE Trans. Power Del., vol. 19, no. 2, pp. 472–480, Apr. 2004.
[8] M. Djuric, Z. Djurisic, "Frequency measurement of distorted signals using Fourier and zero crossing techniques", Electric Power Systems Research Volume 78, Issue 8, pp. 1407-1415, Aug. 2008.
[9] E. Lavopa, P. Zanchetta, M. Sumner, and F. Cupertino, "Real-Time Estimation of Fundamental Frequency and Harmonics for Active Shunt Power Filters in Aircraft Electrical Systems", IEEE Transactions on Industrial Electronics, Vol. 56, No. 8, pp. 2875-2884, Aug. 2009.
[10] AM335xand AMIC110Sitara™ Processors Technical Reference Manual, Literature Number: SPRUH73P, October2011 – Revised March 2017
[11] CMC 356 Reference Manual, Article Number VESD2003 - Version CMC356.AE.

Rotor bars skewing impact on electromagnetic pulsations in cage induction motor

Gojko Joksimović[1], Aldin Kajević[1], Saša Mujović[1], Tatjana Dlabač[2],

Vanja Ambrožič[3], Alberto Tessarolo[4]

***Abstract* — An alternative way of integral skew factor derivation that provides a deep insight into electromagnetic processes in machine and clearly demonstrates the basic idea behind the skewing of rotor bars in cage induction motors is presented in the paper. Detailed analysis of impact of skewing of rotor bars on elimination of stator slot harmonics in rotor bar currents is conducted. Additionally, by using the multiple coupled circuit model of cage induction motor impact of rotor bar skewing on electromagnetic torque pulsations in steady state condition is conducted in natural frame of reference. Impact of skewing of rotor bars on motor starting ability in case of inadequate stator slot/rotor slot combination is illustrated, too.**

***Keywords* — Cage rotor induction motor, Skewing of rotor bars, Skew factor, Electromagnetic torque ripple, Steady-state condition**

I. INTRODUCTION

SKEWING of rotor slots (rotor bars) in a cage induction motor is well known method for attenuation of higher time harmonics in rotor bar currents due to the higher space harmonics in rotating magnetic flux density wave. Unfortunately, authors did not find the answer on question when and where this measure is firstly proposed and applied in commercially available induction motors.

As it is well known, three-phase stator windings produce rotating magnetomotive force (mmf) wave that beside the fundamental harmonic contain higher space harmonics of order $\nu=6k+1$, where $k=0,\pm1,\pm2...$ Fifth and seventh harmonic, phase belt harmonics, are consequence of the trapezoidal like mmf shape. Both of them could be significantly attenuated by appropriate shorting of stator winding coils. The most common solution is shortening of coil pitch for one sixth of the pole pitch. However, this measure has no impact on stator slot harmonics that are the most significant space harmonics in the rotating mmf wave from stator side, [1]. The order of these harmonics is $Q_S/p\pm1$, where Q_S is number of stator slots.

[1]Gojko Joksimović, Aldin Kajević, Saša Mujović – Faculty of Electrical Engineering, University of Montenegro, Cetinjski put b.b., 81000 Podgorica, Montenegro (e-mail: Gojko.Joksimovic@ucg.ac.me).

[2]Tatjana Dlabač – Maritime Faculty, University of Montenegro, Dobrota 36, 85330 Kotor, Montenegro.

[3]Vanja Ambrožič, – Faculty of Engineering, University of Ljubljana, Tržaška cesta 25, 1000 Ljubljana, Slovenia.

[4]Alberto Tessarolo, – Department of Engineering and Architecture, University of Trieste, Via Valerio 10, 34127 Trieste, Italy.

As any other of mmf space harmonics, through action across the assumed uniform air gap, relatively strong stator slot harmonics could induce higher time harmonics in rotor windings i.e. rotor bars. As a result, parasitic pulsation torques will be developed that have as a result higher vibrations of the machine, acoustic noise, higher losses and lower efficiency. In order to eliminate these high frequency components in rotor currents, skewing of rotor bars is commonly used. By this way stator slot harmonics in rotor currents could be significantly attenuated. On the other side, skewing of rotor bars has a small deteriorating effect on fundamental frequency emf induced in rotor bars i.e. weaker electromagnetic coupling of stator and rotor windings. Additionally, skewing of rotor bars as a side effect has appearance of inter-bar currents in cage rotors with uninsulated bars, [2]. Existence of these currents yields worse machine efficiency as well as appearance of axially directed forces on the rotor.

Quantitative measure of skewing of rotor bars is integral skew factor that is ratio of induced emf in skewed rotor bar and emf in straight rotor bar. In many electrical machine books skewing of rotor bars is not mentioned at all. In some of them skew factor is given in its final form, without explanation how this factor is obtained. In some books there exists derivation of skew factor that is based on vectorial summation of emfs along the rotor bar, [3]. Somewhere, skew factor is derived using the definition of winding pitch factor, [4].

Here, derivation of skew factor is given on clear and easy-to-follow physical explanation that is based on fundamental law of electromagnetic induction – flux cutting rule.

II. INTEGRAL SKEW FACTOR DERIVATION

Fig 1. shows frozen instant of time when one skewed rotor bar is beneath the maximum value of magnetic flux density wave from stator side. Without losing generality, it is assumed that magnetic flux density wave is stationary while rotor moves with constant tangential velocity v. For taking into account the most general case, magnetic flux density wave is described by its ν^{th} harmonic in $2p$ pole machine,

$$B(\theta)=B_m\cos(\nu p\theta) \tag{1}$$

where θ is a mechanical angle measured along the machine circumference. From the following proportion,

$$x : 2p\tau = \theta : 2\pi \tag{2}$$

where x is linear distance along the machine circumference and τ is the pole pitch, magnetic flux density could be written down as a function of x:

$$B(x) = B_m \cos\left(\frac{\nu\pi}{\tau}x\right) \tag{3}$$

From the flux cutting law, induced electromotive force (emf) along the infinitesimal length dl of skewed rotor bar is,

$$de_{skew} = d\mathbf{l}\cdot(\mathbf{v}\times\mathbf{B}) \tag{4}$$

i.e.

$$de_{skew} = vB_m \sin\alpha\cos\left(\frac{\nu\pi}{\tau}x\right)dl \tag{5}$$

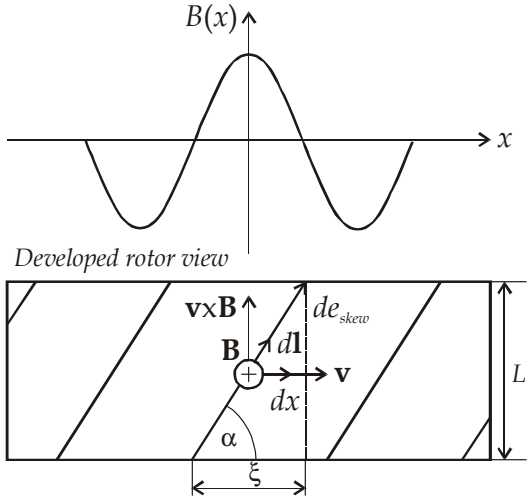


Fig 1. Induced emf in skewed rotor bar

The last expression is equivalent to:

$$de_{skew} = vB_m \tan\alpha\cos\left(\nu\frac{\pi}{\tau}x\right)dx \tag{6}$$

By integration of above expression between the limits x=–ξ/2 and x=ξ/2, all elementary induced emfs are taken into account:

$$e_{skew} = vB_m \tan\alpha \int_{-\xi/2}^{\xi/2} \cos\left(\nu\frac{\pi}{\tau}x\right)dx \tag{7}$$

$$e_{skew} = \frac{v\cdot B_m \tan\alpha}{\nu\pi/2\tau}\sin\left(\nu\frac{\pi}{\tau}\frac{\xi}{2}\right) \tag{8}$$

Finally, as $\tan\alpha = L/\xi$ (Fig 1):

$$e_{skew} = B_m Lv\frac{\sin\left(\frac{\nu\pi\xi}{2\tau}\right)}{\frac{\nu\pi\xi}{2\tau}} \tag{9}$$

As induced emf in straight rotor bar in observed instant of time is $e=B_mLv$, ratio of induced emf in skewed and straight bar is defined as integral skew factor, for ν^{th} harmonic:

$$k_{skew\_\nu} = \frac{\sin\left(\frac{\nu\pi\xi}{2\tau}\right)}{\frac{\nu\pi\xi}{2\tau}} \tag{10}$$

III. CHOOSING APPROPRIATE SKEW ANGLE

With a view to eliminate stator slot harmonic of order $\nu = Q_S/p - 1$, skew factor should be equal to zero for that harmonic, i.e.

$$\left(\frac{Q_s}{p} - 1\right)\frac{\pi\xi}{2\tau} = n\pi \tag{11}$$

where n is an integer. It means that the distance between the start and end point of rotor bar observed from the opposite rotor sides should be (Fig 1):

$$\xi = \frac{2n\tau}{\frac{Q_s}{p} - 1} = \frac{n\pi D}{Q_s - p} \tag{12}$$

For the smallest integer for which the above condition is satisfied, n=1, follows:

$$\xi = \frac{\pi D}{Q_s - p} \tag{13}$$

Similarly could be obtained from the cancellation conditions for second one stator slot harmonic, of order $\nu = Q_S/p + 1$:

$$\xi = \frac{\pi D}{Q_s + p} \tag{14}$$

In order for both the slot harmonics to be maximally attenuated in induced emf, wherein none of them does not eliminate completely, mean of two upper values should be chosen,

$$\xi = \frac{\pi D}{Q_s} = \tau_s \tag{15}$$

that is obviously equal to the stator slot pitch, τ_S, and additionally, does not depend on the number of pole pairs. Previous expression could be given in mechanical radians, so, desirable skew angle is:

$$\xi = \frac{2\pi}{Q_S} \tag{16}$$

By substitution of (15) into (10) skew factor for the ν^{th} harmonic is obtained for special case when skewing of rotor bars corresponds to the one stator slot pitch:

$$k_{skew\_\nu} = \frac{\sin\left(\frac{\nu p\pi}{Q_S}\right)}{\frac{\nu p\pi}{Q_S}} \tag{17}$$

This way, the induced emf in rotor bar due to the fundamental flux density wave in air gap is slightly deteriorated – skew factor for fundamental harmonic has value slightly less then one,

$$k_{skew\_1} = \frac{\sin\left(\frac{p\pi}{Q_S}\right)}{\frac{p\pi}{Q_S}} \tag{18}$$

while for both slot harmonics its value is close to zero:

$$k_{skew\_\nu=\frac{Q_S}{p}\pm 1} = \frac{\sin\left(\pi \pm \frac{\pi p}{Q_S}\right)}{\pi \pm \frac{\pi p}{Q_S}} = \frac{\mp \sin\left(\frac{\pi p}{Q_S}\right)}{\pi \pm \frac{\pi p}{Q_S}} \tag{19}$$

Fig 2. illustrates the effect of rotor bar skewing for one stator slot pitch. Along the skewed rotor bar emfs of opposite polarities are induced due to action of stator slot harmonic in flux density wave so the resultant emf in rotor bar approaches the zero value, as has already been shown analytically.

As an illustration, following case should be observed: four pole motor, p=2, with Q_S=48 stator slots. Skew factor for fundamental harmonic is,

$$k_{skew\_1} = \frac{\sin\left(\frac{2\pi}{48}\right)}{\frac{2\pi}{48}} = \frac{24\sin\left(\frac{\pi}{24}\right)}{\pi} = 0.9971$$

i.e. induced emf of basic frequency is smaller 0.3% due to the skewing of rotor bars. As skew factor for slot harmonics are,

$$k_{skew\_\nu=\left(\frac{Q_S}{p}\pm 1\right)} = k_{skew\_23,25} = \begin{cases} \frac{24\sin\left(\frac{23\pi}{24}\right)}{23\pi} = 0.0434 \\ \frac{24\sin\left(\frac{25\cdot\pi}{24}\right)}{25\cdot\pi} = -0.0399 \end{cases}$$

it is clear that induced emfs due to these harmonics are smaller by approximately 96%, what practically means that these emfs are almost completely eliminated in rotor bars.

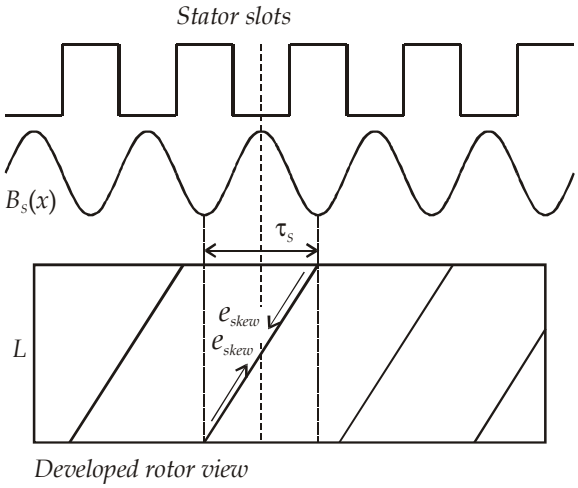


Fig 2. Skewing effect on induced emf in rotor bar for skewing angle of one stator slot pitch

By choosing higher value for n in (12), higher skewing angle will be obtained that is equal to integer multiple of stator slot pitch. The same effect regarding stator slot harmonics will be obtained but fundamental frequency emf in rotor bar will be significantly deteriorated and higher inter-bar currents will results as consequence of that choice. So, skew that achieves the effect of attenuation of stator slot harmonics with the small as a possible side effect is skew for one stator slot pitch.

IV. SKEWING EFFECT ON STATOR-ROTOR MUTUAL INDUCTANCE

In static induction motor model i.e. in standard equivalent single-phase circuit of induction motor skewing has an impact on value of magnetizing reactance [4]. Its value is smaller for integral skew factor in comparison with straight rotor bar case.

Skewing of rotor bars in induction motor numerical models is commonly taken into account by technique of slicing of rotor structure in n segments. In that case every of n segments are observed as segments with straight rotor bar. This technique is also commonly used for calculation of integral skew factor. In such cases it is desirable that number of slices

is as high as possible what on the other side leads to very time consuming software procedures.

A time and computationally effective approach is use of winding function approach as it is demonstrated in [5]. Number of slices could be as high as it has sense without any significant influence on execution time in numerical procedure. The main idea is in the definition of mutual inductance per length between stator phase windings and rotor loops. Rotor loop consists of two side by side rotor bars with corresponding end ring segments from both front sides of the rotor.

By using technique that is described in [5], following results for mutual inductance between stator phase winding and rotor loop is obtained for straight rotor bars, Fig 3, for motor whose details are given in Appendix. Together with mutual inductance dependence, derivative of that function with respect to mechanical angle is shown. Obviously, very intensive spikes in derivative of mutual inductance curve appear, which is resulting in spikes in developed electromagnetic torque versus time, [1].

Fig 4. gives the spectral content of mutual inductance curve from Fig 3. For better visibility of higher harmonics, fundamental harmonic (the second one) is reduced by five times.

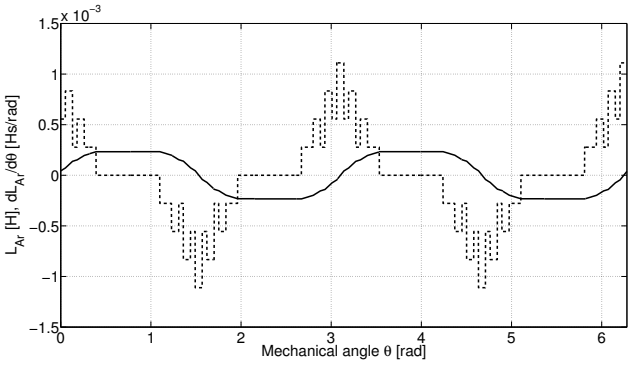


Fig 3. Mutual inductance between stator phase winding and rotor loop (solid line) and its derivative (dashed line) respect to the mechanical angle: straight rotor bars case

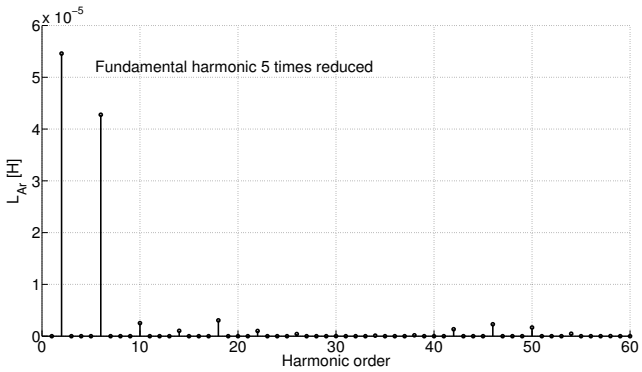


Fig 4. Spectral content of mutual inductance curve from previous figure. Amplitudes of harmonics of interest are: 1st: 0.273mH; 23rd: 2.3047μH; 25th: 1.673μH.

The case when rotor bars are skewed for one stator slot pitch is illustrated on Figs 5. and 6. Results are obtained by slicing of rotor structure axially in one hundred segments. Now, high spikes in derivative of mutual inductance disappear, Fig 5. that is additionally confirmed by spectral content of mutual inductance curve, Fig 6. By dividing the amplitudes of higher harmonics of interest by the amplitude of the main harmonic, following results could be obtained: for fundamental harmonic, 0.9971, for lower stator slot harmonic, 0.04577, for upper stator slot harmonic, 0.0431, which is in good correlation with already calculated integral skew factor for these harmonics.

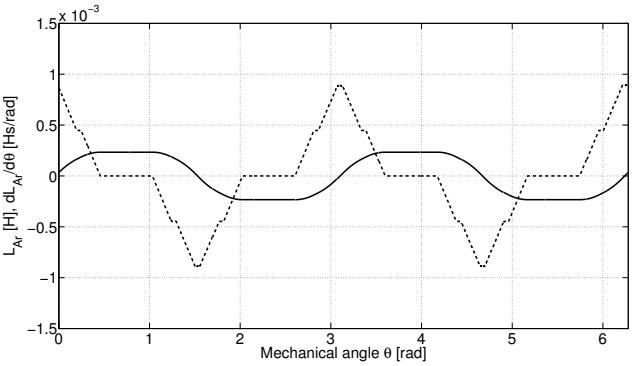


Fig 5. Mutual inductance between stator phase winding and rotor loop (solid line) and its derivative respect to the mechanical angle (dashed line): rotor bars skewed for one stator slot pitch

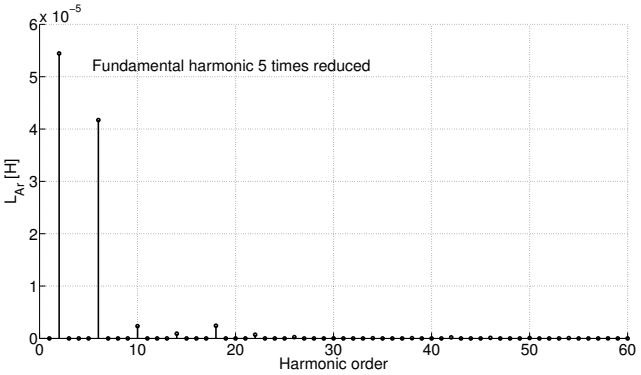


Fig 6. Spectral content of mutual inductance curve from previous figure. Amplitudes of harmonics of interest are: 1st: 0.2722mH; 23rd: 0.1055μH; 25th: 0.0721μH.

V. RESULTS FROM THE NUMERICAL MODEL

In order to illustrate the effect of skewing of rotor bars on pulsations in developed electromagnetic torque, Fig 7. gives results for two cases of rated loaded induction motor at steady state. As it is evident, significant attenuation of pulsations is a result of skewing of rotor bars.

As an additional feature of rotor bars skew is the elimination of so called cogging torque that appears as a consequence of improperly chosen number of stator slots and rotor bars. As one of drastic case of such choice is the same

number of stator slots and rotor bars, Q_S=Q_R. By using recently developed model, [6], developed electromagnetic torque and rotor speed in transient regime of starting of unloaded motor is given on Fig. 8.

As it is evident, motor with straight rotor bars can not start at all, while in case when rotor bars are skewed for one stator slot pitch motor can start without problem.

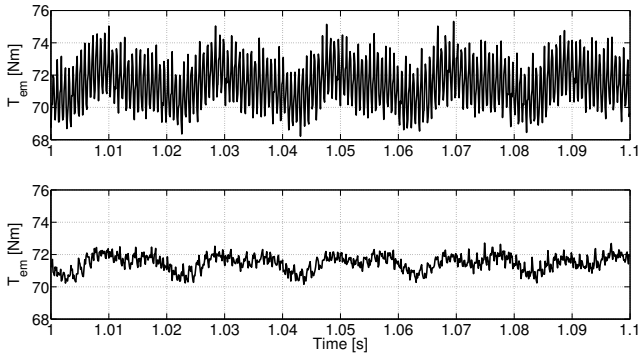


Fig 7. Developed electromagnetic torque in steady state condition: straight rotor bars case, above, and rotor bars skewed for one stator slot pitch, bellow.

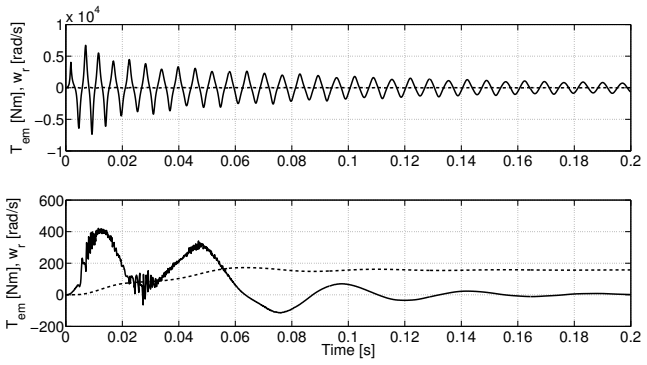


Fig 8. Electromagnetic torque (solid line) and rotor speed (dashed line) during the no-load speed up of the motor with Q_S=Q_R=48. Motor with straight rotor bars is unable to start, above, and motor with skewed rotor bars can start without problem, below.

VI. Conclusions

An alternative way of integral skew factor derivation that provides a deep insight into electromagnetic processes in machine and clearly demonstrates the basic idea behind the skewing of rotor bars in cage induction motors is presented in this paper. Detailed analysis of impact of skewing of rotor bars on elimination of stator slot harmonics in rotor bar currents is conducted. Additionally, by using the multiple coupled circuit model of cage induction motor impact of rotor bar skewing on electromagnetic torque pulsations in steady state condition is conducted in natural frame of reference. Impact of skewing of rotor bars on motor starting ability in case of inadequate stator slot/rotor slot combination is illustrated, too.

VII. Appendix

Motor rated values, electrical and geometrical parameters:

P_r=11kW, U_{LL}=400V, 50Hz, wye, I_r=17.6A, cosφ_r=0.97, η_r=0.93, n_n=1470 rev/min

Q_S=48, Q_R=30, y/τ=10/12, q=4, W_1=112, n_{coil}=7, J=0.068kgm^2

D_{is} [mm]	L [mm]	g [mm]	R_s [Ω]	$L_{\sigma s}$ [mH]	R_b [μΩ]	R_{er} [μΩ]	L_b [nH]
146.36	172.42	0.3975	0.318	2.522	60.636	1.445	401.88

References

[1] J. Faiz, V. Gorbanian, G. Joksimović, „Fault Diagnosis of Induction Motors“, book, IET, 2017.

[2] D. G. Dorrell, P. J. Holik, C. B. Rasmussen, “Analysis and Effects of Inter-Bar Current and Skew on a Long Skewed-Rotor Induction Motor for Pump Applications”, *IEEE Transactions on Magnetics*, Vol. 43, Issue 6, pp. 2534-2536, June 2007.

[3] I. Boldea, L. Tutelea, “Electric machines: steady state, transients and design with MATLAB®”, book, CRC Press, 2010.

[4] J. Pyrhönen, T. Jokinen, V. Hrabovcová, “Design of rotating electrical machines”, book, John Wiley & Sons, Ltd, 2008.

[5] G. Joksimović, M. Đurović, A. Obradović, “Skew and linear rise of MMF across slot modeling – winding function approach”, *IEEE Transactions on Energy Conversion*, vol. 14, no. 3, pp. 315-320, September 1999.

[6] G. Joksimović, “Dynamic model of cage induction motor with number of rotor bars as parameter”, *The Journal of Engineering*, IET, Vol. 2017, Issue 6, pp. 205-211, June 2017.

Mechanical and Electrical Faults Detection in Uncontrolled Drives with AC Motors

Bratislav Trojić, Vladislav Lazić, Uroš Ilić and Milutin Petronijević, *member IEEE*

***Abstract*—This paper shows new and contemporary possibilities for faults detection and operation conditions estimation of electric drives. The implementation of the machine learning and the artificial intelligence is technique that is increasingly represented in the recent times. Here it was dealt with theoretical and practical support for the implementation of machine learning for the purpose of mechanical and electrical faults detection in uncontrolled drives. The sequence of work and data flow is shown through several stages, from data processing through the developing classification technique, i.e. optimal machine learning algorithm to the evaluation of that algorithm. All development phases of the algorithm were made in MATLAB.**

***Keywords*—Machine learning; electric drives; mechanical faults; electrical faults; preprocessing; algorithm.**

I. INTRODUCTION

Electric drives with AC motors are the basis of today's and future industrial applications. The most prevalent motors are still Tesla's asynchronous motors, and for this reason, engineering progress and mass of scientific work are based on these machines. However, faults are a regular occurrence for all machines. They can be roughly divided into mechanical and electrical faults. Mechanical faults are mainly related to the misalignments [1],[2], soft foot and other mechanical unbalances [3], that lead to vibration of the machine [4]. On the other hand, electrical faults apply to short circuits, open circuits, as well as to asymmetrical power supply to the motor [5], which leads to further side effects, but here is the focus on voltage asymmetry [6],[7]. One of the most effective ways to diagnose and monitor drives is to implement machine learning for these purposes. Through the collected literature, a similar topic was dealt in various possible ways, by monitoring the data of vibrations or stator currents, on motor defects and developing algorithms for detection [8]. Some papers deal with fault monitoring in controlled drives [9],[10], and some are specifically based on the improvement of different algorithms [11],[12].

This paper should explain the process of applying the machine learning on electric power drives and to develop a diagnostic system that has the ability to accurately detect and distinguish the condition of electric drive by measuring vibrations of motor. Conditions should be able to define through three operating modes. These states are normal operating mode, mechanical and electrical faults (especially voltage asymmetry). The machine learning algorithm has been developed from representative examples of machine behavior, so that the algorithm could recognize certain modes precisely enough in the future.

The paper is organized as follows. The collecting of large data sets, with vibration measurements of motor is given in chapter II. In chapter III is presented some laboratory results of measuring. The next step, preprocessing data and feature extraction is given in chapter IV. Chapter V presents the building of the machine learning algorithm via training data. After building the model, it is shown algorithm evaluation. If the evaluation gives positive assessment, this means that the algorithm is ready for application on the new data. That is given in chapter VI. The paper ends with conclusions and the used literature.

II. DATA ACQUISITION

The usual approach for faults supervision is monitoring of vibrations or stator currents. Often, because of the nature of mechanics fault, the effect is best showed on the vibration of the motor. A cheaper method is based on the measurement of stator currents [9]. In this paper, the goal is to achieve fault detection by observing vibrations of induction motor. The reason for gathering only vibrations is because the machine learning algorithms showed better performances in distinguishing faults observing only vibrations rather than only stator currents.

The measuring data were obtained in laboratory condition. In fact, various electrical and mechanical failures were recreated, for measuring vibrations condition along vertical and horizontal axis.

A. Drive and Load Setup

Our drive consists from a main motor, whose condition we monitor and an auxiliary motor, coupled to the shaft of the main motor, in order to simulate the load. For the main motor was used induction motor, SIEMENS *1LA7096-4AA10-Z A11*. The motor characteristics are shown in the TABLE I

TABLE I
DATA FOR THREE-PHASE SQUIRREL-CAGE-MOTOR

Motor data	
Rated voltage:	230/400 V
Rated power:	1.5 kW
Rated speed:	1420 1/min
Rated torque:	10.1 Nm

In Fig. 1. is shown the measuring setup which was used in laboratory for data acquisition. The upper mentioned motor is supplied through 2 three-phase autotransformer. The load is emulated with the servo motor *BMH1401P16A1A*. Unlike the induction motor, servo motor is supplied and controlled

Bratislav Trojić is with the Department of Power Engineering, Faculty of Electrical Engineering, University of Nis, Aleksandra Medvedeva 14, 18000 Nis, Serbia (e-mail: bratislav.trojic@gmail.com).
Vladislav Lazić is with the Department of Power Engineering,Faculty of Electrical Engineering, University of Nis, Aleksandra Medvedeva 14, 18000 Nis, Serbia (e-mail: vladislav.lazic@gmail.com).
Uroš Ilić is with the Department of Power Engineering, Faculty of Electrical Engineering, University of Nis, Aleksandra Medvedeva 14, 18000 Nis, Serbia (e-mail: urosilic@elfak.rs).
Milutin Petronijević is with the Department of Power Engineering, Faculty of Electrical Engineering, University of Nis, Aleksandra Medvedeva 14, 18000 Nis, Serbia (e-mail: milutin.petronijevic@elfak.ni.ac.rs).

by the matching drive *LXM32CD30N4*. This motor had to be controlled through drive, so that it could be used as a load emulator. In order to emulate constant load on the shaft of the induction motor, servo motor is operated in **torque mode**.

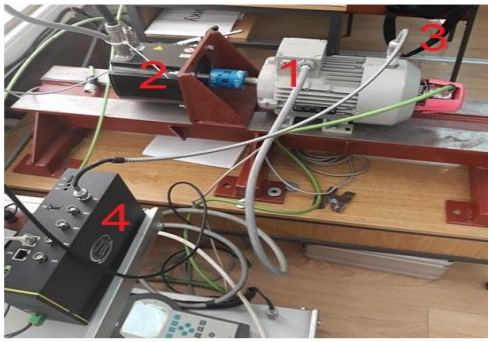


Fig. 1. The measuring setup: 1- Induction motor; 2- Servo motor for load emulation; 3- piezoelectric sensor; 4- Data acquisition device.

B. *Vibrations measuring*

The most commonly used vibration sensors nowadays are based on piezoelectric effect. Piezoelectric sensor accepts vibrations and through the transmission ratio, translates that value into *mV*, with which the device is capable of working and collecting data. The transmission ratio of used sensor is 100 *mV/g*, where is $g = 9.81\ m/s^2$.

The data acquisition device that the sensor is attached to is still under development and presents a prototype of the future series developed for similar purpose. It was programmed in program language Python, so that it could acquire vibrations and writing them in the *.txt* file. Sampling frequency of data collection is 25.6 *kHz*, for 10 *s* time, with 256 000 points per one signal.

C. *Failures Simulation*

In above-mentioned conditions, data of several different motor states were collected. First, electrical undesirable effects on the motor were simulated. Because of the laboratory conditions, it was possible only to simulate the voltage supply asymmetry. These electric faults were simulated by suppling induction motor through 2 three-phase autotransformer. By suppling one or two phases with separate autotransformer, thus, it was simulated the single-phase and two-phase asymmetry of *3%*, and then *5%*. Then, mechanical undesirable effects and faults were simulated on the motor. In our conditions it was possible to achieve soft foot, angular and parallel misalignment. These faults were set by moving motor through horizontal or vertical axis. Every simulated condition was recorded with more various load values.

Based on this recorded states of the induction motor, the data were categorized into three main conditions: normal operation, electrical faults and mechanical faults. In TABLE II is shown recorded failures.

TABLE II
SCHEDULE OF MOTOR OPERATING CONDITIONS

Normal mode	Normal operating mode
Electrical faults	3% asymmetry–single-phase
	5% asymmetry–single-phase
	3% asymmetry–two-phase
	5% asymmetry–two-phase
Mechanical faults	Soft foot
	Angular misalignment
	Parallel misalignment

Data recorded like this, is used for machine learning process. First, these data must be preprocessed, for further training of the algorithm from the data sets.

III. LABORATORY RESULTS

In the below figures, it is shown vibration results measured in laboratory conditions. There were presented signals in smaller range then measured, for better review.

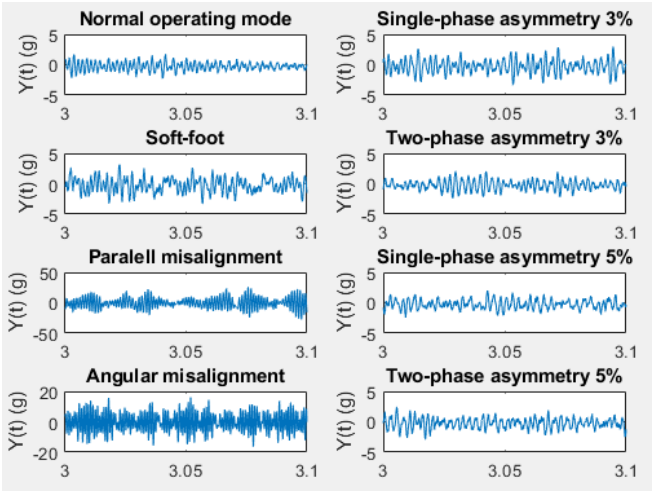


Fig. 2. Vibrations in mentioned operating modes in time domain for vertical axis, where $g = 9.81\ m/s^2$.

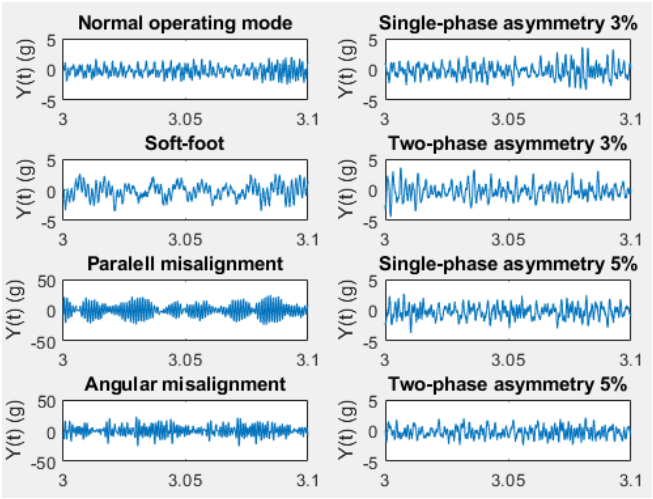


Fig. 3. Vibrations in mentioned operating modes in time domain for horizontal axis, where $g = 9.81\ m/s^2$.

In Fig. 2. from vibration in time domain signals for vertical axis, we can see that there are not much difference between the waveforms of normal operating mode and voltage asymmetries i.e. electric faults. The amplitudes are higher for asymmetry, but that can lead to misclassification sometimes. On the other hand soft foot is not so different from asymmetry waveforms, but parallel and angular misalignments have similar waves to each other, but so different from other motor conditions. In Fig. 3. from horizontal axis, we can notice almost similarly conclusions. Because on the basis on this representation, it is more difficult to make good differentiation between faults, we use preprocessing and features extraction methods like Fourier transform.

IV. PREPROCESSING

Raw data sets may possess missing values, outliers, some noisy parts of data which could have influence on the algorithm, for that reason, the preprocessing is an important

step in data analysis [13]. In this paper were used several important methods for data preprocessing techniques, including the detection and removing of the missing values from time domain signal, next the detection and removing of the outliers, then normalization. After that Fast Fourier Transform was done. The whole development is programmed through the script in MATLAB.

A. *Normalization*

To avoid dependence on the choice of measurement units, the data should be normalized or standardized. This involves transforming the data to fall within a smaller or common range such as [-1, 1] or [0.0, 1.0]. In our case, it was the second range. There are a few normalization techniques, but in this paper is used *min-max normalization*.

Min-max normalization performs a linear transformation on the original data. Suppose that X_{min} and X_{max} are the minimum and maximum values of a signal vector X, respectively. *Min-max normalization* maps a value v_i , of vector X to $v_i^{'}$ in the range [$X_{\min}^{new}$, $X_{\max}^{new}$] by computing:

$$v_i^{'} = \frac{v_i - X_{\min}}{X_{\max} - X_{\min}} \left(X_{\max}^{new} - X_{\min}^{new} \right) + X_{\min}^{new} \qquad (1)$$

By picking new values of maximum and minimum, we determine in what range we want the data to be. *Min-max normalization* preserves the relationship among the original data values [13].

B. *Fast Fourier Transform (FFT)*

Preprocessing of data is usually followed by Fourier transform. Basically, Fourier transform is also one of the methods for dimensionality reduction and feature extraction.

Fast Fourier Transform or *FFT* is almost inevitable algorithm for signal processing analysis of faults. That is because the different faults are expressed differently in the frequency spectrum. Some faults have characteristic peaks at certain frequencies, which makes it easier for algorithm to find characteristic pattern for recognizing certain faults and to distinguish from another. By noticing that, picking the right features makes it easier for detection and training algorithm.

C. *Feature Selection*

The direct measured signals are not suitable for on-line use since short sampling number is deficient for diagnosis, and enough sampling number is a burden for transferring and calculation. So feature selection of the signal is a critical initial step in any monitoring and fault diagnosis system. Its accuracy directly affects the final monitoring results [12].

Many papers are dealt with studying of frequency spectrum signals like in [4]. That is because many fault diagnosis tasks in induction motors depend on feature extraction from the measured signals. The feature characteristics directly affect effectiveness of fault recognition. In the existing literature, many feature extraction methods are suitable for fault diagnosis tasks and in using various training algorithm like neural network for decision [12] and other algorithms [8].

Based on above-mentioned literature, it was concluded that the best solution in feature selection is to take the highest peaks with its characteristic frequencies. In the Fig. 4. and 5. is shown model of peaks selection of the frequency signals by axis. Here were selected four highest peaks of every drive condition with constant load of 6.1 Nm. There can be noticed model by which peaks occur for some examples.

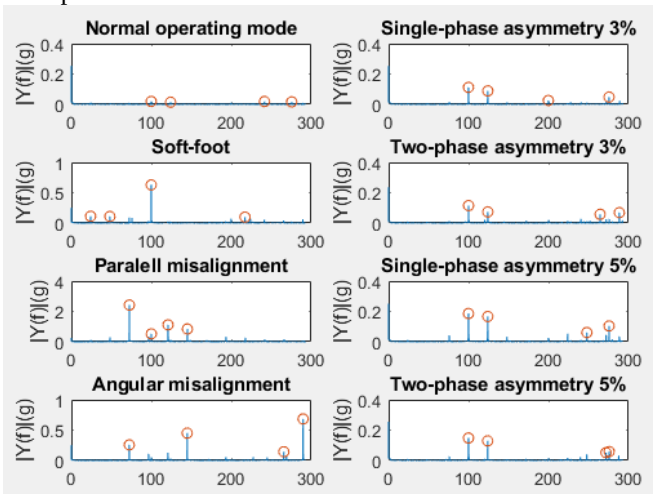


Fig. 4. Characteristic peaks selection on frequency signals of vibrations on vertical axis, where $g = 9.81\ m/s^2$ (before normalization).

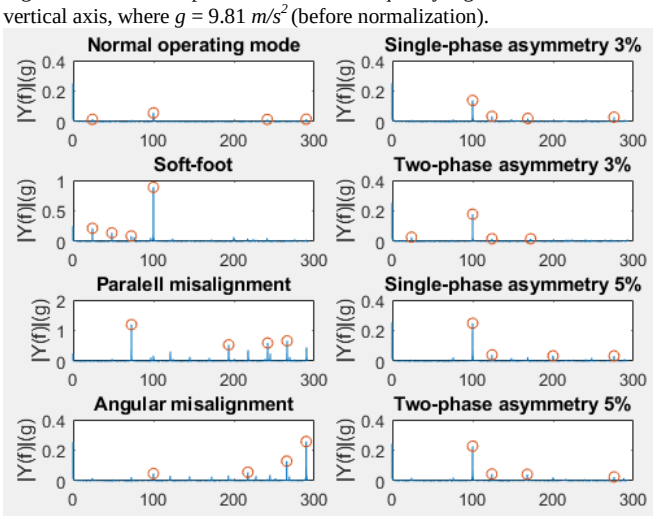


Fig. 5. Characteristic peaks selection on frequency signals of vibrations on horizontal axis, where $g = 9.81\ m/s^2$ (before normalization).

Frequency spectrum for normal operating mode is simple and its peaks are quite similar to low asymmetry levels but amplitude is much smaller. That fact can lead to misclassification of faults. From the frequency spectrum of electrical faults, we can see that the characteristic peaks are occurring in frequencies of approximately 100 and 125 Hz, according to the vertical axis, as far as the horizontal is concerned, the most significant component is in 100 Hz. Based on this, the algorithm can find a clear difference in electrical asymmetry faults in compared to others. On the other hand, mechanical faults do not have such a clear common difference compared to electric faults and normal mode. The soft foot shows the most visible component of 100 Hz, however, it is several times higher than the components of asymmetries at the same frequency, and from that side the difference is easy to conclude, but also sometimes could make trouble. In parallel misalignment, the largest peaks are at approximately 75 and 125 Hz, while for angular misalignment, at some 150 and 290 Hz.

Keeping all this in mind, after preprocessing, features were selected. Because it is difficult to find a similar difference for mechanical faults against other, which could separate them from the electrical ones, and because of the overall reliability of the algorithm, there were selected more features. From the signal in frequency domain, it has been selected 9 amplitudes of characteristic peaks followed by their locations i.e. 9 frequencies on which characteristic

peaks are occurring. It goes up to the frequency of 300Hz, for every signal that represents machine condition or fault, because that is enough range for faults representation. In this way we have 18 values for features i.e. 9 pairs of peak-location value. The selected number for pairs of features was chosen after a certain analysis of algorithms with other features. So number of features is reduced from 256000 to 18. Now every training signal has only 18 features which are adequately prepared for the training algorithm. It is still necessary for the data to be arranged properly before entering the algorithm.

V. Machine Learning Algorithm

Machine learning algorithms use methods for learning from data directly, without relying on a pre-programmed model. The higher the learning database the more accurate algorithm is. In the Fig. 6. we can see flow of data and building machine learning algorithm for faults diagnosis system. This flow chart shows the real application of machine learning in creating an algorithm for these purposes.

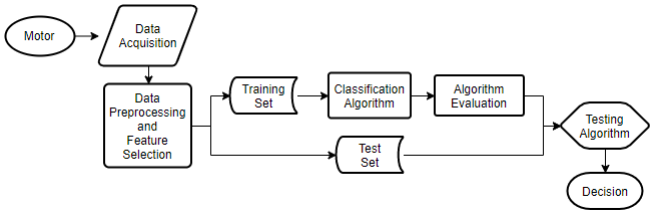


Fig. 6. Architecture of faults diagnosis procedure.

A. Data preparing

Here was used *supervised learning* [13],[14]. This method takes input data sets, obtained from experiment, along with known output data, trains a model that will generate precise predictions for new data. For algorithm training, it was used the *Classification Learner* application, that is part of MATLAB. In order to do this type of training in *Classification Learner*, it is necessary to provide the algorithm with the appropriate sets of input data, and in addition, known responses about what these data represent.

Input in application must be in form of matrix, with column reserved for features of the training signals (in our case 18), and rows reserved for number of training examples (in our case 264). It is better to have as many training examples as possible, because that affects the performances of an algorithm. But that affect is good to a certain point, after that the differences are minor. In addition, the last column is for responses of data in previous columns. In our case, we have only tree classes. Normal operating mode was assigned with class 0, electric failures were assigned with class 1, and mechanical failures with class 2. In this way, the classifier will be able to recognize what are the known data on which the classification is to be made, and what the answers are, on the basis of which the model should be trained. Input matrix is size *264x19*. That is shown on the TABLE III.

TABLE III
Response classes for Motor operating conditions

Normal mode	Normal operating mode	Class
Electrical faults	3% asymmetry–single-phase	1
	5% asymmetry–single-phase	1
	3% asymmetry–two-phase	1
	5% asymmetry–two-phase	1
Mechanical faults	Soft foot	2
	Angular misalignment	2
	Parallel misalignment	2

B. Training Model

Data derived from the preprocessing and preparing, described in the previous chapters are ready to enter the machine learning algorithm. *Classification Learner* application provides the ability to enter the input data (in a given form), then train the model, select the classifier, and finally evaluate the precision of the work and results. An automatic training can also be performed, which allows you to search for the best classification model for the given data. When a model is built, it can be exported to the workspace and used for predictions of new data.

Supervised learning uses methods of classification and regression such as *Support Vector Machine (SVM), K – Nearest Neighbors (KNN), Decision Tree* etc. After training model based on the data, *Classification Learner* gives as an accuracy percentage the particular prediction algorithm. That is shown in the Fig. 7.

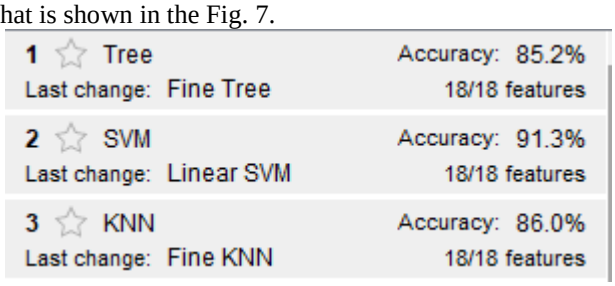


Fig. 7. The accuracy of machine learning algorithms.

By training various models for various classification algorithms, we get accuracy for three main algorithms: *Decision Tree*, *Support Vector Machine (SVM), K – Nearest Neighbors (KNN).* If we evaluate on accuracy, *SVM* would be the best choice. After that it was trained advanced models for *SVM*, and got that better accuracy with *Gaussian SVM*, with value of **92.8%**. However, accuracy will not be used as a solo decision factor and because its small difference, both training algorithms will be evaluated with further investigation methods.

C. Support Vector Machine (SVM)

Support Vector Machine or *SVM* is a supervised machine learning algorithm which is often used for classification problems. This algorithm finds pattern, which classifies data based on position of the them in *n*-dimensional space, where *n* is number of features. The *SVM* classifies data by finding the best *hyper-plane* that separates data points of one class from those of the other class. The best *hyper-plane* for an *SVM* means the one with the largest margin between the two classes. Margin means the maximal distance between nearest data point and *hyper-plane*. The *support vectors* are the data points that are closest to the separating *hyper-plane*. These points present the boundary of a class and they have the most influence on the position and orientation of *hyper-*

plane [13], [15].

This algorithm has technique called kernel. These are functions which take low dimensional input space and transform it to a higher dimensional space for separating non-separable points. Advanced model *Gaussian SVM* use Gaussian distribution.

VI. Algorithm Evaluation

In previous chapter was shown training and development of algorithms in *Classification Learner*. Beside shown accuracy, there are still many unresolved issues, so it is necessary to find the best evaluation procedure and check performances of training *SVM* algorithms.

Performance scores can be classified according to the return values. The results can be presented in a table form (*Confusion Matrix*) or graphically (*ROC Curves*) [11]. Here it will be considered both forms of our two best trained algorithms: *Linear SVM* and *Gaussian SVM*.

A. Confusion Matrix

A Confusion Matrix provides general information about the classifier performance. In our case *Confusion Matrix* (in further text *CM)* is three by three, because here we have three classes. Regular *CM* for *Linear SVM* and *Gaussian SVM* algorithms is shown in the Fig. 8.

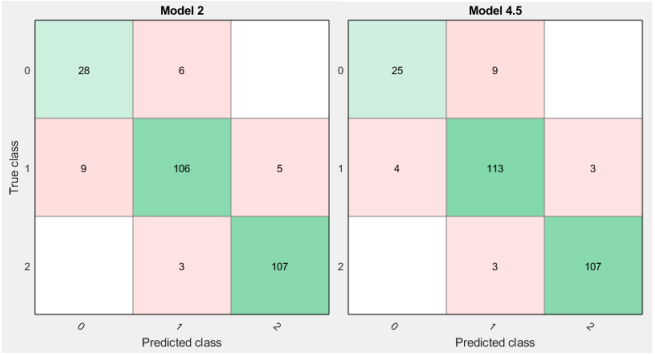


Fig. 8. Regular *CM* for *Linear SVM* (on the left) and *Gaussian SVM* algorithm (on the right)

In Fig. 8. on the left, we can cee that prediction power of every class with *Linear SVM*. For 34 training examples for class 0 (normal operating mode), 6 were predicted wrong, while the rest are predicted correctly. From the *CM* can be observed that the normal mode is confused with the electrical asymmetry, but never with mechanical problems. When it comes to class 1 (electrical faults), prediction is correct in 106 examples from 120. Here we can see that electrical asymmetry can be misclassified with normal mode and mechanical faults. From frequency and time domain waveforms, we can conclude that it is because of similarity of vibrations signals low asymmetry with normal mode and soft foot with asymmetry. That can occur because in some cases locations of peaks are matching for different faults, and only way to differentiate that is by value of amplitude.

In Fig. 8. on the right, we can see regular *CM* for *Gaussian SVM*. This algorithm is dealing with similar issues like *Linear SVM*. The conclusion is almost the same, but here we have better classification for electric faults, and a bit more misclassification for normal operating mode. The reason for this is because we have much less training examples for class 0, then for other 2 classes. Generally it is better to misclassify normal operating mode, then to have fault and label that like normal mode. That can be observed on Fig. 9.

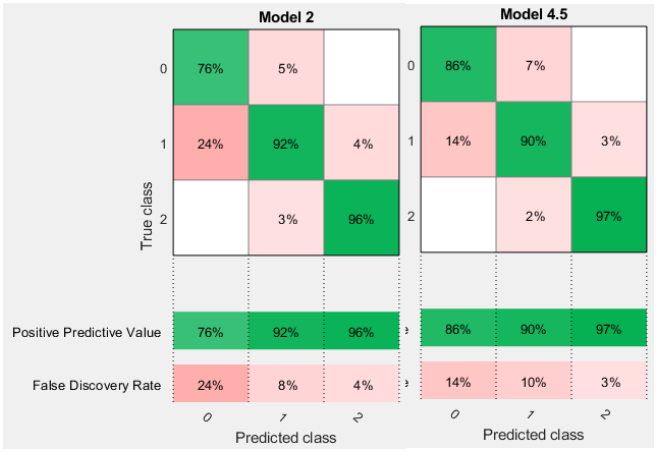


Fig. 9. *CM* with False Discovery Rate

Here we can see that the False Discovery Rate for classes is bigger with *Linear SVM* (on the left), especially with false prediction of normal mode, if it is electric fault condition. It is 24%, and with *Gaussian SVM* (on the right), we have 10% lower. Other 2 classes are mostly similar.

B. ROC Curves

ROC curve is showing true and false positive rates. The curve on the plot shows the values of the *False Positive Rate* (*FPR*), the *True Positive Rate* (*TPR*) and through that the performance of currently selected classifier.

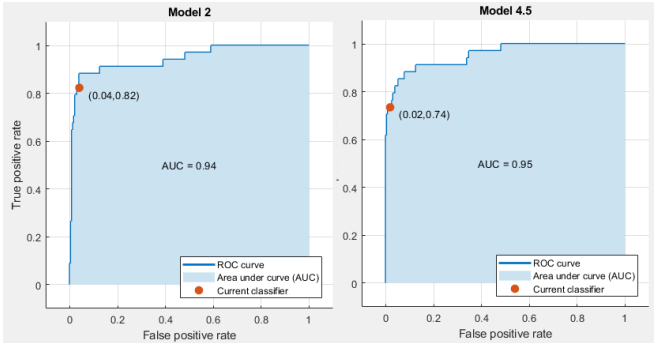


Fig. 10. *ROC* curve classifier 0 against classes 1 and 2.

In Fig. 10. on the left is shown *ROC* curve of *Linear SVM*, for the worst classifier 0, normal operating mode. Here we can see a *FPR* of 0.04 indicates that the current classifier assigns 4% of the observations incorrectly to the positive class (other 2 classes). A *TPR* of 0.82 indicates that the current classifier assigns 82% of the observations correctly to the positive class. The *Area Under Curve (AUC)* number is a measure of the overall quality of the classifier. Larger *AUC* values indicate better classifier performance. In this case it is 0.94 or almost 94%. For other two classes performances are better. In case class 1, we have performances (0.06, 0.88) with *AUC* = 0.95 for *Linear SVM* and (0.08, 0.94) with *AUC* = 0.98 for *Gaussian SVM*. In case class 2, we have (0.03, 0.97) with *AUC* = 0.99 for *Linear SVM* and (0.02, 0.97) with *AUC* = 1 for *Gaussian SVM*. That all indicates that, *Gaussian SVM* shows a bit better performances of an algorithm.

According to all parameters of evaluation, we conclude that these two training algorithms (*Linear SVM* and *Gaussian SVM*) can work with high precision of prediction power. Now this algorithm can be used for prediction of new data.

C. Predictions on New Data

After a good evaluation, the algorithms were exported for using with the new data. In process of data collecting, we

gather data for one more malfunction of electric drive. In addition to all mechanical faults, it was simulated also a rotor unbalance of the induction motor. This additional fault and its data set were used for test set, with which we can test our training algorithms. In Fig. 11. is shown signals of rotor unbalance of horizontal and vertical axis when load of a motor is 6.1Nm. In this figure we can see prominence of characteristic peaks on various frequencies, like 25 and 100Hz. Because the highest peak is in 25Hz, this signal isn't similar to any electrical fault (asymmetry) in our case neither the normal operating mode, but there is some resemblance with soft foot or even parallel misalignment.

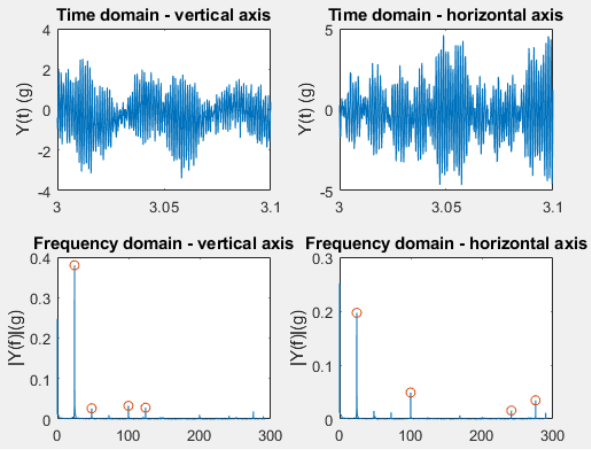


Fig. 11. Rotor unbalance signals in time and frequency domain.

For input new data in existing model, we must prepare that data, in the same way as we prepared all training data. In the end of preprocessing, there were 18 features for every of 30 testing example unbalance condition (for various load). Input matrix of size of *30x18*. After applying our *Linear SVM* and *Gaussian SVM*, we got the same results. Output in both algorithms was a new matrix with size *30x1*, where all elements were number 2. That means that all 30 new test examples belong to class 2, which represent a mechanical fault. From the previous can be concluded that on the same induction motor, previously unseen mechanical fault was successfully classified by both training algorithms. We can also conclude that the algorithms weren't so sensitive to the load change, because all 30 training examples were gathered for various loads. This way it was tested that both algorithms work excellent with detecting mechanical faults.

Regardless of the accuracy, the *Gaussian SVM* algorithm has shown through other evaluation methods that it is a better option for predicting new motor faults.

VII. Conclusion

In this paper, it was presented a study of the data analysis and machine learning implemented for the purpose of predictive maintenance in electric motor drive. Here was shown a development of training algorithm, which can predict malfunction in uncontrollable electric drive and conclude whether the fault is of an electrical or mechanical nature. This can be very helpful with the detection of failures, because the fault can be easily found and removed when we know if the problem is of electrical or mechanical nature.

This paper presents a complete data analysis, from the method of collection, preprocessing, classification and evaluation. It was applied *SVM* algorithm (*Linear* and *Gaussian*). The evaluation showed that algorithms have great efficiency and accuracy. After the trained model is provided with the new data, it was able to successfully classify the type of fault.

The diagnosis of electrical machine's working conditions presented here follows the trend of implementation of artificial intelligence. The artificial intelligence presents an alluring solution for this type of problem, because it enables detection algorithm to improve over time and also to follow the changing parameters of the motors during the exploitation life. Proposed approach could be able to implement in real time faults diagnosis with optimization in data preprocessing and then implementation of classification algorithm in real time processors. The future researches will be devoted to the preparation of these methods for real-time applications.

Acknowledgment

This work was supported by Ministry of Science and Technological Development, Republic of Serbia (Project number: III 44006).

References

[1] José M. Bossio, Guillermo R. Bossio, Cristian H. De Angelo ,,Angular Misalignment in Induction Motors with Flexible Coupling" IECON 2009.

[2] Irvin Redmond ,, Shaft Misalignment and Vibration - A Model" Saudi Aramco, 2002.

[3] Úrsula B. Ferraz, Paulo F. Seixas, Webber E. Aguiar ,, A Simplified Model for Mechanical Loads under Angular Misalignment and Unbalance" International Journal of Mechanical and Mechatronics Engineering, 2013.

[4] Željko Kanović, Dragan Matić, Zoran Jeličić, Milena Petković „ Induction motor faults diagnosis based on vibration analysis – A case study" Journal on Processing and Energy in Agriculture, 2013.

[5] G.R. Bossio, C.H. De Angelo, P.D. Donolo, A.M. Castellino, G.O. Garcia ,,Effects of voltage unbalance on IM power, torque and vibrations" 978-1-4244-3441-1/09 2009 IEEE.

[6] M'hamed Drif, Malika Drif, Jorge O. Estima, A. J. Marques Cardoso ,,The Use of the Stator Instantaneous Complex Apparent Impedance Signature Analysis for Discriminating Stator Winding Faults and Supply Voltage Unbalance in Three-Phase Induction Motors",2013.

[7] M.O Okelolal, O.E Olabode ,, Detection of Voltage Unbalance on Three Phase Induction Motor Using Artificial Neural Network" International Journal of Emerging Trends in Engineering, 2018.

[8] Tian Han, Bo-Suk Yang, Zhong-Jun Yin-,,Feature-based fault diagnosis system of induction motors using vibration signal" *Journal of Quality in Maintenance Engineering* Vol. 13 No. 2, 2007.

[9] Yi L. Murphey, M. Abul Masrur, ZhiHang Chen, BaiFang Zhang-,, Model-Based Fault Diagnosis in Electric Drives Using Machine Learning" Independent Research, USA, 2005.

[10] Andre A. Silva, Ali M. Bazzi, Shalabh Gupta-,, Fault Diagnosis in Electric Drives using Machine Learning Approaches" University of Connecticut, USA, 2014.

[11] Ignacio Martín-Díaz, Oscar Duque-Perez, René Romero-Troncoso, Daniel MorinigoSotelo ,, Supervised Diagnosis of Induction Motor Faults: A Proposed Methodology for an Improved Performance Evaluation" 978-1-4799-7743-7/15, 2015.

[12] S.-Y. Shao, W.-J. Sun, R.-Q. Yan, P. Wang, and R. X. Gao, "A Deep Learning Approach for Fault Diagnosis of Induction Motors in Manufacturing," *Chinese Journal of Mechanical Engineering*, 2017.

[13] Jiawei Han, Micheline Kamber, Jian Pei - ,,Data Mining Concepts and Techniques" Third Edition, Morgan Kaufmann Publishers, USA, 2012.

[14] Peter Tavner, Li Ran, Jim Penman , Howard Sedding-,,Condition Monitoring of Rotating Electrical Machines" The Institution of Engineering and Technology, 2008.

[15] MathWorks - https://www.mathworks.com/ (accessed 15.04.2019.).

Detection of Supply Voltage Unbalance Condition in Induction Motor Using Machine Learning

Vladislav Lazić, Uroš Ilić, Bratislav Trojić, and Milutin Petronijević, *Member, IEEE*

***Abstract*—This paper presents possibilities of detecting supply voltage unbalance condition in induction motor using machine learning algorithms. Supply voltage unbalance condition can be dangerous for motor and can invoke bigger faults. Early detection of this condition is important in order to prevent costly outages. This paper aims to show results and analyze the performance of these algorithms, using MATLAB application Classification Learner. Mechanical vibration and stator current measurements have been taken into consideration, as they are easily accessible for measuring and hold a lot of important information. These signals were preprocessed in an appropriate way in order to gain valuable features, which are then used in the algorithm. Results were analyzed and displayed in the paper.**

***Index Terms*— Supply Voltage Unbalance, Squirrel-Cage Induction Motor, Machine Learning, Condition Monitoring**

I. Introduction

The electrical energy that is used in the industry is in the largest part being converted into mechanical energy [1]. In the center of this conversion is an electric motor. The industry is today more developed than ever and large amounts of goods are produced every second. In order to avoid costly production halts, the motors that drive this production need to be reliable. The high reliability is the very reason why the squirrel-cage induction motor is the most commonly used electric motor in the industry [2].

However, despite the high reliability and other advantages that the induction motor has, it is still subjected to many faults [3-4]. Depending on the severity of these faults, they can cause additional losses in the motor, or even the motor shut down. Many times, it is possible to detect a fault in the initial stage, which is essential in preventing the complete failure of the motor. In accordance with the above, one can conclude that it is necessary to monitor the condition in some way and to extract key features from monitored signals which show us that fault occurred.

Vladislav Lazić is with the Faculty of Electronic Engineering, University of Niš, 14 Aleksandra Medvedeva, 18000 Niš, Serbia (e-mail: vladislav.lazic@ gmail.com).

Uroš Ilić is with the Faculty of Electronic Engineering, University of Niš, 14 Aleksandra Medvedeva, 18000 Niš, Serbia (e-mail: urosilic@elfak.rs).

Bratislav Trojić is with the Faculty of Electronic Engineering, University of Niš, 14 Aleksandra Medvedeva, 18000 Niš, Serbia (e-mail: bratislav.trojic@ gmail.com).

Milutin Petronijević is with the Faculty of Electronic Engineering, University of Niš, 14 Aleksandra Medvedeva, 18000 Niš, Serbia (e-mail: milutin.petronijevic@elfak.ni.ac.rs)

Numerous fault detection methods have been analyzed and presented in the literature. One of the methods, that is becoming more important as technology advances, is a knowledge-based approach [5]. Knowledge-based approaches are based on soft computing techniques and they are used to solve complex calculations with non-linear big data problems. This type of problems requires a probabilistic approach as there is a lot of imprecision and uncertainty in the data. Machine learning techniques fall into this category. In recent times, machine learning approaches, such as K-nearest neighbors, Support vector machine and Decision tree algorithms were suggested as a way of detection of a fault occurrence in the induction motor at the early stages and it showed some viable results [5-8].

This type of fault detection methods gets even more significant as the fourth industrial revolution is on the way. Among other things, the fourth industrial revolution is characterized by mass production and automation of the processes, as well as more efficient use of energy. Machine learning approach has not still found its place as the number one condition monitoring system for induction motors in the industry, but it could be more used in this field in the upcoming years as an answer to the demands of the fourth industrial revolution.

In this paper, it was dealt with the detection of an unbalanced voltage condition in induction motor using machine learning algorithms. In section II, general remarks regarding the supply voltage unbalance condition and the possibilities of its detection are given. This condition has an effect on both stator currents and motor vibrations, so these two quantities were taken into consideration and two models were built using these measurements. The experimental setup for data collection, the strategy for dataset creation and final results are presented in section III. The fifth and the seventh harmonic were present in stator currents and their effect on the detection process is evaluated, as they are the most common harmonics in the grid [9].

II. Supply Voltage Unbalance

The normal working condition of a motor presumes balanced supply voltage. Ideally, magnitudes of all three phase voltages are the same and phases are equally shifted between each other at an angle of 120 degrees. A described three-phase system is completely symmetrical. If either magnitudes or phase angles differ from each other, it comes to

unbalanced voltage condition and certain asymmetry is inserted into the motor.

The causes of the unbalanced condition can either be external, where common reason is an occurrence of load variations and uneven load distribution among the phases, so motor supply voltage becomes unbalanced, or internal if the three-phase winding has some defects and is not symmetrical.

Analysis of the motor operation under the unbalanced condition is usually performed using the method of symmetrical components. This means that the asymmetrical system is divided into three symmetrical systems; direct (positive), inverse (negative) and zero (homopolar). In the ideally symmetrical system, there is only one component, the direct one. When the three-phase stator winding gets connected to the unbalanced voltage supply, asymmetrical stator currents are formed, hence the inverse sequence currents are generated. These currents form a magnetic field that rotates in the opposite direction of the rotor rotation. As a consequence, resulting torque is smaller than the torque motor has under the balanced condition. Also, negative sequence currents cause additional heating in the stator windings, which can lead to isolation deterioration and ultimately to a short circuit in the stator winding.

This is the reason why it is important to detect the presence of the supply voltage unbalance. If this condition is detected at the early stages, isolation may not be completely damaged and it may not come to the short circuit inside the stator windings. Also, it is important that the motor runs more efficiently. If the unbalanced condition is not detected and motor runs under it, then it would run with the smaller torque, increased slip, bigger losses, and overall efficiency factor would be smaller.

In order to detect the presence of the supply unbalance condition, one could focus on detecting the presence of negative sequence currents. This is possible as the negative sequence magnetic field creates a 100 Hz component in the frequency spectrum of the rotor currents [10]. Knowing that the magnetic field of negative sequence components is rotating in the opposite direction from the rotor, we can write

$$s_i = \frac{n_s - (-n)}{n_s} = \frac{n_s + n}{n_s} = 1 + \frac{n}{n_s} = 1 + 1 - s = 2 - s. \qquad (1)$$

In the previous equation, we see that the inverse sequence slip s_i is equal to the $2-s$, where s is the motor slip in the normal working condition. As s is a small value, then $2-s$ is approximately equal to 2. If we write the equation for finding the rotor frequency of the negative sequence electrical values in the rotor, we get

$$f_{ri} = s_i f = (2-s) f. \qquad (2)$$

This shows us that voltage unbalance condition creates frequency components at the double of the supply frequency, while the frequency of the rotor electric values under normal conditions is usually 1-3 Hz. This gives the opportunity to extract this feature using some of the condition monitoring methods and to conclude that inverse sequence components are present, which definitely means that there is a certain asymmetry in the motor.

However, rotor currents are not accessible for measurement in case of a squirrel-cage induction motor. This is why in this paper stator currents and vibrations were used to detect the unbalanced condition.

A. Definitions of the Voltage Unbalance

There are three definitions of voltage unbalance, where three different parameters are calculated [11]. They are NEMA (National Equipment Manufacturer's Association) definition, IEEE definition and True definition, which is adopted by the IEC.

In NEMA and IEEE definitions LVUR and PVUR parameters are calculated, respectively. LVUR – line voltage unbalance rate is calculated as a percentage value of the ratio between the maximum voltage deviation from the average line voltage and the average line voltage value. PVUR - phase voltage unbalance rate is calculated like the LVUR, except the average phase voltage is considered, along with the maximum deviation of one phase voltage from the average.

In this paper, voltage unbalance is calculated using the True definition, where voltage unbalance factor (VUF) is calculated as the percentage value of the ratio between the negative sequence voltage component and the positive sequence voltage component. To perform this calculation one would have to use complex algebra. To avoid this (2) is derived as a good approximation to the True definition [11]:

$$\%VUF = \frac{82 \cdot \sqrt{(V_{ab} - V_{avg})^2 + (V_{bc} - V_{avg})^2 + (V_{ac} - V_{avg})^2}}{V_{avg}} \qquad (2)$$

where V_{avg} is the average line voltage.

NEMA standard defines the derating factor [11]. This factor shows how much overall motor performance deteriorates with the rise of voltage unbalance.

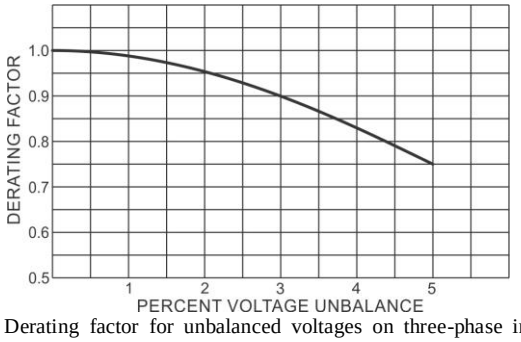


Fig. 1. Derating factor for unbalanced voltages on three-phase induction motor (adapted from [11])

Fig. 1 shows how it is expected that the motor performs well if voltage unbalance is below 1%. Above this value, motor performance starts to slow down. It is not recommended to run a motor under the circumstances of voltage unbalance factor being bigger than 5%.

B. Effect of the Voltage Unbalance on Stator Currents and Vibration

It is already mentioned that stator currents under supply

voltage unbalance are unbalanced as well. Unbalanced stator currents are the source of power and torque oscillation which leads to higher vibration levels than normal. Higher vibration levels can lead to deterioration of the mechanical parts of the motor, which is one more reason that the supply voltage unbalance should be fixed as soon as possible.

Due to the existence of the asymmetry in the motor quantities and thereby the existence of the negative sequence currents, vibrations occur at the double supply frequency in the vibration frequency spectrum. This is the case with the supply voltage unbalance condition [12-14], among some other faults [15-17]. The possibility of detecting the occurrence of vibrations at the double supply frequency in the frequency spectrum is processing the signal via Fast Fourier transformation and extracting key features. This was used in this paper.

According to expectations, there is a larger 100 Hz component in the vibration frequency spectrum when the motor was under unbalanced condition than in a normal motor (Fig.2). A very large 300 Hz component is present in both normal and a faulty case and it is a consequence of the presence of the 5th and the 7th harmonic in stator currents (Fig. 3), as expected in case of harmonics polluted supply voltage [9].

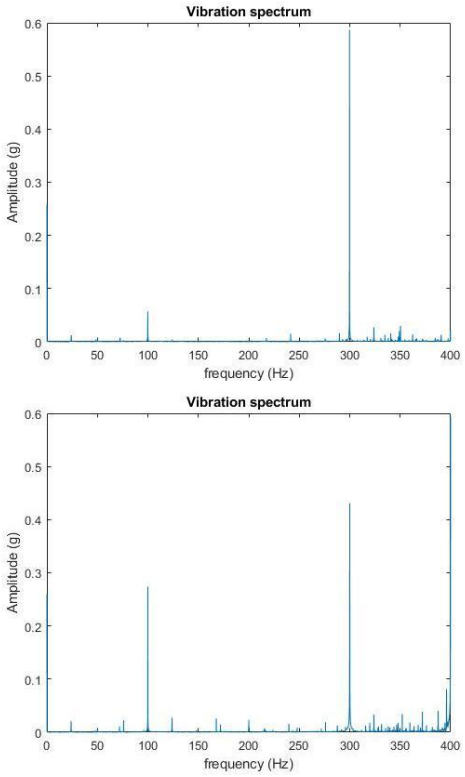


Fig. 2. The vibration spectrum of the motor used in this experiment without voltage unbalance (top) and with voltage unbalance of 5% (bottom) at 75% load

The stator current frequency spectrum is not as conclusive as rotor currents spectrum would be. The frequency spectrum of stator currents in normal and in unbalanced voltage condition is presented in Fig. 3. As can be seen, there is no significant difference in harmonic content between these two cases, as normal motor also has some asymmetry.

It is interesting to see if machine learning algorithms could find the pattern in stator current as well, even though it does not seem plausible by just looking at the stator currents frequency spectrum. Stator currents are easily accessible and it would be good if one could conclude if there is voltage unbalance condition by extracting features from the stator currents.

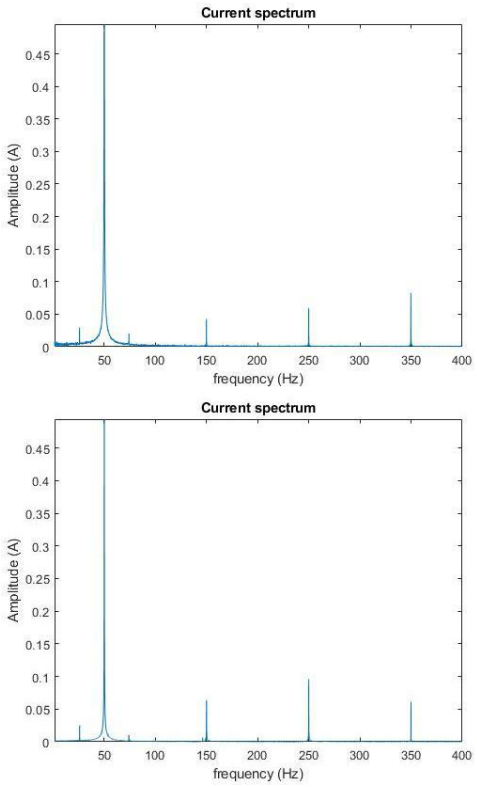


Fig. 3. The current spectrum of the motor used in this experiment without voltage unbalance (top) and with voltage unbalance of 5% (bottom) at 75% load

III. Experimental Results

A. Experimental Setup

In order to check if a machine learning algorithm can adequately detect voltage unbalance, first data had to be collected. The motor used in this experiment was three-phase induction motor Siemens 1LA7096-4AA10-Z. Its nameplate parameters are given in Table 1 below.

TABLE I
MOTOR PARAMETERS

Rated voltage	400 V
Rated current	3.45 A
Rated power	1.5 kW
Rated speed	1420 1/min
Rated torque	10.1 Nm
Power factor	0.81
Efficiency	77.2%

This motor was supplied from the 3x400V, 50Hz grid through two three-phase autotransformers (Fig. 4) in order to

create voltage unbalance while maintaining the same phase angle between phases. In this way, both one-phase and two-phase voltage asymmetry could be created, depending on which autotransformer's output voltages were changed from the grid voltage.

This motor was coupled with Schneider Electric servo motor BMH1401P16A1A, which was controlled with the proper servo drive. In this way, the second motor served as a load and was controlled in such a way to cover 5 states: no load (Siemens motor only had Schneider Electric motor's inertia on the way), 25% of the full load, 50% of the full load, 75% of the full load and full load state.

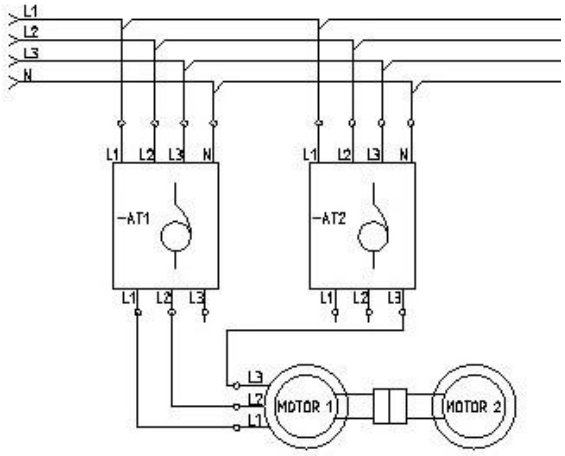


Fig. 4. Schematic diagram of the experimental setup

B. Data Collection

Stator currents and vibration signals were intended to be measured. Stator currents signals were acquired with the current probes and displayed using a four-channel oscilloscope. Each current measurement was recorded for a period of 10 s with the sampling frequency of 10000 Hz. Vibrations were measured in the horizontal and vertical axis. Vibration signal data were acquired using a dedicated device adjusted and programmed to collect data from the vibration sensor. Measurement length was also 10 s with the sampling frequency of 25600 Hz. A total of 90 current measurements and 150 vibration measurements were collected.

The data were collected for three different voltage unbalance conditions: VUF=0-1% (normal conditions), VUF=3% and VUF=5% (one-phase and two-phase unbalance). The main idea investigated in this paper is to determine how well machine learning algorithms deal with the detection of voltage unbalance as a source of vibrations or as a source of unbalanced currents. In this sense, the results are compared between two models where vibration signals and stator currents are used as a dataset. Additionally, it is checked whether algorithms can properly distinguish between different levels of voltage unbalance.

C. Dataset Preparation and Feature Selection

These raw data are preprocessed in order to create adequate datasets. It was already mentioned that valuable information can be extracted from the signals if Fast Fourier transformation was performed.

Important information about measured signals was extracted in a way that 10 highest peaks and their corresponding locations were found in each measurement's frequency spectrum inside of 0-300 Hz range, forming 20 features which machine learning algorithm should use to identify the pattern. In order to retrieve only these peaks, frequency spectrums were cut off outside this range.

The justification for this approach was derived from MATLAB's Neighborhood Component Analysis method for feature selection [18]. The graph in Fig. 5 represents the significance of each feature for having a more accurate distinction between classes in the model based on vibration measurements. The first 20 dots represent 20 highest peaks that were obtained, and the next 20 dots are their corresponding frequencies. It can be seen that higher peaks have a bigger influence on the classification. It can be noted that frequencies of those peaks are also highly influential on the classification process. That is why 10 highest peaks and their corresponding frequencies are used as features for each measurement in creating a prediction model. These 10 peaks have a stronger impact than the remaining 10 peaks and machine learning algorithms work better and faster with fewer features.

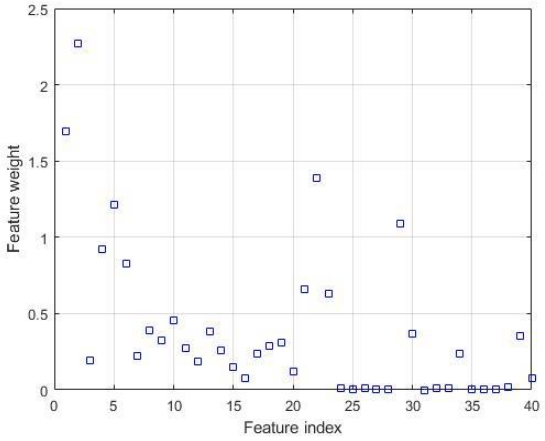


Fig. 5. Relative significance of features for the accurate classification

The dataset matrix is formed in a way that each row consists of 20 features that are selected to represent one measurement. At the end of each row class labels were added, forming the complete dataset, suitable for creating a prediction model in MATLAB Classification Learner toolbox. Classes were labeled as in Table 2. Group of measurements that represented normal conditions were labeled as 0. The signals that were measured under the faulty conditions of 3% and 5% voltage unbalance were labeled as 1 and 2, respectively.

TABLE II
CLASS LABELS

%VUF	Class label
0%	0
3%	1
5%	2

All of this applies to both current measurements and vibration measurements models. The only difference is that in the model based on vibration measurements the accuracy was better when 0-250 Hz range was observed. The reason for this range reduction is the existence of a significant 300 Hz component in the vibration spectrum, which was present in the normal conditions as well as in the faulty condition. This component is induced due to the presence of the 5th and the 7th voltage harmonic [9]. It affected the detection process by lowering the accuracy of the proposed method, and this component is excluded from the model. The effect of mentioned harmonics on the detection process in the vibration model is eliminated in this way.

On the other hand, harmonics in stator currents changed as operation mode changed, in such a way that they were not worsening the voltage unbalance condition detection accuracy in the model based on stator currents. The accuracy was even better when these harmonics were included. Particularly, the 5th harmonic had more impact, so the 0-300 Hz range was chosen. This frequency range provided the best accuracy, but the accuracy was not much lower when both 5th and 7th harmonic were taken into consideration, or when none of them was.

D. Prediction Model Based on Current Measurements

After the prediction model is created, the results were obtained and analyzed. Results will be presented in the form of a Confusion matrix (Fig. 6).

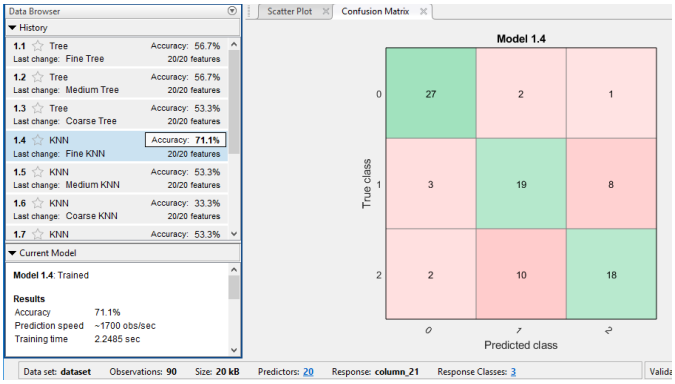


Fig. 6. Confusion matrix of a prediction model based on stator currents dataset

It is shown that Fine KNN had the highest accuracy, 71.1%. Weighted KNN was also close with 67.8%. This accuracy of 71.1% is not high, but it is important to notice that the low accuracy was caused because the algorithms mixed up classes 1 and 2 (3% and 5% voltage unbalance), while they correctly understood in most cases when there was no voltage unbalance.

As can be seen in Fig. 6, Fine KNN correctly predicted normal conditions in 27 out of 30 normal condition measurements. Also, the model predicted wrongly that there is no voltage unbalance, when there was, 5 times in total out of 60 measurements.

Reasons for the mix up between the class 0 and other two classes can be found in the uncertainty in data and not in the capabilities of the algorithms themselves, as normal condition measurements were never symmetrical, especially at the lower load. The reason for this is that autotransformers were not able to pass through fully symmetrical voltages, so the certain asymmetry was inserted into stator currents. It is possible that some of the class 0 measurements were degraded by this in such a manner that they seemed more like an unbalanced condition than a normal one to the algorithm.

As for the distinction between different levels of voltage unbalance, it can be concluded that the algorithm did not perform well, as it mixed up classes 1 and 2 in numerous instances. For the better prediction of voltage unbalance operation conditions, additional modifications in the process of preparation of the dataset are necessary.

E. Prediction Model Based on Vibration Measurements

Confusion matrix of the prediction model based on vibration measurements is given in Fig. 7.

The best accuracy in this model is 78.7%, shown by the Coarse Tree algorithm and SVM had 76% in this case, but other algorithms showed better accuracies as well in comparison with the current measurements model. Again, 27 of 30 predictions were performed correctly for normal motor conditions, but in this case, it was predicted only once out of 120 times that there is no fault, when there was.

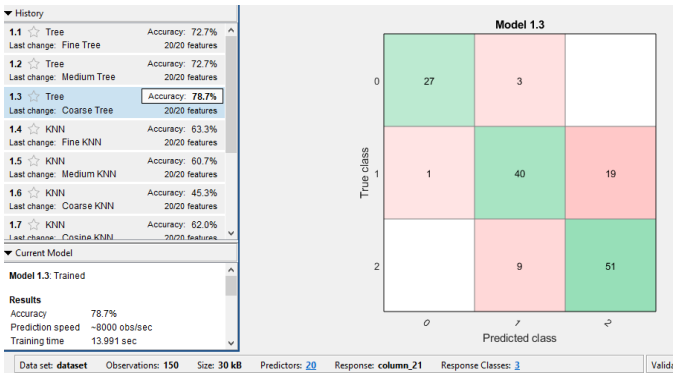


Fig. 7. Confusion matrix of a prediction model based on vibrations dataset

Unlike in the current model, the vibration model did not mix up classes 0 and 2 even once. Also, it showed a better distinction between classes 1 and 2. This leads to the conclusion that the vibration model has presented a better prediction of the voltage unbalance level and better overall prediction of a voltage unbalance condition. This is expected, given that the vibration frequency spectrum seems much more conclusive than the current frequency spectrum. It should be stated here as well that there were more vibration measurements and it might have had a positive influence on the machine learning algorithms performance, as it had more training data.

IV. CONCLUSION

This paper aimed to show and compare the possibilities of machine learning detection of an unbalanced supply condition using current and vibration data and the results were presented. The vibration model displayed better results. The main reason for this could be the fact that the vibration frequency spectrum of the unbalanced supply condition is

clearly different than that of a normal one, which is not the case with the current frequency spectrum. Despite this fact, it is shown that machine learning algorithms managed to find some pattern in the current frequency spectrum as well and to separate the normal condition from the faulty one.

The presented results are not conclusive, as the dataset was not big enough. Also, the results could have been better with additional data preprocessing, especially for the current model. More precise sensors, programmable voltage sources, and additional experimental tests are necessary in order to make further improvements in creating datasets. However, these first results are promising and machine learning algorithms displayed capabilities of finding the pattern that is necessary to distinguish between the normal working conditions and the unbalanced voltage supply condition in both models. On the other hand, they could not differentiate between different levels of voltage unbalance with satisfactory accuracy. The vibration model showed somewhat better results in this case as well.

Further work could include different ways of raw data preprocessing and preparing the dataset, which could provide better results, especially in the distinction between different levels of the voltage unbalance. The effect of the 5th and the 7th harmonic on the prediction models has been mentioned in the paper, but more investigation of the harmonic impact on the supply voltage unbalance detection results should be conducted, as there are also other voltage harmonics which may have an influence. It should also be dealt with the fact that other sources can mask the essential information for voltage unbalance detection, such as some mechanical imperfections which also induce 100 Hz component in the vibration spectrum, and that can potentially affect the detection process. Lastly, this approach should be tried with bigger datasets, as that could help machine learning algorithms to be trained better and display better results.

ACKNOWLEDGMENT

We would like to express gratitude towards our colleagues Filip Filipović and Lazar Sladojević, who kindly helped by sharing their thoughts on this subject. Thanks to company Netico which provided us with the device that was used for vibration measuring. This work was supported by the Ministry of Science and Technological Development, Republic of Serbia (Project number: III 44006).

REFERENCES

[1] T. Javied, T. Rackow, R. Stankalla C. Sterk, J. Franke, "A study on electric energy consumption of manufacturing companies in the German industry with the focus on electric drives," Procedia CIRP 41, pp. 318-322, 2016.

[2] M. H. Rashid, Power Electronics Handbook. Butterworth-Heinemann, 2018.

[3] G. Jose, V. Jose, Author, "Induction motor fault diagnostics – A comparative study," ICEE, At Hyderabad, India, July 2013.

[4] H. A. Tolyat, S. Nandi, S. Choi, H. Mesgin-Kelk – "Electric machines: modeling, condition monitoring, and fault diagnosis," CRC Press Taylor & Francis Group, 2012.

[5] M. Z. Ali, M. N. S. K. Shabbir, X. Liang, Y. Zhang, and T. Hu, "Experimental investigation of machine learning based fault diagnosis for induction motors," Proc. IEEE Ind. Appl. Soc. Annu. Meeting, Portland, OR, USA, Sep. 23–27 2018, pp. 1–14.

[6] K. Madhuri, K. S. Kavitha., V. K. Agrawal – "Supervised Machine Learning Algorithm Used to Detect Fault in an Induction Motor", IRJET, Volume: 04 Issue: 04, April 2017.

[7] V. A. D. Silva, R. Pederiva, "Fault detection in induction motors based on artificial intelligence," Surveillance 7, Institute of Technology, Chartres, France, October 2013.

[8] P. Gangsar, R. Tiwari, "Diagnostics of mechanical and electrical faults in induction motors using wavelet-based features of vibration and current through support vector machine algorithms for various operating conditions," J. Braz. Soc. Mech. Sci. & Eng., February 2019.

[9] P. D. Danolo, G. R. Bossio, C. H. De Angelo, G. O. Garcia, M. Danolo, "Voltage unbalance and harmonic distortion effects on induction motor power, torque and vibrations," EPSR Volume 140, November 2016

[10] Z. Stajić, Đ. Vukić, M. Radić, "Asinhrone mašine," Faculty of Electronic Engineering in Niš, 2012.

[11] P. Pillay, M. Manyage, "Definitions of Voltage Unbalance," IEEE Power Engineering Review, Volume: 21, Issue: 5, May 2001.

[12] M. O. Thurston, "Energy-Efficient Electric Motors," Electrical and Computer Engineering, Marcel Dekker, New York, 2005

[13] G. R. Bossio, C. H. De Angelo, P. D. Donolo, A. M. Castellino, G.O. Garcia, "Effects of voltage unbalance on IM power, torque and vibrations," IEEE International Symposium on Diagnostics for Electric Machines, Power Electronics and Drives, Cargese, France, September 2009.

[14] M. Campbell, G. Arce, "Effect of motor voltage unbalance on motor vibration: Test and evaluation," PCIC, Philadelphia, PA, USA, September 2016

[15] G.H. Bate, "Vibration Diagnostics for Industrial Electric Motor Drives," Bruel & Kjaer, Available at: https://www.bksv.com/media/doc/BO0269.pdf (Accessed: 17. May 2019)

[16] M. Tsypkin, "Induction Motor Condition Monitoring – Vibration Analysis Technique – a Twice Line Frequency Component as a Diagnostic Tool," IEMDC, Chicago, IL, USA, May 2013

[17] I. Hamernick, "Detect Soft Foot with Vibration Analysis," Available at: https://www.reliableplant.com/Read/920/soft-foot-vibration-analysis (Accessed 17. May 2019)

[18] Feature selection using neighborhood component analysis for classification – MATLAB. [Online] Available at: https://www.mathworks.com/help/stats/fscnca.html?s_tid=mwa_osa_a. (Accessed: 17.May 2019)

Classification models of machine learning for vibration analysis of induction motor

Uroš Ilić, Bratislav Trojić, Vladislav Lazić and Filip Filipović

***Abstract* — The main aim of writing this paper is the idea to use machine learning techniques to recognize possible faulty operating modes of induction motor drives using only mechanical vibrations. In this paper, the vibrations of one induction motor were recorded for different operation modes: normal, soft foot, rotor imbalance, angle and parallel misalignment. Recorded data is transferred from the time into a frequency domain using Discrete Fourier Transformation (DFT) to be used as an input for machine learning algorithms. The input into the algorithm consists of ten components of frequency spectra (peaks) with the highest amplitude and corresponding frequencies in frequency range up to 300 Hz. Classification models of machine learning obtained in the end, have a very good ability to easily classify unlabeled data.**

***Index Terms*—induction motor, vibrations, machine learning**

I. Introduction

Cost optimization in industrial drives is a dominant goal that tends in today's time. In industrial plants, there are fixed and variable costs. The first are regular costs that go to the spare parts and maintenance, and significant part of the second is unexpected costs that go to the remediation of defective equipment. Induction motors (IMs) are a part of a drive that can generate the greatest additional loss of money if, for some reason, it enters in a faulty operating mode that could lead to equipment failure. In order to avoid failures in electrical machines, it is necessary to monitor and analyze the condition in which a machine is running, i.e. it requires the implementation of a condition-based maintenance system [1]. Efficiency of drives and savings achieved in this way have led to increased money investing in systems that would have the ability not only to analyze the current operating condition of the machine, but also to predict future events.

Fault detection at their initial stage contributes to increasing the efficiency of the entire drive [2]. Modern technologies have contributed to the development of many techniques for monitoring the operating conditions of machines and early fault detection. Methods of fault detection are based on measurement different variables: mechanical (vibrations, shock pulses, acoustic noise, speed fluctuations), electro-mechanical (currents, partial discharges, leakage fluxes) and temperature, oil analysis, gas analysis, and overall performance monitoring [3]. These techniques are often based on the processing and analyzing of a large number of data. The main problem is extraction of useful information from raw data [4].

Using machine learning in data analysis allows a large number of tools for understanding the data obtained directly from industrial drive. In the paper, three machine learning algorithms were used to form classification models: support vector machine (SVM), k-Nearest Neighbors, decision tree.

Support vector machine was first introduced in 1995 by Cortes & Vapnik [5]. This type of classifier is based on the foundations of statistical learning theory. Using SVM algorithm in classification problem is explained in [6].

K-NN algorithm is among the simplest of all machine learning algorithms. It is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. Application of both algorithms (SVM and k-NN) for classification of IM faults described in [7].

Decision tree is a machine learning algorithm that builds a knowledge-based system by inductive inference from case histories. In [8] vibration data were used for training decision tree algorithm that was used in the detection of failures.

Section II describes the process of vibration analysis: detecting characteristic frequencies, data collection, analysis of the frequency spectrum. Section III describes the process for creating a dataset, while Section IV provides an analysis of the obtained classification models. Finally, the conclusion is drawn based on presented results.

II. Vibration Analysis

Motor vibrations exist even though the motor is fully healthy, so it is necessary to understand the causes that lead to vibration change. Regular vibration monitoring can detect deteriorating or defective bearings, mechanical looseness and worn or broken gears. Vibration analysis can also detect misalignment and unbalance before these conditions result in bearing or shaft deterioration [9]. Vibration analysis is an effective, non-intrusive method to monitor machine condition during start-ups, shutdowns and normal operation. The most important advantage is that the reaction to the changes is instantaneous which makes it possible to use in current, but also occasional analysis of operating conditions. Also, a very important feature of the vibration analysis is the possibility of widespread use of signal processing techniques in order to easily identify the fault indicators that are not visible in the source signal.

A vibration analysis system usually consists of four basic parts:

Uroš Ilić is with the Faculty of Electronic Engineering, University of Nis, Aleksandra Medvedeva 14, 18000 Nis, Serbia (e-mail: urosilic@elfak.rs).

Bratislav Trojić is with the Faculty of Electronic Engineering, University of Nis, Aleksandra Medvedeva 14, 18000 Nis, Serbia (e-mail: bratislav.trojic@gmail.com).

Vladislav Lazić is with the Faculty of Electronic Engineering, University of Nis, Aleksandra Medvedeva 14, 18000 Nis, Serbia (e-mail: vladislav.lazic@gmail.com).

Filip Filipović is with the Faculty of Electronic Engineering, University of Nis, Aleksandra Medvedeva 14, 18000 Nis, Serbia (e-mail: filip.filipovic@elfak.ni.ac.rs).

1. Signal pickup, also called a transducer
2. A signal analyzer
3. Analysis software
4. A computer for data analysis and storage.

A. *Dominant Frequencies of Mechanical Faults*

Mechanical failures where the vibration analysis is the most commonly used tool for detection are rotor unbalance, shaft misalignment (parallel and angle) and soft foot. Characteristic frequencies in the vibrational spectrum on the basis of which mechanical problems can be detected are defined in [10], and shown in Table I.

TABLE I
DOMINANT FREQUENCY IN VIBRATIONAL SPECTRA

Type of failure	Dominant Frequency	Dominant Plane
Unbalance	1 x rpm	Radial
Misalignment (angle)	1 x, 2 x rpm	Radial
Misalignment (parallel)	1 x, 2 x rpm	Radial
Soft foot	1 x, 2 x rpm	Radial

Thus, failures listed in Table I, are manifested by increasing the amplitude of components in the frequency spectrum on a single or double frequency of rotation of the shaft. The component of dual frequency signal appears in the majority of failures. This means that misalignment and soft foot can be distinguished from rotor unbalance if there is a 2 x rpm component, but in case it does not exist, the distinction of failures is very difficult to make.

Experiments made in [11] show that, depending on the coupling method and applications in which the motors are used, misalignment can occur in the frequency spectrum at all frequencies from one to six times the frequency of rotation.

Soft foot usually occurs at a frequency of 1x rpm, but is occasionally present at two and three times rpm [12], which once again confirms that detection of faults only by observing characteristic frequencies does not always lead to clear conclusions.

B. *Data Collection*

The most commonly used transducer for measuring vibration is a piezoelectric accelerometer. The construction of accelerometer is such that the piezoelectric properties of certain crystals and ceramics are used to generate an electric signal proportional to strain. Due to its design, this sensor does not require any additional power source, which makes it easy to transform a mechanical signal into an electrical signal. The wide frequency and dynamic range with good linearity in all ranges represent the most important advantages of this transducer [13].

C. *Frequency Analysis*

The transition from the time domain, in which the data were recorded, to the frequency domain, is done via the Discrete Fourier Transformation (DFT). DFT is obtained by discretizing the Fourier transform of a discrete signal into N frequencies uniformly distributed over an interval of length 2π [14]:

$$\omega_k = \frac{2\pi}{n} k, k = 0,1,2 \ldots N-1. \quad (1)$$

The discrete Fourier transform of the signal $x(n)$ into N points is calculated by the formula (2).

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j\omega_k n}, k = 0,1,2 \ldots N-1. \quad (2)$$

In the frequency domain, analysis of mechanical failure becomes easier, due to knowledge of characteristic frequencies, but mechanical failures can be hardly distinguishable among them self.

III. CREATING A DATASET

All data were recorded at a laboratory drive located at the Faculty of Electronic Engineering in Nis, in the laboratory for Electric Drives. Fig. 2. shows a drive consisting of an induction motor on which the vibration was measured and the permanent magnet motor which is served for load emulation. The drive is coupled with a flexible coupling.

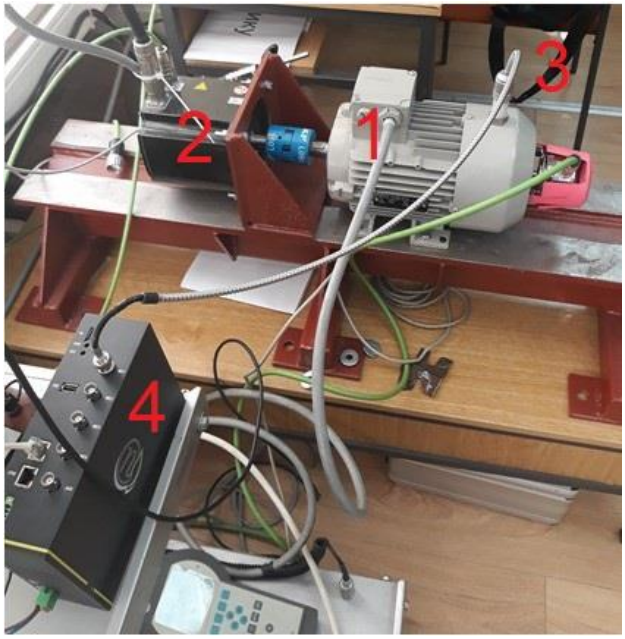


Fig. 2. Laboratory test setup: 1 – IM under test, 2 – PMSM for load emulation, 3 – piezoelectric accelerometer, 4 – DAQ system

Induction motor, 1.5 kW, was connected via an autotransformer to the power grid, which enabled the gradual commissioning. There was no speed control, and the IM worked at a nominal speed of 1420 revolutions per minute. Horizontal and vertical vibrations were measured. The load is changed in the range from 0 to nominal, with a step of 25% of the nominal load. To collect signals of horizontal and vertical vibrations, we used a high-sensitivity (100 mV/g) Metrix accelerometer. Data acquisition is done

using a device that has the ability to collect data from 25600 samples per second. Data collection took 10 seconds for each signal.

In Fig. 3, typical vibration signal without load variations for five operating conditions are displayed. From here it can be concluded that angle and parallel misalignment have far greater amplitude of vibration than normal operating mode, soft foot and rotator imbalance. Also, the soft foot signal behaves variable in time, with a period of 2.5 seconds. This means high frequencies fade in and fade out every 2.5 seconds. Detecting the fault is possible with a window length of 2.5 seconds. However, observing windows of this width can lead to an alternate change in the detected operating conditions and therefore a window with a width of 10 seconds is observed. This provides more information to the machine learning algorithm.

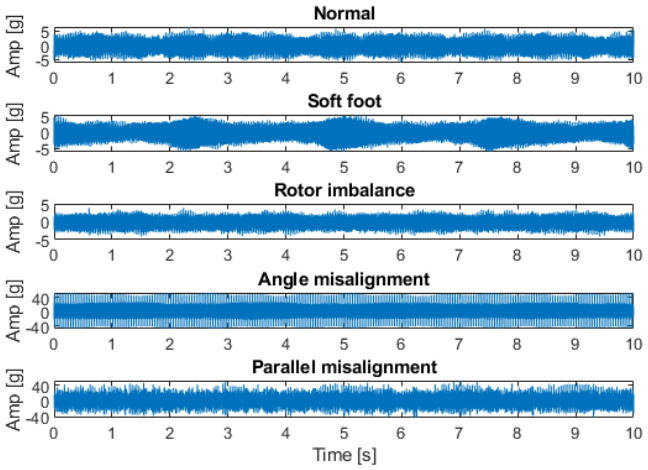


Fig. 3. Vibration signals of operating states in the time domain

The transformation of a signal from a time domain into a frequency domain is performed using algorithm of Fast Fourier transformation. Schematic representation of the data collection process is shown in Fig. 4.

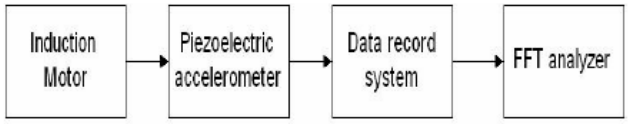


Fig. 4. Schematic representation of the data collection process

As expected, different anomalies result in different frequency vibration spectrum. Literature [10] - [12] suggest usage of frequency up to six times of the shaft rotational frequency (125 Hz in our case). With data inspection, distinguishable signature of different anomalies was observed even on higher frequencies (this can be seen in Fig.5.). Due to the lack of a clear difference between mechanical failures, because most failures occur on the same characteristic frequencies, which are shown in Section II, for this analysis, a wider frequency spectrum (up to 350 Hz) is used.

Fig. 5. shows the frequency spectrum of the signals of all five operating states. To distinguish operating modes, in addition to the characteristic frequency components other components in the frequency spectrum can be observed. This analysis provides more information to machine learning algorithms, which use them to make decisions about the affiliation of the failure class.

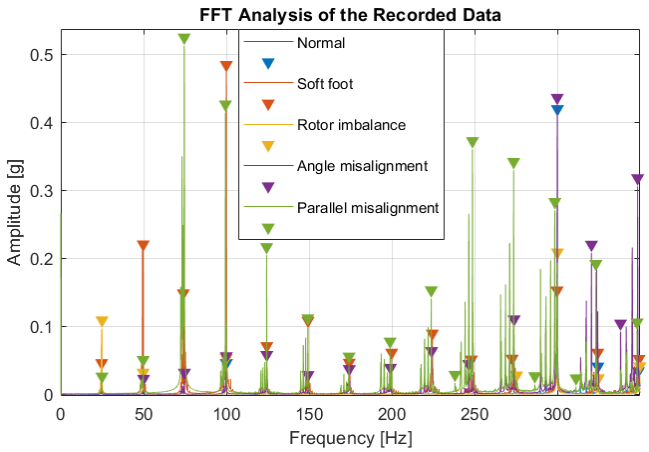


Fig. 5. Frequency spectrum of vibration signals for five operating modes

Components with the largest amplitude, as well as the frequencies at which they occur, are expected to have the greatest influence on the correct detection of the operating mode. In Fig. 5., frequency spectrum of vibration from several examples of all working conditions is presented.

As a tool for justification of selected features, MATLAB's feature selection using neighborhood component analysis for classification is used [15]. For the test in the feature selection algorithm, 30 features in total are used. The first 15 are amplitudes of 15 highest peaks in frequency spectrum in range up to 300 Hz. The other half of the features are their corresponding frequencies. The obtained results are shown in Fig. 6.

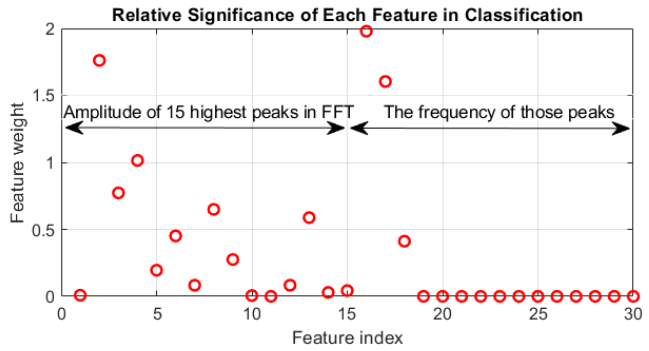


Fig. 6. Impact of features on the correct detection of the operating mode

The results confirm the dominant impact of the components with the highest amplitudes. It also concludes that it is sufficient to use the dataset that consists of ten components of frequency spectra (peaks) with the highest amplitude and corresponding frequencies, because the last five components have no significant impacts. Machine learning algorithms decide about affiliation of the class of failure, based on twenty features. Because of the reduced number of features, classification models are faster in prediction than if all thirty features are used.

Table II gives an overview of the operating conditions of the induction motor and class labels that are necessary for using the dataset in supervised machine learning.

TABLE II
OVERVIEW OF OPERATING MODES AND CLASS LABELS

Operating mode	Class label
Normal operating mode	0
Soft foot	1
Imbalance	2
Angle Misalignment	3
Parallel Misalignment	4

Share of each class in the dataset is presented in Fig. 7.

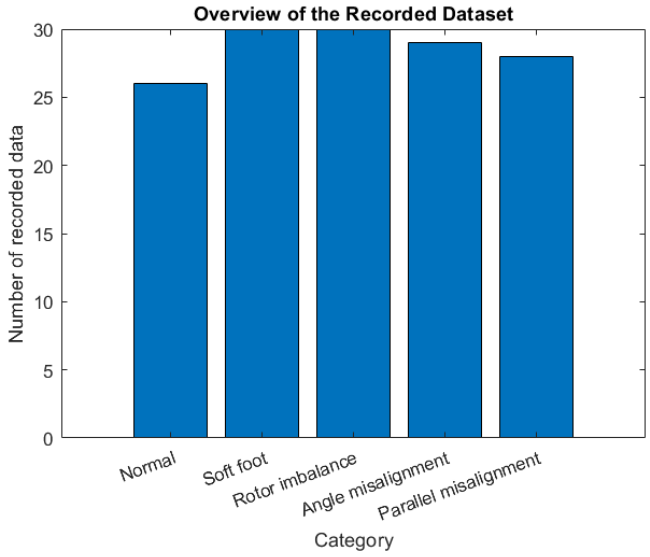


Fig. 7. Overview of recorded dataset

This dataset is loaded into the MATLAB application *Classificaton Learner*. The method of using the MATLAB application *Classification Learner* is described in [16].

IV. ANALYSIS OF CLASSIFICATION MODELS

The measure for assessing the quality of the formed classification model is defined as the ratio of the number of correctly classified examples to the total number of classification examples (*Accuracy*). Confusion matrix and Receiver operating characteristic curve, also called *ROC* curve are tools for analyzing the obtained classification models.

Observing only the *Accuracy* of the classification model for a multi-class problem can lead to misleading information, because there is no insight into the accuracy of the classes. The Confusion matrix is used to analyze the multi-class problem. Each row of the matrix corresponds to a predicted class. Each column of the matrix corresponds to an actual class. Correct and incorrect classifications are then filled in in the table. The total number of correct predictions for a class go into the expected row for that class value and the predicted column for that class value. In the same way, the total number of incorrect predictions for a class go into the expected row for that class value and the predicted column for that class value. At the end, matrix on the main diagonal contains information on properly classified examples. Fields above the diagonal are false positives, and under the diagonal are false negative examples. False positives are examples that are predicted as positive but it is actually negative. False negatives are examples that are predicted as negative but it is actually positive.

The *ROC* curve is the most commonly used way to visualize the performance of a classifier and this is a good way to visualize the performance of a classifier in order to select a suitable operating point, or decision threshold. Area Under Curve (AUC) is the best way to summarize classifier's performance in a single number. The *ROC* curve is a two-dimensional representation that contains data from the confusion matrix, which has a false positive rate on the x-axis of the selected class, and at the y-axis has a true positive rate. A false positive rate is the proportion of negative examples that are wrongly classified as positive, and the true positive rate represents the proportion of correctly classified examples from the total number of examples. Point (0,1) is the perfect classifier: classifies all positive and all negative examples correctly [17].

The first classification model was formed using the decision tree algorithm. The accuracy of prediction of new (unlabeled) data is 89.5%, while the success of data prediction by class is shown in Fig. 8.

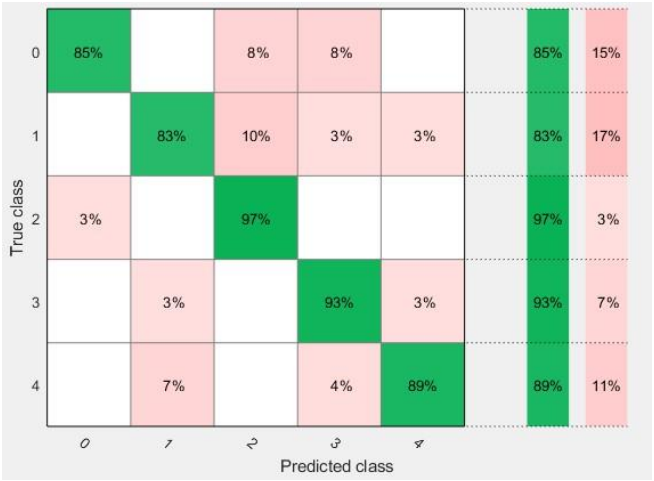


Fig. 8. Confusion matrix – decision tree

From the confusion matrix of this model we can conclude that all data by class is classified with high precision. The results show that the algorithm clearly separates the imbalance rotor from parallel and angle misalignment, and has certain problems with the distinction in classes 0 and 1. This happens because the soft foot at some moments during oscillation has similar characteristics as the rotor imbalance. Also, under load that is close to nominal, vibrations in all operating modes are damped, which affects the possibility of differentiation.

Fig. 9. shows the *ROC* curve of the decision tree algorithm whose x-axis has a false positive rate of Class 2, and on the y-axis, a true positive rate of the same class. This class is shown because it has the highest rate of positively classed examples and the current classifier is closest to the ideal classifier.

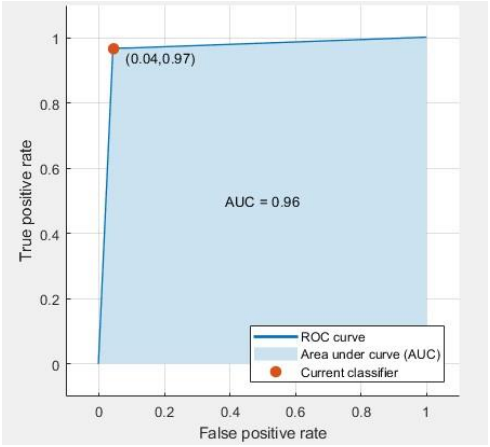


Fig. 9. *ROC* curve for class 2 – decision tree

The second observed model was formed by quadratic support vector machine algorithm (SVM). The accuracy of this model is 90.9%. The confusion matrix is shown in Fig. 10.

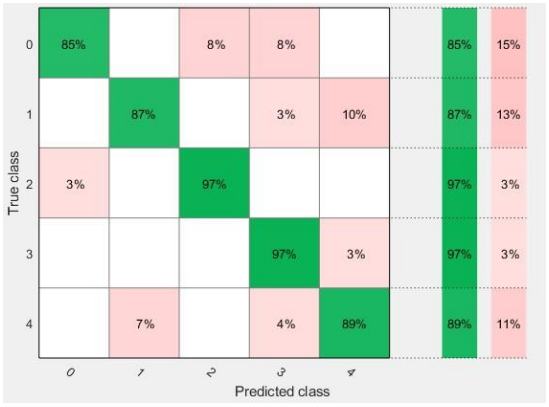


Fig. 10. Confusion matrix – support vector machine

The classification model formed by this algorithm shows the highest accuracy in distinguishing faults, but also in some cases makes mistakes because the characteristic frequencies of faults are similar.

Observing the *ROC* curve of Class 3 shown in Fig. 11, it can be concluded that this class is most precisely classified because the current classifier is very close to the ideal classifier.

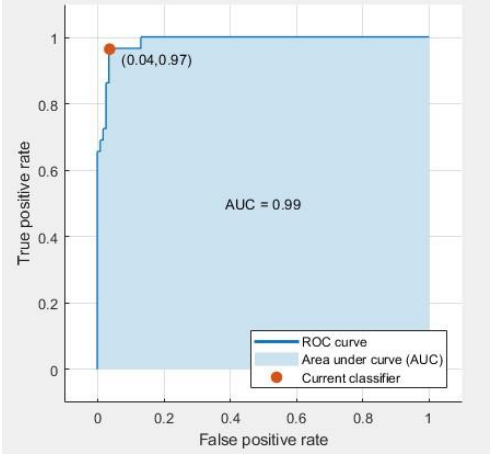


Fig. 11. *ROC* curve for class 23 – support vector machine

The last considered classification model was formed by the k-Nearest Neighbors algorithm (k-NN). A model with the highest precision is obtained when a decision on belonging to the class is made by observing the 10 nearest neighbors. The accuracy in this case is 83.2%. The confusion matrix is shown in Fig. 12.

Fig. 12. Confusion matrix – k-NN

This model has an additional problem with Class 0 and Class 3, which this model does not recommend in the analysis of faults for classes 0 to 2, but to distinguish misalignments compared to other operating modes can be used.

V. Conclusion

The obtained results show very good qualities of classification models trained with decision tree and machine support vectors algorithms. The quality of these models is primarily reflected in the ability to very well distinguish correctly from an incorrect operating condition, but also to classify vibrating signals of faulty operating modes. Software that can predict the future operating conditions of the motor by measuring vibrations would allow undesired operating states not to occur frequently. This would also reduce the variable operating costs of the plant, which was the aim that was set at the beginning. For the development and implementation of such software, it is necessary to create a classification model that will have a large database for training. This is exactly the task for future research. Also, in the future work will be considered motor with two or more defective operating conditions that exist at the same time, in order to obtain a more robust classification model.

Acknowledgement

The authors would like to thank professor Milutin Petronijević and Lazar Sladojević for providing advice throughout writing the paper. This work was supported by the Ministry of Science and Technological Development, Republic of Serbia (Project number: III 44006).

References

[1] Sundin, P.O., Montgomery, N. and Jardine, A. K. S. "Pulp mill on-site implementation of CBM decision support software" Proceedings of International Conference of Maintenance Societies, Melbourne, Australia, 2007.

[2] B. Luo, H. Wang, H. Liu, B. Li, F. Peng. "Early Fault Detection of Machine Tools Based on Deep Learning and Dynamic Identification. IEEE Transactions on Industrial Electronics, V. 66, Issue 1, pp. 509-518, 2018.

[3] V. Thorsen, M. Dalva, "*Methods of Condition Monitoring and Fault Diagnosis for Induction Motors*", European Transactions on Electrical Power 8(5):383 – 395, 2007
[4] D. Hand, H. Mannila, P. Smyth, "*Principles of Data Mining*", A Bradford Book The MIT Press, Cambridge, Massachusetts Institute of Technology, 2001
[5] Cortes, C., & Vapnik, V. (1995). "*Support-vector networks*". Machine learning, 20(3):273-297
[6] A. A. Pinheiro, I. M. Brandao, C. da Costa, "*Vibration Analysis of Rotary Machines Using Machine Learning Techniques*", EJERS, European Journal of Engineering Research and Science Vol. 4, No. 2, pp. 12-15, February 2019
[7] A. Moosavian, H. Ahmadi, A. Tabatabaeefar, "Fault diagnosis of main engine journal bearing based on vibration analysis using Fisher linear discriminant, K-nearest neighbor and support vector machine", Journal of Vibroengineering . Jun 2012, Vol. 14 Issue 2, pp. 894-906.
[8] N. Nguyen, J. Kwon, H. Lee, "*Fault Diagnosis of Induction Motor using Decision Tree with An Optimal Feature Selection*", The 7th International Conference on Power Electronics, pp. 729-732, Exeo, Daegu, Korea, October, 2007
[9] R. B. Randall, *Vibration-based Condition Monitoring, 3th ed.*, Chichester, John Wiley & Sons, Ltd, 2011
[10] G. H. Bate, "Vibration Diagnostics for Industrial Electric Motor Drives", Bruel & Kjaer, link: https://www.bksv.com/media/doc/BO0269.pdf (Accessed: 10. Apr. 2019)
[11] J. Piotrowski, "*Shaft Alignment Handbook*", *3td ed.*, CRC Press Taylor & Francis Group, New York, 2007.
[12] I. Hamernick, "*Detect soft foot with vibration analysis*". Available at: https://www.reliableplant.com/Read/920/soft-foot-vibration-analysis (Accessed 11. Apr. 2019)
[13] Bruel & Kjaer, "*Measuring Vibration*". Available at: https://www.bksv.com/media/doc/br0094.pdf (Accessed: 10. Apr. 2019).
[14] Spektralna analiza audio signala, Available at: http://dsp.etfbl.net/multimediji/2017/08a_audio_spektralna_analiza.pdf (Accessed 12. Apr. 2019.)
[15] *Feature selection using neighborhood component analysis for classification* - MATLAB. [Online]. Available: https://www.mathworks.com/help/stats/fscnca.html?s_tid=mwa_osa_a. (Accessed: 12. May 2019).
[16] U. Ilić, "Primena klasifikacionih algoritama mašinskog učenja za analizu mehaničkih kvarova u električnim mašinama," IEEESTEC student project conference, pp. 321-324, Niš, 2018.
[17] S. Raschka, "*Python Machine Learning*", 1st ed., Packt Publishing Ltd, Birmingham, UK, 2015.

Design and Analysis of the Droop Control Method for Parallel Inverters Operation in the Autonomous Microgrid

Bojan Banković, Nebojša Mitrović, Milutin Petronijević, Filip Filipović and Vojkan Kostić

***Abstract*—This paper evaluates the stability of the autonomous low inertia low voltage microgrid. Tested microgrid consists of two inverters connected to the load via power lines of different impedance. The decentralized control of the inverters is realised through the application of the conventional and opposite droop method. For both control methods dynamic phasor model is derived for stability analysis. Validation of initially selected and then retuned droop control parameters for both methods is done by simulation in order to show credibility of stability analysis method. Dynamic and steady state active and reactive power sharing is shown along with the roots of the characteristic equation. Simulation of the autonomous microgrid operation with applied droop control strategies is performed in MATLAB/Simulink software.**

***Index Terms*— Autonomous microgrid, conventional droop control, opposite droop control, parallel inverter operation.**

I. INTRODUCTION

THE contribution of the Renewable Energy Sources (RES) to the total electricity power production is increasing, taking over the majority share in relation to traditional energy sources [1]. The systems based on local energy production increase the reliability of the power supply, efficiency and cost effectiveness. Such systems integrate a large number of low-power energy sources in different locations forming microgrids with Distributed Energy Sources (DES). The most of the energy sources in the microgrid are connected through the power electronic devices in its various topologies [2]. These devices are connected in parallel bringing the new challenges in the power delivery, system stability, reliability and energy efficiency comparing to the traditional power systems [3]. Substantial efforts are made in order to find more efficient, reliable and robust control algorithms for its control.

Bojan Banković is with the Faculty of Electronic Engineering, University of Niš, 14 Aleksandra Medvedeva, 18000 Niš, Serbia (e-mail: bojan.bankovic@ elfak.ni.ac.rs).

Nebojsa Mitrović is with the Faculty of Electronic Engineering, University of Niš, 14 Aleksandra Medvedeva, 18000 Niš, Serbia (e-mail: nebojsa.mitrovic@ elfak.ni.ac.rs).

Milutin Petronijević is with the Faculty of Electronic Engineering, University of Niš, 14 Aleksandra Medvedeva, 18000 Niš, Serbia (e-mail: milutin.petronijevic@ elfak.ni.ac.rs).

Filip Filipović is with the Faculty of Electronic Engineering, University of Niš, 14 Aleksandra Medvedeva, 18000 Niš, Serbia (e-mail: filip.filipovic@ elfak.ni.ac.rs).

Vojkan Kostić is with the Faculty of Electronic Engineering, University of Niš, 14 Aleksandra Medvedeva, 18000 Niš, Serbia (e-mail: vojkan.kostic@ elfak.ni.ac.rs).

A definition of the microgrid is given in [4] as a group of interconnected loads and distributed energy resources within clearly defined electrical boundaries, that acts as a single controllable entity with respect to the grid. Some rural areas due to lack of the technical possibility for connection to the main power grid, operate as an autonomous microgrid. Coordination of power sources in microgrid is done through multistage management and control structures, followed by an optional communication infrastructure [5]. A decentralized control of each generation unit with communication-less control algorithm is performed in order to secure reliability, power flow control and safe operation, with the application of the droop control concept. The droop control is implemented in grid-supporting inverters to regulate the exchange of active and reactive power with the grid in order to keep the voltage frequency and amplitude in desired boundaries [6].

In literature, droop control strategy is usually selected according to the ratio of line resistance and inductance. For a grid with dominantly inductive lines, active power control is linked to the frequency and reactive power is linked to the voltage (P-ω/Q-V). Basic droop control that utilises this dependence is called conventional, or direct droop. For a low voltage grid with a dominant active resistance, the voltage amplitude depends mainly on the active power, while the voltage frequency is affected by the reactive power (P-V/Q-ω). If this basic droop control concept is used, it is called opposite droop. The problem with opposite droop is inability of proportional active power sharing, thus the tendency for the load to be supplied by the closest inverter in the grid [7].

The purpose of this paper is examination of the microgrid system stability and load sharing capabilities among the inverters for different designed droop control strategies. The testing was done on a model of two inverters in autonomous low voltage microgrid with low inertia, supplying the load through different line impedances. Along with opposite droop recommended for this case, direct droop strategy is also tested with coefficients calculated according to [9]. Validity of selected coefficients is verified using dynamic phasor model (DPM) analysis [10]. The transient response in the load step change and steady state power sharing performance are presented. The microgrid system of the two DES is modelled and simulated in MATLAB/Simulink R2018b software. Inverter, power lines, load and measurements are built using generic components from Simulink Simscape library, while the complete control subsystem is developed by authors.

II. Basic Droop Control Methods

The microgrid system examined in this paper is presented in Fig. 1. It consists of two DESs with the Voltage Source Inverters (VSI) along with their LCL filters, connected in parallel and supplying consumers concentrated at one connection point. The LCL filter is used to mitigate VSI's switching harmonics. The consumers are at different distances from DESs. This implies the difference in power line impedance between two inverters ($\underline{Z}_{l1}$ and $\underline{Z}_{l2}$).

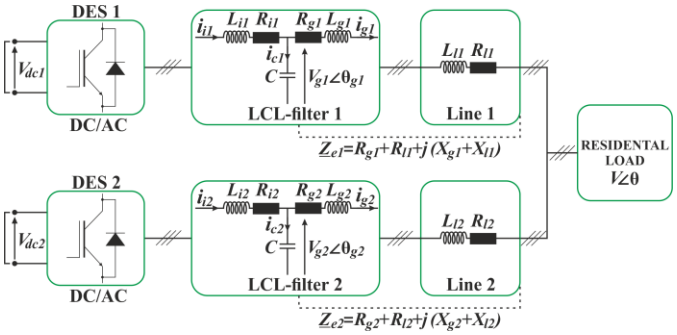


Fig. 1. The autonomous microgrid system with two DES.

The inverter output active and reactive power are given by [10]:

$$P_n = \frac{3}{R_{en}^2 + X_{en}^2}\left(R_{en}V_{gn}^2 - R_{en}V_{gn}V\cos(\delta_n) + X_{en}V_{gn}V\sin(\delta_n)\right), \quad (1)$$

$$Q_n = \frac{3}{R_{en}^2 + X_{en}^2}\left(X_{en}V_{gn}^2 - X_{en}V_{gn}V\cos(\delta_n) - R_{en}V_{gn}V\sin(\delta_n)\right), \quad (2)$$

where subscript n denotes n-th inverter branch information, δ_n (power angle), represent the phase angle difference between the capacitor voltage V_{gn} and load voltage V. Namely, the capacitor voltages are selected as output voltage control variables. In above equations, the LCL filter grid side impedance $\underline{Z}_{gn}$ is added to the power line impedance $\underline{Z}_{ln}$ and equivalent impedance $\underline{Z}_{en}=\underline{Z}_{gn}+\underline{Z}_{ln}=R_{en}+jX_{en}$. Typical line parameters for different voltage levels are given in Table I according to [6] and [8].

TABLE I
Line Parameter for Different Voltage Levels

Type of Line	R [Ω/km]	X [Ω/km]	Ratio, R/X
Low Voltage	0.642	0.083	7.7
Medium Voltage	0.161	0.190	0.85
High Voltage	0.06	0.191	0.31

A. Direct Droop Method

A P-ω/Q-V droop method has been introduced for microgrid control in [10]. This control method represents the simple communication-less/decentralized control with plug and play feature for new DES installation.

For a microgrid with dominantly inductive impedance, line active resistance can be neglected, thus simplifying the expression (1) and (2) in the form presented in (3) and (4):

$$P_n = 3\frac{V_{gn}V}{X_{en}}\sin\delta_n, \quad (3)$$

$$Q_n = 3\frac{V_{gn}\left(V_{gn} - V\cos\delta_n\right)}{X_{en}}. \quad (4)$$

Strong active power-power angle and reactive power-inverter output voltage correlation can be observed from the equations (3) and (4). Using direct droop method, the inverter output frequency ω and the inverter output voltage V_g can be controlled by means of the droop characteristics:

$$\omega_n = \omega_n^* - k_{pn}(P_n - P_n^*), \quad (5)$$

$$V_{gn} = V_{gn}^* - k_{qn}(Q_n - Q_n^*), \quad (6)$$

where k_{pn} and k_{qn} are the frequency and voltage magnitude droop coefficients of the n-th inverter. In the expressions (5) and (6) ω_n^* and V_{gn}^* presents capacitor voltage frequency and phase amplitude setpoint, while ω_n and V_{gn} are measured values of the same variables. The active power P and reactive power Q are filtered references from power calculation block in droop controller presented in Fig. 2.

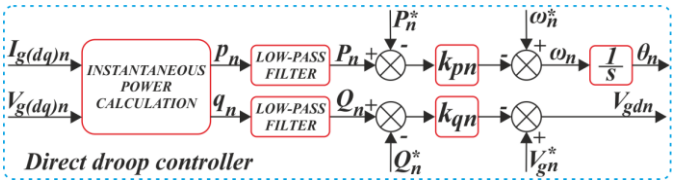


Fig. 2. Direct droop control structure.

The power sharing via direct droop control techniques is highly affected by the value of the line impedance and grid elements like transformers between the load's connection and the power supply side [5].

B. Opposite Droop Method

The modification of direct droop method suitable for operation in grid with dominant active resistance in lines is described in [7]. By neglecting the inductance in lines, equations (1) and (2) can be presented in the following form:

$$P_n = 3\frac{V_{gn}\left(V_{gn} - V\cos\delta_n\right)}{R_{en}}, \quad (7)$$

$$Q_n = -3\frac{V_{gn}V}{R_{en}}\sin\delta_n. \quad (8)$$

Opposite of direct droop, here can be seen strong active power-inverter output voltage and reactive power-power angle correlation. Using opposite droop method, the inverter output frequency ω and the inverter output voltage V_g are controlled

by means of the opposite characteristics:

$$\omega_n = \omega_n^* + k_{qon}(Q_n - Q_n^*), \tag{9}$$

$$V_{gn} = V_{gn}^* - k_{pon}(P_n - P_n^*), \tag{10}$$

where the k_{qon} and k_{pon} are voltage frequency and amplitude opposite droop coefficients of the n-th inverter. The advantages of the opposite droop method compared to the conventional droop control are presented in [11]. The opposite droop control structure is shown in Fig. 3.

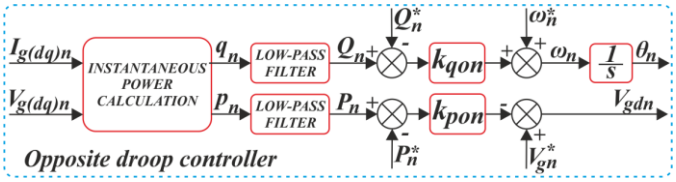


Fig. 3. Opposite droop control structure.

C. Additional Considerations

For the parallel connected VSIs, the power distribution has been achieved according to proper droop coefficient selection for each inverter. Their initial selection is done according to the desired voltage frequency and amplitude change for the maximum appropriate power component change ((11) for direct and (12) for opposite droop).

$$k_{pn} = \frac{\Delta\omega_n}{\Delta P_{\max,n}} \wedge k_{qn} = \frac{\Delta V_{gn}}{\Delta Q_{\max,n}}, \tag{11}$$

$$k_{pon} = \frac{\Delta V_{gn}}{\Delta P_{\max,n}} \wedge k_{qon} = \frac{\Delta\omega_n}{\Delta Q_{\max,n}}. \tag{12}$$

If the load share among the inverters proportional to their rated power is desired, conditions (13) or (14) must be fulfilled for direct or opposite control, respectively:

$$\Delta P_1 k_{p1} = \ldots = \Delta P_n k_{pn} \wedge \Delta Q_1 k_{q1} = \ldots = \Delta Q_n k_{qn}, \tag{13}$$

$$\Delta P_1 k_{po1} = \ldots = \Delta P_n k_{pon} \wedge \Delta Q_1 k_{qo1} = \ldots = \Delta Q_n k_{qon}. \tag{14}$$

Output voltage amplitude and frequency of each inverter is generated according to its calculated active and reactive power. If instantaneous output active (p) and reactive power (q) are estimated in synchronous (dq) frame, equations for power calculation are given in (15):

$$p = \frac{3}{2}\left(V_{gd}I_{gd} + V_{gq}I_{gq}\right) \wedge q = \frac{3}{2}\left(V_{gd}I_{gq} - V_{gq}I_{gd}\right). \tag{15}$$

Instantaneous active and reactive power are not used per se in the droop control algorithm. These quantities are averaged over certain time period, usually using a first order low pass filter.

A complete structure of the droop controller for one inverter with inner current control loop and outer voltage control loop is presented in Fig. 4, according to [6].

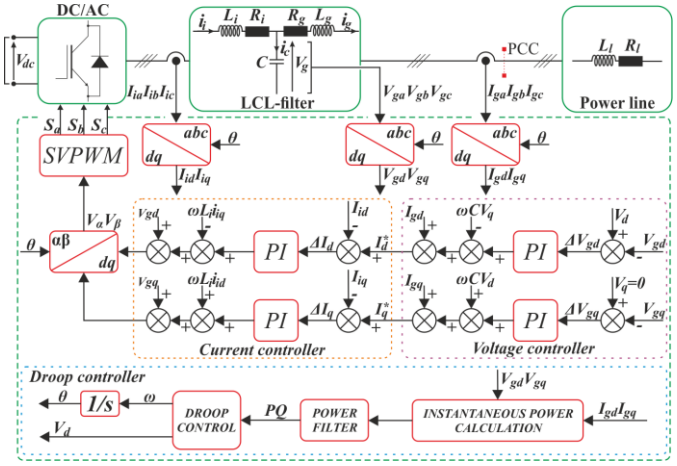


Fig. 4. Applied droop control structure for one inverter.

The output quantities from droop control block are voltage reference V_d and frequency ω. The angle θ for ($abc \rightarrow dq$) transformation is obtained through frequency integration.

III. Droop Coefficients Tuning for Low Voltage Grid

For the system stability analysis, dynamic phasor modelling is used [10]. Impact of the power low pass filter on the system stability cannot be neglected due to relatively large time constant. The influence of the current and voltage control structures is neglected in this stability analysis. In order to obtain characteristic equation of the system using DPM, equations (1) and (2) have to be rewritten and power low pass filter transfer function is induced as:

$$P_n = \frac{3\left(L_{en}s + R_{en}\right)\left(V_{gn}^2 - V_{gn}V\cos(\delta_n)\right)}{\left(L_{en}s + R_{en}\right)^2 + \left(\omega L_{en}\right)^2} + \frac{3\left(\omega L_{en}\right)V_{gn}V\sin(\delta_n)}{\left(L_{en}s + R_{en}\right)^2 + \left(\omega L_{en}\right)^2}, \tag{16}$$

$$Q_n = \frac{3\left(\omega L_{en}\right)\left(V_{gn}^2 - V_{gn}V\cos(\delta_n)\right)}{\left(L_{en}s + R_{en}\right)^2 + \left(\omega L_{en}\right)^2} - \frac{3\left(L_{en}s + R_{en}\right)V_{gn}V\sin(\delta_n)}{\left(L_{en}s + R_{en}\right)^2 + \left(\omega L_{en}\right)^2}, \tag{17}$$

$$P = \frac{\omega_f}{s + \omega_f}p \wedge Q = \frac{\omega_f}{s + \omega_f}q, \tag{18}$$

where s denotes Laplace variable and ω_f is a crossover frequency of first order low pass filter. For small disturbances around equilibrium point (δ_e, V_{ge}, V_e), linearized version of equations (16) and (17) can be obtained along with appropriate coefficients as:

$$\Delta P_n = k_{pen}\Delta V_{gn} + k_{pdn}\Delta\delta_n, \tag{19}$$

$$\Delta Q_n = k_{qen}\Delta V_{gn} + k_{qdn}\Delta\delta_n, \tag{20}$$

$$k_{pen} = \frac{3\left(L_{en}s + R_{en}\right)V_{gn}}{\left(L_{en}s + R_{en}\right)^2 + \left(\omega L_{en}\right)^2}, \tag{21}$$

$$k_{pdn} = \frac{3\omega L_{en}V_{gn}^2}{\left(L_{en}s + R_{en}\right)^2 + \left(\omega L_{en}\right)^2}, \tag{22}$$

$$k_{qen} = \frac{3\omega L_{en}V_{gn}}{\left(L_{en}s + R_{en}\right)^2 + \left(\omega L_{en}\right)^2}, \tag{23}$$

$$k_{qdn}=\frac{-3\left(L_{en}s+R_{en}\right)V_{gn}^2}{\left(L_{en}s+R_{en}\right)^2+\left(\omega L_{en}\right)^2}. \quad (24)$$

Equations (19)-(24) have to comply in direct and opposite method. Derivation of the model used for dynamic phasor analysis will be conducted separately for direct and opposite droop control in the next two sections.

A. Direct Droop Method Dynamic Phasor Model

Application of direct droop control on the specified low voltage grid would be considered inadequate following general recommendations. It is shown in [8] that direct droop control has indeed desired behaviour in grid with high active resistance, if droop coefficients from certain range are selected. In this paper droop coefficients will be selected according to the stability analysis from the DFM. Linearized version of equations (5) and (6) can be written in form:

$$s\Delta\delta_n=\Delta\omega_n^*-k_{pn}\Delta P_n+k_{pn}\Delta P_n^*, \quad (25)$$

$$\Delta V_{gn}=\Delta V_{gn}^*-k_{qn}\Delta Q_n+k_{qn}\Delta Q_n^*. \quad (26)$$

Combination of equations (18)-(20) and (25)-(26) provides following relations:

$$\Delta\omega_n\Big|_{\Delta\omega_n^*=\Delta P_n^*=0}=\frac{-k_{pn}\omega_f}{s+\omega_f}\left(k_{pen}\Delta V_{gn}+k_{pdn}\Delta\delta_n\right), \quad (27)$$

$$\Delta V_{gn}\Big|_{\Delta\omega_n^*=\Delta P_n^*=0}=\frac{-k_{qn}\omega_f}{s+\omega_f}\left(k_{qen}\Delta V_{gn}+k_{qdn}\Delta\delta_n\right). \quad (28)$$

Characteristic equation is obtained with combination (18)-(20) and (25)-(28) in the form:

$$as^5+bs^4+cs^3+ds^2+es+f=0, \quad (29)$$

where appropriate coefficient *a* to *f* can be found in [10].

B. Opposite Droop Method Dynamic Phasor Model

Linearized version of equations (9) and (10) is derived in equations (30) and (31) as:

$$s\Delta\delta_n=\Delta\omega_n^*+k_{qon}\Delta Q_n-k_{qon}\Delta Q_n^*, \quad (30)$$

$$\Delta V_{gn}=\Delta V_{gn}^*-k_{pon}\Delta P_n+k_{pon}\Delta P_n^*. \quad (31)$$

With combination of (18)-(20) and (30)-(31) relations (32) and (33) are provided as:

$$\Delta\omega_n=\frac{k_{qon}\omega_f}{s+\omega_f}\left(k_{qen}\Delta V_{gn}+k_{qdn}\Delta\delta_n\right), \quad (32)$$

$$\Delta V_{gn}=\frac{-k_{pon}\omega_f}{s+\omega_f}\left(k_{pen}\Delta V_{gn}+k_{pdn}\Delta\delta_n\right). \quad (33)$$

From previous two equations, characteristic equation of the form presented in (29) can be obtained, with the coefficients:

$$a=L_{en}^2, \quad (34)$$

$$b=2R_{en}L_{en}+2\omega_f L_{en}^2, \quad (35)$$

$$c=R_{en}^2+\omega^2L_{en}^2+4\omega_f R_{en}L_{en}+\omega_f^2L_{en}^2+3\omega_f L_{en}V_{gn}k_{pon}, \quad (36)$$

$$d=2\omega_f R_{en}^2+2\omega_f\omega^2L_{en}^2+2\omega_f^2R_{en}L_{en}+3\omega_f L_{en}V_{gn}^2k_{qon}+ +3\omega_f^2L_{en}V_{gn}k_{pon}+3\omega_f R_{en}V_{gn}k_{pon}, \quad (37)$$

$$e=\omega_f^2R_{en}^2+\omega_f^2\omega^2L_{en}^2+3\omega_f^2R_{en}V_{gn}k_{pon}+ +3\omega_f R_{en}V_{gn}^2k_{qon}+3\omega_f^2L_{en}V_{gn}^2k_{qon}, \quad (38)$$

$$f=3\omega_f^2R_{en}V_{gn}^2k_{qon}+9\omega_f^2V_{gn}^3k_{pon}k_{qon}. \quad (39)$$

C. Initial Coefficient Selection and Validity

Selection of droop control coefficients has a direct impact on stability of the microgrid. Their initial selection is done according to the desired voltage frequency and amplitude deviations. In this case, initial coefficient selection is made with desired voltage amplitude and frequency droop of 10 % and 1 % for rated appropriate power, respectively. Details of the analysed microgrid system are shown in Table II. It can be observed that although low voltage lines are dominantly resistive, combination of inverter inductance and lines provides dominantly inductive equivalent impedance.

TABLE II
PARAMETER OF TWO PARALLEL INVERTERS AND DISTRIBUTION LINES IN SIMULATION

DES 1 = DES 2	
DC voltage V_{dc1}, V_{dc2}	650 V
Rated active power P_1, P_2	7.5 kW
Rated reactive power Q_1, Q_2	3.75 kVAR
Inverter switching frequency f_{sw1}, f_{sw1}	10 kHz
Inverter filter resistance R_{i1}, R_{i2}	0.02 Ω
Inverter filter inductance L_{i1}, L_{i2}	4 mH
Inverter filter capacitance C_1, C_2	3.5 μF
Grid side filter resistance R_{g1}, R_{g2}	0.015 Ω
Grid side filter inductance L_{g1}, L_{g2}	2.4 mH
LINE PARAMETERS	
Line 1 $\underline{Z}_{l1}$ (l≈50 m)	(0.0321+j0.00415) Ω
Line 2 $\underline{Z}_{l2}$ (l≈250 m)	(0.1605+j0.02075) Ω
LOAD PARAMETERS	
Load 1 (pure active load R)	P=8 kW
Load 2 (RL load cosφ=0.8)	S=7.5 kVA

Initial value of coefficients according to desired droop are k_p=4.18·10^{-4} rad/(Ws), k_q=8.7·10^{-3} V/(VAR), k_{po}=4.3·10^{-3} V/(W), k_{qo}=8.37·10^{-4} rad/(VARs). Coefficients are calculated on maximum value of phase voltage. Crossover frequency of the low pass filter is held constant at ω_f=62.8 rad/s.

In Fig. 5, locations of characteristic equation poles are shown for the direct droop method. It can be observed that initial coefficients result in unstable system. If power sharing balance is to be maintained, coefficients k_p and k_q have to be coordinated selected so that all solutions of characteristic equation of every inverter have negative real part and ratio presented in (13) is maintained.

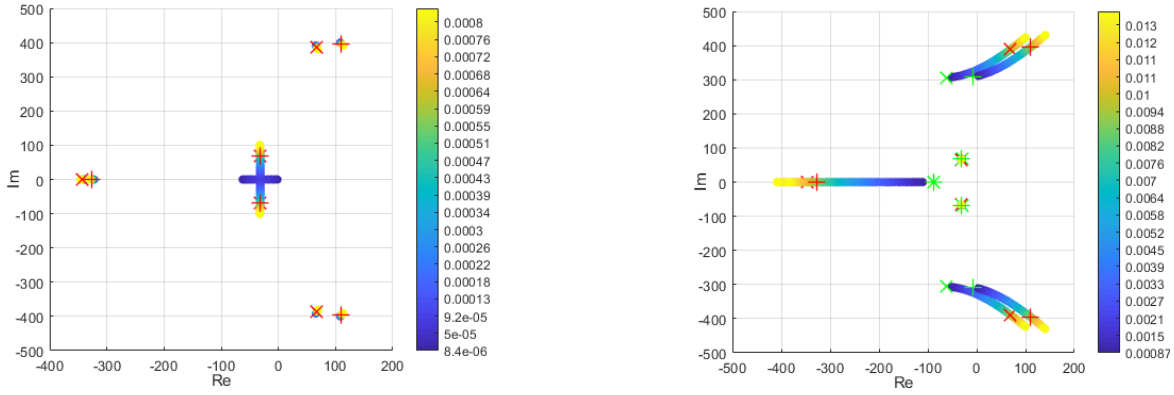


Fig. 5. Direct droop method: location of characteristic equation poles with variation of k_p (left) and k_q (right) of both inverters. Colour bar denotes value of the appropriate coefficient in the specific point. Red (green) marks represent poles of the initial (retuned) system. Inverter with shorter power line is denoted with "+", while inverter with longer supply line is denoted with "x". Selected values are k_p=4.18·10^{-4} rad/(Ws), k_q=4.3·10^{-4} V/(VAR).

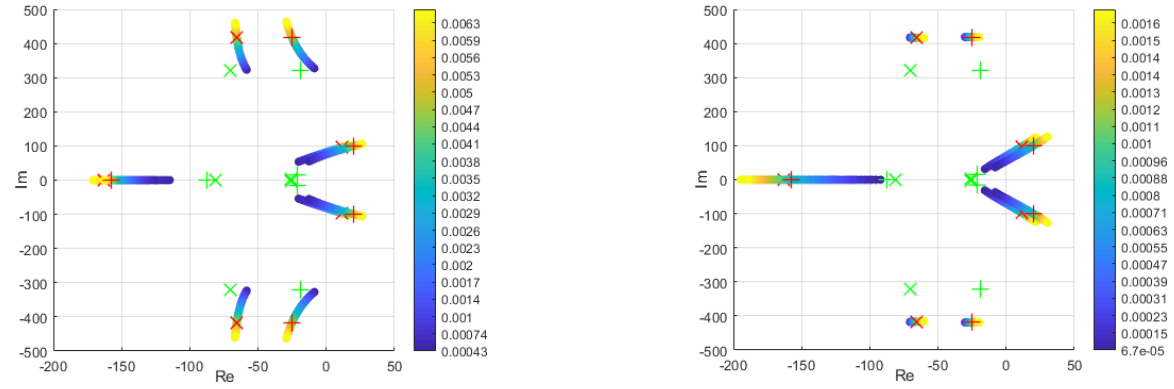


Fig. 6. Opposite droop method location of characteristic equation poles with variation of k_{po} (left) and k_{qo} (right) of both inverters. Colour bar denotes value of the appropriate coefficient in the specific point. Red (green) marks represent poles of the initial (retuned) system. Inverter with shorter power line is denoted with "+", while inverter with longer supply line is denoted with "x". Selected values are k_{po}=2.17·10^{-4} V/(W), k_{qo}=1.67·10^{-4} rad/(VARs).

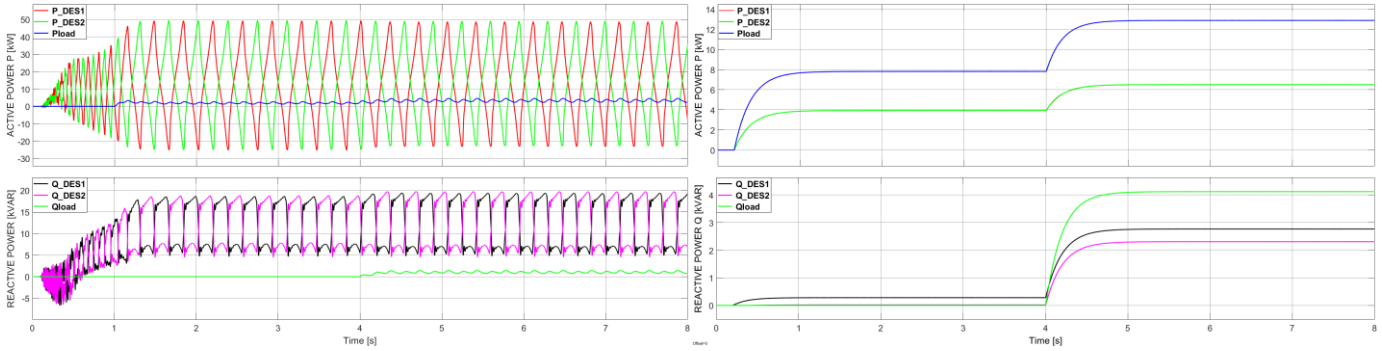


Fig. 7. Direct droop method active and reactive power sharing with initial droop coefficients (left) and with retuned coefficients (right).

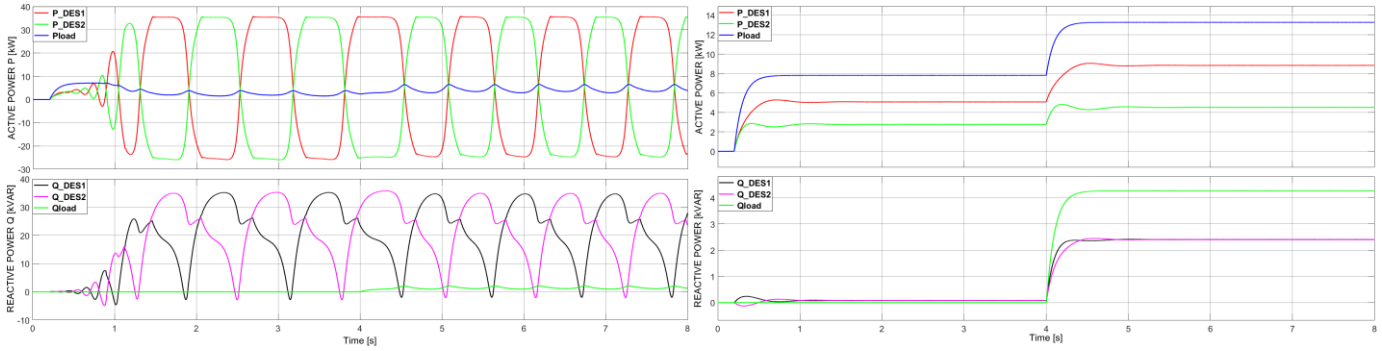


Fig. 8. Opposite droop method active and reactive power sharing with initial droop coefficients (left) and with retuned coefficients (right).

From the Fig 5. can be observed that variation of k_p has low impact on stability for selected k_q. System is retuned with only k_q variation.

Fig. 6. shows location of characteristic equation poles for k_{po} and k_{qo} variation in opposite droop method. Initial coefficients values result in unstable system. Stabilization is done with change of both parameters.

The detailed model of the system from Fig. 1 is done in MATLAB/Simulink, with the implementation of the control structure from Fig. 3 for each inverter. At the start of the simulation both inverters are operating in parallel, without load connected to the microgrid up to t=0.2 s. In that moment an active load R, power of 8 kW is connected as a consumption up to the time t=4 s. In the t=4 s the RL load of apparent power 7.5 kVA and cosφ=0.8 is added as a consumption.

Power sharing capability of inverters is presented in Figs. 7 and 8. Left graphs show active and reactive power sharing for initial droop setup and it can be observed that system is unstable. For a stable system on right graph in the same figure, characteristic power sharing can be seen. With the direct droop method, active power is shared proportionally among inverters, while the share of reactive power is disproportional. Opposite droop method enables equal share of reactive power, while closer inverter provides larger share of active power.

IV. Conclusion

Stability and load sharing performance of the conventional and opposite droop control is analysed in this paper for the load step change in the case of unequal line parameters of the microgrid. Basic direct and opposite droop control methods are presented along with typical control structure of the inverter. Validation of selected droop coefficients is done through stability analysis using dynamic phasor model of autonomous microgrid.

Direct droop method is applicable in this case in low voltage grid. Although power lines have dominantly resistive character, combination of inverter inductance and line impedance results in equivalent impedance that is predominately inductive. Active power is shared as desired, while reactive power has a sharing disbalance. Further improvements of basic droop method that address the issue of reactive power sharing will be considered in future work.

Opposite droop method, when properly tuned, enables desired sharing of reactive power, with a drawback of disproportional active power share. As expected, closer inverter supplies load with higher share of active power.

Dynamic phasor method provides credible results for stability analysis of both direct and opposite droop control strategy. Using it, autonomous microgrid is modelled via characteristic equation or fifth order, that is relatively small compared to full model analysis. This small order of equation is possible because dynamic of inverter control is neglected. Dynamic of current and voltage control loops are neglected, thus implying that obtained stability results need to be taken in consideration with that in mind.

Acknowledgment

This work was supported by the Ministry of Science and Technological Development, Republic of Serbia (Project number: III 44004 and III 44006).

References

[1] "World Energy Outlook," *WEO*. [Online]. Available: https://www.iea.org/weo/. [Accessed: 29-Apr-2019].

[2] F. Blaabjerg, K. Ma, Y. Yang, "Power electronics for renewable energy systems-status and trends," *CIPS 2014, 8th International Conference on Integrated Power Electronics Systems*, pp. 1-11, 2014.

[3] P. Palensky and D. Dietrich, "Demand Side Management: Demand Response, Intelligent Energy Systems, and Smart Loads," *IEEE Transactions on Industrial Informatics*, vol. 7, no. 3, pp. 381–388, 2011.

[4] "Microgrid Definitions," *Microgrid Definitions | Building Microgrid*. [Online]. Available: https://building-microgrid.lbl.gov/microgrid-definitions. [Accessed: 29-Apr-2019].

[5] Y. Han, H. Li, P. Shen, E. A. A. Coelho, and J. M. Guerrero, "Review of Active and Reactive Power Sharing Strategies in Hierarchical Controlled Microgrids," *IEEE Transactions on Power Electronics*, vol. 32, no. 3, pp. 2427–2451, 2017.

[6] J. Rocabert, A. Luna, F. Blaabjerg, and P. Rodríguez, "Control of Power Converters in AC Microgrids," *IEEE Transactions on Power Electronics*, vol. 27, no. 11, pp. 4734–4749, 2012.

[7] N. Hatziargyriou, *Microgrids: Architectures and Control*. John Wiley & Sons, 2014.

[8] X. Hou, Y. Sun, W. Yuan, H. Han, C. Zhong, and J. Guerrero, "Conventional P-ω/Q-V Droop Control in Highly Resistive Line of Low-Voltage Converter-Based AC Microgrid," *Energies*, vol. 9, no. 11, p. 943, 2016.

[9] X. Guo, Z. Lu, B. Wang, X. Sun, L. Wang, and J. M. Guerrero, "Dynamic Phasors-Based Modeling and Stability Analysis of Droop-Controlled Inverters for Microgrid Applications," *IEEE Transactions on Smart Grid*, vol. 5, no. 6, pp. 2980–2987, 2014.

[10] K. D. Brabandere, B. Bolsens, J. V. D. Keybus, A. Woyte, J. Driesen, R. Belmans, and K. Leuven, "A voltage and frequency droop control method for parallel inverters," *IEEE 35th Annual Power Electronics Specialists Conference, IEEE Cat. No.04CH37551*. 2004.

[11] J. Guerrero, N. Berbel, J. Matas, J. Sosa, J. Cruz, and A. Alentorn, "Decentralized control for parallel operation of distributed generation inverters using resistive output impedance," *2005 European Conference on Power Electronics and Applications*, 2005.

Образовна лабораторијска поставка пумпног система са могућношћу регулације притиска и протока

Војислав Вујичић, Марко Шућуровић, Милош Божић, Марко Росић, Мирослав Бјекић

***Апстракт*—У овом раду представљен је пумпни систем реализован на Факултету техничких наука у Чачку. Систем се састоји од резервоара, трофазне вишестепене центрифугалне пумпе, фреквентног претварача, преливног (бајпас) и пригушног вентила. Помоћу сензора притиска и протока који су постављени на цевоводу врши се аквизиција на основу које се могу одредити параметри пумпног система, и помоћу којих се може вршити регулација протока и притиска. Процедура извођења лабораторијских вежби и добијени резултати са коментарима приказани су у експерименталном делу рада.**

***Кључне речи*—пумпни систем, мерење протока, мерење притиска, аквизиција, LabVIEW, регулација**

I. Увод

Основни закони везани за хидрауличне и пумпне системе се једноставно могу потврдити кроз едукативне лабораторијске поставке. За њихову реализацију довољан је резервоар, пумпа са електро мотором, цевовод и вентили. Снимање карактеристика пумпе или хидрауличног система могуће је постављањем мерача протока и притиска на цевовод. Ови уређаји могу бити аналогни или дигитални, повезани са мерно-аквизиционим системом који бележи податке на рачунару. Примери одређивања карактеристика хидрауличног система и пумпи дати су у [1]-[2].

Пумпни системи се могу користити за одређивање карактеристика компоненти које се налазе у цевоводу као што су турбине, вентили, колена, рачве итд. На овај начин се могу упоредити теоријске карактеристике са каталошким и експериментално добијеним карактеристикама одређене компоненте. За одређивање хидрауличне снаге (или губитка снаге) неке од хидрауличних компоненти потребно је пратити улазни и излазни притисак на испитиваној компоненти као и проток кроз ту компоненту. Лабораторијска поставка развијена на Факултету техничких наука у Чачку има могућност праћења наведених параметара што отвара могућност за извођење различитих испитивања и мерења на пумпном систему.

У раду је приказана поставка пумпног система, реализоване лабораторијске вежбе за одређивања карактеристика система и пумпе, као и регулационе структуре притиска и протока на овом систему. Зависност притиска и протока (p-Q) које описују систем и центрифугалну пумпу при различитим брзинама пумпе експериментално су добијене и графички су приказане кроз две реализоване лабораторијске вежбе. Регулација протока и притиска у систему реализована и приказана кроз друге две лабораторијске вежбе. Кроз реализацију наведених лабораторијских вежби постиже се боље разумевање и усвајање теоријских знања из хидраулике као и из регулације електромоторних погона.

II. Притисак, проток, и снага у пумпном систему

За одређивање радне тачке пумпног система потребно је одредити зависност притиска и протока пумпе у датом цевоводу. Измерене вредности притиска и протока се могу искористити за одређивање зависности: брзине пумпе (n) од протока (Q), притиска (p) од излазне снаге пумпе (P). Теоријска основа која описује утицај брзине пумпе на проток, притиска на снагу су дате у једначинама (1) [3]-[4].

$$\begin{aligned} Q &= Q_{\mathrm{n}}\left(n / n_{\mathrm{n}}\right) \\ p &= p_{\mathrm{n}}\left(n / n_{\mathrm{n}}\right)^2 \\ P &= P_{\mathrm{n}}\left(n / n_{\mathrm{n}}\right)^3 \end{aligned} \tag{1}$$

У претходној једначини називне вредности су дате са индексом n. Хидраулична снага центрифугалне пумпе се може одредити коришћењем измерених вредности притиска и протока применом једначине:

$$P = 100 Q \cdot p \tag{2}$$

где је: P – хидраулична снага [W], Q –проток [l/s], p –притисак [bar].

Војислав Вујичић – Факултет техничких наука у Чачку, Универзитет у Крагујевцу, Светог Саве 65, 32000 Чачак, Србија (e-mail: vojislav.vujicic@ftn.kg.ac.rs).

Марко Шућуровић – Факултет техничких наука у Чачку, Универзитет у Крагујевцу, Светог Саве 65, 32000 Чачак, Србија (e-mail: marko.sucurovic@ftn.kg.ac.rs).

Милош Божић – Факултет техничких наука у Чачку, Универзитет у Крагујевцу, Светог Саве 65, 32000 Чачак, Србија (e-mail: milos.bozic@ftn.kg.ac.rs).

Марко Росић – Факултет техничких наука у Чачку, Универзитет у Крагујевцу, Светог Саве 65, 32000 Чачак, Србија (e-mail: marko.rosic@ftn.kg.ac.rs).

Мирослав Бјекић– Факултет техничких наука у Чачку, Универзитет у Крагујевцу, Светог Саве 65, 32000 Чачак, Србија (e-mail: miroslav.bjekic@ftn.kg.ac.rs)

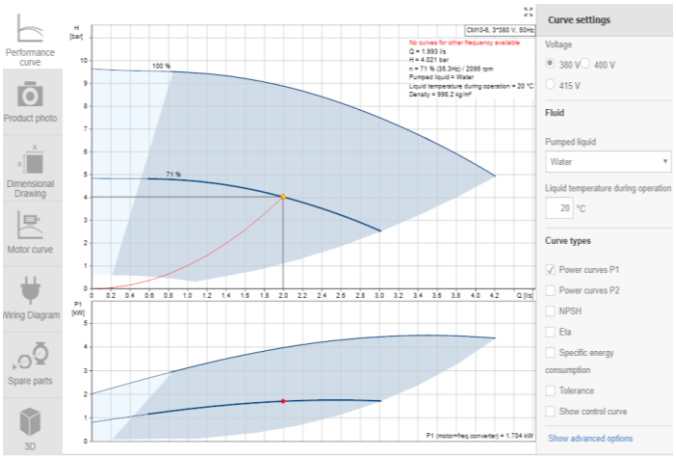
III. Експериментална поставка

У лабораторији за Процесну технику на Факултету техничких наука у Чачку, постављен је пумпни систем намењен за образовни и научно-истраживачки рад [5]. На слици 2 приказан је физички изглед пумпног система, док је на слици 3 приказана блок шема уграђених хидрауличних компоненти. Максималан притисак који се може постићи на овом систему је 6 бара, и максималан проток је 5 l/s.

Пумпни систем на слици 2 се састоји од отвореног резервоара капацитета 400 литара (1) из ког се систем напаја водом. Цевовод је направљен од 6/4" PVC цеви, које су спојене коленима и вентилима. Пумпа (3) је са шестостепеним центрифугалним радним колом које покреће асинхрони мотор. Произвођач пумпе је фирма *Grundfoss* тип CM10-6 A-R-I-E-AQQE F-A-A-N и постављена је у најнижој тачки (испод дна резервоара) [6]. Каталошки подаци одабране пумпе графички су приказани на слици 1.

Назначени подаци пумпе су: проток Q = 10 m^3/h (2,78 l/s), висина воденог стуба H = 78,3 m (притисак ≈7,7 bar), брзина обртања мотора n = 2900 min^{-1} и минимални индекс ефикасности MEI ≥ 0,52.

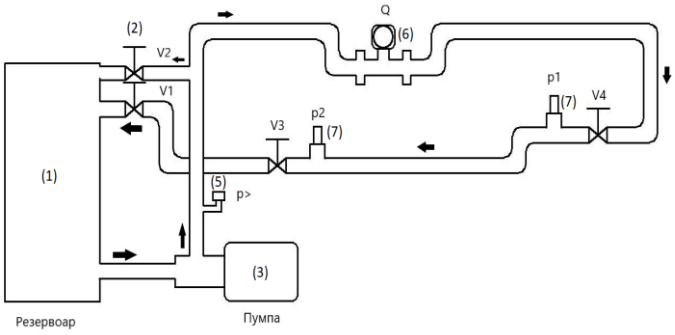
Пумпу покреће трофазни асинхрони мотор снаге 4 kW, фактора снаге 0,87-0,84, и степена искоришћења 85,5% (IE2 class).



Сл.1. Каталошки подаци произвођача [7]

За промену притиска и протока у систему (променом брзине обртања пумпе) користи се Danfoss VLT Aqua Drive FC 202 фреквентни претварач [8]. Овај претварач поседује додатне функције намењене раду са пумпним системима. Дефинисање референтне брзине пумпе се врши помоћу потенциометра (8) који се налази на вратима ормара у коме се налазе све електричне компоненте потребне за управљање системом.

Сл. 2. Изглед пумпног система у лабораторији



Сл. 3. Блок шема компоненти пумпног система

За мерење протока у цевоводу користи се електромагнетни мерач протока (6). Мерни опсег мерача протока је до 60 m^3/h (16,6 l/s). Мерач поседује екран осетљив на додир на коме се може пратити тренутна вредност протока. Проток се мери употребом струјног излаза 4 – 20 mA.

Мерење притиска врши се употребом два сензора притиска (7). Мерни опсег ових сензора је 0 – 10 бара, са струјним излазом 4 – 20 mA. Између два сензора притиска (10) налази се део цевовода који се може заменити са другим елементима који се испитују. Ти елементи могу бити вентили, колена, рачве, турбине, пумпе. У овом раду нису вршена испитивања других компоненти. За потребе предложених лабораторијских вежби између два сензора притиска постављена је цев, тако да ће притисак на оба сензора показивати исту вредност. На овај начин се одређују карактеристике пумпе и карактеристике система, као и регулација притиска и протока.

Струјни сигнали са мерача протока и притиска повезани су на NI 6009 картицу за снимање аналогних сигнала. Сигнали су прво преведени из струјних у напонске. Картица је галвански изолована од рачунара да би се избегле сметње при очитавању. Снимљени сигнали приказани су у LabVIEW програму у облику нумеричких података, а и у форми графика.

Пригушни вентил (2) се користи за дефинисање радне тачке система. Променом положаја вентила (затварањем вентила) може се снимити карактеристика пумпе и система. Лептир вентил има ручицу на себи и обележене положаје. Овим вентилом се такође омогућава уношење поремећаја у систем када се врши регулација притиска или протока.

Пресостат (5) је постављен на цевовод и има заштитну функцију. Овај електро-механички прекидач је постављен да би се спречило оштећење цевовода или неког другог елемента на цевоводу услед преоптерећења. У случају прекорачења притиска долази до прекидања струјног круга који напаја фреквентни претварач и искључује мотор. Поред пресостата налази се и аналогни барометар (4) који се користи за визуелни надзор притиска у систему.

IV. Реализоване лабораторијске вежбе

На описаном пумпном систему реализоване су следеће лабораторије вежбе:

А. снимање *p-Q* карактеристике пумпе,

Б. снимање карактеристике хидрауличног система,

В. регулација протока у систему, и

Г. регулација притиска у систему.

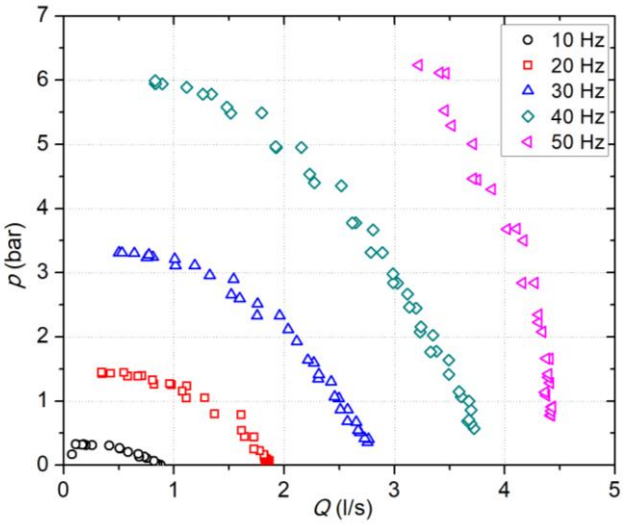
У наставку су дати задаци, поступци и примери резултат мерења – снимања карактеристика у стационарном стању и при регулацији притиска и протока.

А. Снимање p-Q карактеристике пумпе

Задатак вежбе: снимање *p-Q* карактеристика центрифугалне пумпе при различитим брзинама обртања.

Карактеристика пумпе се добија при одређеној брзини обртања, променом карактеристике хидрауличног система. Заправо, променом радне тачке врши се снимање карактеристике пумпе и то тако што се постепено затвара пригушни вентил. Снимање карактеристика пумпи врши се за више различитих брзина обртања пумпе (фреквенције прикључног напона асинхроног мотора).

На слици 4 дати су примери резултата снимања карактеристика пумпе. Приказане су карактеристике за пет различитих фреквенција прикључног напона асинхроног мотора.

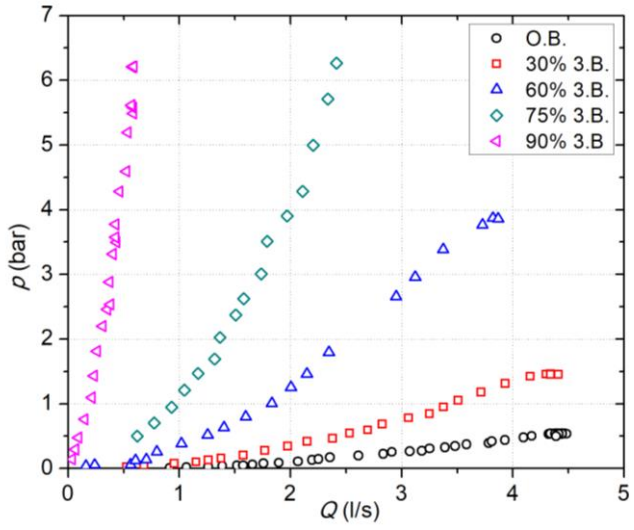


Сл. 4. *p-Q* карактеристика центрифугалне пумпе снимљена при различитим фреквенцијама (брзинама обртања)

Б. Снимање p-Q карактеристике хидрауличног система

Задатак вежбе: Снимање *p-Q* карактеристика хидрауличног система.

Карактеристике хидрауличног система добијају се при промени положаја радне тачке. У овом случају то је реализовано променом брзине радног кола пумпе. На слици 5 приказане се пет различитих карактеристика система које су добијене различитим положајима пригушног вентила: отворен вентил (О.В.), 30%, 60%, 75%, и 90% затвореног вентила (З.В.).



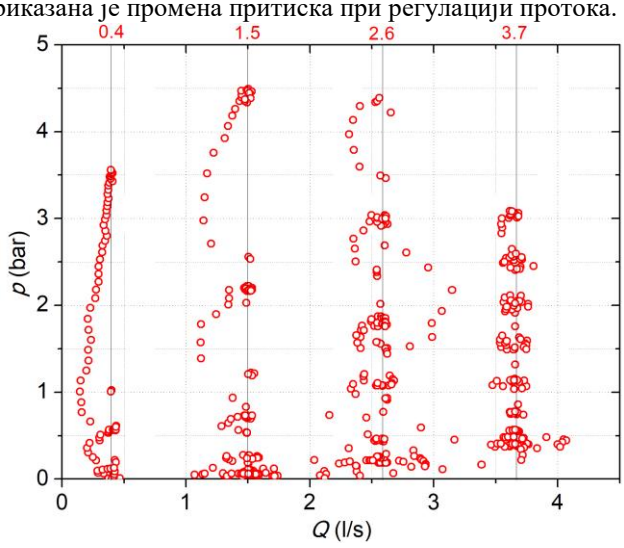
Сл. 5. Карактеристика *p-Q* хидрауличног система снимљена при различитим положајима пригушног вентила

B. Регулација протока у систему

Задатак вежбе: Реализовати систем са затвореном повратном спрегом по протоку. Снимити *p-Q* дијаграм при затварању пригушног вентила и при томе снимити промену фреквенције.

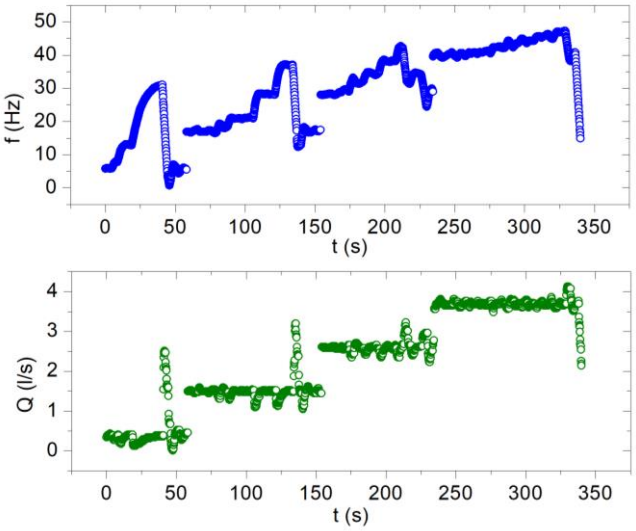
У овој вежби потребно извршити модификацију управљачког система додавањем повратне спреге. Потребно је сигнал са мерача протока довести на улаз за повратну спрегу на фреквентном претварачу. Када се изврше модификације управљачког кола, изврши активирање повратне спреге и дефинише опсег сигнала на фреквентном претварачу, потребно је покренути пумпу и унети поремећај у систем.

Поремећај се изазива затварањем или отварањем пригушног вентила. При задатој вредности протока уколико се нпр. у одређеном проценту затвори вентил, долази до тренутног пада протока и промене улазног сигнала на фреквентном претварачу. То доводи до повећања фреквенције и убрзавања пумпе. На слици 6 приказана је промена притиска при регулацији протока.



Сл. 6. Промена *p-Q* радних тачака при регулацији по протоку

Дати су примери за 4 различита референтна протока од 0,4; 1,5; 2,6 и 3,7 l/s. Са слике 4 је могуће уочити да поремећај у виду затварања вентила смањује проток, што изазива повећање фреквенције и постепено повећање притиска при поновном успостављању задатог протока. На слици 7 приказане су промене фреквенције и протока у времену при регулацији по протоку. За сваку референтну вредност протока након поремећаја у виду постепеног затварања вентила на крају је извршено отварање вентила чиме се радна тачка враћа у првобитан положај.



Сл. 7. Промена фреквенције и протока при уношењу поремећаја у систем са регулацијом протока

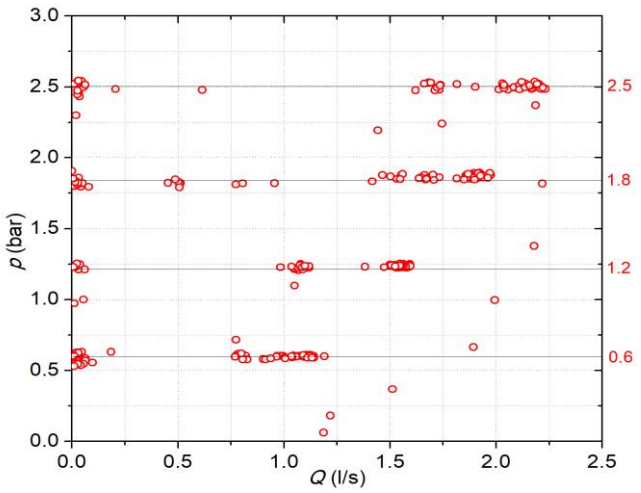
Г. Регулација притиска у систему

Задатак вежбе: Реализовати систем са затвореном повратном спрегом по притиску. Снимити промене притисака и фреквенције при отварању преливног (бајпас) вентила.

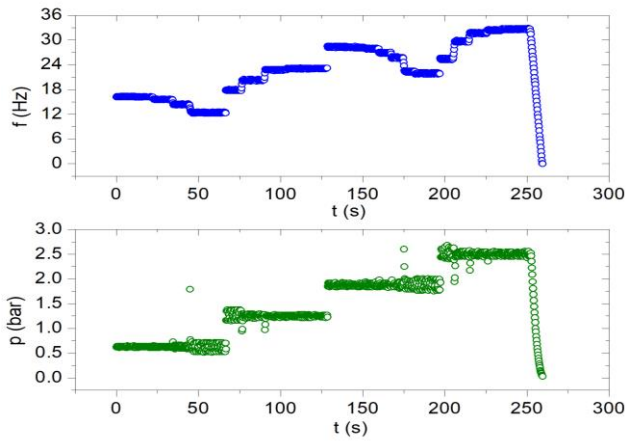
За ову лабораторијску вежбу потребно је модификовати управљачки систем додавањем повратне спреге, сигнала са сензора притиска. Потребно је сигнал са мерача притиска довести на улаз за повратну спрегу на фреквентном претварачу. Када се модификује управљачко коло, потребно је активирати опцију и подесити параметре за повратну спрегу на фреквентном претварачу. Након овог поступка може се приступити пуштању система у рад и вршити поремећај система.

Вежба се изводи тако што се пумпа пусти у рад, затвори се преливни вентил, којим се вода усмерава кроз цевовод. На фреквентном претварачу се подеси жељени притисак, пригушни вентил се постави у један од положаја (нпр. затвори се 30%). Када се достигне задати притисак, поремећај на систему се врши отварањем преливног вентила, чиме долази до пада притиска у систему. Повећањем брзине пумпе долази до пораста протока кроз систем и до враћања притиска на задату вредност. Поступак се понавља за неколико вредности задатог притиска. Дијаграм промене протока при

регулацији притиска дат је на слици 8 за примере референтних вредности од 0,6; 1,2; 1,8 и 2,5 бара. При регулацији притиска долази до промене протока, што је проузроковано повећањем брзине обртања пумпе тј. порастом фреквенције. На слици 9 приказани су дијаграми промене фреквенције и притиска у времену за различита оптерећења система при регулацији притиска.



Сл. 8. Промена *p-Q* радних тачака при регулацији по притиску



Сл. 9. Промена фреквенције и притиска при уношењу поремећаја у систем са регулацијом притиска

V. Закључак

Овај рад описује пумпни систем који ради у лабораторијским условима као и могућности његове примене у настави. Поред коришћења у истраживачке сврхе овај систем даје могућност примене у практичној настави кроз лабораторијске вежбе из области пумпних система, електромоторних погона, аутоматске регулације. Примена центрифугалне пумпе са мотором и фреквентним претварачем даје могућност снимања *p-Q* карактеристика пумпе и хидрауличног система. Поред тога, кроз резултате мерења у виду вежби показано је да коришћењем фреквентног претварача са повратном спрегом је могуће успешно остварити регулацију по протоку и притиску. Погодности рада на овој лабораторијској поставци је боље разумевање основних законитости пумпних система, електромоторних погона и регулацијоних структура притиска и протока. Студенти се радом на једном оваквом систему могу упознати са начином функционисања и подешавањем фреквентног претварача намењеног за пумпне системе. Уз све то могу директно уочити добити коришћења фреквентно регулисаних пумпних система у односу на класичну регулацију вентилима.

Захвалница

Овај рад је резултат пројекта ТР33016, чији је носилац Факултет техничких наука у Чачку и који је подржан од Министарства просвете, науке и технолошког развоја Републике Србије.

Литература

[1] Đ. Čantrak, M. Banjac, N. Janković, D. Ilić, "Pump system in the energy manager training center at the Faculty of mechanical engineering University of Belgrade", Proc. of regional conference: IEEP in South Eastern European countries, Zlatibor, Serbia, 21-24. jun 2017.
[2] Лабораторија за електромоорнепогоне ЕТФ Београд. Доступно на: http://www.pogoni.etf.bg.ac.rs/pregledEP.htm (6. вежба)
[3] Centrifugal Pump Handbook, Sulzer Pumps, 3ed edition, Elsevier, 2010.
[4] The Centrifugal Pump, Grundfos Доступно на: http://machining.grundfos.com/media/16620/the_centrifugal_pump.pdf
[5] M. Bozic, M. Sucurovic, M. Rosic, V. Vujicic, M. Bjekic, "Laboratory setup for measurements basic pump system characteristics", Proceedings of International scientific conference - UNITECH 2018, Vol. III, pp. 154-158, ISSN: 1313-230X, Gabrovo, Bulgaria, 16-17. Nov. 2018.
[6] Grundfos Pump, data sheet. Доступно на: https://www.lenntech.com/uploads/grundfos/97644355/Grundfos_CM10-6-A-R-I-E-AQQE.pdf
[7] Каталошки подаци одабране центрифугалне пумпе. Доступно на: https://product-selection.grundfos.com/product-detail.product-detail.html?custid=GMA&productnumber=99057080&qcid=561677969
[8] Danfoss VLT® AQUA Drive FC 202. Data sheet. Доступно на: https://www.danfoss.com/en/products/ac-drives/dds/vlt-aqua-drive-fc-202/#tab-overview

Abstract

In this work, pump system realized on the Faculty of technical sciences, is presented. System consists of the reservoir, three-phase multi-stage centrifugal pump, frequency converters, bypass and throttle valve. Parameters of the pump system can be determined by acquiring signals from the pressure and flow sensors placed in the pipeline. Using the acquired signals, flow and pressure regulation is performed. Procedure of laboratory exercise realization and the obtained results with comments are presented in the experimental part of the work.

Educational laboratory setup of the pump system with pressure and flow regulation

Vojislav Vujičić, Marko Šućurović, Miloš Božić, Marko Rosić, Miroslav Bjekić

Optimizacija primene V2G tehnologije u mikromreži sa obnovljivim izvorima energije

Dario Javor, *Member, IEEE,* Nebojša Raičević, *Member, IEEE*

***Apstrakt*—U radu je prikazan postupak optimizacije funkcije troškova električne energije u slučaju korišćenja obnovljivih izvora energije u mikromreži povezanoj s mrežom za napajanje. Mikromreža ima mogućnosti punjenja i pražnjenja električnih vozila tj. omogućena je efikasna upotreba njihovih baterija za skladištenje energije. U slučajevima korišćenja vetrogeneratora i fotonaponskih panela ostvaruju se različite uštede troškova. Za rešavanje optimizacionih problema korišćen je program Lingo.**

***Ključne reči*—Fotonaponski paneli; vetrogenerator; električna vozila; optimizacija; troškovi energije.**

I. Uvod

PROBLEMI zagađenja okoline, globalnog zagrevanja i emisije gasova koji izazivaju efekat staklene bašte (GHG), povezani su sa proizvodnjom električne energije koja je neophodna za razvoj čovečanstva. Da bi razvoj bio održiv, intenzivirano je korišćenje obnovljivih izvora energije. Zbog intermitentne prirode proizvodnje električne energije vetrogeneratorima i fotonaponskim panelima, potrebno je korišćenje i drugih raspoloživih resursa kao što su različiti sistemi za skladištenje energije. Ubrzani razvoj u ovoj oblasti su dodatno omogućili: primena savremenih informacionih i komunikacionih tehnologija, koncept pametnih mreža i pametnih gradova, kao i poboljšanja sistema za upravljanje električnom energijom. Poseban izazov za istraživanja predstavljaju i električna vozila, kao i optimalno korišćenje V2G (vehicle-to-grid) tehnologije [1]-[6].

Korišćenje baterija električnih vozila za skladištenje električne energije, naročito u slučaju velikog broja vozila i dugih vremenskih intervala raspoloživosti za punjenje ili pražnjenje na parkinzima ili u garažama, pruža mogućnosti optimizacije troškova za električnu energiju. Moguće je da optimizacija ima i više ciljeva, kao što su smanjenje troškova investicija, povećanje efikasnosti energetskog sistema i smanjenje emisija GHG.

U ovom radu se razmatra problem minimizacije troškova energije u mikromreži koja sadrži vetrogeneratore, solarne panele i flotu električnih vozila koja ima mogućnost punjenja ili pražnjenja, izuzev u vremenskom intervalu kad se koristi za prevoz. Matematički modeli ovakvih problema mogu se predstaviti funkcijom cilja koju treba minimizirati, uz ograničenja, i zatim rešiti numerički. Radi primene nekog od optimizacionih metoda, moguće je koristiti programe kao što su MATLAB, Yalmip, Excel Solver, Lingo [7] itd., u kojima se različite metode mogu izabrati za rešavanje istog problema.

Dario Javor – Elektronski fakultet, Univerzitet u Nišu, Aleksandra Medvedeva 14, 18000 Niš, Srbija (e-mail: dariojavor@mts.rs).

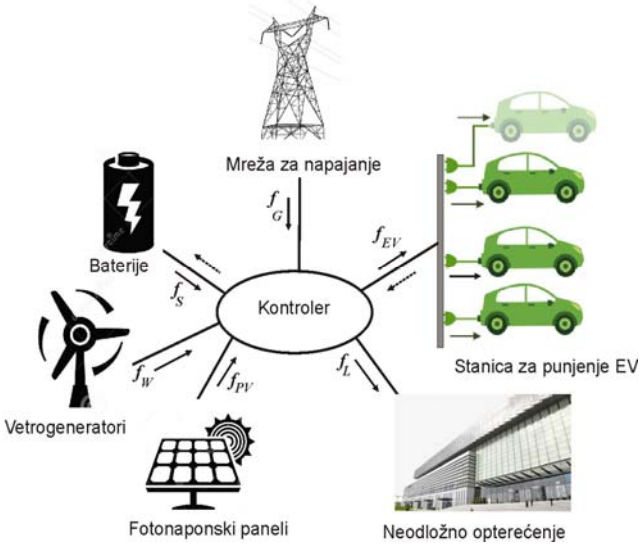
Nebojša Raičević – Elektronski fakultet, Univerzitet u Nišu, Aleksandra Medvedeva 14, 18000 Niš, Srbija (e-mail: nebojsa.raicevic@elfak.ni.ac.rs).

II. Optimizacioni problem

Pametne mreže i mikromreže sa distribuiranim izvorima energije imaju mogućnost bidirekcionog toka snage ka mreži za napajanje. Ovo omogućava uštede troškova energije, a može biti veoma značajno za očuvanje čovekove okoline. Ako se raspolaže dnevnim dijagramima očekivane proizvodnje energije solarnim panelima i vetrogeneratorima i ako su na raspolaganju baterije za skladištenje energije, mogu se minimizirati troškovi i proračunati optimalni kapaciteti ovih resursa i njihova isplativost.

Obnovljivi izvori energije (OIE) u 2018. god. predstavljali su čak 83% od instaliranih elektroenergetskih kapaciteta, čime su dostigli oko trećinu proizvodnje električne energije u svetu. Mikromreže koje koriste OIE mogu instalirati industrijske kompanije, univerzitetski kampusi i različite vrste korisnika. U ovom radu je uzet primer mikromreže (Sl.1) koja raspolaže sledećim resursima:

- vetrogeneratorima izabrane snage 33kW za koji je dnevni dijagram očekivane proizvodnje energije procenjen na osnovu brzine vetra u Banatu merene 24.04.2011. god. na svakih 10 minuta na visini 10m, usrednjene WAsP softverom (Wind energy industry standard software) [8] za satne vrednosti i određenu visinu turbine [9];



Sl. 1. Šema mikromreže (sa neodložnim opterećenjem, vetrogeneratorom, fotonaponskim panelima, baterijama i stanicom za punjenje električnih vozila) priključene na mrežu za napajanje.

- fotonaponskim panelima snage 33kW za koje je dnevni dijagram očekivane proizvodnje energije procenjen na osnovu solarne iradijacije i temperature vazduha merene u Južnom Banatu u opštini Kovin na lokaciji Bavanište za prosečan dan 2009. god. [10];
- stanicom koja ima dovoljan broj priključaka snage P_{ch}=3.3kW za punjenje/pražnjenje flote sa n_V vozila istog tipa čija je baterija pojedinačnog kapaciteta en_{EV}=30kWh;
- baterijama čije punjenje ili pražnjenje može dodatno umanjiti troškove energije.

Može se obaviti proračun i za veći broj vetrogeneratora n_W, kao i za veći broj n_{PV} fotonaponskih panela. Poznat je i dijagram kupovnih BP(i) i prodajnih cena SP(i) za svaki sat i=1,...,24. Dijagram BP(i) je dobijen na osnovu Hungarian Power Exchange podataka [11], a uzeto je SP(i)=0.75BP(i).

Ciljna funkcija C_{COST} koju treba minimizirati predstavlja sumu troškova za energiju preuzetu iz mreže za napajanje u toku jednog dana,

$$C_{COST} = \min\left\{\sum_{i=1}^{24} P_{ch}\cdot n_V\cdot[\mathrm{BP}(i)\cdot x_B(i)\text{-}\mathrm{SP}(i)\cdot x_S(i)]\right.$$
$$\left.\text{-}\mathrm{SP}(i)\cdot n_{PV}\cdot P_{PV}(i)\text{-}\mathrm{SP}(i)\cdot n_W\cdot P_W(i)+\mathrm{BP}(i)\cdot P_L(i)\right\}, \quad (1)$$

s obzirom na procenjeni dijagram neodložnog opterećenja u toku dana $P_L(i)$, i=1,...,24.

Promenljiva $x_B(i)$ ima vrednost 1 ako se kupuje energija iz mreže u i-tom satu, vrednost 0 ako se ne kupuje, dok $x_S(i)$ ima vrednost 1 ako se prodaje energija mreži za napajanje u i-tom satu, 0 ukoliko se ne prodaje. Umesto ove dve promenljive, može se definisati jedna promenljiva $x(i)$ koja ima vrednost 1 kada $x_B(i)$ ima vrednost 1, vrednost -1 kada je $x_S(i)$ jednako 1, a 0 ako bilo koja od ove dve promenljive ima vrednost 0 (kada je vozilo priključeno na stanici, ali se niti puni niti prazni, ili je u vožnji tj. nije priključeno na stanici). Ako je i redni broj satnog intervala u toku dana, stanje baterije električnog vozila mora biti uvek u opsegu od 20% do 100%, pa je definisano ograničenje:

$$20 \le SOC(i) \le 100, \qquad i = 1,...,24. \quad (2)$$

Stanje napunjenosti baterije $SOC(i)$ se određuje diskretno, na početku svakog vremenskog intervala i=1,...,24, na osnovu stanja iz prethodnog intervala, procenta punjenja/pražnjenja baterije na sat (P_{ch}/en_{EV}) i promenljive $x(i$-1) koja označava punjenje/pražnjenje u prethodnom intervalu.

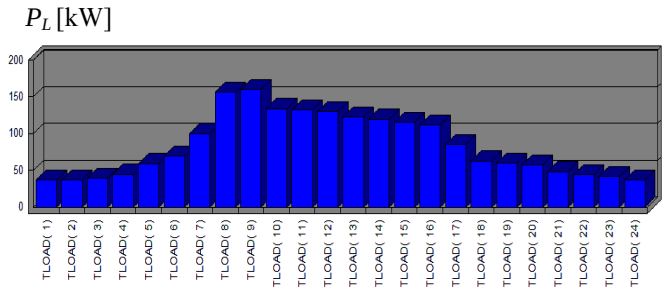
$$SOC(i) = SOC(i-1) + x(i-1)\cdot P_{ch}/en_{EV}\cdot 100. \quad (3)$$

Na početku vremenskog intervala od 6 sati kada vozila voze, baterije su napunjene na $SOC0$=100, a na kraju 67%. Izabrano je malo električno vozilo Nissan Leaf koje sa napunjenom baterijom može da pređe od 100 do 200km. U ovom radu, procenjeno je da vozilo može da pređe 167km sa 100% napunjenom baterijom, odnosno, da svako od vozila prelazi 55km za 6 sati vožnje, kao u [14]. SOC opada po 5.5% na sat kada se automobil vozi, a smanji se od 100% na 67% u vremenskom intervalu počev od 11-tog, a zaključno sa 16-tim. U preostalih 18 sati vozila su na raspolaganju na stanici. Kad se vozilo puni ili prazni na stanici, brzina punjenja/pražnjenja je 11% na sat, pošto je P_{ch}/en_{EV} =0.11. Usvojen je isti koeficijent efikasnosti i za punjenje i za pražnjenje.

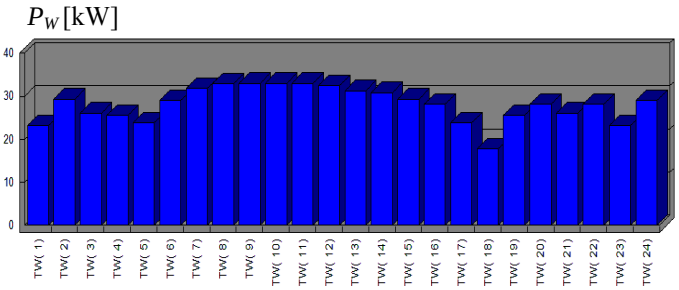
Dodatno ograničenje može biti maksimalna vrednost snage koja se može prenositi vodovima, kao i snaga P_{Gmax} koja se može zahtevati od mreže za napajanje. Za i=1,...,24, bilans snaga dat je sa

$$P_G(i) = P_L(i) - P_{PV}(i) - P_W(i) + P_{EV}(i). \quad (4)$$

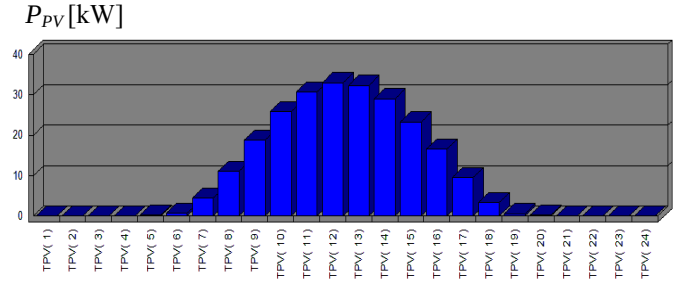
Na Sl.2 dat je dnevni dijagram neodložnog opterećenja kao u [12], [13]. Na Sl.3 prikazan je dijagram proizvodnje energije vetrogeneratora maksimalne snage 33kW, a na Sl.4 fotopanela maksimalne snage 33kW u toku dana. U Tabeli I su kupovne/prodajne cene električne energije [6], dok je dnevni dijagram cena na tržištu električne energije [11] dat na Sl.5.



Sl. 2. Dnevni dijagram snage neodložnog opterećenja u kW.

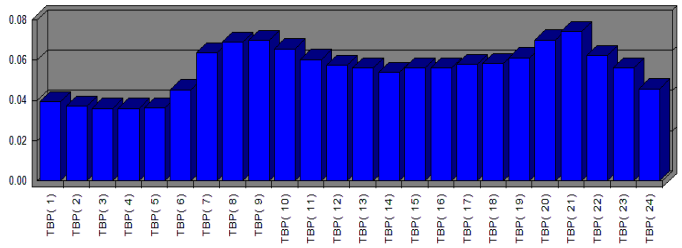


Sl. 3. Dnevni dijagram proizvodnje vetrogeneratora max snage 33kW.



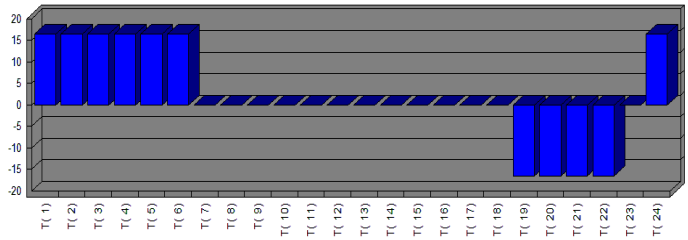
Sl. 4. Dnevni dijagram proizvodnje fotonaponskih panela max snage 33kW.

BP [€/kWh]



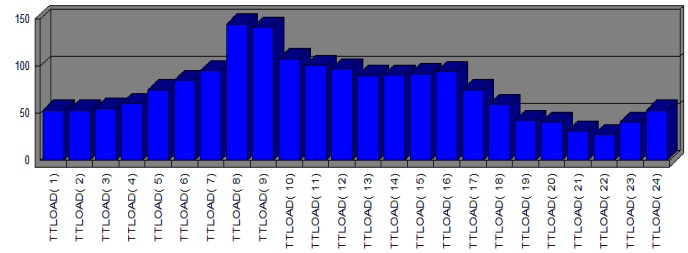
Sl. 5. Dnevni dijagram tržišnih kupovnih cena energije BP u € po kWh.

P_{EV} [kW]



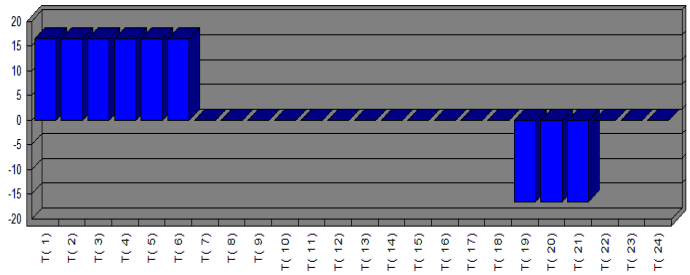
Sl. 6. Optimizovani dnevni raspored punjenja/pražnjenja vozila za n_V=5, P_{PV}=33kW, P_W=0, BP sa Sl.5 i SP=0.75BP.

P_{tot} [kW]



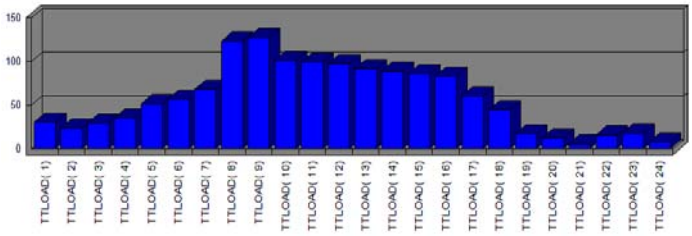
Sl. 7. Dnevni dijagram ukupne potrošnje u slučaju n_V=5, P_{PV}=33kW, P_W=0, BP sa Sl.5 i SP=0.75BP.

P_{EV} [kW]



Sl. 8 Optimizovani dnevni raspored punjenja/pražnjenja vozila za n_V=5, P_{PV}=0, P_W=33kW, BP sa Sl.5 i SP=0.75BP.

P_{tot} [kW]



Sl. 9. Dnevni dijagram ukupne potrošnje u slučaju n_V=5, P_{PV}=0, P_W=33kW, BP sa Sl.5 i SP=0.75BP.

III. Rezultat optimizacije primene V2G tehnologije

Izabran je scenario sa n_V=5 električnih vozila koja su na raspolaganju za punjenje/pražnjenje na stanici, odnosno za korišćenje njihovih baterija, u intervalu od početka 17-tog jednočasovnog intervala u toku dana do početka 11-tog intervala narednog dana. Ostalih 6 časova vozila nisu na raspolaganju, a troše energiju svojih baterija za vožnju. Za Scenario 1 i 2 izabrane su varijabilne cene električne energije kao na Sl.5, a za Scenario 3 i 4 cene iz Tabele I. Program Lingo ima ugrađene solvere za različite tipove optimizacionih problema i bira onaj koji je pogodniji za formulisani problem.

Scenario 1: rade fotopaneli snage 33kW, a vetrogeneratori ne rade. Poznat je dnevni dijagram kupovnih cena BP(i) za svaki sat, kao na Sl.5, i prodajnih cena SP(i)=0.75BP(i). Na osnovu optimizacije funkcije (1), uz ispunjene uslove (2) i (3), dobija se raspored punjenja/pražnjenja flote vozila kao na Sl.6 za minimalni dnevni trošak energije. Na Sl.7 prikazan je dijagram snage $P_G(i)$ koju mikromreža kupuje od mreže za napajanje (4).

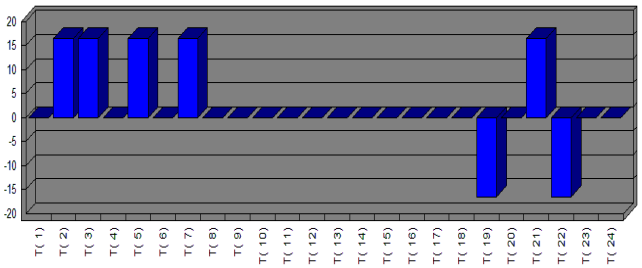
Scenario 2: rade vetrogeneratori snage 33kW, a fotopaneli ne rade. Na osnovu optimizacije funkcije (1), uz uslove (2) i (3), dobija se raspored punjenja/pražnjenja flote vozila kao na Sl. 8. Troškovi su 16.36% manji nego za Scenario 1. Na Sl.9 je prikazan dijagram $P_G(i)$. Bidirekcioni tok snage obezbeđen je ka mreži za napajanje, ali su za Scenario 1 i 2 snaga neodložne potrošnje $P_L(i)$ i snaga punjenja vozila $P_{EV}(i)$ veće od proizvodnje vetrogeneratorima $P_W(i)$ i fotopanelima $P_{PV}(i)$ u svakom satu, što se vidi i iz grafika na Sl.7 i Sl.9.

Scenario 3: rade fotopaneli snage 33kW, vetrogeneratori ne rade. Kupovne i prodajne cene BP(i) i SP(i) za svaki sat poznate su iz Tabele I. Na osnovu optimizacije funkcije (1), uz (2) i (3), dobija se raspored punjenja/pražnjenja flote vozila kao na Sl.10. Na Sl.11 prikazan je dijagram snage $P_G(i)$ koju mikromreža kupuje od mreže za napajanje.

TABELA I
KUPOVNE I PRODAJNE CENE ENERGIJE U TOKU DANA [6]

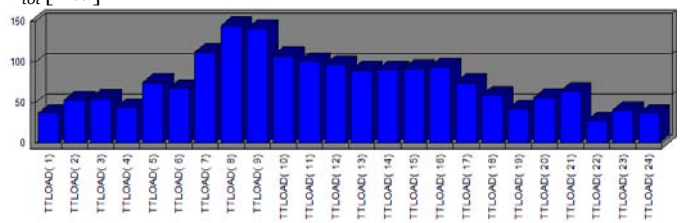
Kupovna (BP) i prodajna cena (SP)	Cena [€/kWh]
BP od 1-9. i od 20-24. sata	0.2
BP od 10-19. sata	0.26
SP od 1-24. sata	0.15

P_{EV} [kW]



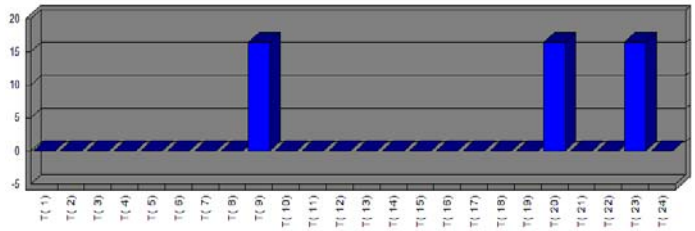
Sl. 10. Optimizovani dnevni raspored punjenja/pražnjenja vozila za n_V=5, P_{PV}=33kW, P_W=0 i BP/SP iz Tabele I.

P_{tot} [kW]



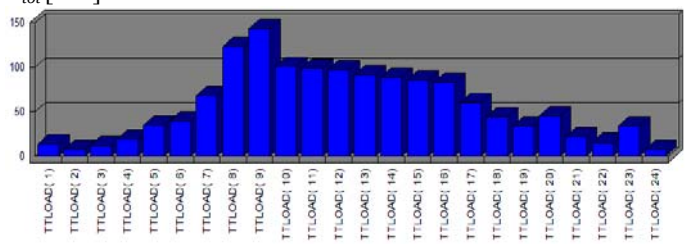
Sl. 11. Dnevni dijagram ukupne potrošnje u slučaju n_V=5, P_{PV}=33kW, P_W=0 i BP/SP iz Tabele I.

P_{EV} [kW]



Sl. 12. Optimizovani dnevni raspored punjenja/pražnjenja vozila za n_V=5, P_{PV}=0, P_W=33kW i BP/SP iz Tabele I.

P_{tot} [kW]



Sl. 13. Dnevni dijagram ukupne potrošnje u slučaju n_V=5, P_{PV}=0, P_W=33kW i BP/SP iz Tabele I.

Scenario 4: rade vetrogeneratori snage 33kW, fotopaneli ne rade. Kupovne i prodajne cene BP(i) i SP(i) za svaki sat poznate su iz Tabele I. Na osnovu optimizacije funkcije troškova dobija se raspored punjenja/pražnjenja flote vozila kao na Sl.12. Troškovi su 14.3% manji nego za Scenario 3. Na Sl.13 prikazan je dijagram snage $P_G(i)$.

Na osnovu proračuna određeno je da dodatne baterije kojima se može obezbediti veća ušteda energije za Scenario 1 i 2 treba izabrati sa snagom punjenja/pražnjenja 85kW, dok je za Scenario 3 i 4 dovoljno 70kW.

IV. Zaključak

Na osnovu rezultata datih u radu može se zaključiti da varijabilne cene električne energije iz sata u sat u toku dana omogućavaju veći procenat uštede. U radu je određeno i koliko instalirani kW snage vetrogeneratora omogućava veće uštede nego instalirani kW snage fotonaponskih panela, s obzirom na pogodniji dnevni dijagram proizvodnje električne energije od vetrogeneratora, ali su investicioni troškovi u tom slučaju znatno veći.

U radu su razmatrana električna vozila kao flota, ali bi drugačiji pojedinačni raspored vožnje i punjenja/pražnjenja u toku dana omogućio i dodatne uštede. Takođe, mogu se uzeti u obzir i troškovi korišćenja baterija u električnim vozilima.

Osim smanjenja troškova energije, moguće je proračunati i benefit od smanjene emisije GHG. Korišćenjem dobijenih dijagrama, može se odrediti i optimalni kapacitet baterija za skladištenje energije u datoj mikromreži.

Zahvalnica

Autori se zahvaljuju Ministarstvu prosvete, nauke i tehnološkog razvoja za finansijsku podršku rada u okviru projekta TR33008.

Literatura

[1] L. Liu, F. Kong, X. Liu, Y. Peng, and Q. Wang, "A review on electric vehicles interacting with renewable energy in smart grid", Renewable and Sustainable Energy Reviews, vol. 51, pp. 648-661, 2015.
[2] F. Laureri, L. Puliga, M. Robba, F. Delfino, and G. Bulto, "An optimization model for the integration of electric vehicles and smart grids: Problem definition and experimental validation", 2016 IEEE International Smart Cities Conference (ISC2), 2016.
[3] Z. Yang, K. Li, and A. Foley, "Computational scheduling methods for integrating plug-in electric vehicles with power systems: A review," in Renewable & Sustainable Energy Reviews, vol. 51, pp. 396-416, 2015.
[4] B. Kim, S. Ren, M. van der Schaar, and J. Lee, "Bidirectional energy trading and residential load scheduling with electric vehicles in the smart grid", IEEE Journal on Selected Areas in Communications, vol. 31, no. 7, pp. 1219-1234, 2013.
[5] H. Yang, H. Pan, F. Luo, J. Qiu, Y. Deng, M. Lai, and Z. Dong, "Operational planning of electric vehicles for balancing wind power and load fluctuations in a microgrid", IEEE Transactions on Sustainable Energy, vol. 8, no. 2, pp. 592-604, 2017.
[6] G. Ferro, F. Laureri, R. Minciardi, and M. Robba, "An optimization model for electrical vehicles scheduling in a smart grid", Sustainable Energy, Grids and Networks, vol. 14, pp. 62-70, 2018.
[7] https://www.lindo.com (pristupljeno 10.03.2019.)
[8] http://www.wasp.dk (pristupljeno 17.04.2019.)
[9] M. Mirković, "Eksploataciona karakteristika obnovivih izvora energije", Diplomski rad, ETF, Beograd, pp.32, 2011.
[10] I. Babić, "Modelovanje uticaja vremenskog profila solarnog zračenja na efekte rada fotonaponskih sistema u elektroenergetskom sistemu", Doktorska disertacija, ETF, Beograd, pp.66, 2016.
[11] https://hupx.hu/en, MC prices, Hungarian Power Exchange, 17.04.2019.
[12] Elektroprivreda Srbije 2015, "Odluka o izmenama pravila o radu distributivnog sistema, 6.16 Profili potrošnje", pp. 11, 2015.
[13] I. Anastasijević, and A. Janjić, "Electric vehicles charging optimization: reducing operational costs of small companies", 6th Int. Conference on Transport and Logistics TIL 2017, pp. 109-113, 2017.
[14] D. Javor, A. Janjić, and N. Raičević, "Reducing energy costs by using optimal electric vehicles scheduling and renewable energy sources", 18th Int. Symposium, Infoteh-Jahorina 2019, Mart 2019.

Abstract

The paper presents the procedure of optimizing the function of the cost of electricity in case of using renewable energy sources in a microgrid connected to the main grid. The microgrid has possibilities of charging and discharging electric vehicles, that is, the efficient use of their batteries as storage units is enabled. In cases of using wind generators and photovoltaic panels different cost savings are achieved. Lingo program is used for solving these optimization problems.

Optimization of V2G Technology Application in the Microgrid with Renewable Energy Sources

Dario Javor, Nebojša Raičević

Osetljivost greške dinamičke estimacije stanja na promene pojedinih parametara Kalmanovog filtra

Dragan Ćetenović, Aleksandar Ranković

Fakultet tehničkih nauka u Čačku

***Apstrakt*— U kvazistacionarnom režimu rada distributivne mreže matrica kovarijansi grešaka dinamičkog modela Q obično se modeluje kao vremenski nepromenljiva u dijagonalnoj formi sa svim elementima na dijagonali međusobno jednakim, što dovodi do jednoparametarskog modela matrice. U tom slučaju se podešavanje matrice kovarijansi Q svodi na podešavanje jednog jedinog parametra q. Tačnost dinamičke estimacije stanja uslovljena je izborom vrednosti ovog parametra, ali i izborom inicijalnog rešenja, gde se pod inicijalnim rešenjem podrazumeva inicijalni vektor stanja x_0^+ i njemu pripadajuća matrica kovarijansi P_0^+. Cilj ovog rada je detaljno ispitati simultano dejstvo parametra q i inicijalnog rešenja na kvalitet dinamičke estimacije. Analize su sprovedene na modifikovanom IEEE distributivnom test sistemu sa 13 i 123 čvora korišćenjem EKF i UKF algoritma dinamičke estimacije stanja.**

***Ključne reči*—matrica kovarijansi grešaka dinamičkog modela; dinamički estimator; Kalmanov filtar.**

I. UVOD

Sa porastom stepena integracije distribuiranih resursa u distributivnu mrežu rastu i zahtevi za aktivnijim učešćem u monitoringu i upravljanju distributivnom mrežom. Takođe, rastu i zahtevi za boljim kvalitetom estimacije stanja kako bi se operateru distributivnog sistema omogućilo da preuzme odgovarjuće upravljačke akcije pravovremeno.

S druge strane, mali broj telemetrisanih merenja prisutnih u distributivnoj mreži vodi ka lošijem kvalitetu estimacije u poređenju sa prenosnom mrežom. Povećanje broja telemetrisanih merenja iziskuje nova ulaganja u mernu i telekomunikacionu infrastrukturu, pa ne predstavlja popularnu meru. Stoga, ideja je da se izbegnu dodatna ulaganja, a da se promeni algoritam estimacije stanja.

Estimacija stanja u savremenim upravljačkim centrima zasniva se na upotrebi statičkih estimatora stanja. Upotreba dinamičkih estimatora stanja pruža brojne prednosti [1], od kojih je jedna da se kvalitet estimacije može popraviti [2]. Međutim, njihova upotreba donosi nove probleme: koji Kalman filter odabrati, kako inicirati algoritam dinamičke estimacije stanja, koji model za opisivanje dinamike sistema upotrebiti, kako podesiti matricu kovarijansi kojom se obuhvataju neizvesnosti dinamičkog modela itd.

Neki od odgovora na prethodna pitanja su već dati u [2], gde je matrica kovarijansi grešaka dinamičkog modela predstavljena u sledećoj formi:

$$\boldsymbol{Q} = 10^{q} \cdot \boldsymbol{I}_n \tag{1}$$

gde je $\boldsymbol{I}_n$ $n-$dimenziona jedinična matrica (matrica identiteta), a n broj promenljivih stanja u sistemu. Pokazano je da ukoliko je vrednost parametra q adekvatno podešena, izbor inicijalnog rešenja, kao i tipa Kalmanovog filtra, ima zanemarljiv uticaj na kvalitet estimacije stanja [2]. Međutim, ukoliko vrednost ovog parametra nije adekvatno podešena, kvalitet estimacije stanja može značajno da varira zavisno od izbora inicijalnog rešenja i tipa Kalmanovog filtra.

Glavni cilj ovog rada je ispitati razloge zbog kojih je greška dinamičke estimacije osetljiva na promene pojednih parametara na način na koji je to prikazano u [2]. Rad sumira zaključke do kojih se došlo istraživanjem u okviru doktorske disertacije [3].

Rad je organizovan na sledeći način: u sekciji II predstavljeni su EKF i UKF algoritam dinamičke estimacije stanja, u sekciji III prikazana je osetljivost greške dinamičke estimacije stanja na promene pojedinih parametara Kalmanovog filtra, u sekciji IV izvedena je detaljna analiza dobijenih rezultata, da bi u sekciji V bili dati glavni zaključci.

II. ALGORITMI DINAMIČKE ESTIMACIJE STANJA

EKF algoritam se može opisati setom jednačina [4-6]:

$$\boldsymbol{x}_{k+1}^{-} = \boldsymbol{F}_k \boldsymbol{x}_k^{+} + \boldsymbol{g}_k \tag{2}$$

$$\boldsymbol{P}_{k+1}^{-} = \boldsymbol{F}_k \boldsymbol{P}_k^{+} \boldsymbol{F}_k^{\mathrm{T}} + \boldsymbol{Q}_k \tag{3}$$

$$\boldsymbol{z}_{k+1}^{-} = \boldsymbol{h}\left(\boldsymbol{x}_{k+1}^{-}\right) \tag{4}$$

$$\boldsymbol{T}_{k+1} = \boldsymbol{H}_{k+1} \boldsymbol{P}_{k+1}^{-} \boldsymbol{H}_{k+1}^{\mathrm{T}} \tag{5}$$

$$\boldsymbol{\nu}_{k+1} = \boldsymbol{z}_{k+1} - \boldsymbol{z}_{k+1}^{-} \tag{6}$$

$$\boldsymbol{S}_{k+1} = \boldsymbol{T}_{k+1} + \boldsymbol{R}_{k+1} \tag{7}$$

$$\boldsymbol{K}_{k+1} = \boldsymbol{P}_{k+1}^{-} \boldsymbol{H}_{k+1}^{\mathrm{T}} \boldsymbol{S}_{k+1}^{-1} \tag{8}$$

$$\boldsymbol{x}_{k+1}^{+} = \boldsymbol{x}_{k+1}^{-} + \boldsymbol{K}_{k+1} \boldsymbol{\nu}_{k+1} \tag{9}$$

$$\boldsymbol{P}_{k+1}^{+} = \boldsymbol{P}_{k+1}^{-} - \boldsymbol{K}_{k+1} \boldsymbol{S}_{k+1} \boldsymbol{K}_{k+1}^{\mathrm{T}} \tag{10}$$

gde su $\boldsymbol{x}^{-}$ i $\boldsymbol{x}^{+}$ predviđeni i estimirani vektor stanja, respektivno, $\boldsymbol{F}$ i $\boldsymbol{g}$ tranzicioni parametri dinamičkog

Dragan Ćetenović – Fakultet tehničkih nauka u Čačku, Univerzitet u Kragujevcu, Svetog Save 65, 32102 Čačak, Srbija (e-mail: dragan.cetenovic@ftn.kg.ac.rs).

Aleksandar Ranković – Fakultet tehničkih nauka u Čačku, Univerzitet u Kragujevcu, Svetog Save 65, 32102 Čačak, Srbija (e-mail: aleksandar.rankovic@ftn.kg.ac.rs).

modela, respektivno, $\boldsymbol{P}^-$ i $\boldsymbol{P}^+$ matrica kovarijansi predviđenog i estimiranog vektora stanja, respektivno, $\boldsymbol{z}^-$ i $\boldsymbol{T}$ predviđeni vektor merenja i njemu pripadajuća matrica kovarijansi, respektivno, $\boldsymbol{h}$ nelinearna vektorska funkcija merenja, $\boldsymbol{H}$ Jakobijeva matrica, $\boldsymbol{v}$ i $\boldsymbol{S}$ vektor inovacija i njemu pripadajuća matrica kovarijansi, respektivno, $\boldsymbol{z}$ i $\boldsymbol{R}$ vektor merenja i njemu pripadajuća matrica kovarijansi, respektivno, $\boldsymbol{K}$ matrica Kalmanovog pojačanja i k vremenski trenutak.

UKF algoritam se može predstaviti sledećim setom jednačina [7-9]:

$$\boldsymbol{Y}_k^+ = \boldsymbol{x}_k^+ \cdot \boldsymbol{1}^{\mathrm{T}} + \sqrt{n+\lambda_{ut}}\left[\boldsymbol{0} \quad \sqrt{\boldsymbol{P}_k^+} \quad -\sqrt{\boldsymbol{P}_k^+}\right] \tag{11}$$

$$\hat{\boldsymbol{X}}_{k+1} = \boldsymbol{F}_k \boldsymbol{Y}_k^+ + \boldsymbol{g}_k \cdot \boldsymbol{1}^{\mathrm{T}} \tag{12}$$

$$\boldsymbol{x}_{k+1}^- = \hat{\boldsymbol{X}}_{k+1} \boldsymbol{w}_m \tag{13}$$

$$\boldsymbol{P}_{k+1}^- = \hat{\boldsymbol{X}}_{k+1} \boldsymbol{W} \hat{\boldsymbol{X}}_{k+1}^{\mathrm{T}} + \boldsymbol{Q}_k \tag{14}$$

$$\boldsymbol{Y}_{k+1}^- = \boldsymbol{x}_{k+1}^- \cdot \boldsymbol{1}^{\mathrm{T}} + \sqrt{n+\lambda_{ut}}\left[\boldsymbol{0} \quad \sqrt{\boldsymbol{P}_{k+1}^-} \quad -\sqrt{\boldsymbol{P}_{k+1}^-}\right] \tag{15}$$

$$\hat{\boldsymbol{Z}}_{k+1}^- = \boldsymbol{h}\left(\boldsymbol{Y}_{k+1}^-\right) \tag{16}$$

$$\boldsymbol{z}_{k+1}^- = \hat{\boldsymbol{Z}}_{k+1}^- \boldsymbol{w}_m \tag{17}$$

$$\boldsymbol{T}_{k+1} = \hat{\boldsymbol{Z}}_{k+1}^- \boldsymbol{W} \left[\hat{\boldsymbol{Z}}_{k+1}^-\right]^{\mathrm{T}} \tag{18}$$

$$\boldsymbol{\nu}_{k+1} = \boldsymbol{z}_{k+1} - \boldsymbol{z}_{k+1}^- \tag{19}$$

$$\boldsymbol{S}_{k+1} = \boldsymbol{T}_{k+1} + \boldsymbol{R}_{k+1} \tag{20}$$

$$\boldsymbol{C}_{k+1} = \boldsymbol{Y}_{k+1}^- \boldsymbol{W} \left[\hat{\boldsymbol{Z}}_{k+1}^-\right]^{\mathrm{T}} \tag{21}$$

$$\boldsymbol{K}_{k+1} = \boldsymbol{C}_{k+1} \boldsymbol{S}_{k+1}^{-1} \tag{22}$$

gde je $\boldsymbol{Y}^+$ matrica sigma tačaka dobijenih aproksimacijom raspodele verovatnoće estimiranog vektora stanja, $\boldsymbol{1}$ jedinični vektor, λ_{ut} parametar skaliranja, $\boldsymbol{0}$ nula veltor, $\hat{\boldsymbol{X}}$ matrica sigma tačaka preslikanih pomoću dinamičkog modela, $\boldsymbol{w}_m$ i $\boldsymbol{W}$ vektor i matrica težinskih faktora, respektivno, $\boldsymbol{Y}^-$ matrica sigma tačaka dobijenih aproksimacijom raspodele verovatnoće predviđenog vektora stanja, $\hat{\boldsymbol{Z}}^-$ matrica sigma tačaka preslikanih pomoću modela merenja, $\boldsymbol{C}$ matrica unakrsnih kovarijansi stanja i merenja.

III. Osetljivost greške dinamičke estimacije na promene pojedinih parametara Kalamnovog filtra

Ukupna greška estimacije stanja dobija se kao:

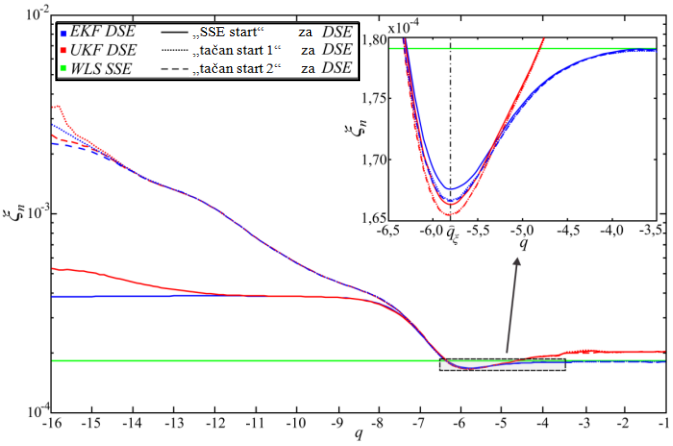
$$\xi_n(q) = \frac{1}{K}\sum_{k=1}^{K} \xi_{n,k}(q) \tag{23}$$

$$\xi_{n,k}(q) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{i,k}^+(q) - x_{i,k}^{true}\right)^2} \tag{24}$$

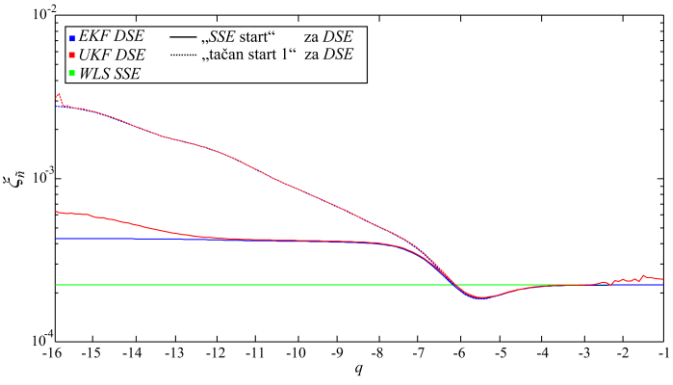
Ukupna greška estimacije stanja izračunata je na modifikovanom distributivnom IEEE 13 [2,10] i IEEE 123 [10] test sistemu korišćenjem EKF i UKF algoritma dinamičke estimacije i to za tri različita scenarija inicijalizacije:

1. „SSE start“: Inicijalne vrednosti $\boldsymbol{x}_0^+$ i $\boldsymbol{P}_0^+$ dobijene su korišćenjem *WLS* algoritma statičke estimacije,
2. „tačan start 1“: Inicijalna vrednost vektora stanja $\boldsymbol{x}_0^+$ dobijena je na osnovu proračuna tokova snaga i predstavlja tačan vektor stanja $\boldsymbol{x}_0^{true}$ u trenutku $k=0$, dok je inicijalna matrica kovarijansi grešaka estimacije podešena na $\boldsymbol{P}_0^+ = \boldsymbol{0}$
3. „tačan start 2“: Dinamička estimacija stanja takođe je inicirana s tačnim vektorom stanja $\boldsymbol{x}_0^{true}$, ali je matrica $\boldsymbol{P}_0^+$ sada podešena kao dijagonalna, tako da je $\boldsymbol{P}_0^+ = p_0 \cdot \boldsymbol{I}_n$ gde je p_0 skalapni broj konačno male vrednosti ($p_0 = 10^{-15}$). Ovaj scenario nije analiziran na IEEE 123 test sistemu.

Izračunata je i ukupna greška estimacije sprovedene statičkim WLS estimatorom. Na slici 1 je prikazana greška esimacije u funkciji parametra q za a) modifikovani IEEE 13 i b) IEEE 123 test sistem.



a)



b)

Slika 1. Ukupna greška estimacije za a) modifikovani IEEE 13 i b) IEEE 123 test sistem u funkciji parametra q za različite senarije inicijalizacije i različite algoritme estimacije

S obzirom da izbor parametra q i scenarija inicijalizacije na isti način utiče na grešku dinamičke estimacije stanja u slučaju oba test sistema, u nastavku će detaljno biti analizirani samo rezultati dobijeni na modifikovanom IEEE 13 test sistemu.

IV. ANALIZA REZULTATA

Na osnovu rezultata prikazanih na slici 1, prethodno definisani opseg parametra q može se podeliti na uže opsege (intervale) radi lakše analize rezultata:

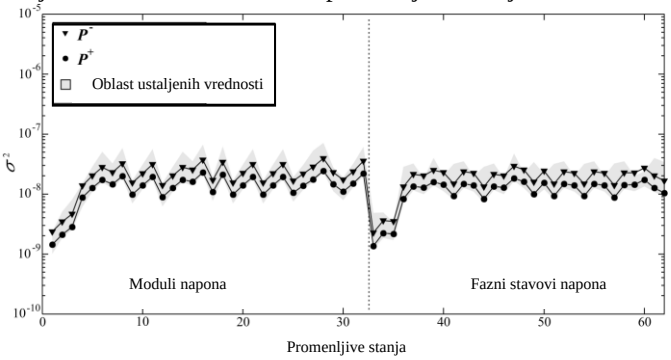
- $-16 \le q \le -8$, gde greška dinamičke estimacije dostiže izuzetno visoke vrednosti, pri čemu izbor scenarija inicijalizacije ima ogroman uticaj na kvalitet estimacije; na ovom intervalu greška dinamičke estimacije stanja je nerastuća funkcija parametra q za bilo koji scenario inicijalizacije,
- $-8 \le q \le \hat{q}_{\xi}$, gde se greška dinamičke estimacije i dalje smanjuje s porastom vrednosti parametra q, ali u ovom slučaju izbor scenarija inicijalizacije ne utiče značajnije na kvalitet estimacije,
- $q > \hat{q}_{\xi}$, gde greška estimacije raste s porastom vrednosti parametra q, pri čemu je izbor scenarija inicijalizacije i dalje zanemarljiv.

Interval vrednosti parametra $-16 \le q \le -8$

Za vrednosti parametra q rangirane na ovom intervalu greška dinamičke estimacije može biti značajno veća od greške statičkog estimatora, što zavisi od vrednosti parametra q, ali i od scenarija inicijalizacije (pogledati sliku 1). S obzirom da smanjenje vrednosti dijagonalnih elemenata matrice kovarijansi $\boldsymbol{Q}$ može dovesti do porasta prioriteta koji filtar daje rezultatima predviđanja u odnosu na sama merenja, porast greške dinamičkog estimatora na ovom intervalu je i očekivan. U kojoj meri će pri smanjivanju vrednosti parametra q rasti nivo prioriteta koji filtar daje rezultatima predviđanja zavisi od izbora scenarija inicijalizacije. Tačnost predviđenog vektora stanja $\boldsymbol{x}_{k+1}^{-}$ zavisi od tačnosti dinamičkog modela, ali i od tačnosti vektora stanja $\boldsymbol{x}_{k}^{+}$ estimiranog u prethodnom trenutku. Ukoliko je greška dinamičkog model zanemarljiva u odnosu na grešku koja je pri estimiranju stanja u prethodnom trenutku, onda će greška u predviđanju vektora stanja u najvećoj meri zavisiti od greške prethodno estimiranog vektora stanja. Sa matematičkog stanovišta, odnos grešaka dinamičkog modela i prethodno estimiranog vektora stanja opisuje se odnosom vrednosti dijagonalnih elementa matrice $\boldsymbol{Q}$ i matrice $\boldsymbol{P}^{+}$, koji predstavljaju varijanse grešaka dinamičkog modela i estimiranog vektora stanja, respektivno.

Scenario „SSE start“

Sa slike 1 vidi se da u slučaju scenarija „SSE start“ promena parametra q nema nikakvog uticaja na grešku estimacije EKF algoritma sve dok parametar ima vrednost manju od $q \approx -8{,}5$. Na tom intervalu vrednosti parametra q su takve da su varijanse grešaka dinamičkog modela i za nekoliko redova veličine manje od varijansi prethodno estimiranih vrednosti promenljivih stanja. Varijanse inicijalnih vrednosti promenljivih stanja su u opsegu $3{,}73 \cdot 10^{-9}$ do $1{,}06 \cdot 10^{-7}$ za IEEE 13 test sistem. S obzirom da su ove varijanse procenjene na osnovu varijansi merenja, koje se ne menjaju značajnije u kvazistacionarnom režimu, može se zaključiti da se i vrednosti elemenata matrice $\boldsymbol{P}^{+}$ neće značajnije menjati tokom vremena. Blage promene su jedino posledica blagih varijacija opterećenja. Sa stanovišta samog filtra može se reći da je matrica $\boldsymbol{P}^{+}$ dostigla ustaljenu (stvarnu) vrednost. Može se zaključiti i da će vrednosti elemenata matrice $\boldsymbol{P}^{-}$ biti približno jednake vrednostima odgovarajućih elemenata matrice $\boldsymbol{P}^{+}$, pri čemu će promene opterećenja imati isti efekat na matricu $\boldsymbol{P}^{-}$ kao što imaju i na matricu $\boldsymbol{P}^{+}$. Na slici 2 prikazane su varijanse predviđenih i estimiranih vrednosti svih promenljivih stanja (dijagonalni elementi matrice $\boldsymbol{P}^{-}$ i $\boldsymbol{P}^{+}$, respektivno) za proizvoljno izabran trenutak $k = 10$. Oblast ustaljenih vrednosti označena je sivom bojom. Ona prestavlja oblastu kojoj se kreću varijanse estimiranih vrednosti promenljivih stanja tokom dana.



Slika 2. Varijanse predviđenih i estimiranih vrednosti promenljivih stanja aproksimirane EKF algoritmom dinamičke estimacije stanja na trening periodu u trenutku $k = 10$

Iz prethodnih razmatranja mogu se izvesti sledeći zaključci:

1. Odgovarajući elementi matrica kovarijansi predviđenog vektora merenja $\boldsymbol{T}$ i grešaka merenja $\boldsymbol{R}$ biće približno jednaki, što znači da će nivoi prioriteta koje filtar daje predviđenim vrednostima merenja i trenutnom preseku merenja biti otprilike isti.

2. Predviđene vrednosti promenljivih stanja ulaze u proces filtriranja s istim nivoom prioriteta bez obzira koliko je q, pa se iz tog razloga greška dinamičke estimacije ne menja s promenama vrednosti parametra q.

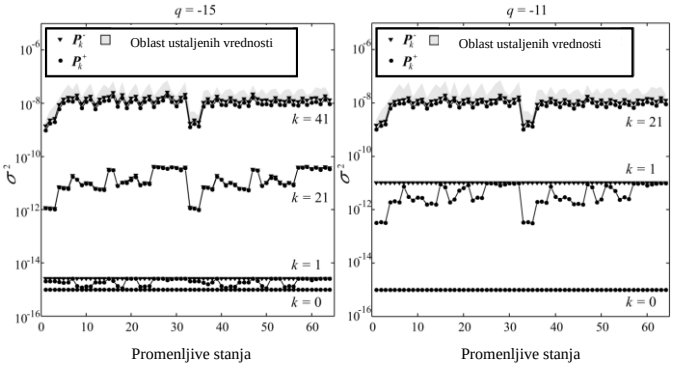
Scenario „tačan start 2“

U slučaju scenarija „tačan start 2“ inicijalne vrednosti promenljivih stanja imaju znatno manje varijanse, koje iznose 10^{-15}. Za vrednost parametra $q = -15$, varijanse grešaka dinamičkog modela i estimiranih promenljivih stanja u početku imaju podjednak udeo u aproksimiranim vrednostima

varijansi predviđenih promenljivih stanja. Zbog zoga su varijanse predviđenih vrednosti merenja znatno manje od varijansi grešaka merenja, pa filtar daje znatno veći prioritet predviđenim vrednostima nego trenutnom preseku merenja. Ovo će dovesti do izuzetno velike greške estimacije ξ_n, kao što se vidi sa slike 1. S druge strane, zbog malih varijansi predviđenog vektora stanja u trenutku $k=1$ i sama matrica kovarijansi $\boldsymbol{P}_1^+$ imaće vrednosti istog reda veličine. Ovde se dolazi do kontradikcije: rezultati estimacije su loši, a matrica kovarijansi estimiranog vektora stanja ima izuzetno niske vrednosti, što znači da bi rezultati estimacije trebalo da budu dobri. Treba imati u vidu da filtar jednačinom (10) samo aproksimira stvarnu matricu kovarijansi $\boldsymbol{P}^+$ (takođe, i matrica $\boldsymbol{P}^-$ data jednačinom (3) je aproksimacija). Razlog za lošu aproksimaciju u ovom slučaju je to što je greška dinamičkog modela neadekvatno modelovana, odnosno loše podešenje parametra q. Greška dinamičkog modela u ovom slučaju je potcenjena, što znači da je pretpostavljena varijansa greške znatno manja od stvarne.

Kako su Holtovi parametri izravnanja odabrani tako da dijagonalni elementi tranzicione matrice $\boldsymbol{F}$ imaju vrednosti nešto veće od 1, varijanse predviđenih promenljivih stanja će postepeno rasti tokom vremena. S obzirom da približna jednakost odgovarajućih elemenata matrica $\boldsymbol{P}^-$ i $\boldsymbol{P}^+$ važi i ovde, s porastom varijansi predviđenih promenljivih stanja rašće i varijanse estimiranih promenljivih stanja ka svojoj ustaljenoj (stvarnoj) vrednosti. To znači da će s vremenom opadati nivo prioriteta koji filtar daje rezultatima predviđanja, a rasti nivo prioriteta koji filtar daje merenjima. Ulaskom matrice $\boldsymbol{P}^+$ u oblast ustaljenih vrednosti ova dva nivoa prioriteta postaće približno isti.

Povećanjem varijansi grešaka dinamičkog modela ubrzaće se ulazak matrice $\boldsymbol{P}^+$ u oblast ustaljenih vrednosti, zbog čega se s porastom vrednosti parametra q greška estimacije ξ_n smanjuje. Na slici 3 prikazana je progresija dijagonalnih elemenata matrica $\boldsymbol{P}^-$ i $\boldsymbol{P}^+$ za dva različita podešenja parametra q. Vidi se da pri $q=-11$ matrica kovarijansi $\boldsymbol{P}^+$ dostiže svoju ustaljenu vrednost skoro dvostruko brže nego pri podešenju $q=-15$.



Slika 3. Progresija varijansi predviđenih i estimiranih vrednosti promenljivih stanja aproksimiranih EKF algoritmom dinamičke estimacije stanja na trening periodu za dva različita podešenja parametra q

Sa smanjenjem vrednosti parametra q ispod -15 opada i udeo koji imaju varijanse grešaka dinamičkog modela u aproksimiranim vrednostima varijansi predviđenih promenljivih stanja. Sa slike 1 vidi se da sa smanjenjem vrednosti parametra q ispod -15, greška estimacije ξ_n postaje sve manje osetljiva na promene parametra q.

Scenario „tačan start 1“

Za scenario „tačan start 1“ mogu se izvesti slični zaključci kao i za prethodno analizirani scenario, s tim da su varijanse inicijalnog rešenja jednake 0, pa filtar tretira inicijalni vektor stanja kao apsolutno tačan. Zbog toga varijanse grešaka dinamičkog modela u početnom trenutku imaju glavni udeo u aproksimiranim vrednostima varijansi predviđenih promenljivih stanja, pa je greška estimacije ξ_n osetljiva na promene parametra q i za vrednosti $q<-15$. U tom slučaju će u početku varijanse predviđenih promenljivih stanja biti još manje nego kod scenarija „tačan start 2“, što će produžiti vreme konvergencije matrice $\boldsymbol{P}^+$ ka svojoj stvarnoj vrednosti i time rezultovati još većom greškom estimacije ξ_n. Za scenario „tačan start 2“ važi da, s porastom vrednosti parametra q iznad -15, varijanse grešaka dinamičkog modela dobijaju, u početnom trenutku, sve veći udeo u aproksimiranim vrednostima varijansi predviđenih promenljivih stanja, zbog čega greške estimacije ξ_n dobijene za scenarija „tačan start 1“ i „tačan start 2“ teže da se izjednače. Drugim rečima, ako su dijagonalni elementi matrice $\boldsymbol{Q}$ samo za red veličine veći od 10^{-15}, onda je nebitno da li su dijagonalni elementi matrice $\boldsymbol{P}_0^+$ jednaki 0 ili 10^{-15}.

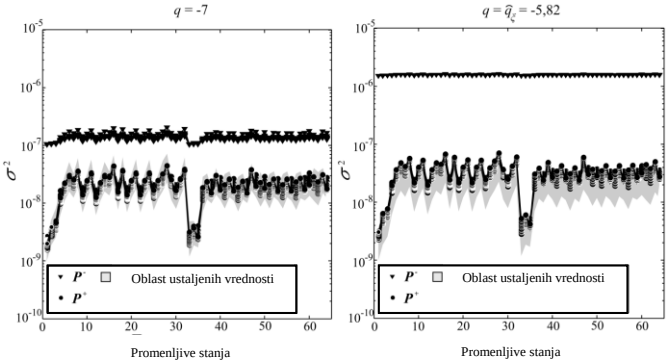
Slično važi i za UKF algoritam na ovom intervalu vrednosti parametra q, samo što on postaje osetljiviji na promene parametra q pri ekstremno niskim vrednostima.

Interval vrednosti parametra $-8 \le q \le \hat{q}_\xi$

Iako scenario „SSE start“ u intervalu vrednosti parametra $-16 \le q \le -8$ značajno popravlja kvalitet dinamičke estimacije u odnosu na preostala dva scenarija, greška dinamičke estimacije je i dalje prilično veća od greške statičkog estimatora (pogledati sliku 1). To znači da treba smanjiti nivo prioriteta koji filtar daje predviđenim vrednostima, a to se može postići povećanjem vrednosti parametra q. Već pri kraju prethodno navedenog intervala varijanse grešaka dinamičkog modela postepeno zalaze u oblast ustaljenih vrednosti varijansi estimiranih promenljivih stanja. Kako vrednost parametra q dalje raste, nivo prioriteta koji filtar daje rezultatima predviđanja opada, a nivo prioriteta koji filtar daje merenjima istovremeno raste, što smanjuje grešku dinamičke estimacije ξ_n. Ključno pitanje je koliko treba povećati vrednost parametra q da bi odnos u nivoima prioriteta koje filtar daje rezultatima predviđanja (s jedne strane) i merenjima (s druge strane) bio optimalan? U slučaju analiziranog test sistema to je vrednost $q=\hat{q}_\xi=-5{,}82$.

Čim vrednost parametra q poraste toliko da varijanse grešaka dinamičkog modela izađu iz oblasti ustaljenih vrednosti dijagonalnih elementa matrice $\boldsymbol{P}^+$, varijanse grešaka dinamičkog modela dobijaju dominantan uticaj pri aproksimaciji varijansi predviđenih promenljivih stanja. Zbog toga se, pri vrednostima većim od $q \approx -7{,}5$, rezultati estimacije sa scenarijom „SSE start“ gotovo identično poklapaju s rezultatima estimacije dobijenim s preostala dva scenarija (pogledati sliku 1).

S porastom vrednosti parametra q, sve više se narušava jednakost odgovarajućih elemenata matrica $\boldsymbol{P}^-$ i $\boldsymbol{P}^+$, što se vidi na slici 4. Veća vrednost parametra q modeluje grešku dinamičkog modela s većom varijansom. Usled toga se povećava neizvesnost u rezultatima predviđanja, a posledično, kroz proces filtriranja, dolazi i do povećanja varijansi estimiranih promenljivih stanja.



Slika 4. Varijanse predviđenih i estimiranih vrednosti promenljivih stanja aproksimirane EKF algoritmom dinamičke estimacije stanja na celom trening periodu za vrednosti parametra (a) $q = -7$ i (b) $q = \hat{q}_\xi = 5{,}82$

Sa slike 1 može se zapaziti da dinamička estimacija stanja ima manju grešku od statičke ukoliko je matrica kovarijansi $\boldsymbol{Q}$ adekvatno podešena. Bez obzira na scenario inicijalizacije ili tip Kalmanovog filtra, greška dinamičke estimacije stanja dostiže minimum za gotovo identičnu vrednost parametra $\hat{q}_\xi$. Osim toga, u okolini optimuma scenariji „tačan start 1“ i „tačan start 2“ daju apsolutno iste rezultate (krive se poklapaju). Iako su rezultati dobijeni s ova dva scenarija u okolini optimuma bolji nego oni dobijeni sa scenarijom „SSE start“, razlike u grešci estimacije su zanemarljivo male, što je značajno iz dva razloga:

1. Ako je matrica kovarijansi $\boldsymbol{Q}$ podešena optimalno, izbor inicijalnog vektora stanja i njemu pripadajuće matrice kovarijansi nije od velikog značaja.

2. Scenariji „tačan start 1“ i „tačan start 2“ su hipotetički, u smislu da su zasnovani na pretpostavci da je poznato tačno stanje u sistemu, što se može ostvariti samo u uslovima simulacije. Tačan vektor stanja nije moguće poznavati u praksi. S druge strane, scenario „SSE start“ može se koristiti u praktičnim aplikacijama. Male razlike u rezultatima ukazuju na to da se u praktičnim aplikacijama stanje u sistemu može estimirati s gotovo najvećim mogućim stepenom tačnosti.

Interval vrednosti parametra $q > \hat{q}_\xi$

Već na početnom delu ovog intervala vrednosti grešaka estimacije dobijene pomoću dva različita algoritma dinamičke estimacije stanja počinju da se razilaze, pri čemu greška UKF algoritma počinje naglo da raste (pogledati sliku 1).

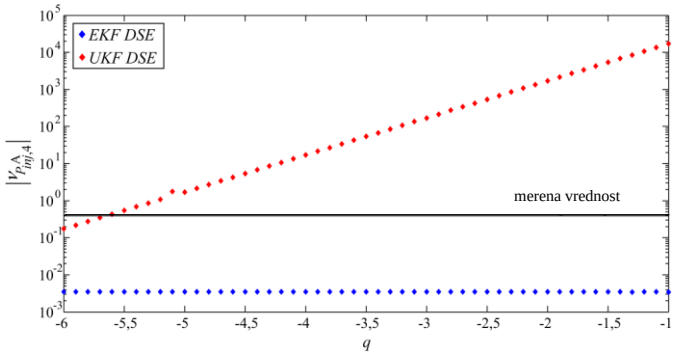
EKF algoritam koristi matricu kovarijansi $\boldsymbol{Q}$ samo pri aproksimiranju matrice kovarijansi predviđenog vektora merenja $\boldsymbol{T}$, ali ne i pri proračunu predviđenog vektora merenja $\boldsymbol{z}^-$ (pogledati jednačine (4) i (5)). Kako u ovom intervalu filtar daje veći nivo prioriteta merenjima nego rezultatima predviđanja, predviđeni vektor merenja se neće značajnije menjati s promenom parametra q. Osim toga, predviđene vrednosti merenja biće generalno blizu vrednosti samih merenja (biće istog reda veličine). S porastom vrednosti parametra q nivo prioriteta koji filtar daje predviđenim vrednostima merenja biće sve manji i manji. Može se zaključiti da greška estimacije ξ_n na ovom intervalu raste jer filtar ne daje dovoljan nivo prioriteta vrednostima predviđenim pomoću dinamičkog modela. Drugim rečima, greška dinamičkog modela je precenjena u smislu da je pretpostavljena varijansa greške veća od stvarne. Indikativno je da greška ξ_n, dobijena EKF algoritmom dinamičke estimacije stanja, teži grešci statičkog estimatora kako parametar q raste po vrednosti. To je i opravdano jer će se, pri velikim vrednostima parametra q, prioritet koji filtar daje rezultatima predviđanja svesti na toliko mali nivo da se može reći da filtar estimira samo na osnovu trenutnog preseka merenja.

Za razliku od EKF algoritma, kod UKF algoritma matrica kovarijansi $\boldsymbol{Q}$ ima udela i pri proračunu predviđenog vektora merenja $\boldsymbol{z}^-$ (pogledati jednačine (17) i (18)). Na prethodna dva intervala vrednosti parametra q taj udeo je bio zanemarljiv, ali na aktuelnom intervalu dolazi do izražaja. S porastom vrednosti parametra q rastu varijanse predviđenih vrednosti merenja (i to s istim trendom rasta kao kod EKF algoritma), ali ujedno raste i odstupanje predviđenih vrednosti merenja u odnosu na sama merenja, tj. rastu inovacije merenja (što kod EKF algoritma nije izraženo). Porast varijansi predviđenih vrednosti merenja odražava se na porast varijansi inovacija (jednačina (20)).

Na početnom delu intervala (oko optimuma) porast varijansi inovacija uspeva da prati porast inovacija. Međutim, već pri vrednostima većim od $q \approx -5$ inovacije počinju da rastu znatno brže tako da varijanse inovacija ne stižu da adekvatno isprate taj porast. Za pojedina merenja, predviđene vrednosti se pri $q = -5$ razlikuju od izmerenih vrednosti za red veličine. Iako pri ovom podešenju parametra q filtar daje veći nivo prioriteta merenjima, izvestan nivo prioriteta daje se i rezultatima predviđanja. Kako su za pojedina merenja razlike između predviđenih i izmerenih vrednosti velike, rezultati predviđanja će estimirane vrednosti „udaljiti“ od merenja više nego što je to slučaj kod EKF algoritma, a samim tim više i od tačnog vektora stanja. Zbog toga greška estimacije ξ_n raste. Zbog nesrazmernosti u brzini rasta inovacija i varijansi

inovacija, greška estimacije najpre naglo raste (do $q \approx -3$), da bi se daljim porastom parametra q ova nesrazmernost počela polako da opada i nestaje, zbog čega greška prestaje da raste. Osim toga, pri kraju ovog intervala vrednosti parametra q su takve da razlike između predviđenih vrednosti i samih merenja mogu biti enormne, što UKF algoritam potencijalno može uvesti u numeričke probleme [11].

Na slici 5 prikazano je kako izgleda apsolutna vrednost inovacije ν telemetrisanog merenja aktivne snage injektiranja u fazi A čvora 4 u trenutku $k = 15$ pri različitim podešenjima parametra q, dobijena pomoću dva različita algoritma dinamičke estimacije stanja. Takođe, na slici 5 je prikazana vrednost merenja u trenutku $k = 15$, kako bi se stekao utisak koliko iznosi relativno odstupanje predviđene vrednosti u odnosu na izmerenu.



Slika 5. Apsolutna vrednost inovacije telemetrisanog merenja aktivne snage injektiranja u fazi A čvora 4 dobijena EKF i UKF algoritmom dinamičke estimacije stanja na trening periodu u trenutku $k = 15$ pri različitim podešenjima parametra q

V. Zaključak

U ovom radu sprovedena je detaljna analiza osetljivosti greške dinamičke estimacije stanja na promene inicijalnog rešenja (inicijalno estimirani vektor stanja i njemu pripadajuća matrica kovarijansi) i matrice kovarijansi grešaka dinamičkog modela.

Utvrđeni su razlozi specifične osetljivosti greške dinamičke estimacije stanja na promenu analiziranih parametara Kalmanovog filtra, čime je načinu na koji se greška menja data čvrsta teorijska potvrda. Može se zaključiti da se osetljivost greške na promene ovih parametara ne može zanemariti čak ni u kvazistacionarnom režimu. Stoga je pravilan izbor ovih parametara, a pre svega parametra matrice kovarijansi grešaka dinamičkog modela, od presudnog značaja kada je u pitanju kvalitet dinamičke estimacije stanja u distributivnim mrežama.

Pokazano je da promena analiziranih parametara utiče na gotovo identičan način na grešku dinamičke estimacije stanja bez obzira koji tip Kalmanovog filtra je u upotrebi. Odstupanja postoje jedino pri izuzetno niskim i izuzetno visokim vrednostima parametra matrice kovarijansi grešaka dinamičkog modela. U tom slučaju EKF pokazuje bolje karakteristike u pogledu filtriranja.

Izbor inicijalnog rešenja i matrice kovarijansi grešaka dinamičkog modela odražava se na aproksimiranje neizvesnosti u predviđenom vektoru merenja, kao i vektoru inovacija, koje imaju presudnu ulogu u fazi detekcije i identifikacije anomalija kod ovakve vrste estimatora. Razumevanje načina na koji Kalmanov filtar aproksimira pojedine matrice kovarijansi je od velikog značaja u fazi detekcije i identifikacije anomalija. Stoga ovaj rad predstavlja dobru osnovu za buduća istraživanja novih metoda za detekciju i identifikaciju anomalija u dinamičkoj estimaciji stanja distributivnih mreža.

Literatura

[1] J. B. Zhao, A. Exposito, M. Netto, L. Mili, A. Abur, V. Terzija, I. Kamwa, B. Pal, A. K. Singh, J. Qi, Z. Huang, A. P. Sakis Meliopoulos, "Power System Dynamic State Estimation: Motivations, Definitions, Methodologies and Future Work," IEEE Trans. Power Systems, 2019.

[2] D. Ćetenović, A. Ranković, „Optimal parameterization of Kalman filter based three-phase dynamic state estimator for active distribution networks“, Int. J. Electral Power & Energy Systems, Vol. 101, No. 1, pp. 472-481, 2018.

[3] D. Ćetenović, „Dinamička estimacija stanja nesimetričnih elektrodistributivnih mreža i optimalno podešavanje parametara Kalmanovog filtra“, Doktorska disertacija, Univerzitet u Kragujevcu, Fakultet tehničkih nauka u Čačku, 2019.

[4] M. B. Do Coutto Filho, J. C. S. de Souza, „Forecasting-Aided State Estimation - Part I: Panorama“, IEEE Trans. Power Systems, Vol. 24, No. 4, pp. 1667-1677, 2009.

[5] L. Zanni et al, „A Prediction-Error Covariance Estimator for Adaptive Kalman Filtering in Step-Varying Processes: Application to Power-System State Estimation“, IEEE Trans. Control Systems Technology, Vol. 25, No. 5, pp. 1683-1697, 2017.

[6] S. Särkkä, „Bayesian Filtering and Smoothing“, Cambridge University Press, 2013.

[7] G. Valverde, V. Terzija, “Unscented Kalman filter for power system dynamic state estimation,” IET Gener., Transm. Distrib., vol. *5*, no. *1*, pp. *29-37*, January, 2011.

[8] S. J. Julier, J. K. Uhlmann, „Unscented filtering and nonlinear estimation“, Proc. IEEE, Vol. 92, No. 3, pp. 401-422, 2004.

[9] S. Särkkä, „Recursive Bayesian inference on stochastic differential equations“, Doctoral dissertion, Helsinki University of Technology, Laboratory of Computational Engineering, 2006.

[10] http://sites.ieee.org/pes-testfeeders/resources/

[11] J. Qi et al, „Dynamic State Estimation for Multi-Machine Power System by Unscented Kalman Filter with Enhanced Numerical Stability“, IEEE Trans. Smart Grid, Vol. 9, No. 2, pp. 1184-1196, 2016.

Abstract

When distribution network operates in steady state conditions the process noise covariance matrix $\boldsymbol{Q}$ is commonly modeled as time invariant in diagonal form with diagonal elements all equal, which leads to a one-parameter matrix model. In this case the assessment of process noise covariance matrix $\boldsymbol{Q}$ comes down to the assessment of a single parameter q. The accuracy of dynamic state estimation is conditioned by the choice of this parameter, but also by the choice of the initial solution, where initial solution encompasses initial state vector $\boldsymbol{x}_0^+$ and its error covariance matrix $\boldsymbol{P}_0^+$. The aim of this paper is to investigate in detail the simultaneous effect of parameter q and initial solution on the accuracy of dynamic state estimation. The analysis are carried out on modified IEEE distribution test system with 13 buses using the EKF and UKF dynamic state estimation algorithm.

The sensitivity of dynamic state estimation error to changes in Kalman filter parameters

Dragan Ćetenović, Aleksandar Ranković

MRTD Measurements Role in Thermal Imager Quality Assessment

Dragana Perić, *Member, IEEE*, Branko Livada, *Member, SPIE*

Abstract— **The MRTD is defined as thermal imager basic parameter that could be theoretically modelled and experimentally assessed in laboratory. The subjective MRTD measurement procedure that is practically implemented is presented. The MRTD theoretical models, possibilities for objective measurements, and alternative MRTD measurement methods are discussed. MRTD measurement results for selected thermal imagers and measurement results post processing for thermal imager range determination are presented**.

***Index Terms* — Thermal imager; Minimum resolvable temperature difference; MRTD subjective measurements; thermal imager range.**

I. INTRODUCTION

Modern long range electro-optical surveillance systems [1], [2] applications involve infrared focal plane based imager channel as a must. The current infrared detector technology development [3] provides that this technology is not exclusively military, but is also available for civilian surveillance/reconnaissance applications. In these applications the basic requirements are connected with knowledge of how far one can see the object of interest – target, and distinguish useful visual data about target.

Various measures and techniques are now available to characterize spatial information transfer, sensitivity and noise in thermal imagers that could help the system architect to predict imager range properties. Also, various figures of merit and measurement techniques were developed to assess the overall device performance. One of the most frequently used and relevant parameters is Minimum Resolvable Temperature Difference – MRTD [4]. MRTD represents thermal imager figure of merit that could be modelled [5] - [9], measured and used for range prediction calculations according to the selected visual data perception criteria. MRTD measurement procedures generally involve application of the selected target test pattern using controlled temperature blackbody source, projection collimator and generated image analysis (perception). The standard MRTD procedure involves human observer in the measurement loop.

In this article we described MRTD measurement procedure realized in the Vlatacom Electro-optical laboratory, we also presented selected measurement results and their post processing which is done in order to determine imagers range according to selected visual data perception criteria.

The subjective MRTD measurement methodology is described after more detailed MRTD definition. Also, we discuss the potential and issues with objective MRTD measurements and new alternative measurement methods using different test targets. At the end we are presenting measurement results for selected thermal imagers together with range predictions using MRTD measurement results.

Dragana Perić is with Vlatacom Institute, Bulevar Milutina Milankovica 5, 11070 Belgrade, Serbia (e-mail: dragana.peric@vlatacom.com).

Branko Livada is with Vlatacom Institute, Bulevar Milutina Milankovica 5, 11070 Belgrade, Serbia (e-mail: branko.livada@vlatacom.com).

II. MRTD DEFINITION

The MRTD is the minimum temperature difference which allows an observer to resolve a test pattern in accordance with a given criterion. MRTD is a function of spatial frequency of the test pattern. It depends on measurement temperature and may depend on the orientation of the test pattern.

Historically, the basic MRTD test pattern as shown in Fig.3 is selected according to Johnson's criteria [10] for visual data perception. Following results from experimental research of the visual data perception by human observer, Johnson determined the number of the periodic line pairs over target critical dimension that provide related level of the visual perception (detection - D, recognition - R, identification - I). His results were accepted as an industry standard. The spatial frequencies mean values for different visual data perception levels as defined by Johnson's criteria are listed in TABLE I. These criteria are derived for visual data transformation process that keep the same contrast ratio in basic and transformed image and perception level do not depends on of the signal to noise ratio, assuming 50% perception probability.

Typical test targets following D, R, I criteria used in laboratory testing are presented in Fig. 1.

TABLE I
VISUAL PERCEPTION LEVELS AGAINST CRITICAL SPATIAL FREQUENCY, AS DEFINED BY JOHNSON'S CRITERIA

Critical Spatial frequency	***Visual perception level***		
	D	**R**	**I**
lp/target size	**1,±0,25**	**4±0,8**	**6,4±1,5**

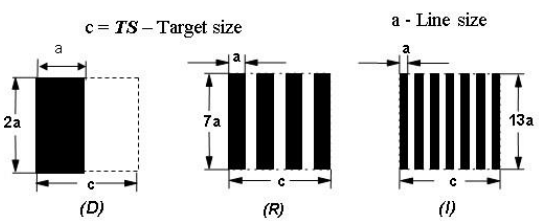


Fig. 1. Commonly used test target patterns following D, R, I visual perception criteria

As defined in the standard for the subjective MRTD test, human observer should determine and indicate when target is resolved. There are efforts to use automatic objective target resolution criteria but current implementations haven't reached yet required level of fidelity, because they require complex image processing techniques to be developed.

Also, there are efforts to use different target (Triangle orientation – TOD) that shows good results and eliminate some of the disadvantages of the concept with four bar target. Both, objective and alternative MRTD measurement methodology will be discusses separately.

In our approach we use the four bar target and subjective methodology that is described in details.

III. SUBJECTIVE MRTD MEASUREMENT METHODOLOGY

According to measurement best practices [10] – [13] and standards [14] – [17], MRTD measurement set up comprises IR reflexive collimator and differential blackbody in conjunction with 4 bar test pattern, as illustrated in Fig. 2 and Fig.3.

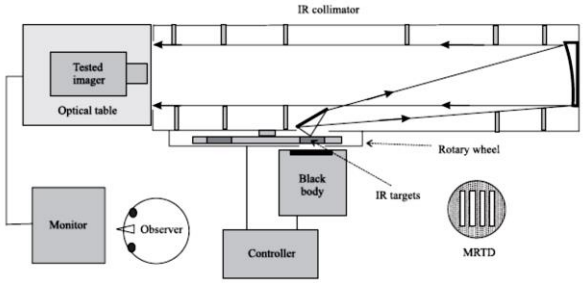


Fig. 2. MRTD Measurement set-up

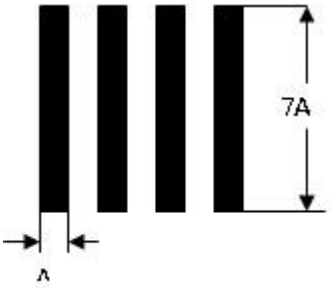


Fig. 3. MRTD four bar test pattern structure

During the MRTD measurements laboratory conditions and other measurement conditions should be established, as follows:

A. Laboratory conditions

Unless otherwise specified in the thermal imager specification, the ambient temperature during the test is set at the normal room temperature (20ºC ± 2ºC).

Relative humidity does not need to be controlled.

The illumination in the test area should be set as close as operational illumination providing that observer could set it as most convenient.

B. Observer

Observer shall have normal visual acuity (post-correction defects less than ±0,25 diopter, and good color vision.

Observer should be experienced (trained) in this type of measurements.

Observer is allowed to adjust room illumination, and the distance of his eye from display to provide optimal image perception.

C. IR collimator

It is recommended to use reflective off-axis collimator with known transmittance in the imager's spectral sensitivity region. The collimator transmittance should be known and could be used in measurement results correction, if necessary.

D. Test pattern

Four bar test pattern structure is illustrated in Fig 3.

Test pattern is placed in the collimator focal plane. In that case one bar angular size is:

$$\alpha_t[rad] = 2 \cdot arctg\frac{A}{2f} \approx \frac{A}{f} \quad (1)$$

Where A is bar width and f is collimator focal length.

Test target spatial frequency f_t^*, could be calculated as:

$$f_t^*\left[\frac{lp}{rad}\right] = \frac{2}{\alpha_t} \quad (2)$$

E. Differential blackbody

The emissivity of the test pattern surface and black body surface should be at least 0,95.

The blackbody must have the capability of achieving temperature difference of ±10 K. The temperature measurement accuracy must be better than 0,5% generally and 0,01 K, for temperature differences between 0 K and ±2K.

Test pattern and blackbody temperature measurement accuracy TR, temporal stability TS and spatial uniformity TU requirements are related to device under test (DUT) sensitivity limit, such as Noise Equivalent Temperature Difference NETD. These parameters should have the given ratio respectively: NETD ≥ 5TR, NETD ≥ TS and NETD≥5TU.

F. Monitor

Monitor characteristics (type, size and resolution) should be selected in correspond with the specification of the imager under test. Monitor should not be readjusted during measurements, it should keep initial settings.

G. Measurement Procedure

MRTD Measurement Preparation

STEP 1: Nyquist frequency (f_0) determination for device under test (DUT)

Using available DUT specification or measurement results Nyquist frequency should be determined, using one or all listed methods:

- According to DUT design data and related model application results, calculate Instantaneous Field of View (IFOV) and Nyquist frequency accordingly.
- Using USAF1951Test Chart determine DUT resolution frequency.
- Using DUT calibration procedure determine IFOV and

DUT Field of View (FOV).

STEP 2: Test target selection

To generate MRTD curve against spatial frequency, at least three targets with related spatial frequencies should be used. The selection of targets in the test depends on availability.

It is recommended to use targets with following spatial frequencies: ~f_0 , 0,8f_0, 0.5f_0 , 0,2f_0. In the case that DUT requirements define other test frequencies the additional targets with required spatial frequencies should be provided. The target spatial frequencies must be within ±5%, of the nominal value.

MRTD Measurement Execution

- Position the target with the bars oriented vertically to obtain the horizontal MRT or horizontally to measure the vertical response.
- Verify that four bars are visible when limiting test temperature difference (≤ 2K) is set.
- Using test set-up controller establish a positive sub-MRTD temperature differential and slowly increment the blackbody temperature differential.
- Allow the observer to continually adjust DUT controls to optimize the image for given focal length.
- Record the temperature difference (MRTD+) at which the observer can detect all four bars with 50% probability.
- Establish a negative sub-MRTD temperature differential and slowly decrement the blackbody temperature differential. Wait until temperature difference stabilized before apply next decrement step.
- Record the temperature difference (MRTD-) at which the observer can detect all four bars with 50% probability.
- Average the absolute values of positive and negative temperature differential recordings to obtain the MRTD.

$$MRTD(f^*) = \frac{MRTD_+ - MRTD_-}{2} \tag{3}$$

- Record measured values in the prepared measurement results template (it is recommended to use the separate measurement results EXCEL spreadsheet containing data processing and measurement report sheet).
- If a target cannot be resolved, record TNR.
- Repeat for other spatial frequencies.

NOTE 1: *Blackbody temperature differential controls*

MRTD measurements are time consuming. To economize measurement time it is recommended to follow best practice in selection of the starting temperature difference and temperature change step value. Some good recommendation could be applied.

- Following expected MRTD values and knowledge about DUT NETD, select starting temperature difference to minimize measurement time (Example: set starting temperature difference value to be 5 NETD).
- Select temperature difference step value accordingly to starting value following recommendations from table below:

ΔT [K]	Proposed Step
Below 0,5 K	0,01 K
0,5 K to 1,0 K	0,02 K
1,0 K to 2,0 K	0,05K

NOTE 2: *50% probability of detection criteria*

50% probability of detection during MRTD test could be derived from following recommendations.

- All four bars are visible during 50% of the observation time.
- All four bars are visible all the time but with height equal or higher than 50% of nominal height.
- In the case that several observers are involved in measurement the 50% detection probability is achieved if at least 50% of observers agree that target is resolved.

MRTD measurement results should be listed as table and plotted as MRTD measurement chart.

IV. Objective MRTD measurement issues

MRTD provides connection between thermal imager sensitivity represented by NETD (noise equivalent temperature difference), image transformation represented by MTF (modulation transfer function) and response represented by SNR (signal to noise ratio) [21], as:

$$MRTD(f) = \frac{SNR \cdot NETD}{MTF(f)} \cdot \boldsymbol{F}_{TIS}(f, DP) \tag{4}$$

Where: F_{TIS}(f, DP) – is thermal imager related design function depending on system design parameters - DP. This MRTD model could be proved by laboratory measurements.

The main goal of objective MRTD measurements [22] is to deliver results comparable with subjective method, but less costly and much faster. The key issue is how to automatically represent visual perception of the human observer and displayed image. Different techniques were applied but still nono of them delivered result which is considered good enough.

V. Alternative MRTD measurement methods

There are two general weaknesses with standard MRTD testing [23], [24]: (i) subjective test (depends of the observers that lead to higher inaccuracies in the test results obtained in different labs); (ii) the test target is a 3-bar or 4-bar pattern with a specific spatial frequency so limited bracket of test patterns is applied. Because of that some new test methods and concepts are considered.

MTDP (Minimum Temperature difference Perceived) [25] that uses the traditional static 3 or 4 bar target but allows that under-sampled thermal imagers could be evaluated above Nyquist frequency.

A new test should at least have the following properties [26], [27]: lab testing is objective and easy, the measure is representative for field performance, and modelling (sensor and human) the test should be relatively easy. Several alternative test methods and test patterns have already been proposed. An example is the TOD method [28] that uses non-periodic test patterns

The TOD (triangle orientation discrimination) method use test patterns (oriented equilateral triangles) that represent features or the relation between features of real targets. Triangle size is related to spatial frequency. If an observer is able to discriminate between the different triangle orientations, he also has information about the target features

necessary to identify a target. Observer is not able to determine the correct orientation of the original test target if: (i) the test pattern cannot be detected (because the SNR is too low); (ii) its shape cannot be determined (because corners and edges disappeared due to blur or sampling); or (iii) the shape is incorrectly judged (relative positions are disturbed due to under-sampling or to phase shifts introduced by electronics).

TOD metric has short observation window, provides better capability for evaluating sensors that exhibit non-negligible uniformity drift, and could be more effective for sensor dynamic evaluation. In addition, same methodology is applicable for IR and visible image sensors.

VI. MRTD MEASUREMENT RESULTS AND PROCESSING

MRTD measurement results could be used to calculate related perception range using predefined criteria and procedure [15], [16].

Nominal static range performances are determined for target size (TS=2,3m x 2,3m) and target effective temperature difference to background ΔT=2K. The atmospheric attenuation influence is calculated using

$$\tau(R) = e^{-\sigma \cdot R} \tag{5}$$

Where - R is distance in kilometers, σ is atmospheric transmission, that is that has following values:

σ_{gtc} = 0,2 km^{-1} , for good transmission conditions

σ_{ltc} = 1,0 km^{-1} , for limited transmission conditions

MRTD(*f*) must be transformed into MRTD(*R*) (R=R_D, R_R, R_I) using resolution criteria according to Johnson's criteria (Detection- N_D=1 lp/target size; Recognition - N_R = 3 lp/target size; Identification - N_I = 6 lp/target size)

Spatial frequency axis is transformed into related range axis using relation:

$$R[km] = \frac{TS \cdot f^*[lp/mRad]}{N} \tag{6}$$

Where N is resolution criteria value applied.

Both MRTD(R) and apparent temperature difference ΔTA = ΔT(R) should be plotted in the same chart The range coordinate of the cross section point of these two represent related perception range.

The example of the measurement results obtained in our laboratory are presented in Fig.4 (USAF 1951 – test chart showing thermal imager's limiting resolution, Fig. 5 showing MRTD raw measurement results, and Fig.6, Fig.7, Fig.8 are showing visual perception range processing results for standard target and atmospheric conditions in cases of detection, recognition and identification respectively. Three graphs show: apparent temperature difference (ATD – in case of detection, ATP - recognition, ATI - identification), measured values for the selected targets (MRTD_MEAS), and MRTD-FIT which is obtained from values MRTD-MEAS, in order to provide results for all values of range R. MRTD-FIT is calculated according to the function (7):

$$MRTD(f^*) = A + B \cdot f^* \cdot \exp(C \cdot f^{*2}) \tag{7}$$

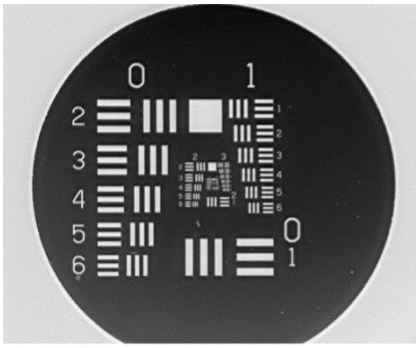


Fig. 4. USAF 1951 Test target image, G3E5 resolved

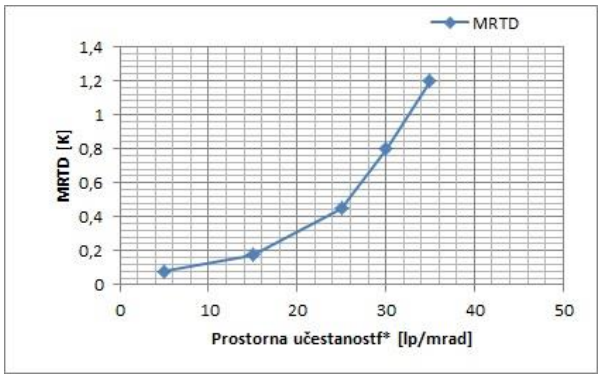


Fig. 5. MRTD measurement results chart

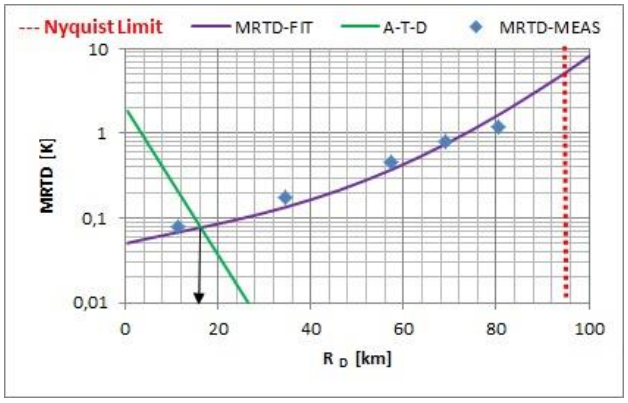


Fig. 6. MRTD Chart with target apparent temperature used for range prediction in case of Detection - D

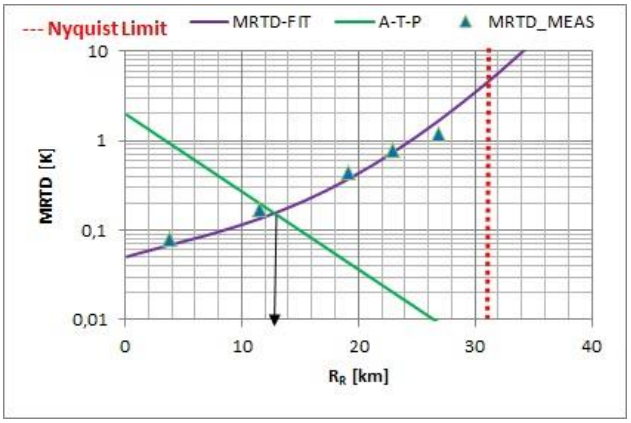


Fig. 7. MRTD Chart with target apparent temperature used for range prediction in case of Recognition - R

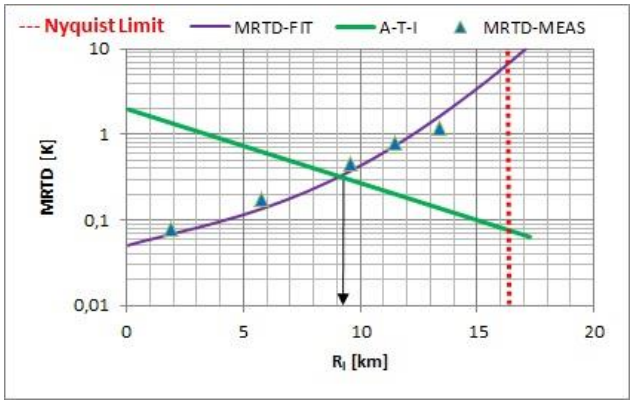


Fig. 8. MRTD Chart with target apparent temperature used for range prediction in case of Identification - I

VII. CONCLUSION

MRTD is complex parameter that combines thermal imager response, image transformation and sensitivity limit, providing possibility for imager range prediction. Because of that it is one of the most important thermal imager parameters. MRTD is imagers figure of merit that measure end-to-end sensor performance including the human observer.

Subjective measurement methods use predefined test patterns and human observer is involved in the measurement. Subjective MRTD measurements are time consuming and costly but provides reliable and reasonably repeatable measurement results

MRTD laboratory test is used for several purposes: (i) to verify if a sensor is working properly; (ii) to compare competing sensor systems; (iii) to verify if a new type of sensor meets the expectations based on its design – compare measurement results with model predictions, and (iv) to predict field performance (target acquisition task performance as a function of distance).

Objective MRTD measurements methods are still being developed and main issue is to criteria definition for human observer perception of the displayed information.

TOD methodology main advantage is that it is applicable without limitation to all the sensor types.

The standard MRTD methodology is successfully realized and applied in Vlatacom EO laboratory for quality assessment of the thermal imager channel in our long range surveillance systems. Our next goal is to extend standard MRTD methodology for evaluation of the SWIR and visible imaging channels.

REFERENCES

[1] Bahram Javidi (editor), *Optical Imaging Sensors and Systems for Homeland Security Applications*, Springer Science + Business Media, Inc., New York, 2006

[2] Carlo S. Regazzoni, Gianni Fabri, Gianni Vernazza (Editors): *Advanced video-based surveillance Systems*, (Kluwer Academic Publishers), Springer Science + Business Media, New York, 1999

[3] B. Livada, D. Perić:, "Imaging Detector Technology: A short insight in history and future possibilities", *7th International Scientific Conference OTEH 2016*, Belgrade, 06-07, October 2016

[4] Branko Livada, "Minimal resolvable temperature (MRT) the basic parameter of thermal imaging equipment", (in Serbian) XLIII Jugoslovenska Konferencija ETRAN 99, Zlatibor, 20-22 septembra 1999

[5] M. Menat, Derivation of various Temperature detection limits for thermal imagers under field conditions, , *Infrared Phys.* Vol. **22,** pp. 175 to 179. 1982

[6] Gaillan H. Abdullah, Ruaa Adel Abas, Ali Hassan Jabur, Study relationship between (MRTD) for different targets with contrast in image, *International Refereed Journal of Engineering and Science (IRJES*), Vol.**2**, Issue 10, pp. 21-25, 2013

[7] J. G. Vortman, A. Bar-Lev, Improved minimum resolvable temperature difference model for infrared imaging system,*Optical Engineering*, Vol. **26** No. 6, pp 492-498, 1987

[8] Krzysztof Chrzanowski, A minimum resolvable temperature difference model for simplified analysis, *Infrared Physics* Vol. **31**. No. 4, pp. 313-318, 1991

[9] Piet Bijl, J. Mathieu Valeton and Maarten A. Hogervorst, A critical evaluation of test patterns for EO system performance characterization, *Proc. SPIE*, Vol. **4372,-** Infrared Imaging Systems: Design, Analysis, Modeling, and Testing XII; 2001

[10] Johnson J., Analysis of image forming systems, *Image Intensifier Symposium*, Fort Belvior Oct. 6-7, USA, pp. 249-273, 1958

[11] Gerald Holst, *Electro-Optical System Performances*, SPIE--The International Society for Optical Engineering, Belingham, USA, 2000

[12] Wolfe W.L., Zissis G.J (ed.), *The Infrared Handbook*, Office of Naval Reseach, Department of Navy, Washington, 1978

[13] Krzysztof Chrzanowski: *Testing Thermal Imagers - Practical guidebook*, Military University of Technology, Warsaw, Poland, 2010

[14] Krzysztof Chrzanowski, Xianmin Li, Configuration of systems for testing thermal imagers, *OpticaApplicata*, Vol. **XL**, No. 3, pp. 727-736, 2010

[15] STANAG 4347 - Definition of static range performance for thermal imaging systems, NATO, 1995

[16] STANAG 4349, Adition 1 - Measurements of the Minimum Resolvable Temperature Difference (MRTD) of thermal cameras, NATO, 1995

[17] Experimental Assessment Parameters and Procedures for Characterization of Advanced Thermal Imagers, NATO RTO TECHNICAL REPORT 75(II), 2003

[18] ASTM E1213-97, Standard Test Method for Minimum Resolvable Temperature Difference for Thermal Imaging Systems, 1997.

[19] J. Barela, M. Kastek K. Firmanty and P. Trzaskawka, Accuracy of measurements of basic parameters of observation thermal cameras, *13th International Conference on Quantitative Infrared Thermography 2016, July 4-8, QIRT 2016*. Gdańsk, Poland, pp.87-94, 2016

[20] J. Bareła, M. Kastek, K. Firmanty, P. Trzaskawka, R. Dulski, J. Kucharz, Determination of range parameters of observation devices, *Proc. of SPIE* Vol. **8541**, 2012

[21] Ratches J A., Lawson R.W., Obert P,L., Bergemann R.J., Cassidy T.W., Swenson M.J., *Night vision laboratory static performance for thermal viewing systems*, NTIS *AD-A011 212*, (U.S. Army Electronics Command, Night Vision Laboratory), 1975

[22] Thomas L. Williams, Update on objective MRTD measurement, *Proc. SPIE 1309*, Infrared Imaging Systems: Design, Analysis, Modeling, and Testing, (1 October 1990);

[23] Piet Bijl. Alexander Toet, J. Mathieu Valeton, "Electro-Optical Imaging System Performance Measurement", in *Encyclopedia of Optical Engineering*, pp. 443-452, Marcel Dekker Inc., New York, 2003

[24] Glenn D. Boreman "Imaging IR sensors: future directions for test and evaluation", Proc. SPIE 10269, Optical Technologies for Aerospace Sensing: A Critical Review, 16 November 1992

[25] Wolfgang Wittenstein: "Minimum temperature difference perceived— a new approach to assess undersampled thermal imagers", *Optical Engineering* **38**(5) pp.773–781 (May 1999)

[26] Piet Bijl, J. Mathieu Valeton, and Maarten A. Hogervorst "Critical evaluation of test patterns for EO system performance characterization", *Proc. SPIE* 4372, Infrared Imaging Systems: Design, Analysis, Modeling, and Testing XII, (10 September 2001

[27] Stephen W. McHugh, Alan Irwin, J. M. Valeton, P. Bijl "TOD Test Method for Characterizing Electro-Optical System Performance", *Proceedings of SPIE* **4373**, September 2001

[28] Joseph Kostrzewa, John Long, John H. Graff, and John David Vincent "TOD versus MRT when evaluating thermal imagers that exhibit dynamic performance", *Proc. SPIE* **5076**, Infrared Imaging Systems: Design, Analysis, Modeling, and Testing XIV, August 2003

FPGA Implementation of Selective Pseudo Coloring of Thermal Image

Petar Marin, Igor Beracka, Nikola Latinović, *Member, IEEE*, Miloš Pavlović, *Member, IEEE*, Miroslav Perić, *Member, IEEE*

***Abstract*—In this paper we describe Field Programmable Gate Array (FPGA) implementation of a new method for selective pseudo coloring of thermal image. Pseudo coloring is a digital image processing method of coloring gray scale images, imparting a visual appeal and highlighting specific features in the image. Our selective pseudo coloring method colors only features with thermal emissivity in certain range that is controlled within thresholds. Since pseudo coloring is memory intensive task for many embedded video processing platforms our method of pure FPGA implementation has the advantage in execution speed and reduction of processing delay. Experimental results on selected scene have shown improvement in operator ability to spot objects of interest on image when using our pseudo coloring method.**

***Index Terms*— FPGA, pseudo coloring, video, thermal imaging, digital image processing, embedded**

I. Introduction

Thermal imaging is a technique of acquisition and image processing of thermal radiation emitted by hot objects [1]. Any objects at a temperature higher than absolute zero (i.e. T>0K) emits electromagnetic radiation [2]. The Stefan–Boltzmann law describes the power radiated from a black body in terms of its temperature. Human body temperature has peak of thermal radiation at infrared part of electromagnetic spectrum. The human eye cannot see this type of radiation therefore infrared measuring devices are required to acquire and process this information [3].

There are several types of detectors of medium wave infrared radiation. The most basic one is pyrometer which has single detector capable of detecting temperature of a single spot. Most advanced devices include an array of sensors to produce a detailed infrared image of the scene.

The main advantage of thermal over optical imaging is that illumination of objects is not needed because of emissive nature of hot objects in infrared part of electromagnetic spectrum.

Pseudo coloring, also known as false coloring, is a method of artificially coloring an image in effort to reveal otherwise hidden details and textures from original grayscale image. Various pseudo coloring methods and color palettes are developed [4].

II. System Description

System consist of thermal imaging camera, our frame grabber implemented on Xilinx Kintex 7™ [5] series FPGA and NVIDIA Jetson TX2™ [6] ARMv8™ based video processing engine with NVIDIA Compute Unified Device Architecture (CUDA™) cores for graphics acceleration, all of which is integrated on Vlatacom Video Signal Processing (vVSP™) platform [7].

Output video format from thermal camera is VGA (640x480) pixels at 60 frames per second. FPGA handles Camera Link™ to HDMI™ [8] interface conversion and real time pseudo coloring of video. One of the two HDMI™ interfaces is connected to NVIDIA Jetson TX2™ and the second one is used for displaying uncompressed video on monitor.

Video is compressed on NVIDIA Jetson TX2 ™ embedded video processing system.

PC grabs compressed network stream via Real Time Streaming Protocol (RTSP) [9] and displays compressed video footage on monitor.

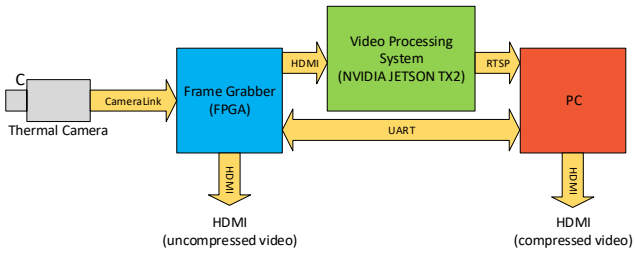


Fig. 1. Functional block diagram of system used for real time selective pseudo coloring of video frame from thermal imaging camera.

III. Selective Pseudo Coloring With Thresholds

Selective pseudo coloring is our approach to emphasize objects with thermal emissivity by choosing two thresholds. Process of forming Look-up Table (LUT) is shown in figure 2. There are two thresholds that user can choose manually. Grayscale LUT is described with equation (1). *R(i),G(i), B(i)* are values of red, green and blue color channel respectively. Parameter "*i*" is intensity of pixel from thermal camera.

$$R_a(i) = G_a(i) = B_a(i) = i,\ i \in [0{,}255] \in N_0 \quad (1)$$

All described LUT palettes from figure 2. are derived from "Image J" [10], an open source project for scientific image analysis. It is possible to use all of defined pseudo

Petar Marin is with the Institute Vlatacom, 5 Bulevar Milutina Milankovića, 11020 Belgrade, Serbia (e-mail: petar.marin@ vlatacom.com).

Igor Beracka is with the Institute Vlatacom, 5 Bulevar Milutina Milankovića, 11020 Belgrade, Serbia (e-mail: igor.beracka@ vlatacom.com).

Nikola Latinović is with the Institute Vlatacom, 5 Bulevar Milutina Milankovića, 11020 Belgrade, Serbia (e-mail: nikola.latinovic@ vlatacom.com).

Miloš Pavlović is with the Institute Vlatacom, 5 Bulevar Milutina Milankovića, 11020 Belgrade, Serbia (e-mail: milos.pavlovic@ vlatacom.com).

Dr. Miroslav Perić is with the Institute Vlatacom, 5 Bulevar Milutina Milankovića, 11020 Belgrade, Serbia (e-mail: miroslav.peric@ vlatacom.com).

coloring palettes for described selective pseudo coloring method.

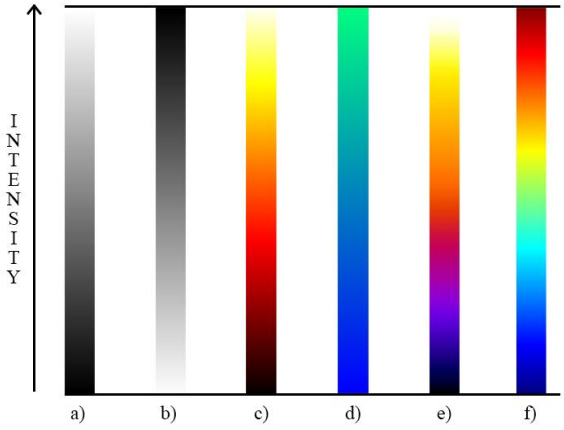


Fig. 2. Graphical representation of all available pseudo color palettes for selective pseudo coloring. a) direct LUT palette, b) inverse LUT palette, c) hot LUT palette, d) cold LUT palette, e) inferno LUT palette, f) jet LUT palette

Formally selective pseudo coloring with threshold and non-expanded LUT is described with equations (2), (3) and (4). User chooses functions $R_b(i),\ G_b(i),\ B(i)$ by choosing pseudo color palettes from Graphic User Interface (GUI) on Personal Computer (PC). Those palettes are shown on figure 2.

$$R_c(i) = \begin{cases} i,\ i < th1,\ \ i \in [0,255] \\ R_b(i),\ i \geq th1, i \leq th2\ ,\ \ i \in [0,255] \\ i,\ i > th2,\ \ i \in [0,255] \end{cases} \quad (2)$$

$$G_c(i) = \begin{cases} i,\ i < th1,\ \ i \in [0,255] \\ G_b(i),\ i \geq th1, i \leq th2,\ i \in [0,255] \\ i,\ i > th2,\ \ i \in [0,255] \end{cases} \quad (3)$$

$$B_c(i) = \begin{cases} i,\ i < th1,\ \ i \in [0,255] \\ B_b(i),\ i \geq th1, i \leq th2\ ,\ \ i \in [0,255] \\ i,\ i > th2,\ \ i \in [0,255] \end{cases} \quad (4)$$

Formally selective pseudo coloring with threshold and expanded LUT is described with equations (5), (6), (7) and (8).

$$th_2 - th_1 \neq 0 \quad (5)$$

$$R_d(i) = \begin{cases} i,\ i < th1, i \in [0,255] \\ R_b\left(\left[\frac{255}{th_2 - th_1}\right]\right),\ i \geq th1, i \leq th2\ , i \in [0,255] \\ i,\ i > th2, i \in [0,255] \end{cases} \quad (6)$$

$$G_d(i) = \begin{cases} i,\ i < th1, i \in [0,255] \\ G_b\left(\left[\frac{255}{th_2 - th_1}\right]\right),\ i \geq th1, i \leq th2, i \in [0,255] \\ i,\ i > th2, i \in [0,255] \end{cases} \quad (7)$$

$$B_d(i) = \begin{cases} i,\ i < th1, i \in [0,255] \\ B_b\left(\left[\frac{255}{th_2 - th_1}\right]\right),\ i \geq th1, i \leq th2, i \in [0,255] \\ i,\ i > th2, i \in [0,255] \end{cases} \quad (8)$$

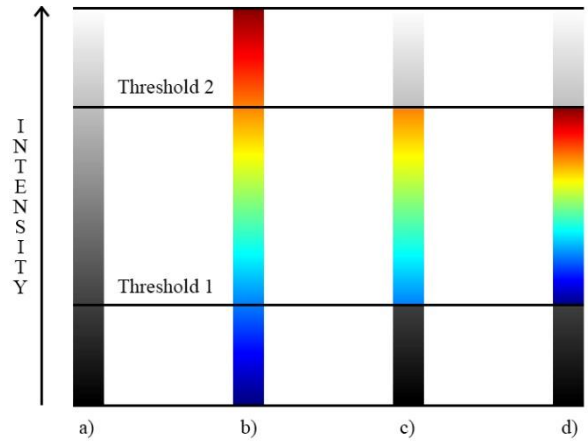


Fig. 3. Visual representation of generating selective pseudo coloring LUT palette. a) Grayscale LUT palette, b) "Jet" pseudo color LUT palette, c) Selective pseudo coloring with two thresholds and non-expanded LUT palette, d) Selective pseudo coloring with two thresholds and expanded LUT palette.

IV. Practical Implementation

FPGA frame grabber is used for grabbing frames from Camera Link™ interface and processing of captured frame. Frame grabber consist of video preprocessor module, video buffer, pseudo coloring module, HDMI™ driver module and MicroBlaze™ microcontroller implemented on FPGA. Video preprocessor module consists of multiplexers for choosing bit-width of input stream from Camera Link™ interface and RGB to Y converter. Possible configurations are: 8, 10, 12, 14 and 16 bits.

Purpose of video buffer is to synchronize input and output frame rates from different interfaces i.e. clock domains. It can store one whole frame.

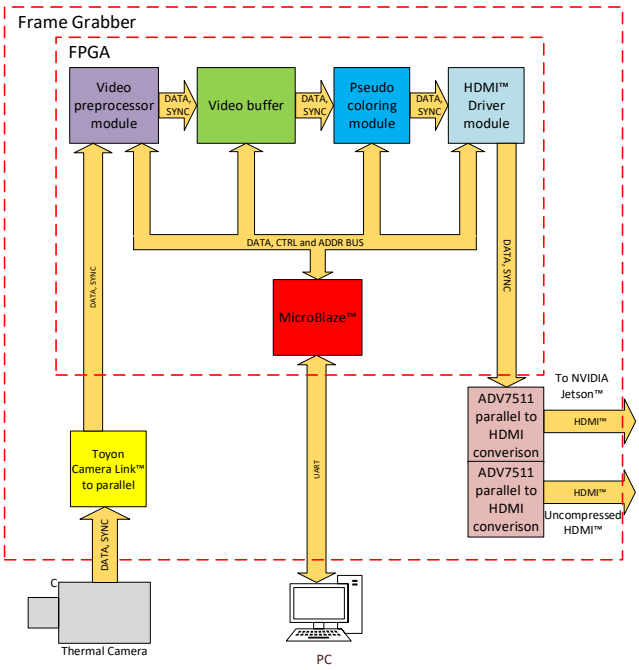


Fig. 4. Simplified block diagram of FPGA frame grabber

Pseudo coloring module is part of a frame grabber implemented on FPGA platform. This is most relevant module in this paper so it will be thoroughly described. Grayscale data is 8 bit wide signal that carries information from pixel that currently draws on HDMI™ interface. It is registered to reduce fan-out of logic that drives this signal because it's used for several other modules. Registered grayscale signal goes to read address input of Dual-ported RAM (DPRAM) and output data is used as output from LUT. Write port of DPRAM is used for loading LUT from MicroBlaze™ microcontroller. It's not possible to read content of this LUT from MicroBlaze™. In this configuration DPRAM act as a configurable LUT. Multiplexer after DPRAM is used for bypassing LUT when DPRAM is not initialized e.g. after system reset. There are three DPRAMs for each color channel. DPRAM has depth of 256x8bit words. Output from multiplexers are registered for reducing fan-out of multiplexer logic.

Output delay from this module is described with equation (12).

$$\tau_{delay} = \tau_{reg} + \tau_{DPRAM} + \tau_{reg} \cong 3T_{CLK} \qquad (12)$$

Frame valid, data valid and line valid signals are synchronization signals needed for generating HDMI™ output signals. Synchronization signals must be delayed for τ_{delay} to be in synchronization with output data from pseudo coloring module.

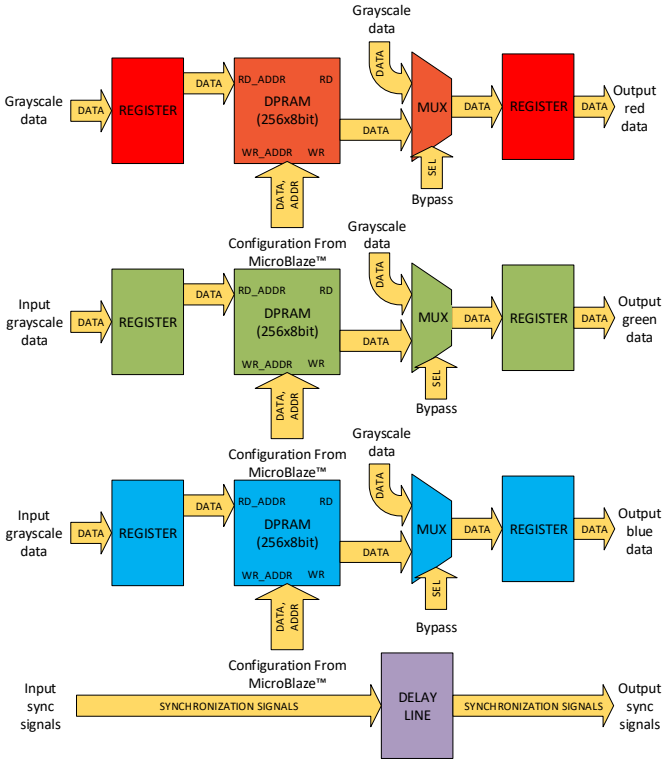


Fig. 5. Block diagram of implemented pseudo coloring module on FPGA platform.

HDMI™ driver module generates address signals for reading from video buffer and synchronization signals for driving external parallel to HDMI™ conversion Integrated Circuit (IC). HDMI™ driver module can be configured with MicroBlaze™ to change resolution and framerate of HDMI™ interface.

MicroBlaze™ is a Soft Intelectual Property (IP) core microcontroller used to control mentioned modules. It uses 64 KB of internal block RAM implemented in logic fabric of FPGA for storing UART terminal application which receives commands from MATLAB™ GUI application. LUTs are created in GUI on PC and loaded in pseudo coloring module via UART interface on MicroBlaze™.

V. Experimental Results

Experimental system consist of thermal imaging camera RP Optics C330™ [11] with electronically controlled optics from PC, Vlatacom vVSP module and two UART connections on PC. One is used for configuring camera and the other one is for communication with MicroBlaze™ microcontroller on Vlatacom vVSP module. Ethernet connection is used for watching video stream via RTSP on local PC.

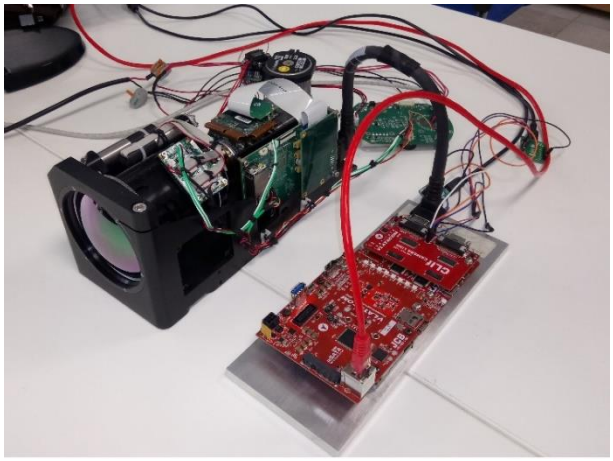


Fig. 6. Setup used for capturing images for examples. RP Optics C330 thermal imaging camera is on the left side of the picture and Vlatacom VSP module is on the right side of picture (red PCB).

On figure 7. scene is observed in visible part of electromagnetic spectrum. There are several objects of interest of this picture like hot cup of water, bottle of cold water and a subject holding a hot soldering iron.

Fig.7. Observed scene captured in visible part of electromagnetic spectrum

Frame in figure 8. is captured with thermal imaging camera. In this picture it is clearly visible that cup of water and soldering iron are indeed hot. Reflection of soldering iron on a glass wall is clearly visible in this picture.

Fig. 8. Grayscale frame with resolution 640x480 pixels from thermal camera.

After the pseudo coloring is applied for a specified thresholds subject, hot cup of water, and a soldering iron are emphasized in figure 9.

Fig. 9. Selective pseudo coloring of frame from thermal camera with th1=0 and th2=87 without LUT expansion. LUT palette is on the right side of the picture.

Contrast of objects of interest are further increased when LUT expansion is applied in figure 10.

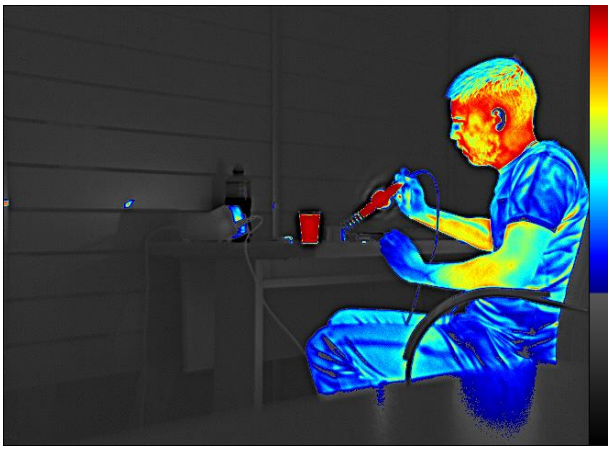


Fig. 10. Selective pseudo coloring of frame from thermal camera with th1=0 and th2=87 with LUT expanded. LUT color palette is on the right.

It is also possible to isolate cold background for empirically derived threshold levels as it was shown on figure 11.

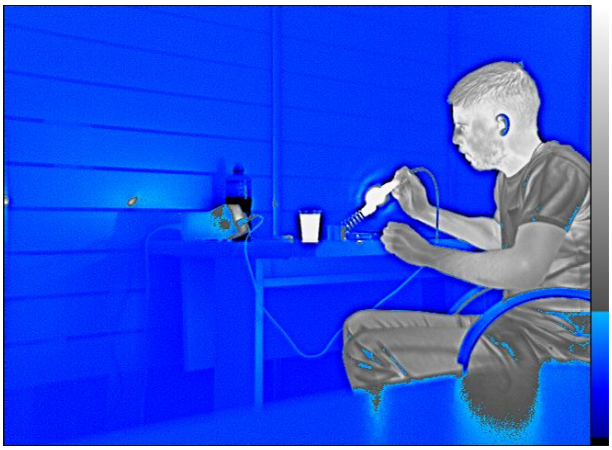


Fig. 11. Selective pseudo coloring of frame for parameters th1=5 and th2=79. LUT color palette is on the right side of the picture.

VI. Conclusion

Implementation has proven that acceleration of pseudo coloring process is indeed possible with the use of FPGA. Selective pseudo coloring gave good results as visibility of object of interest in image is greatly increased. Visibility of halo effect from sharpening filter of camera internal video processing engine is also increased.

Increasing captured frame dynamic range and bit depth of pseudo coloring module is expected to give even better results.

References

[1] X. P. Maldague, “Introduction to Thermal Emission” in “Theory and Practice of Infrared Technology for Nondestructive Testing”, New York, USA, Wiley, 2001, ch. *4*, sec. *1*

[2] S. Blundell, K. Blundell, “Rods, bubbles and magnets” in “Concepts in Thermal Physics”, Oxford, UK, OUP, 2006, ch. *17*, sec. *3*, *190*.

[3] M. Vollmer, K.P. Mollmann, “Fundamentals of Infrared Thermal Imaging” in “Infrared Thermal Imaging”, New Jersey, JW.A, USA, 2010, ch. *1*, sec. *1*, *1*.

[4] P. A. Jacobs, “A Review on Recent Pseudo-Coloring Techniques”, IJSTE, vol. *1*, no. *11*, *344*. 2349-784X, may, 2015.

[5] “https://www.xilinx.com/support/documentation/data_sheets/ds180_7Series_Overview.pdf“, “7 Series FPGAs Data Sheet: Overview”, Xilinx, 2018 [Online].

[6] “https://directory.ifsecglobal.com/40/product/95/41/68/327494_DS_Jetson_TX2_Module_A4_fnl_WEB.pdf”, “Jetson TX2 supercomputer on module for AI on the edge”, NVIDIA Corporation, 2017 [Online].

[7] M. Trifunović, I. Popadić, V. Lukić, M. Perić, “Unified interfacing solution in video processing platforms based on FPGA”, IcETRAN, no 4, pp. EKI3.4.1-6, june 2017.

[8] “https://www.hdmi.org/manufacturer/specification.aspx”, “HDMI Specification Version 1.4a”, HDMI Forum, 2010 [Online].

[9] “https://www.ietf.org/rfc/rfc2326.txt”, “Real Time Streaming Protocol (RTSP)”, IETF, 1998 [Online].

[10] “https://imagej.net/Welcome”, “An open platform for scientific image analysis ImageJ”, [Online].

[11] “http://rp-optical-lab.com/wp-content/uploads/2016/04/MWIR-LRC3Z-r9-1.pdf”, “RP Optics C330 thermal imaging camera system specification”, RP Optical Lab, [Online].

Real-Time Dead Pixels Removal in Thermal Imaging

Miloš Pavlović, *Member*, *IEEE*, Nataša Vlahović, *Member*, *IEEE*, Miroslav Perić, *Member*, *IEEE*, Aleksandar Simić and Srđan Stanković

***Abstract*— Video enhancement techniques are very important part of long-range multi-sensor surveillance systems. Infrared sensors are typically affected by nonresponsive pixels, or "dead pixels." These dead pixels can severely degrade the quality of images and often have to be replaced before subsequent image processing and display of the imagery data. Conventional dead pixel replacement methods (methods for 'salt' noise reduction) take a lot of processor time and cannot be employed in real-time applications. This paper proposes a real-time infrared imaging dead pixels removal algorithm, applicable to SWIR sensor. Dead pixels reduction is based on *Inverse Distance Weighting* algorithm applied only on positions of dead pixels. Also, paper provides comparison of *Inverse Distance Weighting* algorithm with the most commonly used techniques in literature - median filtering. All the techniques are tested on experimental data.**

***Index Terms*— Dead pixels, Real - Time, IDW, Median filtering, Thermal imaging.**

I. Introduction

Infrared (IR) thermal cameras have been used in military and civil applications for many years. An IR thermal camera provides a picture of the electromagnetic energy radiated from an object in the IR spectral band. The IR spectrum consists of the reflected band which is dependent upon material reflectivity and thermal IR band, which is dependent upon the temperature and emissivity of the material. A number of sensor technologies have been designed and optimized for imaging in the IR spectral band, each posing unique design challenges. In various atmospheric conditions that make visible light sensors partially or completely "blind", infrared sensors provide ability of effective view.

The IR band is further divided into four bandwidths: Near IR (NIR) - 0.74-1.4μm wavelength range, Short-Wave IR (SWIR - 1.4-3μm), Medium-Wave IR (MWIR - 3-8μm) and Long-Wave infrared (LWIR - 8-14μm) [1]. The main advantage of thermal sensors is that they are nearly invariant to changes in illumination conditions and provide visual information in low visibility environments and conditions. Emitted thermal energy is less affected by scattering and the absorption by smoke, fog or dust [2].

Despite these advantages and robustness to illumination variations, infrared sensors also suffer from disturbances - most commonly from blur and noise originating from sensors and their environment, sensor dead pixels, … [1], [3].

All thermal sensors are affected by the common problem of unresponsive pixels - termed dead pixels [4], [5]. Dead pixels affect the quality of both, the visual image and the elementary data. True scene that is being measured and dead pixels measurement do not have any correlation, and in that way, dead pixels intensely degrade the quality of measured images. Dead pixels measurements need to be replaced with more appropriate values to improve image quality to an acceptable level.

In situations when servicing and calibration of thermal cameras is not possible or takes too much time, image processing algorithm for the repair of dead pixels must be implemented.

Problem of dead pixels in image is a type of 'salt' noise problem. Standard approaches to salt noise removal from literature are mostly based on nearest-neighbor or median filtering [6], [7], [8]. In applications for surveillance, monitoring, especially for ultra-long-range target detection, recognition and identification in full zoom, applying filtering on whole image reduces the noise from the image very successfully, but with the cost of blurring edges. This effect greatly degrades the image quality. Also, applying filtering on the whole image with other video processing algorithms such as digital video stabilization, enhancement and tracking is very time-consuming, especially for real-time applications, such as 25-30 fps and high image resolution. So first, it is necessary to find the positions of dead pixels and then replace their values with appropriate values. It is important to note that determining whether the pixel is dead or not is a significant topic in itself.

This paper proposes a real-time infrared image enhancement pipeline, based on dead pixels reduction (edge preserving smoothing), independent of type of scene content. Dead pixels position detection is based on calculating average pixels value using set of frames while dead pixels reduction is based on Inverse Distance Weighting (IDW) [9] algorithm applied only on positions of dead pixels. Also, this paper

Miloš Pavlović is with Vlatacom Institute of High Technologies, Blvd. Milutina Milankovića 5, 11070 Belgrade, Serbia (e-mail: milos.pavlovic@vlatacom.com).

Nataša Vlahović is with Vlatacom Institute of High Technologies, Blvd. Milutina Milankovića 5, 11070 Belgrade, Serbia (e-mail: natasa.vlahovic@vlatacom.com).

Dr Miroslav Perić is with Vlatacom Institute of High Technologies, Blvd. Milutina Milankovića 5, 11070 Belgrade, Serbia (e-mail: miroslav.peric@vlatacom.com).

Aleksandar Simić is with Vlatacom Institute of High Technologies, Blvd. Milutina Milankovića 5, 11070 Belgrade, Serbia (e-mail: aleksandar@vlatacom.com).

Prof. dr Srđan Stanković, is with School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia and Vlatacom Institute of High Technologies, Blvd Milutina Milankovića 5, 11070 Belgrade, Serbia; (e-mail: srdjan.stankovic@vlatacom.com).

provides comparison of processing time per frame for median filtering applied on positions of dead pixels and on the whole image (as the most commonly used technique in literature) and IDW algorithm.

The paper is organized as follows. Section II describes the dead pixels detection algorithm. Section III proposes algorithms for dead pixels replacement – Median filtering and IDW. Section IV describes the experimental work including the experimental setup, results and discussion. Section V lists conclusions and indicates directions for future work in this research area.

II. Dead Pixels Detection

An attempt to find dead pixels as pixels that do not change their value through frames is not possible. Although dead pixels shown in image are actually sensor's dead pixels, the dead pixel intensity on raw images is not constant on all video frames. Dead pixels change values from frame to frame, depending on the scene, as a result of histogram equalization that is performed on the sensor. Also, dead pixels do not have either maximum or minimum intensity value, so their intensities only cannot be indicator of dead pixels positions.

An image contains a number of distinct gray levels of the pixels with difference compared to their neighborhood pixels. If the gray levels of the pixels in an image and their neighbors are mapped in such a way that the difference in the gray levels of the neighbor with the pixel is noticeable, then dead pixels can be detected and mapped by estimating average pixel value for each pixel using set of frames contaminated with dead pixels – Eq. (1).

$$I(i,j) = \frac{1}{N_{frames}} \sum_{k=1}^{N_{frames}} F_k(i,j) \qquad (1)$$

Fig. 1 presents average pixel values obtained using Eq. (1).

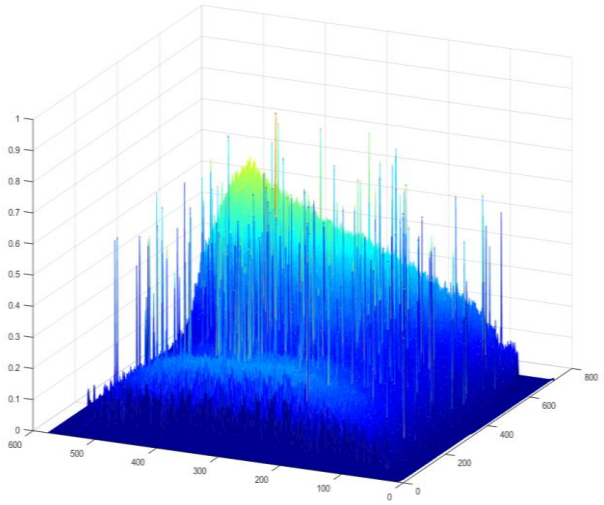
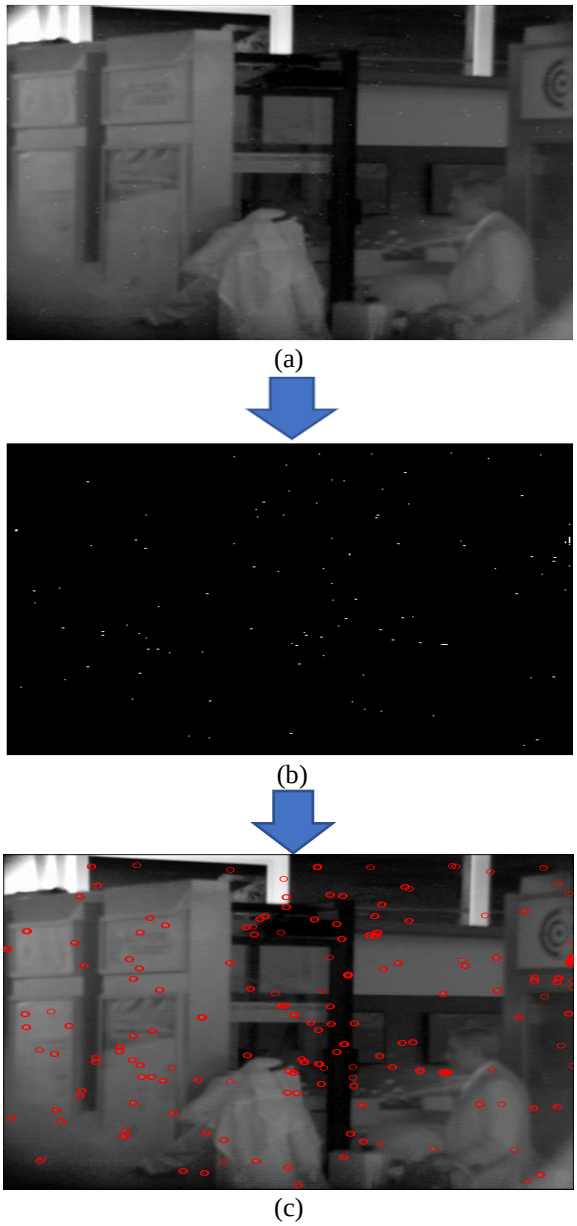


Fig. 1 Average pixel values

Existing image pixel values are mapped to new values. By mapping, pixel values in the image are more or less equal or rather uniform, except for the dead pixels. Mapping of the pixels by estimating the average value for each pixel results in similar pixel values within the correct image pixels and augments the obviousness of dead pixels. Dead pixels are detected as those whose average value is higher than defined threshold. For better detection, image is divided in cells, and estimating average value for pixels in different cells is calculated using different sets of frames and thresholds.

Fig. 2 presents an image contaminated with dead pixels and map of detected dead pixels.



(a)

(b)

(c)

Fig. 2 Image contaminated with dead pixels (a), map of detected dead pixels(b), original image with detected dead pixels(c)

In Fig. 2, there are a significant number of dead pixels that are observed. Also, it can be observed that in the SWIR camera, the visibility of dead pixels depends on the scene

structure, object temperature, as well as the object surface properties. For the purposes of dead pixel replacement, a binary image of the same dimensions as the original image is obtained, such that a dead pixel is indicated with 1 and a properly functioning pixel is indicated with 0 – Fig. 2 (b). This binary image is called the dead pixel map and it is assumed that this map already exists in replacement strategies discussions in Section III.

III. Proposed Algorithms

Image filtering is one of the most fundamental techniques for modifying or enhancing an image. Image is filtered to emphasize certain features or remove other features. Image processing operations implemented with filtering include smoothing, sharpening, and edge enhancement. Filtering is a neighborhood operation - the value of any given pixel in the filtered image is determined by applying some algorithm to the values of the pixels in the neighborhood of the corresponding input pixel. A pixel's neighborhood is set of pixels, defined by their locations relative to that pixel.

A. Median Filtering

Median filtering [7], [10] is a non-linear method used to remove noise from images. It is widely used in digital image processing as it is very effective at removing noise while preserving edges. Particularly, it is effective at removing 'salt and pepper' type noise.

The main idea of the median filtering is moving through the image pixel by pixel, replacing each value with the median value of neighboring pixels. The pattern of neighbors is called the "window", which slides, pixel by pixel over the entire image. First, all the pixel values from the window are sorted into numerical order. The window usually has an odd number of pixels, so value of the central window pixel is simply replaced with the middle (median) pixel value. In processing the boundary values, there are other approaches that have different properties (shown in Fig. 3):

- Avoid processing the boundaries (with or without cropping the image boundary afterwards).
- It is possible to select entries from a far horizontal or vertical boundary.
- Changing the window size near the boundaries, so that every window is full.

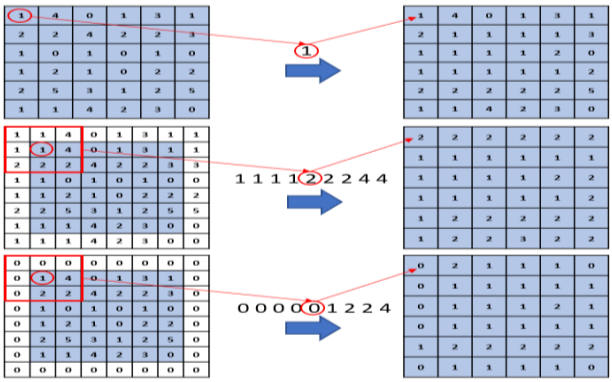


Fig. 3 Processing the boundary values in median filtering

B. Inverse Distance Weighting (IDW)

Inverse Distance Weighting - IDW algorithm is one of the simplest interpolation algorithms. IDW algorithm is widely used in many applications such as [9], [11], [12] and is operated as the spatial interpolation method. In this paper IDW is applied to remove dead pixels.

The interpolation of new value for each dead pixel can be calculated using the uncorrupted surrounding pixels. For the interpolation pixels, a window of the eight nearest pixels surrounding the current dead pixel is used. The new value of the dead pixel is obtained as:

$$Y_{DP} = \frac{\sum_{i=1}^{8} DPM_i \cdot \frac{1}{d_i} \cdot Y_i}{\sum_{i=1}^{8} DPM_i \cdot \frac{1}{d_i}} \qquad (2)$$

In Eq. (2):
Y_{DP} - pixel value at the DP position
Y_i - neighbor pixel value at the position i
DPM_i - dead pixel indicator used for interpolation (1 – correct, 0 – dead pixel)
d_i- Euclidian distance from the DP to the interpolation pixel

In case that none of the given eight pixels is correct, the wider field is searched, but only by width and height until the first correct pixel is reached or it does not reach the image boundary. In that case, the new value of the dead pixel is obtained as:

$$Y_{DP} = \frac{\sum_{i=1}^{4} DPM_i \cdot \frac{1}{d_i} \cdot Y_i}{\sum_{i=1}^{4} DPM_i \cdot \frac{1}{d_i}} \qquad (3)$$

If there is no correct pixel, the Y_{DP} value is 0.5.

IV. Experimental Work and Results

The system setup used in this paper is based on vMSIS (Vlatacom Multi-Sensor Imaging System) [13]. Integrating visible and infrared imaging sensors and providing ultra-long-range target detection, recognition and identification, vMSIS is a state-of-the-art monitoring and surveillance system.

The algorithm pipeline is implemented for real-time performance for SWIR sensor, with 500mm continous zoom lens, image resolution 576×720 pixels, and 25 fps frame rate, using specialized GPU based hardware, C++ and OpenCV and CUDA libraries.

Detection of dead pixels is implemented offline, and based on map of dead pixels, further dead pixels replacement is implemented in real time.

A. Image Database

For the purpose of the proposed algorithm pipeline evaluation of three SWIR video footages containing 750 frames each are used. The frame's resolution is 576×720 pixels. All video footages are obtained with Vlatacom Multi-Sensor Imaging System.

B. Results and Discussion

Median filtering applied on whole image is a method independent on positions of dead pixels. However, some image texture could be lost, which results in small objects disappearing from an image, which is not allowed, especially in applications for ultra-long range monitoring in full zoom. Namely, image pixels which define small objects in the image can be replaced by the median value of surrounding pixel values during the filtering process. Because of that, filtering must be applied only on pixels detected as dead pixels. Another limiting factor is the processing time, since dead pixels removal is a preprocesing method for other video processing algorithms, such as digital video stabilization or object tracking that are very time-consuming. In order to reach frequency of 25 fps, processing time needed for dead pixels removal has to be as short as possible. Table I presents processing time per frame measured for three methods: median filtering (full image), median filtering (based on dead pixels position) and IDW (based on dead pixels position). Based on these results, IDW algorithm applied on dead pixels only is the best solution, in terms of processing time per frame. Average filtering technique applied to dead pixels' position gives same effect as IDW algorithm, but only in situation when none of the local neighborhood of 3x3 pixels is correct. For calculating the new value of the dead pixel using the uncorrupted pixels from longer distance form observed dead pixel, it is important to take into account the distance of those pixel pixels from the dead pixel.

TABLE I
PERFORMANCE MEASURE – TIME

Algorithm	Time [s]
Median (full image)	0.003446
Median	0.001807
IDW	0.000913

Visual results for subjective quality assessment (original frames contaminated with dead pixels and corresponding filtered frames) for SWIR video sequences and IDW algorithm are presented in Fig. 4. The conclusion from the observer stand point is that image clarity is better, and that the quality is enhanced, without deteriorating image quality on other pixel's positions.

V. CONCLUSION

The real-time enhancement algorithm pipeline for infrared imaging in multi-sensor imaging systems proposed in this paper is designed for solving SWIR camera imaging problem – dead pixels presence. The algorithm is designed for detection and replacement of dead pixels. Also, this method is applicable to LWIR and MWIR sensor types. The proposed method is implemented in a real-time environment.

The given results of SWIR sensor testing indicate that the proposed IDW method gives satisfying performance in terms of processing time per frame, without dropping frame rate. Image quality is subjectively better without replacing values of any other image pixel except dead pixels.

Except of dead pixels problem, thermal images have very different nature of noise. In outdoor applications, IR imaging is sensitive to changes in environment's temperature and especially at long distance, the object temperature can be drastically reduced by atmospheric loss. Thermal image noise analysis and enhancement techniques applied to thermal images are certainly planned to be part of the future research to provide better quality of thermal images, more convenient for information extraction.

REFERENCES

[1] J. Lloyd, Thermal imaging systems, Springer Science & Business Media, 2013.

[2] D. Sadot, G. Kitron, N. Kitron and N. Kopeika, "Thermal imaging through the atmosphere: atmospheric modulation transfer function theory and verification," *Optical Engineering,* vol. 33, no. 3, pp. 881-889, 1994.

[3] N. Kopeika, "A system engineering approach to imaging," *Bellingham: SPIE Optical Engineering Press,* pp. 517-520, 1998.

[4] B. M. Ratliff, J. S. Tyo, J. K. Boger, W. T. Black, D. L. Bowers and M. P. Fetrow, "Dead pixel replacement in LWIR microgrid polarimeters," *Optics express,* vol. 15, no. 12, pp. 7596-7609, 2007.

[5] H. Budzier and G. Gerlach, "Calibration of uncooled thermal infrared cameras," *Journal of Sensors and Sensor Systems 4.1,* pp. 187-197, 2015.

[6] A. D. Restrepo Giron and H. Loaiza Correa, "A new algorithm for detecting and correcting bad pixels in infrared images," *Ingeniería e Investigación,* vol. 30, no. 2, pp. 197-207, 2010.

[7] S. S. Beagum, N. Hundewale and M. M. Sathik, "Improved adaptive median filters using nearest 4-neighbors for restoration of images corrupted with fixed-valued impulse noise," in *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, 2015.

[8] A. E. Mudau, C. J. Willers, D. Griffith and P. F. le Roux, "Non-uniformity correction and bad pixel replacement on LWIR and MWIR images," in *2011 Saudi International Electronics, Communications and Photonics Conference (SIECPC)* , 2011.

[9] G. Y. Lu and D. W. Wong, "An adaptive inverse-distance weighting spatial interpolation technique," *Computers & geosciences,* vol. 34, no. 9, pp. 1044-1055, 2008.

[10] Y. Lee and S. Kassam, "Generalized median filtering and related nonlinear filtering techniques," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 33, no. 3, pp. 672-683, 1985.

[11] P. M. Bartier and C. Keller, "Multivariate interpolation to incorporate thematic surface data using inverse distance weighting (idw)," *Computers,* vol. 22, no. 7, pp. 795-799, 1996.

[12] Z. Li, X. Li, C. Li and Z. Cao, "Improvement on inverse distance weighted interpolation for ore reserve estimation," *Fuzzy Systems and Knowledge Discovery (FSKD),* vol. 4, p. 1703–1706, 2010.

[13] Vlatacom Instituute, "Vlatacom Institute Border Protection – Land," [Online]. Available: https://www.vlatacominstitute.com/border-protection-land.

Fig. 4 SWIR image dead pixels removal results: left - originals, right - corrected images

A Novel Approach for Pan/tilt Drift Detection in Gyro Stabilized Systems Using IMU Sensors

Petar Milanović, Marko Nerandžić, Medhat Abdelrahman Mohamed Mostafa, Ilija Popadić, *Member, IEEE*, Miroslav Perić, *Member, IEEE*

***Abstract*—It is known that drift occurs when the pan/tilt gyro stabilization is turned on. On multi-sensor camera systems, drift is the reason why after a while the observed object is no longer in the centre of the screen and it has a tendency to disappear from the screen. The conclusion is that gyro stabilization is unusable when the field of view is narrow. In this paper, a method for pan/tilt drift detection and its measurement is developed. IMU sensor, which consists of a 3-axis MEMS gyroscope and a 3-axis MEMS accelerometer is used for measuring the system's angular position. Kalman filter is implemented to compensate for the MEMS inertial sensor flaws by combining a gyroscope and accelerometer data.**

***Index Terms* — Pan/tilt, Drift, Gyro stabilization, IMU, Gyroscope, Accelerometer, Kalman filter.**

I. Introduction

Pan/tilt positioner is a servo-driven two-axis device which is designed for long-distance video surveillance systems. Pan/tilt can carry heavy long range electro-optical surveillance sensors that require azimuth and elevation rotation with very high accuracy and angular velocity. The range of movements of pan/tilt that is used is n x 360° per azimuth and ±40° per elevation, and the rotation speed goes from 0.005°/s to 120°/s per azimuth and up to 100° per elevation.

Fig. 1. Pan/tilt with 2 cameras

Petar Milanović is with the Vlatacom Institute of High Technologies, Blvd. Milutina Milankovića 5, 11070 Belgrade, Serbia (e-mail: petar.milanovic@vlatacom.com).

Marko Nerandžić is with the Vlatacom Institute of High Technologies, Blvd. Milutina Milankovića 5, 11070 Belgrade, Serbia (e-mail: marko.nerandzic @vlatacom.com).

Medhat Abdelrahman Mohamed Mostafa is with the European University of Belgrade, Carigradska 28, 11000 Belgrade, Serbia

Dr Ilija Popadic is with the Vlatacom Institute of High Technologies, Blvd. Milutina Milankovića 5, 11070 Belgrade, Serbia (e-mail: ilija.popadic@vlatacom.com).

Dr Miroslav Perić is with the Vlatacom Institute of High Technologies, Blvd. Milutina Milankovića 5, 11070 Belgrade, Serbia (e-mail: miroslav.peric@vlatacom.com).

Stabilization is essential for most long-range multi-sensor camera systems with pan/tilt, because the effect of system shake is more pronounced when using larger focal lengths or narrower field of view – FOV [1]. For that purpose, fibre optic gyroscope or MEMS gyroscope, which measures each movement of the system is engaged. When the gyroscope detects a movement, the command is sent to the pan/tilt engines, so they can resist disturbances. Gyro stabilization performance depends on the gyroscope accuracy, system delays and pan/tilt engine accuracy and speed.

After the stabilization has been turned on, the user gets a stable image, but as a negative effect the system drift occurs. After a while the observed object is no longer in the centre of the screen and it has a tendency to disappear from the screen. The time after which the drift is observed depends on the FOV. If the FOV is narrower, the drift is more pronounced. For narrow FOV, gyro stabilization is unacceptable and after just a few seconds user don't observe the same scene. Because of this, it is very important to detect the drift, to measure its value and to send the command to pan/tilt engine for its compensation.

The pan/tilt drift appears because gyro stabilization does not take into the account the angular position of the system. Fitness function does not take into the account its angular position per azimuth and elevation. In systems that are used, drift is significantly more pronounced per elevation than per azimuth, so in this paper the drift detection will be done only per elevation. This is due to the pan/tilt construction and the position and weight of the cameras placed on it. The cameras can be positioned so the system is not perfectly balanced and the centre of gravity does not match ideally with the pan/tilt centre. Also, pan/tilt manufacturers probably ignore errors that are less than 0.1° per minute, which is very important when observing an image with narrow FOV.

For the purpose of drift detection in this paper, IMU sensor which consists of tri-axis gyroscope and tri-axis accelerometer is used. Both MEMS sensors, the gyroscope and the accelerometer, have flaws [2]. The angle that is calculated by gyroscope measurement drifts. Bias drift estimation is explained in [3]. The accelerometer is too sensitive on disturbances and its signal has a lot of noise. Sensors and algorithm for signal fusion is used in order to achieve reliable estimate about the angular position. A Kalman filter is implemented to yield the reliable information. By monitoring the movement of the system when the gyro stabilization is activated, drift can be detected and measured.

II. IMU SENSORS – FUNCTIONAL DESCRIPTION AND SIGNAL ACQUISITION

The system is composed of a tri-axis MEMS gyroscope and a tri-axis MEMS accelerometer. The sampling rate is 10 samples per second. The gyroscope resolution is 16 bits and the sensitivity is 7.8125 mdps/LSB. The accelerometer resolution is 14 bits and the sensitivity is 0.244mg/LSB.

The gyroscope measures the angular velocity, and the rotation angle can be computed by time-integrating of gyroscope output. Trapezoidal integration method is used.

$$\int_a^b f(x)dx = (b-a)f(a) + \frac{1}{2}(b-a)[f(b) - f(a)] \quad (1)$$

The computed result drifts over time and after approximately 300 seconds it drifts about 5 degrees (Fig. 2). The explanation for this is that integration accumulates the noise over time and turns noise into the drift, which yields unacceptable results. The problem is unpredictable bias due to temperature effect, calibration errors, flicker noise and random walk due to thermo-mechanical white noise [4].

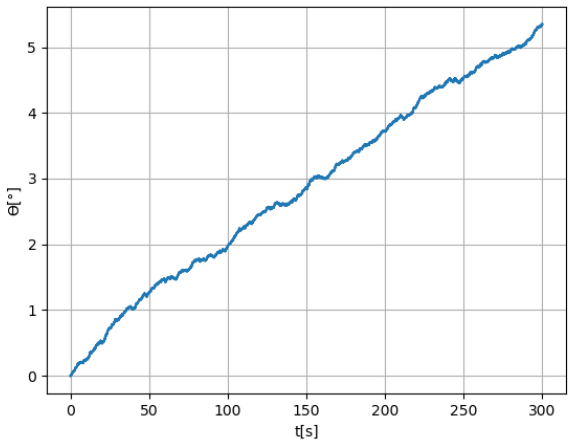


Fig. 2. Gyroscope output, sensor is still

Accelerometers are sensitive to both linear acceleration and local gravity field. In the absence of linear acceleration, which is the case here, the accelerometer output is rotational gravity field vector and it can be used to determine pitch and roll orientations:

$$roll = arctan\left(\frac{G_{py}}{G_{pz}}\right) \quad (2)$$

$$pitch = arctan\left(\frac{-G_{px}}{\sqrt{{G_{py}}^2 + {G_{pz}}^2}}\right) \quad (3)$$

where: G_{px}, G_{py}, G_{pz} are measurement of acceleration along *x*, *y*, *z* axis, respectively [5].

The great advantage of such output from the accelerometer is that integration isn't performed in order to get a position.

III. ALGORITHM FOR SYSTEM POSITION ESTIMATION

In order to get better estimation of system angular position, a combination of measurement from both sensors is used. In this paper Kalman filter is implemented as a sensor fusion algorithm in order to compensate for the weakness of each sensor by utilizing other sensors.

The system should be described in a state space form:

$$x[k+1] = Ax[k] + Bu[k] + w[k], \quad (4)$$

$$y[k] = Cx[k] + v[k], \quad (5)$$

where: *x* is the state vector, *A* is the state transition matrix, *B* is matrix which is called the control-input model, *w*is state transition noise which is Gaussian distributed with a zero mean and with covariance *Q*, *y* is measurement, *C* is the observation matrix and *v* is the measurement noise with a zero mean and *R* as covariance.

The Kalman filter is a recursive state *x(k)* estimator based on the known model, parameters *A*, *B* and *C* and measurement *y(k)*, which is optimal in terms of minimization of variance of the estimation error [6]. The estimation of the state is obtained only on the previously estimated values *x(k-i),i>0* and the last measurement. Kalman filter is theoretically ideal for fusion of the noisy data [7].

In our system, the states are angle Θ and gyro bias $\dot{\Theta}_b$.

$$x_k = \begin{bmatrix} \Theta \\ \dot{\Theta}_b \end{bmatrix}_k \quad (6)$$

Matrices *A*, *B*, *C* are defined as follows:

$$A = \begin{bmatrix} 1 & -\Delta t \\ 0 & 1 \end{bmatrix} \quad (7)$$

$$B = \begin{bmatrix} \Delta t \\ 0 \end{bmatrix} \quad (8)$$

$$C = [1\ 0] \quad (9)$$

Control input *u* is the gyroscope $\dot{\Theta}_k$ measurement in degrees per second. Covariance matrices *Q* and *R* should be calculated before implementing the algorithm.

Algorithm:

1. Set initial values,
   $P_0 = 0$, $\hat{x}_0 = 0$ (10)
2. State prediction
   $\hat{x}_{k|k-1} = A\hat{x}_{k-1|k-1} + B\,u_k$ (11)
3. Calculating error covariance matrix prediction
   $P_{k|k-1} = AP_{k-1|k-1}A^T + Q_k$ (12)
4. Kalman gain computation
   $K_k = P_{k|k-1}C^T(CP_{k|k-1}C^T + R)^{-1}$ (13)
5. State estimation
   $\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k(y_k - C\hat{x}_{k|k-1})$ (14)
6. Error covariance matrix computation
   $P_{k|k} = (I - K_kC)P_{k|k-1}$ (15)
7. Loop to step 2.

IV. IMPLEMENTATION IN THE REAL SYSTEM

Communication scheme for data acquisition is shown in Fig. 3. Control communication between host machine (PC) and embedded machine Nvidia Jetson TX2 is established using ssh protocol. Jetson TX2 reads data form sensor periodically using I2C serial communication. Sampling period is 100ms. Data is processed on Jetson TX2 in order to get the system angular position information. Position estimation is sent to PC using UART in order to be graphically displayed.

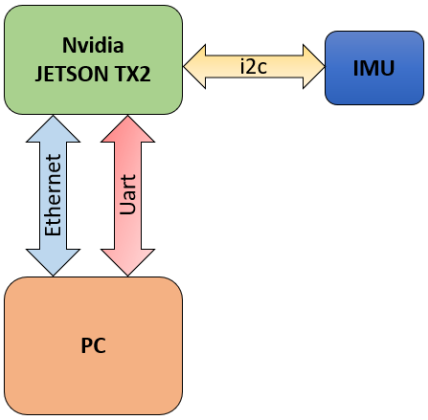


Fig. 3. Communication scheme

A schematic view of the complete procedure for determining the angular position of the system is shown in Fig. 4

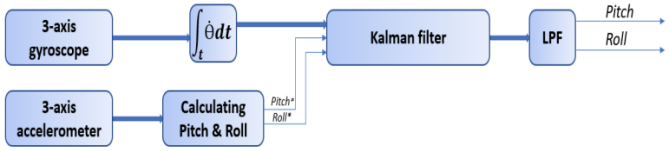


Fig. 4. A complete data processing scheme

As a first step, the initial testing of Kalman filter is done. The sensors were moved from the initial position by 90 degree to one side, then returned and moved by 90 degree to another side and returned to the initial position again. After that sensors were exposed to disturbances. Result is shown in Fig. 5.

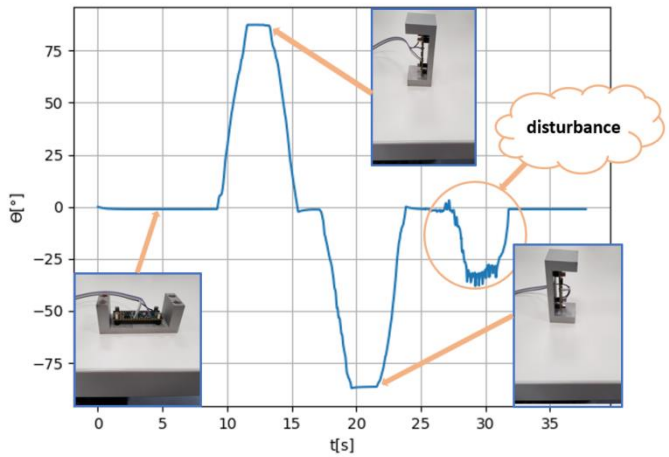


Fig. 5. Estimated system angular position (elevation), initial test

Multi-sensor camera systems with pan/tilt are placed on the shaker. Jetson TX2, sensors and laser are placed on camera sunshield. The laser is directed towards grid paper. Jetson TX2, IMU and host machine are connected as shown in Fig. 3. The described system is shown in Fig. 6 and Fig. 7.

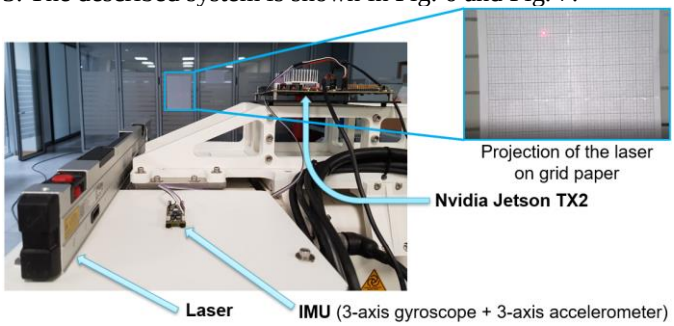


Fig. 6. Measuring equipment

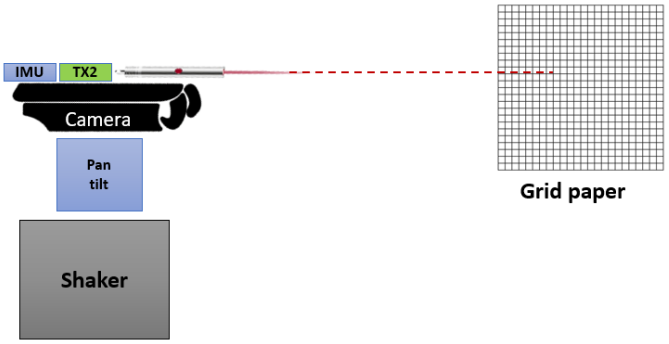


Fig. 7. Scheme of the testing procedure

The shaker is running and the camera is recording movement of laser on grid paper. The shaker oscillates with frequency of 0.25 Hz and brings disturbances per elevation. The disturbances per azimuth are negligible. The purpose of the shaker is to simulate wind-related disturbances. Peak-peak amplitude of laser movement by elevation is around 12cm which translated to degrees is around 1 degree. After starting gyro stabilization, peak-peak amplitude is decreased to approximately 1.2cm (0.1 degree), but drift appears. This drift is very small and projection of laser moves only 2.5cm (0.21 degrees) per elevation after 275s. That does not seem much, but in case where this is observed under narrow FOV (which is 1 degree on some cameras), movement would be significant.

First 25 seconds system works without gyro stabilization. After that gyro stabilization turns on for 275 seconds. Based on video of movement of laser projection on grid paper it is calculated that after 275 seconds drift is 0.21 degree.

Information whether the gyro stabilization is turned on or turned off is known in advance. The assumption is that the system is still during the gyro stabilization, meaning that commands for moving pan/tilt per azimuth and per elevation are not set. System movement should be monitored after the gyro stabilization is turned on.

In Fig. 8, system movement by elevation calculated by gyroscope and accelerometer is shown. Accelerometer measurements have a lot of noise. Gyroscope has accumulated drift due to its uncompensated bias and due to pan/tilt gyro stabilization.

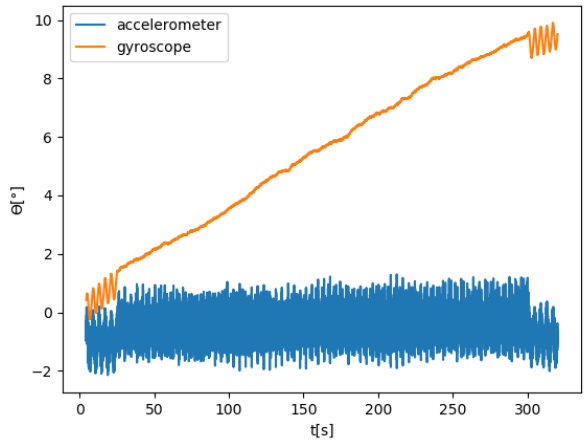


Fig. 8. Gyroscope and accelerometer measurement

System movement per elevation, obtained after the whole procedure of data processing, and the detected drift are depicted in Fig. 9.

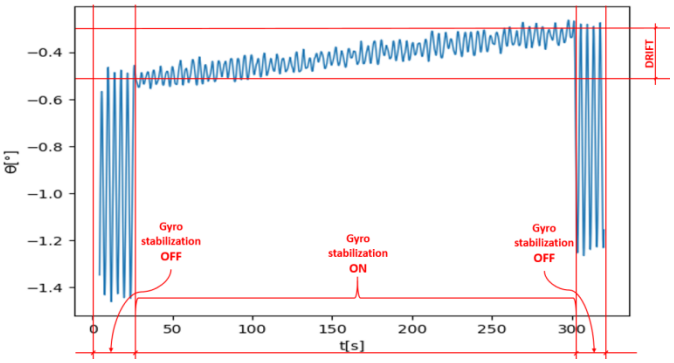


Fig. 9. Detected drift

To eliminate the remaining oscillations in order to determine the exact value of the drift, averaging with window function with length N has been done, as depicted on the image below (Fig. 10).

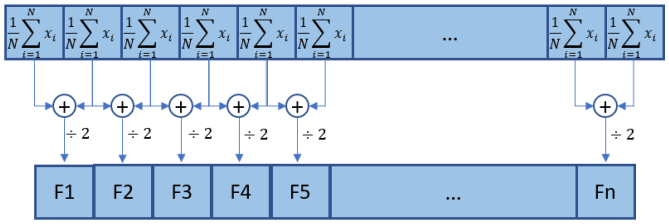


Fig. 10. Extraction drift

x_i in Fig. 10 is data after Kalman filter, and F_i is the final result. Based on the obtained result, it is possible to detect drift of 0.01 degree.

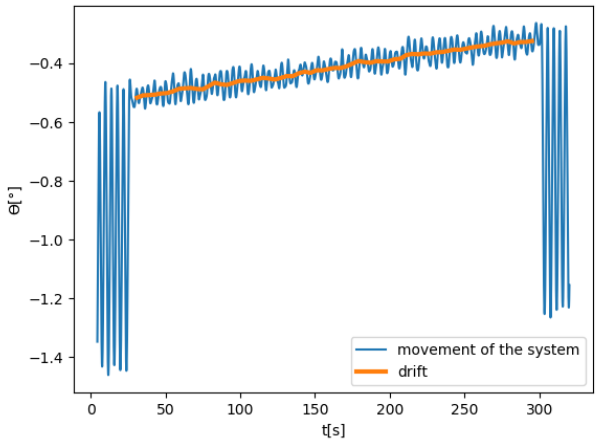


Fig. 11. Signal of system movement and drift

V. CONCLUSION

This paper presents results of the pan/tilt drift detection obtained on the basis of measurements with 2 MEMS sensors, gyroscope and accelerometer. Kalman's filter has been implemented with the aim to get the best out of both sensors and compensate for the flaws which the sensors have when they are used independently. It should be mentioned that some of the existing solutions for the needs of surveillance applicationdo not have the mentioned detection, therefore it is not possible to perform the drift compensation, which leads to the unusable gyro stabilization when the FOV is narrow. It has been shown that it is possible to detect drift of only 0.01 allowing for its almost unnoticeable compensation even in case of the narrow FOV.

Further improvements could be achieved by introducing magnetometer in the system. Using the magnetometer measurements a reliable estimation of the system position per azimuth could be calculated [7].

REFERENCES

[1] Infiniti Electro-Optics, "Gyro Stabilization for Lon-Range Survillance Cameras | Infiniti Electro-Optics," Infiniti Electro-Optics, [Online]. Available: https://www.infinitioptics.com/technology/gyro-stabilization. [Accessed 10 April 2019].

[2] S. Beeby, G. Ensell, M. Kraft and N. White, MEMS Mechanical Sensors, Artech House, 2004.

[3] R. Cechowicz, "Bias Drift Estimation for MEMS Gyroscope Used in Inertial Navigation," *DE GRUYTER*, pp. 104-100, 2017.

[4] O. J. Woodman, "An introduction to inertial navigation," University of Cambridge, 2007.

[5] M. Pedley, "Tilt Sensing Using a Three-Axis Accelerometer," 2013.

[6] SIGNALS & SYSTEMS DEPARTMENT, School of Electrical Engineering, University of Belgrade , "Stochastic Systems and Estimation," [Online]. Available: http://automatika.etf.bg.ac.rs/images/FAJLOVI_srpski/predmeti/obavezni_kursevi_OS/OS3SSE/materijali/12_kalman.pdf.

[7] F. Abyarjoo, A. Barreto, J. Cofino and F. R. Ortega, "Implementing a Sensor Fusion Algorithm for 3D Orientation Detection with Inertial/Magnetic Sensors," in *CISSE*, 2012.

Requirements Analysis for ADAS Perception in Bad Visibility Conditions

Nedeljko Padjen, Nikola Latinović, Dragana Perić, *Member, IEEE*,
Milan Milosavljević, *Member, IEEE*

***Abstract*— In this paper we elaborate on requirements that are imposed on the vision system that should be used for ADAS (Advanced Driver-Assistance Systems) perception. In order to fulfill its task of perception under all visibility conditions, we are taken into consideration bad visibility scenarios that can be met due to fog, rain, snow etc. We analyze the range and the resolution issues of solution based on sensor fusion of different type of sensors like LWIR, SWIR imagers, SWIR LIDARs supported by MIMO Radar. The goal of this analysis is to set the guidelines for selection of appropriate system components.**

***Index Terms*—ADAS perception, sensor fusion, gated SWIR**

I. Introduction

One of the major concerns and challenges in car industry in the last few decades was to increase safety of the passengers. Introduction of seatbelts, airbags and crumple-zones, as a form of passive safety systems, has reduced the severity of injuries, but it was clear they are not sufficient to make a bigger impact on the safety aspect [1]. The overall safety was further improved with the introduction of active safety systems, which were rapidly implemented in the vehicles, following the general technology progress, reflected on decrease size and price of different type of elements implemented in cars. As the name suggests, active systems are actively involved in preventing the accidents, instead of only reducing the severity of accident outcomes.

ADAS, as an electronic system consisted of different sensors, represents one form of the vehicle active safety systems, offering different levels of support to the driver. This support can go from the basic level of only informing the driver on some events ahead of him, to the level where ADAS can take the full control of the vehicle, in order to avoid some imminent danger [2]. The SAE automation level 5 is vehicle with a completely automated navigation system, while SAE level 0 are the vehicle without any ADAS system [3]. Regardless of the level of the control, these driver assistance systems share the same goal – to prevent the accident by giving proper and timely information to the driver/system, or at least to reduce the consequences of it [4]. In other words, their goal is to decrease driver decision time, but also detection time [5].

Nedeljko Padjen is with the Vlatacom Institute of High Technologies, 5 Milutina Milankovića Blvd.,11070 Belgrade, Serbia (e-mail: nedeljko.padjen@vlatacom.com).

Nikola Latinović is with the Vlatacom Institute of High Technologies, 5 Milutina Milankovića Blvd.,11070 Belgrade, Serbia (e-mail: nikola.latinovic@vlatacom.com).

Dragana Perić is with the Vlatacom Institute of High Technologies, 5 Milutina Milankovića Blvd.,11070 Belgrade, Serbia (e-mail: dragana.peric@vlatacom.com).

Milan Milosavljevic is with the Singidunum university, 32 Danijelova Str., 11000 Belgrade, Serbia (e-mail: univerzitet@singidunum.ac.rs).

Selection of the appropriate sensors for ADAS system which can bring a substantial benefit to the driver safety is a complex task, requiring many different inputs and analysis. The system should be designed having in mind the drivers real needs and requirements, but also his/hers attitude, that depend significantly on the region of the world where he comes from [6]. Some of the inputs are controversial; such is one study in Greece which has shown the decreased number of accidents during the rain, while the most of the other studies show the opposite [5].

At the end, the deployment of the sensors and level of introduced automatism should be carefully taught, since they can result in negative tendencies – due to increased safety equipment, the driver can decrease his/hers overall awareness and compliance with traffic regulations, resulting in a higher driving speed. This can result in fatal outcomes since vehicles, regardless on the number and quality of the installed active safety systems, still obeys the laws of the physics, where the car stoppage time is inversely dependent on its speed.

Within this variety of the aspects influencing the design of the ADAS, we have chosen to analyze the effects of the adverse weather condition on the driver and try to define the requirement for the adequate choice of the sensor equipment.

Chapter two of this paper will list the sensors which can be used in one ADAS, pointing out their advantages/disadvantages. Chapter three will give the system requirements in terms of sensor range, and field of view (FOV). Chapter four will present the current equipment available on the market. In chapter five we will discuss the process of integration of the different sensors to a functional system. As this paper is one of the steps in designing, but also assembling and testing of a real system, chapter 6 will give an overview of the potential tests for real performance analysis (laboratory, controlled environment, field test in real scenario). Conclusion on the raised topics and directions for further research will finally be given in chapter 7.

II. Sensor Types

Adverse weather conditions, as the most analysis has shown, have negative affect on the driver's safety. Fatalities (deaths caused by accidents) represent 10-15% of the total number of accidents [5]. While the ADAS for the good weather condition would be mostly based on some color camera(s), the weather such as fog, mist, dust, dark, rain and snow requires the introduction of different type of sensors, which will contribute in the collection of information from the road. Some technologies will assist in detecting the

pedestrians at night; some others will give best results in detection of the other vehicles [7]. It is important to underline that there is no single sensor to perfectly fit all needs, but their fusion, with the usage of appropriate algorithms, can hide their downsides and combine their advantages. The following lines will list the most promising type of sensors that can enhance the perception of ADAS in adverse weather conditions [8]:

- RADAR (RAdio Detection And Ranging) – The well-known concept can be used for detection, but also direction and speed of several targets, using the arrays of micro-antennas installed in the vehicle bumpers. These Millimeter Micro Wave (MMW) radars are generally performing well in all weather conditions, and as such, they are good for detecting the close targets. The downsides are the reduced Field of View (FOV) and low precision.
- LiDAR (Light Detection And Ranging) – Using the same principle as radar, the distance to the object is measured by the equation:

$$d = \frac{c}{2} \times ToF \tag{1}$$

where c is the velocity of the light and ToF is the time of the return trip of the light emitted by lidar.

Solid state lidars, consisting of the arrays of photodiodes, can obtain the 3D object map, thus they are very useful for object identification and avoidance. Downside of this sensor is that its performance is degraded in the bad weather conditions, where the light beams are diffracted on the rain, snow or dust drops and particles.

- Visible (VIS) camera – Capturing the wavelengths in visible spectrum (400 nm to 780 nm), the resulted high-resolution images are the major source of information in good weather conditions. Using more than one camera avails stereoscopy vision, which is additionally giving depth to the pictures. Unfortunately, this type of camera is highly affected by the weather conditions, leading to complete degradation of the performance, for example, during the cloudy nights.
- Longwave Infrared Camera LWIR – Sensitive to emitted thermal radiation in spectral range from 8 µm to 12 µm. For this application we are referring to uncooled micro-bolometer cameras, operating in mentioned spectral range. LWIR cameras provide night vision, and also perform better than visible cameras in case of fog, Sun glare, sudden illumination variations (tunnel entering, headlights from cars in the opposite lane etc.). Object identification range of thermal cameras is not related to illumination level, therefore it can be much larger than visibility range of VIS cameras.
- Shortwave Infrared Camera SWIR– Sensitive to reflected light in spectral range from 0.9 µm to 1.7 µm. Due to different reflectivity of the materials in this spectral range, SWIR offers imagery richer in details than visible camera. SWIR camera can be used as passive imager and as active imager, with laser illuminator. Active, range gated SWIR cameras can provide very good identification capability in all lighting conditions (complete darkness, fog, smoke, etc.). By choosing the delay of the gate impulse, the depth of the scene is regulated.

The following table summarizes the advantages and disadvantages of the listed sensors, grading them to four different levels, being 0 the worst and 3 the best ranked [7]

TABLE I
SENSOR CHARACTERISTICS

FEATURE	RADAR	LIDAR	VIS	GATED SWIR	LWIR
FOV	2	2	3	3	3
RANGE	3	3	2	2	3
ACCURACY	2	3	3	3	2
FRAME RATE	2	2	2	2	3
RESOLUTION	1	2	3	2	1
COLOR PERCEPTION	0	2	3	1	1
WEATHER AFFECTION	3	2	1	2	3

The same results are also presented on the following charts:

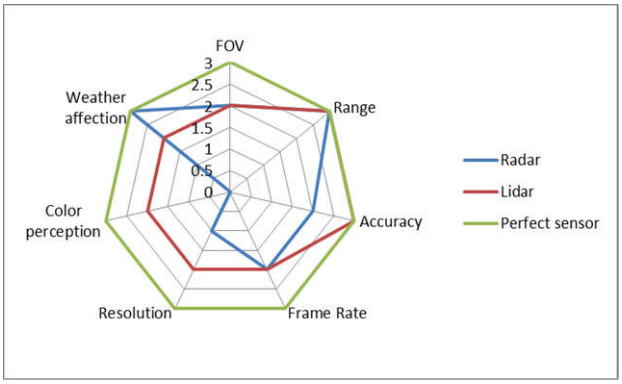


Fig. 1. Main characteristics of radar and lidar sensors

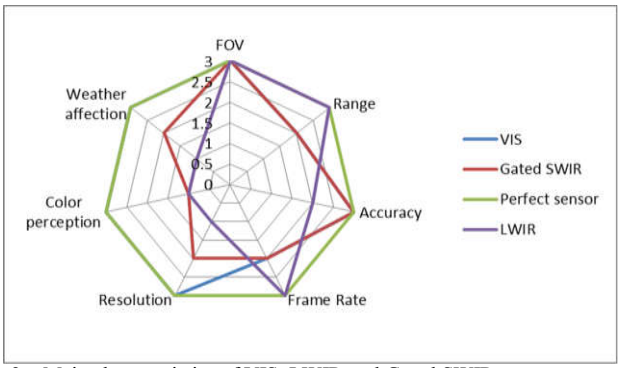


Fig. 2. Main characteristics of VIS, LWIR and Gated SWIR cameras

III. SYSTEM REQUIREMENTS

Triangle concept of the road safety recognizes three major "stakeholders": human, environment and vehicle, with bidirectional relations between them during the course of driving [9]. Any analysis of the concepts that can improve overall safety on the roads must include all three aspects into consideration. In line with that, the reaction of the driver-vehicle system can be divided in three components:

- Decision time (td) – Time from the recognition of some situation, to the moment the driver's brain indicates something should be done. This time is longer in adverse weather condition.

- Initiation time (ti) – Time from "should be done", to doing it. The weather does not affect this time length.
- Action time (ta) – Time from the driver acts, to the moment vehicle has stopped (either safely, or by crashing to the obstacle). Adverse weather is decreasing the vehicle-road adhesion and, as a result, increasing the action time.

Thus, the total vehicle stopping distance is the sum of distances travelled during these three time components, and it is given by the following equation:

$$s = v_0 \times t_r + d_b = v_0 \times t_d + v_0 \times t_i + \frac{v_0^2}{2 \times g \times (f + G)} \quad (2)$$

where:
s - Vehicle stopping distance (m)
v_0 – Vehicle initial speed (m/s)
d_b – Breaking distance (m)
g – Gravity acceleration (m/s^2)
f – Coefficient of deceleration (m/s^2)
G – Gradient (slope of the road)

Based on this equation, we can calculate the residual speed distance at the distance x from the point of detection:

$$\begin{aligned} v(x) &= \sqrt{v_0^2 - 2gf(x - v_0 \times t_r)} && x_d + x_i < x \le s \\ v(x) &= v_0 && \text{, if } x \le x_d + x_i \\ v(x) &= 0 && x \ge s \end{aligned} \quad (3)$$

Where:
x_d – distance driven during decision time,
x_i – distance driven during initiation time,

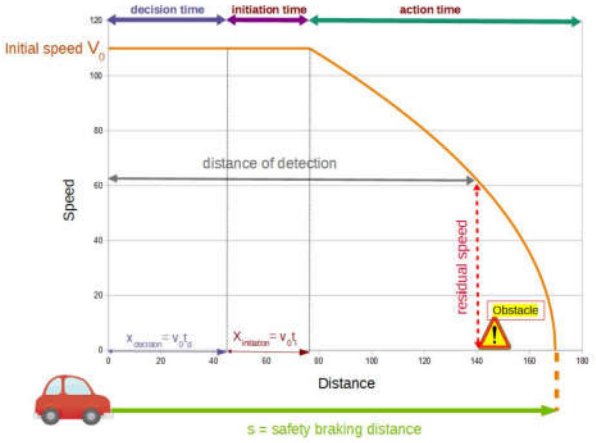


Fig. 3. Decomposition of the total breaking distance [5]

Based on these equations, one European study [5] has calculated the residual speeds, based on initial speed and different weather conditions on the road. The following tables are giving the results of the calculations:

TABLE II
SAFETY BREAKING DISTANCES BREAK-DOWN

		Equation (2) input parameters				
	Unit	Clear weather	Rain		Fog	Snow
Initial speed	km/h	130	130	110	90	67
Decision time distance	m	43.3	57.8	48.9	55	/
Initiation time distance	m	10.8	10.8	9.2	7.5	/
Braking distance	m	93.1	163	116.7	78.1	/
Safety braking distance	m	147.2	231.6	174.8	140.6	<140

TABLE III
RESIDUAL SPEED FOR DIFFERENT WEATHER CONDITIONS

Distance of detection	Residual speed in km/h at detection distance				
	Clear weather	Rain		Fog	Snow
50	130	130	110	90	67
140	36	97	60	8	0
145	20	95	56	0	0
150	0	92	51	0	0
175	0	77	0	0	0
200	0	57	0	0	0

Taking into account some other studies which are showing that residual speed higher than 70 km/h are usually ending with fatality, it is safe to say that we should target the sensor which can detect the obstacle on at least 200 meters.

Beside the sensor range, the important sensor parameter is the field of view (FOV), i.e the angle of camera view. The Horizontal FOV (HFOV) can be expressed as the function of the visible field (D, expressed in meters), on different distances from the camera (s, in meters), by the following equation:

$$FOV = 2 \times \arctan \frac{D/2}{s} \quad (4)$$

If our requirement is that our system should have D=10meters (approximately 3 lanes of the road), we are getting the following table:

TABLE IV
CAMERA FOV AT DIFFERENT DISTANCES, FOR D = 10M

Distance from camera **s** (meters)	Required visible field **D** (meters)	Camera **FOV** (degrees)
5	10	90.000
10	10	53.130
25	10	22.620
100	10	5.724
200	10	2.864
250	10	2.292

The following chapter will analyze some of the cameras on the market to verify if they conform with these requirements.

IV. Equipment Available on the Market

As previous analysis has shown, ADAS system that would fulfill requirements for vision enhancement in adverse weather conditions should contain LWIR and SWIR imagers because of their favorable features. In this work we will focus on part of the system that combines these two sources of information and analyze the currently available

equipment available on the market that can be used in order to realize the subsystem.

A. LWIR camera

When selecting LWIR camera, we choose its parameters with following objectives:

- Detection range – distance at which the camera will spot an object
- Field of view – Width and height of the scene in front of the vehicle that camera will observe.

According to previous analysis, requirements for detection range should be selected to provide sufficient time to the driver to stop the vehicle. Previous analysis show that 200 m is the minimum distance for braking, so we shall select the camera detection range accordingly. Camera's field of view should be wide enough to provide good covering of the road. For this parameter selection we will refer to requirements given in Table IV.

In Table V we presented parameters of two LWIR cameras suitable for this application.

Calculated ranges of pedestrian and car detection, recognition and identification are given for each camera. Performances of thermal cameras are expressed in detection, recognition and identification (DRI) ranges that are calculated following STANAG 4347 procedure [10]. DRI ranges are given for a pedestrian (1.8m x 0.5m) and for a car (2.3m x 2.3m). These three numbers give the information of the level of object perception (detection is defined as 1 lp/object size, recognition is defined as 3 lp/object size and identification is defined as 6 lp/object size).

TABLE V
LWIR CAMERAS PARAMETERS

	Camera 1			Camera 2		
Spectral range	7.5 µm – 13.5 µm			8.0 to 14.0 µm		
Pixel size	12 µm			17 µm		
Resolution	640 x 512			640 x 480		
Frame rate	60 Hz			50 Hz		
Thermal sensitivity	<50 mK			<50 mK		
Lens focal length	4.9 mm			10 mm		
Horizontal FOV	76 º			57 º		
Vertical FOV	64 º			44 º		
Pedestrian DRI	204m	51m	25m	294m	73m	36m
Car DRI*	626m	156m	58m	901m	225m	112m

*DRI – detection, recognition, identification

B. SWIR camera

SWIR cameras can be used either in passive or in active mode. First we will give the characteristics of two types of SWIR cameras in passive mode and then additional details regarding active SWIR will be given.

As in the case of LWIR camera, also for SWIR camera we have to make a compromise with selection of detection range and wide field of view. To select optimum lens focal length, we use geometrical calculation. Field of view should be close to the one already defined in LWIR camera case, in order to provide better alignment of two cameras that will observe the same scene.

In Table VI we present parameters of two SWIR cameras suitable for this application.

TABLE VI
SWIR CAMERAS PARAMETERS

	Camera 1	Camera 2
Spectral range	0.9 µm – 1.7 µm	0.9 µm – 1.7 µm
Pixel size	15 µm	20 µm
Resolution	640 x 512	640 x 512
Frame rate	up to 230 Hz	max 100 Hz
Integration time	10 µs to 10 ms	1 µs - 40 ms
Noise high gain	50 e^-	60 e^-
Noise low gain	270 e^-	400 e^-
Lens focal length	8 mm	8 mm
Horizontal FOV	61 º	77 º
Vertical FOV	51 º	65 º

Active, gated imaging provides the ability to image a specific depth slice of a scene. Applications are in observations through obscurants (severe weather conditions), estimation of distance and localization of obstacles. Imaging devices must be fast enough to cope with the reflected light.

According to the technical specifications of the Camera 1, it is stated that this camera can be used also in gating mode with exposure time from 100ns to 9µs. To provide complete gated vision system, SWIR camera should be coupled with pulsed laser source that will illuminate the scene. Camera gating should be synchronized with laser pulses. More details of the gated system realization can be found in [11].

C. Displaying the images to the driver

We propose the presentation of collected imagery from the ADAS to the driver.

First solution would combine videos from both cameras simultaneously and present them on the same screen, on left and right halves. The driver can switch passive SWIR to active gated SWIR if suspected object appears on LWIR image.

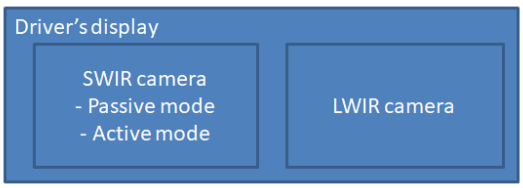


Fig. 4. Driver's display with two images

Second solution would require additional image processing and forming of a single image using both cameras as inputs. In that case image blending or image fusion should be performed. This solution would require minimum execution time of image processing algorithms, as ADAS require no-latency imagery to be presented to the driver, which is a key in providing the with real situational awareness. Advantage of this approach would be in unique screen from which the driver would follow the information.

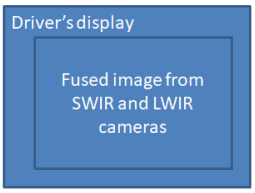


Fig. 5. Driver's display with fused image

V. CAMERAS SIGNAL PROCESSING

As described in the previous chapter, the proposed solution combines the imaging of two cameras on the same presentation medium (display). This will require the complex process of synchronized signal processing, which will not be described in details in this paper. Instead, we will list the main challenges of this process and give the main directions for the design.

The proposed cameras for the integration in the ADAS system are originating from different manufacturers, with different characteristics and parameters (camera interfaces, frame rate, resolution). In order to achieve the equalization of the frame rates and image resolution adoption, but also the advanced image processing such as image stabilization & enhancement, target tracking, or image fusion, the signals from all cameras should be converted to the same format [12].

Our approach will be to develop a dedicated hardware platform which will house the main elements of this signal processing unit:

- Interface board, which will collect the signals from both cameras and forward it to FPGA
- FPGA processing unit, for converting camera input interfaces to common HDMI signal [13]. Unlike most of applications where a single FPGA is processing signal from only one camera, in this realization both camera signals will be processed by the same FPGA. The signals will be forwarded to the main processor board
- Nvidia Jetson TX2, as a main digital signal processing (DSP) unit. This powerful hardware structure will provide the output to vehicle display, by means of Ethernet cable.

Although the previous experience with the proposed design and equipment is indicating the smooth work of the system, the detailed tests must be performed, including the measurements of the overall signal processing delay, which must be taken into calculations for system detection time.

VI. PERFORMANCE TEST ENVIRONMENT

As any other system with the ambition to be recognized as ADAS system, the series of rigorous testing must be performed. We are proposing three-level system testing:

A. Laboratory tests

The initial test will be performed in electro-optical laboratory equipped with collimator test-station [14], with the appropriate radiation sources and different test targets. The test results should give a clear picture of the real system (camera) characteristics and performance, which is a prerequisite for the further testing.

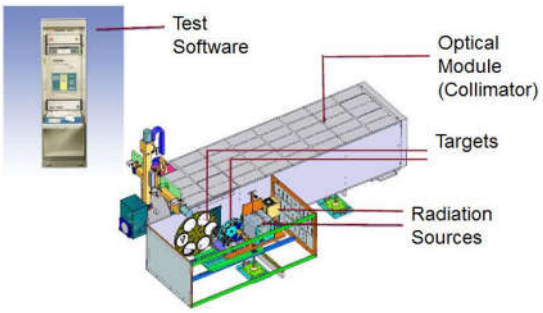


Figure 6. Electro-optical modular test station [14]

B. Tests in the controlled conditions

These tests should verify the system performance under controlled adverse weather condition. Although it would be ideal to test the system behavior in all weather conditions (fog, mist, snow, rain, etc..), our focus will be only on rain and low visibility conditions, since the rain is identified as the biggest challenge for safety breaking (Table 2, Ch. 2).

Fig. 6. Example of tunnel and sprinkle system [7]

As per the world's best practices [7], we will establish a form of tunnel 250 meters long, 10 meter in width (to match the requirements from the Chapter 3 of this paper) and. The rain should be produced by sprinkles with different size of nozzles (for producing different raindrop size), that can produce different rain intensity. The required height of the tunnel will be defined together with sprinkle selection, so the rain distribution on the camera level (around 1 meter from the ground) is even.

C. Real environment testing

At the end, the equipment should be tested in the real environmental conditions. This testing will last for several months where, ideally, the equipment will be tested on several locations known for higher probability of some adverse weather condition. The test location will also be equipped with a small meteorological station, which should be equipped with luxmeter (for measuring ambient light), rain gauge, and temperature and humidity sensors [15].

VII. CONCLUSION

Advanced driver-assistance systems are the group of sensors which main purpose is to decrease the chance of the road accidents. In this paper we have listed the types of sensors suitable for deployment in ADAS, with the focus on adverse weather condition. We have identified two major requirements for these systems, being the detection range and field of view, and confirmed that SWIR and LWIR cameras are in-line with these requirements. Since the key of this system is in providing the timely information to the driver, we have also suggested the components for the signal

processing unit which should introduce the minimal delay. All this information gave us the major guidelines for the future work, which will be to assemble a real system with some of the proposed cameras, and test it in various laboratories, controlled and real environmental conditions. The results of these tests will surely give us new inputs directions in further development of a reliable ADAS system.

REFERENCES

[1] S.Tsugawa, 'Trends and issues in safe driver assistance systems", *IATSS Research,* Vol.30 No.2, 2006

[2] A. Lindgren, F. Chen, P. W. Jordan, & Zhang, H. (2008). Requirements for the design of advanced driver assistance systems - The differences between Swedish and Chinese drivers. *International Journal of Design, 2*(2), 41-54

[3] SAE International and J3016, Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems,https://web.archive.org/web/20170903105244/https://www.sae.org/misc/pdfs/automated_driving.

[4] A. Lindgren, F. Chen, (2007), State of the art analysis: An overview of advanced driver assistance systems (ADAS) and possible human factors issues. In C. Weikert (Ed.), *Proceedings of the Swedish Human Factors Network (HFN) Conference* (pp. 38-50). Linköping: Swedish Network for Human Factors.

[5] EU Project DENSE report, "D2.2 System needs and benchmarking", 2017

[6] B. Shneiderman, C. Plaisant, (2005). Designing the user interface. Boston: Pearson/ Addison Wesley.

[7] EU Project DENSE report, "D2.1 Characteristics of adverse weather conditions", 2017

[8] F. Rosique, P.J. Navarro, C. Fernandez, A. Padilla, "A systematic review of perception system and simulators for autonomous vehicles research," *Sensors* 2019, *19*, 648; doi:10.3390/s19030648.

[9] V. Eksler, The role of structural factors in road safety. Retrieved from ectri org.:1–11.

[10] North Atlantic Treaty Organization, *Definition of nominal static range performance for thermal imaging systems,* Military agency for standardization, STANAG No. 4347, 1995

[11] D. Peric, B. Livada, M. Peric, "Analysis and Selection of Components for Active SWIR/NIR Vision Systems", Proceedings of 4th International Conference on Electrical, Electronics and Computing Engineering, IcETRAN 2017, Kladovo, Serbia, June 05-08, ISBN 978-86-7466-692-0, pp. EKI3.1.1-5

[12] N. Latinovic, I. Popadic, P. Milanovic, M.Peric, M. Veinovic, "Multisensor imaging system video interface implementation in FPGA", Proceedings of Singidunum University International Scientific Conference, Sinteza 2019, Novi Sad, Serbia, April 20, 2019

[13] S. Dhanani, M. Parker , *Digital Video Processing for Engineers: A Foundation for Embedded Systems Design*, Newnes: 1st edition (October 24, 2012)

[14] Modular Electro-Optical Test System, CI Systems, Cat. No. 607-5610H, 2012.

[15] Nicolas Pinchon, Olivier Cassignol, Frédéric Bernardin, Adrien Nicolas, Patrick Leduc, et al.. Allweather vision for automotive safety: which spectral band?. AMAA 2018, Advanced Microsystems for Automotive Applications, Sep 2018, Berlin, Germany. pp. 3-15, ff10.1007/978-3-319-99762-9_1ff. ffhal-01975285f

Adaptive Kalman Filtering Using M-robust Dynamic Stochastic Approximation Combined with Robust Median Estimation

Zoran Banjac, Željko Đurović, Branko Kovačević, *Senior Member, IEEE*

***Abstract*—One of the most significant achievement of the linear estimation theory is the Kalman filtering. In this paper, the problem of robustifying the Kalman filter in the presence of unknown noise statistics has been considered starting from the equivalence between the linear estimation problem and a specific form of dynamic stochastic approximation, M-robust statistical approach is used to robustifying the Kalman filter. The proposed approach includes the modified M-robust performance index based on the mean square optimal prediction of the expected changes in the system states, together with the given output measurements. The M-robust dynamic stochastic approximation algorithm is derived from step-by-step minimization of the adopted criterion. In order to improve the convergence rate, the gain matrix of the algorithm is derived from step-by-step minimization of the prespecified mean square error criterion. In addition, robust median estimations are derived for adaptive estimation of the unknown state and observation noise statistics, simultaneously with the system states. A real life examples of maneuvering target tracking is presented to demonstrate the practical robustness of the proposed adaptive robustifying Kalman filter.**

***Index Terms*—Adaptive filtering, impulsive noise, Kalman filtering, non-Gaussian noise, nonlinear filters, robust estimation**

I. Introduction

AN optimal estimate of a random vector, or a random vector process, from noise-corrupted data, is one which minimizes an appropriate functional of the estimation error. Examples of such criteria are the maximum likelihood, least squares, mean-square error, minimum variance, etc. [1-4]. However, optimal estimation criteria often depend upon assumptions about the statistical characteristics of the random variable which has to be estimated and its associated random data. Particularly, one of the most significant achievements in the linear estimation is the Kalman filtering. Except of the theoretical interest, it is also important from the practical applications, [3, 4]. However, if the system dynamics and its observations are dominated by nonlinear effects that cannot be accurately approximated by linearization, or if the noise processes are non-Gaussian, the corresponding optimal estimation algorithm is often too complex to mechanize. In such circumstances, it is sometimes possible, through the use of a class of time-varying stochastic approximation methods, to obtain a sequence of estimates that possess a statistically bounded error [5]. Thus, a dynamic stochastic approximation algorithm offers a reasonable alternative to optimal estimation techniques in many applications, such as the problems in estimation, optimization and pattern classification [6-8].

Unfortunately, in many practical applications the real noise distribution differs from the supposed normal one by heavier tails. This, in turn, generates the spiky noise realizations named outliers [9]. Since the Kalman filter is a linear function of observations, it is susceptible to outliers. Therefore, it is of great practical interest to designing robustified Kalman filtering techniques that are able to cope with impulsive noise or outliers. In such circumstances, robust statistical methods provide suitable tools to spot bad data points and suppress their effects [9-11]. Particularly, the M-robust statistical approach is widely used in practice, since it represents an approximation of maximum likelihood estimation that is easy to implement, [10]. In recent literature, there exist a number of articles devoted to robustifying the Kalman filter using the M-robust approach [12-19]. Moreover, the solution proposed in [17, 18] is based on the similarity between the Kalman filter and a weighted least squares parameter estimation, as well as the application of M-robust approach to solving the parameter estimation problem robustly. However, the so obtain solution requires complex computations.

The proposed solution in [19] has the predictor-corrector structure, as the original Kalman filter, but the M-robust dynamic stochastic approximation algorithm is used for the measurement update, instead of the original Kalman filtering relations. In addition, M-approach is conservative, and it not only has a lower efficiency in Gaussian noise, but also may degrade otherwise [17, 18]. Therefore, some kind of adaptation to the underlying noise condition is necessary. An approach to overcome this difficulty based on a weighted least-squares approximation of the M-robust estimates, as well as the M-robust estimation of the unknown noise statistics, simultaneously with the system states, has been proposed in [18]. An alternative approach based on M-robust dynamic stochastic approximation with the parallel adaptation of unknown noise statistics using robust median estimation, has been considered in this paper.

Zoran Banjac is with Institute Vlatacom, 5 Milutina Milankovića Blvd, 11070 New Belgrade, Serbia (e-mail: zoran.banjac@vlatacom.com).

Željko Đurović is with the School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail: zdjurovic @etf.bg.ac.rs).

Branko Kovačević is with the School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail: kovacevic_b@etf.bg.ac.rs).

II. Problem Formulation

Let us assume that an abstract linear discrete stochastic system is given by the state-space model

$$x(k+1) = F(k)x(k) + G(k)w(k) \quad (1)$$

$$y(k) = H(k)x(k) + v(k) \quad (2)$$

Here $x(k)$ is the random state vector, $y(k)$ is the observation or measurement vector, $w(k)$ is the state noise or disturbance, and $v(k)$ is the additive measurement noise, at the discrete time indexed by k. Furthermore, $w(k)$ and $v(k)$ are zero-mean noises that are uncorrelated by itself and mutually, satisfying

$$E\left\{\begin{bmatrix} w(j) \\ v(j) \end{bmatrix}\begin{bmatrix} w^T(k) \\ v^T(k) \end{bmatrix}^T\right\} = diag\left\{Q(k)\delta_{kj}, R(k)\delta_{kj}\right\} \quad (3)$$

with $E\{\cdot\}$ being the mathematical expectation, δ_{kj} denotes the Kronecker's delta symbol ($\delta_{kj} = 0$ if $k \neq j$ and $\delta_{kk} = 1$), and $diag\{\cdot,\cdot\}$ represents the block-diagonal matrix. In addition, $F(k)$ is the state-transition matrix, $G(k)$ is the state-noise or disturbance matrix, and $H(k)$ is the measurement or observation matrix. If $\hat{x}(k|l)$, $l = k-1, k$, is the linear least squares type estimate of $x(k)$ when the measurements $\{y(j), j \le l\}$ are given, while $P(k|l)$ denotes the underlying estimation error covariance matrices, then the classical Kalman filter equations are the following, [1-4]

Time update (prediction stage):

$$\hat{x}(k+1|k) = F(k)\hat{x}(k|k) \quad (4)$$

$$P(k+1|k) = F(k)P(k|k)F^T(k) + G(k)Q(k)G^T(k) \quad (5)$$

Measurement update (estimation or correction stage):

$$\varepsilon(k) = y(k) - H(k)\hat{x}(k|k-1) \quad (6)$$

$$K(k) = \frac{P(k|k-1)H^T(k)}{\left[H(k)P(k|k-1)H^T(k) + R(k)\right]} \quad (7)$$

$$\hat{x}(k|k) = \hat{x}(k|k-1) + K(k)\varepsilon(k) \quad (8)$$

$$P(k|k) = \left[I - K(k)H(k)\right]P(k|k-1) \quad (9)$$

with I being the identity matrix. The measurement residual or innovation, $\varepsilon(k)$, represents the zero-mean, uncorrelated random sequence with covariance matrix $S(k)$, [3, 4]

$$S(k) = H(k)P(k|k-1)H^T(k) + R(k) \quad (10)$$

In most applications the elements of the observation vector, $y(k)$, are independent, so that they may be analysed one at a time. This means that $y(k)$ and $\varepsilon(k)$ are zero-mean scalar real random variables, having the variances $R(k)$ and $S(k)$, respectively. The state disturbance, $w(k)$, is also taken to be zero-mean scalar real random variable having the variance $Q(k)$, [1-4].

Unfortunately, if the noise statistics are non-Gaussian or undefined, the underlying optimal estimate may be indeterminate [18, 20-22]. Thus, the Kalman estimator is susceptible to departures of the noise statistics from the adopted Gaussian distribution, or it is not robust [12-18]. Formal mathematical definitions of robustness are given in the statistical literature [9-11], but these are inconvenient for practical workers. In this sense, simple intuitive and data-oriented practical definitions are more suitable. Particularly, the two types of such definitions are popular in practice, the so-called resistant and efficiency robustness. Thus, an estimation procedure is robust in the resistant sense if it stays finite when some of the observations become too large in positive or negative direction. On the other hand, an estimator is efficiency robust if it possesses a high efficiency under the assumed Gaussian distribution of observation data, but it also remains acceptably high efficiency when the real distribution differs from the Gaussian one by heavier tails. This, in turn, generates the outlying data points contaminating the normally distributed observations. Thus, a practically robust estimator has to satisfy the resistant property, together with the efficiency one. A possible approach for designing a simple but practically robust version of the standard Kalman filtering technique has been proposed in [19]. The brief review of this approach is presented in the next chapter.

III. Brief Review of Kalman Filtering based on M-robust Dynamic Stochastic Approximation

Most robust statistical procedures represent a nonlinear problem of estimating a set of constant parameters arranged in a vector satisfying a scalar linear measurement [8-11]. This approach can be extended to the time-varying vector variables, with the observations that are linearly dependent upon the variable to be estimated, [8-11]. Particularly, for the discrete time state vector estimation in question, the posed problem reduces to the step-by-step minimization of the modified M-robust performance index, [10]

$$J_k(\bar{x}_k) = E\left\{\rho\left[\varepsilon_k(\bar{x}_k)\right] \middle| \bar{x}_k, Y^k\right\};\ \varepsilon_k = y_k - H_k\bar{x}_k \quad (11)$$

where $E\{\cdot|\cdot\}$ denotes the conditional mathematical expectation of the nonlinearly transformed measurements residuals, $\varepsilon_k(\cdot)$, given the predicted state vector, $\bar{x}_k$, of the unknown systems state vector, x_k, at the discrete time, k, together with the measurement sequence $Y^k = \{y_1, y_2, ..., y_k\}$. Moreover, $\rho(\cdot)$ is a robust score, or loss, function that has to suppress the influence of outliers. Because it is frequently supposed that the measurement noise is confined to the normal distribution, a high efficiency at the normal noise is

also required. This is the feature of efficiency robustness. Thus, $\rho(x)$ should be look like the quadratic function, x^2, for small and moderate values of the argument, x. Furthermore, as a rule, the measurements data contain 5 to 10 percentage of outliers, [9]. For this reason, it is desirable that the first derivative, $\psi(\cdot)=\rho'(\cdot)$, called the influence function, has the bounded and continuous characteristics, [10, 11]. Bounded feature obeys the requirement that a particular outlier may not have a significant influence on estimates. On the other hand, continuity feature ensures that grouped, or patchy, outliers will not produce for large impact. This is the requirement for resistant robustness. The both requirements correspond to the saturation type nonlinearity, the so-called Huber's score function [10]. The corresponding influence function is the first derivate of the Huber's function, that is

$$\psi(x)=\min\left(|x|,\Delta\right)\operatorname{sgn}(x);\ \Delta=1.5 \tag{13}$$

Furthermore, the dynamic stochastic approximation algorithm can be applied for step-by-step minimization of the M-robust criterion (11), where the standard Kalman filter time-update recursions, (4), (5), are used to predict the system states at each stage k, yielding

$$\hat{x}_k=\bar{x}_k-\Gamma_k\nabla_{\bar{x}}J_k(\bar{x}_k);\ \nabla J_k\sim\frac{1}{d_k}\psi\left(\frac{\varepsilon_k}{d_k}\right)\Gamma_k H_k^T \tag{14}$$

with Γ_k being the matrix gain sequence, where the unknown gradient vector in (14) is replaced by the simple sample realization. In addition, the scale factor, d_k, represents some estimate of the standard deviation of measurement residual, ε_k, in (11), and a natural choice is to use the standard Kalman filtering relation (10) for its calculation. Finally, the gain matrix, Γ_k, in (14) represents a free quantity that influence the convergence property of the proposed robust estimator. Starting from the desired fast-tracking performances, Γ_k may be obtained by the minimization of an additional criterion

$$J_1(\Gamma_k)=TraceP_k;\ P_k=E\left\{(x_k-\hat{x}_k)(x_k-\hat{x}_k)^T\right\} \tag{15}$$

where *Trace* denotes the trace operation on the matrix. The derivation of the recursive relation for Γ_k requires convenient approximations, due to both the dynamic, multivariable, time-varying nature of the system and a nonlinear character of the robust state estimator by itself. The approximate optimal solution is given by, [19]

$$\Gamma_k=M_k=E\left\{(x_k-\bar{x}_k)(x_k-\bar{x}_k)^T\right\} \tag{16}$$

Thus, the proposed M-robust Kalman filter obeys the recursive predictor-corrector structure of the standard Kalman filtering. However, the prediction stage is defined by the standard Kalman filtering time-update recursions, (4), (5), while the correction stage is given by the M-robustified dynamic stochastic approximation algorithm, (10), (11)-(16). The corresponding recursions are given by, [19]

Time update, (4), (5):

$$\bar{x}_k=F_{k-1}\hat{x}_{k-1}+m_w(k);\ m_w(k)=E\{w(k)\} \tag{17}$$

$$M_k=F_{k-1}P_{k-1}F_{k-1}^T+G_{k-1}Q_{k-1}G_{k-1}^T \tag{18}$$

Measurement update, (10), (11)-(16):

$$\varepsilon_k=y_k-H_k\bar{x}_k+m_v(k);\ m_v(k)=E\{v(k)\} \tag{19}$$

$$d_k=S_k^{1/2};\ S_k=H_kM_kH_k^T+R_k \tag{20}$$

$$\omega_k=\begin{cases}\dfrac{\Psi\left(\frac{\varepsilon_k}{d_k}\right)}{\frac{\varepsilon_k}{d_k}} & \text{for } \varepsilon_k\neq 0 \text{ and } d_k\neq 0\\ 1 & \text{for } \varepsilon_k=0 \text{ and/or } d_k=0\end{cases} \tag{21}$$

$$K_k=\omega_kM_kH_k^TS_k^{-1} \tag{22}$$

$$\hat{x}_k=\bar{x}_k+K_k\varepsilon_k \tag{23}$$

$$P_k=(I-K_kH_k)M_k \tag{24}$$

As mentioned before, the M-robust approach is conservative and requires further adaptation to the underlying noise statistics.

IV. Adaptive Robustifying Estimation of Noise Statistics Using M-robust Approach Combined with Robust Median Filtering

Starting from the observation model (2), the measurement noise, v_k, can be estimated as

$$r_k=y_k-H_k\hat{x}_k \tag{25}$$

here $\hat{x}_k$ is the M-robust estimate (17)-(24). Taking into the account the assumptions in (3), together with the assumption of adequate initialization of the robust filter, resulting in the unbiased state estimates, one obtains from (2) and (25)

$$m_r(k)=E\{r(k)\};\ m_v(k)=E\{v(k)\};\ m_r(k)=m_v(k) \tag{26}$$

Additionally, the variance of the random varijable, r_k, is given by

$$V_r(k)=E\left\{\left[r(k)-m_r(k)\right]^2\right\}=H_kP_kH_k^T+R_k \tag{27}$$

where P_k is generated by (24). Furthermore, if the first and second order statistics, $m_r(k)$ and $V_r(k)$, are somehow estimated, say by $\hat{m}_r(k)$ and $\hat{V}_r(k)$, one can estimate further the unknown measurement noise statistics, $m_v(k)$ and R_k, from (26) and (27). In this way one obtains the following estimates

$$\hat{m}_v(k) = \hat{m}_r(k);\ \hat{R}_k = \hat{V}_r(k) - H_k P_k H_k^T \tag{28}$$

Similarly, the state noise sample, w_k, can be estimated from the state equation (1), where the unknown states, x_{k+1} and x_k are replaced by the two successive M-robust estimates, $\hat{x}_{k+1}$ and $\hat{x}_k$. A such obtained state noise estimate represents a scalar real random variable

$$q_k = T_k(\hat{x}_{k+1} - F_k\hat{x}_k);\ T_k = \left(G_k^T G_k\right)^{-1} G_k^T \tag{29}$$

The mean value of their random variable is given by

$$m_q(k) = E\{q_k\} = m_w(k);\ m_w(k) = E\{w_k\} \tag{30}$$

while its variable is defined by the relation

$$\begin{aligned} V_q(k) &= E\left\{\left[q_k - m_q(k)\right]^2\right\} \\ &= T_k\left(-P_{k+1} + F_k P_k F_k^T\right)T_k^T + Q_k \end{aligned} \tag{31}$$

with Q_k being the variance of the state noise sample, w_k. In deriving the expressions (30) and (31) are used the assumptions (3), under which the measurement noise, $\{v_k\}$, and the state noise, $\{w_k\}$, are uncorrelated stochastic sequences, that are mutually uncorrelated and are also uncorrelated with the random initial state, x_0. In addition, it is supposed that the adequate robust Kalman filter initialization produce the unbiased state estimates, $\{x_k\}$. Thus, if $\hat{m}_q(k)$ and $\hat{V}_q(k)$ denote some known estimates of $m_q(k)$ and $V_q(k)$, one can estimate the unknown statistics of the state noise, w_k, using relations (30) and (31). In this way, ones obtains the following estimates

$$\hat{m}_w(k) = \hat{m}_q(k);\hat{Q}_k = \hat{V}_q(k) - T_k\left(-P_{k+1} + F_k P_k F_k^T\right)T_k^T \tag{32}$$

The practical applications of the noise statistics estimations in (28) and (32) requires that the first and second order statistics, (26), (27), (30) and (31), of the random variables r_k and q_k, in (25) and (29), respectively, are estimated in advance. There are at least three approaches to perform this task. The first one is to use conventional sample mean and sample variance estimates, which generate the unbiased and consistent estimates of the mean value and variance under the condition that the random samples, r_k and q_k, are produced by the normal distribution, [3, 4]. However the sample mean and the sample variance lack robustness towards the outliers, [9-11]. A better robust solution is to use M-estimate of the location parameter of the unknown probability distribution, from which the data samples are generated, as an alternative to the sample mean. Moreover, an alternative to the sample variance should be based on the asymptotic formula for the variance of M-robust location parameter estimate, [10]. In this paper, we suggest the third approach based on the median estimators of the mean value and the variance of the unknown probability distribution to which the data samples are confined. Although ad hoc, these estimators are commonly used in engineering practice, due to their simplicity and efficiency, [9]. Thus, if $\{s_i,\ i = 1,...,L\}$, denotes the sample of the size L, from the underlying probability distribution, then the sample median represents the central sample within the rearranged data sequence in increasing order, if the sample size, L, is odd positive integer. On the other hand, if L is an even positive number, the median is given by the arithmetic mean of the two central samples within the sorted data sequence. Under the assumption that the underlying probability distribution function is symmetrical, the corresponding mean value, m, can be estimated by sample median, $\hat{m}(L)$, that is, [9-11]

$$\hat{m}(L) = median\{s_i,\ i = 1,...,L\} \tag{33}$$

In addition, the variance, V, of the distribution in question, can be estimated by $\hat{V}(L)$, representing the median of the absolute median deviations, or MAD estimator, [9, 10]

$$\hat{V}(L) = \left(\frac{median\{s_i - \hat{m}(L);\ i = 1,...,L\}}{0.6745}\right)^2 \tag{34}$$

The factor 0.6745 in (34) is used, because if the sample size, L, tends to infinity, and the data are generated by Gaussian distribution, then the statistics (34) converges towards the real unknown variance of the underlying normal distribution, [9, 10]. In this way, the robust median filters, (29), (34), can be applied for estimation of the unknown quantities, $m_r(k)$ and $V_r(k)$, in (26) and (27), using the sliding data frame, $\{r_i, i = k, k-1,...,k-L+1\}$, where the samples, r_i, are generated by (25). Then, the unknown measurement noise statistics, $m_v(k)$ and R_k, can be estimated by (28). In addition, the unknown quantities, $m_q(k)$ and $V_q(k)$, in (30) and (31), can be estimated by the median estimators (33) and (34), respectively, using the sliding data window,

$\{q_i, i = k, k-1, ..., k-L+1\}$, with q_i being generated from (29). Then, the unknown state noise statistics are calculated from (32).

In summary, the estimators for the first and second-order moments of the measurement and state noises are based upon the last L of the state, q_k, and observation, r_k, noise statistics at discrete time, k, generated by the M-robust Kalman filter (17)-(24). The process requires some extra storage, as well as sorting and shifting operations on the noise samples. However, since a common choice in practice is $L \in [3,10]$, there are no significant additional requirements on the computer time and memory storage. On the other hand, the second approach based on the M-robust estimation of location parameter of the underlying probability distribution entails much more work at each step, but usually not through to preclude the consideration, [17, 18]. Moreover, the noise variance estimators may become negative in numerical applications, especially at the beginning steps, and these quantities have to be reset to the absolute values of their estimates. Additionally, the noise samples (25) and (29) are poor indicators of the noise environment during the robust Kalman filter, (17)-(24), initialization. This difficulty, may be overcome by using the fading memory approach in which the successive noise samples are multiplied by a growing weighting factor

$$g_k = (k-1)(k-2)...(k-\gamma)/k^{\gamma};\ k = 1,2,... \tag{35}$$

In this way, the invalid noise samples are delayed for the first γ-stages, and g_k tends to 1 when k tends to infinity, [18].

V. Experimental Results

The proposed noise-adaptive M-robustified Kalman filter has been derived on the basis of heuristic reasoning, requiring further practical verifications. The figure of merit of the so obtained adaptive robust state estimator is presented by simulations. In the following characteristics examples, using the three-state track model with only position measurement is presented. Due to simplicity, the one-dimensional radar tracking problem, where the target is moving in the x-direction with the constant accelerator, is considered. The system dynamics is given by the kinematic equations of motion, [23].

$$\begin{aligned} s_t &= s_k + z_k(t-t_k) + a_k(t-t_k)^2/2 \\ z_t &= z_k + a_k(t-t_k) \\ a_t &= a_k \end{aligned} \tag{36}$$

where, s_k, z_k and a_k indicate the target position, velocity, and acceleration at discrete time, $t_k = kT$, $k = 0,1,2,...$, with T being the sample period ($T = 4s$), and $t_k \le t \le t_{k+1}$.

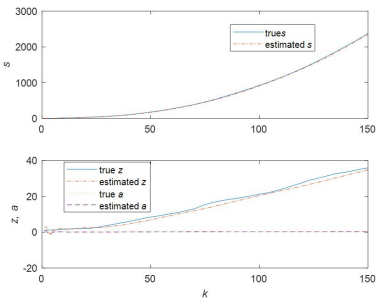


Fig. 1. True and estimated state obtained by the algorithm A1

Since by assumption only the target position is measured, the observations are produced by the equation

$$y_k = x_k + v_k \tag{37}$$

where the additive measurement noise, v_k, is generated from the heavy-tailed Gaussian pdf

$$\begin{aligned} &f(v) = (1-\varepsilon)N(v|0,1) + \varepsilon N(v|0,\sigma_0^2) \\ &0 < \varepsilon < 1;\ \sigma_0^2 \gg 1 \end{aligned} \tag{38}$$

with $N(\cdot|a,b)$ being Gaussian pdf with mean a and variance b. In monopulse radars this heavy-tailed behavior is presented because of the target glint, [23]. By using $t_k = t_{k+1}$ in (36), together with (37), the filter equations are given by

$$x_{k+1} = \begin{bmatrix} 1 & T & T^2/2 \\ 0 & 1 & T \\ 0 & 0 & 1 \end{bmatrix} x_k + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} w_k = Fx_k + Gw_k \tag{39}$$

$$y_k = [1 \quad 0 \quad 0]x_k + v_k = H_k x_k + v_k \tag{40}$$

where $x_k = [s_k, z_k, a_k]^T$ is the state vector and w_k is the disturbance input (state or process noise), which compensates for unmodelled system dynamics in (36). Namely, in the presence of target maneuver the state model (36) is inadequate, and a such model error is compensated by zero-mean white Gaussian noise term with unit variance, $Q = 1$. The performance of the adaptive robustified Kalman filter, denoted as A1, are compared with the standard Kalman filter, denoted as A2. Simulation results are compared in terms of the estimated noise statistics and the cumulative estimation error (CEE) criterion

$$CEE(k) = \frac{1}{k}\sum_{i=1}^{k}\frac{\|\hat{x}_i - x_i\|}{\|x_i\|} \tag{41}$$

where $\|\cdot\|$ is the Euclidean norm. Figure 1 depicts the true, x, and the estimated, $\hat{x}$, states obtained by the algorithm A1.

The comparison of the algorithms A1 and A2 through the *CEE* simulation results for different noise statistics are presented in Fig. 2.

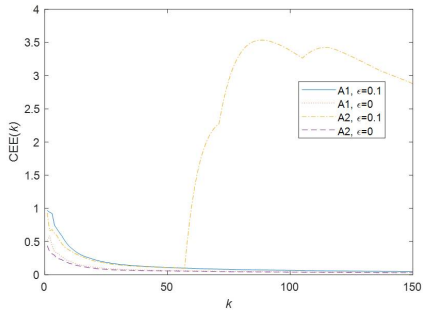


Fig. 2. Cumulative estimation error (*CEE*) of algorithms Alg1 and Alg2 for different noise statistics

Estimated noise statistics are given in Fig. 3. The simulation results have shown that proposed adaptive robustified Kalman filter algorithm, A1, can improve the state estimation performance of the standard Kalman filter, A2, when a priory statistics of the measurement noise are erroneous and there exist a significant dynamic model error, i.e. disturbance input or state noise, owing to the target maneuver.

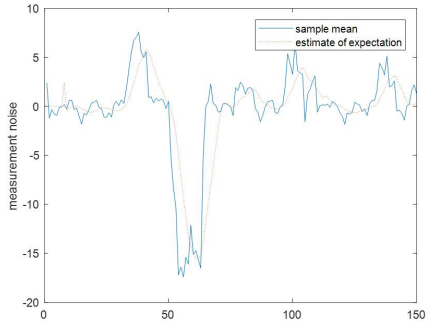


Fig. 3. Estimated noise statistics

VI. Conclusion

In this paper, a feasible Kalman filer modification that performs quite well in the presence of disturbance uncertainty and unmodeled system dynamics has been considered. It has been shown in a realistic simulation task of tracking maneuvering target that the proposed noise-adaptive robustified Kalman filter, based on M-robust dynamic stochastic approximation state estimation combined with robust median estimation of noise-statistics, simultaneously with the system states, can improve the state estimation performance in the presence of erroneous a priory noise statistics and significant dynamic model errors. The proposed adaptive robustified estimator may offer a desirable alternative to other adaptive approaches, particularly those that include state vector estimator of the unknown noise distribution. Finally, the emphasized adaptive robustified procedure in the one that, with some modifications, can be used whenever the least-squares or Kalman filter are used.

Acknowledgment

This paper is an outcome of activities under project TR32038 supported by Serbian Ministry of Education, Science and Technological Development,

References

[1] J. V. Candy, *Model-Based Signal Processing,* Hoboken, NJ: John Wiley & Sons, 2006.

[2] M. Verhaegen, V. Verdult, *Filtering and System Identification: a Least Squares Approach*, Cambridge: Cambridge University Press, 2012.

[3] B. D. Kovačević, Ž. Đurović, *Fundamentals of Stochastic Signals, Systems and Estimation Theory with Worked Examples*, Berlin, Springer, 2011.

[4] M. S. Grewal, A. P. Andrews, *Kalman Filtering: Theory and Practice using MATLAB*, Hoboken, NJ: Wiley, 2015.

[5] A. E. Albert, L. A. Gardner, *Stochastic Approximation and Nonlinear Regression*, Cambridge Mass: M.I.T. Press, 1967.

[6] G. Saridis, Z. Nikolic, and K. Fu, "Stochastic approximation algorithms for system identification, estimation, and decomposition of mixtures," *IEEE Transactions on Systems Science and Cybernetics*, vol. 5, no. 1, pp. 8–15, 1969. DOI: 10.1109/tssc.1969.300238

[7] J. M. Mendel, *Adaptive, Learning and Pattern Recognition Systems: Theory and Applications*," New York: Acad. Press, 2012.

[8] Ya.Z. Tsypkin, *Fundamentals of the Theory of Learning Systems*, Nauka, Moscow, 1970.

[9] V. Barnett, T. Lewis, *Outliers in Statistical Data,* Chichester: Wiley, 2000.

[10] P. J. Huber, E. Ronchetti, *Robust Statistics,* Hoboken, N.J: Wiley, 2009.

[11] R. R. Wilcox, *Introduction to Robust Estimation and Hypothesis Testing,* Amsterdam: Academic Press, 2012.

[12] M. A. Gandhi, L. Mili, "Robust Kalman filter based on a generalized maximum-likelihood-type estimator," *IEEE Transactions on Signal Processing*, vol. 58, no. 5, pp. 2509–2520, 2010. DOI: 10.1109/tsp.2009.2039731

[13] K. Li, B. Hu, L. Chang, and Y. Li, "Robust square-root cubature Kalman filter based on Huber's M-estimation methodology," *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering*, vol. 229, no. 7, pp. 1236–1245, 2014. DOI 10.1177/0954410014548698

[14] Y. Zou, S. Chan, and T. Ng, "Robust M-estimate adaptive filtering," *IEE Proceedings - Vision, Image, and Signal Processing*, vol. 148, no. 4, pp. 289-294, 2001. DOI 10.1049/ip-vis:20010316

[15] H. S. Kim, J. S. Lim, S. J. Baek, K. M. Sung, "Robust Kalman filtering with variable forgetting factor against impulsive noise," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol.E84-A, no. 1, 363-366, 2001

[16] T. Yang, J. Lee, K. Y. Lee, and K.-M. Sung, "On robust Kalman filtering with forgetting factor for sequential speech analysis," *Signal Processing*, vol. 63, no. 2, pp. 151–156, 1997. DOI 10.1016/s0165-1684(97)00150-3

[17] B. Kovačević, Ž. Đurović, and S. Glavaški, "On robust Kalman filtering," *International Journal of Control*, vol. 56, no. 3, pp. 547–562, 1992. DOI 10.1080/00207179208934328

[18] Ž. M. Đurović, B. D. Kovačević, "Robust estimation with unknown noise statistics," *IEEE Transactions on Automatic Control*, vol. 44, no. 6, pp. 1292–1296, 1999. DOI: 10.1109/9.769393

[19] Z. Banjac, Ž. Đurović , B. Kovačević, "Approximate Kalman filtering using robustified dynamic stochastic approximation method", *Telecommunications Forum (TELFOR)*, 2018 26th, Belgrade, Serbia, pp 353 –356, 20-21 Nov. 2018,

[20] P. D. Hanlon, P. S. Mayback, "Characterization of Kalman filter residuals in the presence of mismodeling," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 36, no. 1, pp. 114-131, 2000. DOI 10.1109/7.826316

[21] C. Price, "An analysis of the divergence problem in the Kalman filter," *IEEE Transactions on Automatic Control*, vol. 13, no. 6, pp. 699–702, 1968. DOI 10.1109/tac.1968.1099031

[22] T. Young, R. Westerberg, "Error bounds for stochastic estimation of signal parameters," *IEEE Transactions on Information Theory*, vol. 17, no. 5, pp. 549–557, 1971. DOI 10.1109/tit.1971.1054696

[23] S. S. Blackman, R. Popoli, *Design and Analysis of Modern Tracking Systems,* Norwood, MA: Artech House, 1999

Design, Analysis, Validation, and Reporting of Continuous-Time Systems Using CAS

Miroslav Lutovac, *Senior Member, IEEE*, Maja Lutovac-Banduka, and Aleksandra Pavlović

***Abstract*—The automated symbolic manipulations are presented according to user preferences in such a way that all representations (mathematical, graphical, as net-list, software code, or time-domain and frequency domain responses) are obtained from the same visual system description using computer algebra system (CAS) as add-ons for extending software environment. The paper is devoted to researchers and scientist using basic electrical engineering tasks, so that time-consuming tasks are automated in software and the properties of the systems can be discovered, and optional conditions or discovered properties can be used for synthesis, verifications, simulations, and optimization with real parameters. All derived properties are available as closed-form relations, which can help the faster design of robust systems.**

***Index Terms* — Computer algebra systems, electrical engineering, graphical user interface, symbolic processing.**

I. Introduction

IN the process of designing contemporary electric systems usually we start with system specification and deal with system modeling, analysis, software and hardware synthesis and implementations, simulation, verification, tolerance centering, fault detections, design space optimization, and user interface (more details are available in [1] and [2]). We can use numeric or symbolic software environments to design and implement electronic systems. The symbolic software environment as CAS (Computer Algebra System) [3] is the best initial step because expressions from textbooks or published papers can be simply redrawn or rewritten in CAS environment. The programed functions appear in the same traditional forms as in textbooks. Also, the system represented by formulas can be transformed into realizable program or hardware chips, the hardware realization can be transformed into programs and expressions. It is easily to go between hardware descriptions, software implementations, and traditional mathematical expressions. Optimization, verification, and testing the accuracy are available at each design step between the system specification and implementation.

The unified hardware/software approach is using symbolic processing for more than a decade for the design, analysis, verification, and synthesis ([1], [4], and [5]). In spite of the superiority of the symbolic processing, the prevalent software for the design, analysis, verification, and synthesis is still numeric processing.

The purpose of this paper is to exemplify the benefits of symbolic approach. This paper is the third paper in series on this topic: the first part [6] is presenting the development process of GUI (Graphical User Interface) in Mathematica using modern trends in visual programing and some new functions of Mathematica, following the main features of Wolfram language; the second paper [7] is presenting derivation of mathematical representations from schematic description in a form of frequency responses, or as a set of equations and software for the system synthesis. Instead of using textual description of the system, the visual programming technique is used.

This paper starts with derived expressions in [7] and explains analysis and optimization steps.

It should be noticed that the usage of some of the popular graphical programming packages for teaching electrical engineering was presented in [8]. The importance of the visualization of programing in robotics was explained in [9].

II. GUI in CAS environment

GUI in CAS environment is used for visual programing in such a way that the system specification is generated as lists of element specifications by clicking the appropriate buttons. New element can be added into the lists as new function that has all drawing command for all possible node positions. One, two, and three port elements can be added into schematic by clicking only one or two points on the drawing table (see Figures 1, 2, and 3 in [6]). The distinctive feature of the GUI is that the third node is automatically added assuming the most frequently used element position. The default scale is automatically chosen so that element looks as in the most textbooks. The connection lines between the element and the circuit nodes are the minimal. Tooltip function displays label as a tooltip while the mouse pointer is in the area where the element is drawn; label contains element name and coordinates of all ports.

It is possible to draw smaller elements by specifying the element scale property, but not larger than that allowed by element node positions. For testing all positions of the multi-port elements, a special test-code was used with all possible positions in the drawing table.

For different irregular element positions, when mouse is

Miroslav Lutovac is with The School of Electrical and Computer Engineering of Applied Studies Belgrade, 283 Vojvode Stepe, 11000 Belgrade, Serbia, (e-mail: mlutovac@viser.edu.rs).

Maja Lutovac-Banduka is with RT-RK Institute for Computer Based Systems, Novi Sad, Serbia (e-mail: majalutovac@yahoo.com).

Aleksandra Pavlović is with The School of Electrical and Computer Engineering of Applied Studies Belgrade, 283 Vojvode Stepe, 11000 Belgrade, Serbia, (e-mail: aleksandrap@viser.edu.rs).

over that element in the drawing table, the corresponding tooltip shows what is wrong.

III. Transformation of Schematics with Multi-Port Elements into Schematic with Basic Elements

In [7], it is demonstrate the transformation of multi-port elements into a combination of connected two-port elements, and automated generation of system of equations that can be solved using the basic Ohm's law and Kirchhoff's current and voltage laws. As an example schematic is used the schematic that is successfully implemented in many applications as robust solution; that is the general-purpose second order KHN filter section (Kerwin-Huelsman-Newcomb). The well-known schematic is modified so that a new opamp is added for implementing transfer function zero. The new properties of the modified filter section are not documented in the textbooks, and the designer should solve the system. The GUI is used for visual programing (actually for drawing the schematic), the multi-port elements are replaced with standard two-port elements, and the system of equation is automatically derived for setting up circuit equations.

The derivation of the transfer function, of filter section presented in Figure 1, can be a serious problem even for experienced designers when the amplification A is of finite value or given as an expression. The whole procedure is explained in [7]. A part of the example schematic specification is as follows:

```
mySchematic = {
   {"VoltageSource",{{3,7},{3,9}},vin,""},...
   {"Resistor",{{14,11},{17,11}},r5,""}, ...
   {"Capacitor",{{17,13},{21,13}},c6,""},...
   {"OpAmplifier",{{17,9},{17,11},{21,11}},A,a3},...
   {"Node",{21,11},VB,"",TextOffset→{1,-1}},...
   {"Node",{17,11},V17p11,"",TextOffset→{1,1}},...
   {"Node",{17,9},V17p9,"",TextOffset→{1,-1}},...,
   {"GroundE",{17,9},0,"",ElementSize→{1.5,1.5}},...
   {"Output",{35,11},Yx,"",ElementSize→{2.5,2.5}},...
   {"Line",{{17,11},{17,13}}},
   {"Line",{{21,11},{21,13}}},...
};
```

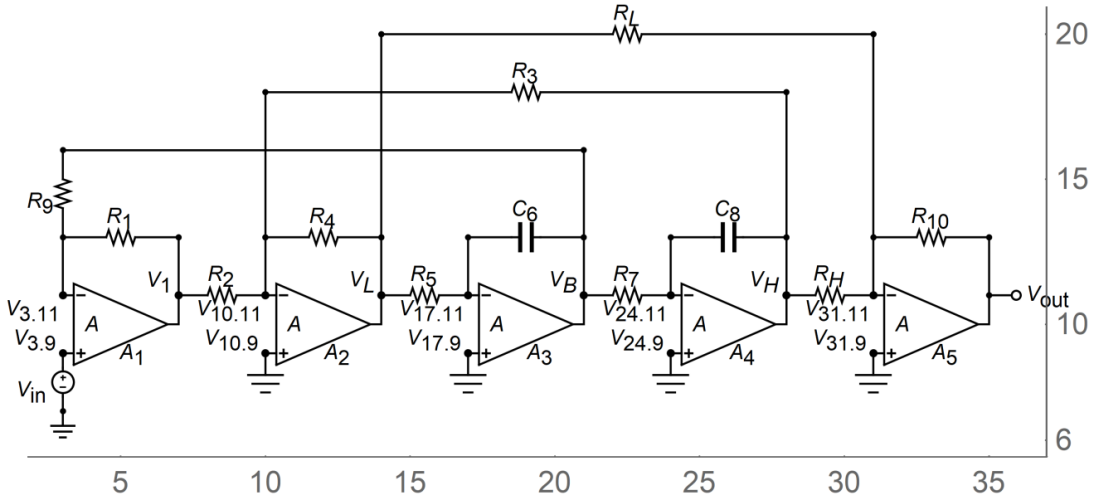


Fig. 1. Modification of KHN filters section.

Coefficients of transfer function denominator	
s^0	1
s^1	$\frac{C_8 R_7 \left(R_1 R_4 \left(A^3 R_3 + R_2 + R_3\right) + (A+1) R_1 R_2 R_3 + (A+1) R_9 \left((A+1) R_2 R_3 + (R_2 + R_3) R_4\right)\right) + C_6 R_5 \left(R_2 \left(A R_3 + R_3 + R_4\right) + R_3 R_4\right) (A R_9 + R_1 + R_9)}{R_9 \left(R_2 \left(A^3 R_4 + A R_3 + R_3 + R_4\right) + R_3 R_4\right) + \left((A-1) A + 1\right) R_1 (R_2 + R_3) R_4 + R_1 R_2 R_3}$
s^2	$\frac{(A+1) C_6 C_8 R_5 R_7 \left(R_2 \left(A R_3 + R_3 + R_4\right) + R_3 R_4\right) (A R_9 + R_1 + R_9)}{R_9 \left(R_2 \left(A^3 R_4 + A R_3 + R_3 + R_4\right) + R_3 R_4\right) + \left((A-1) A + 1\right) R_1 (R_2 + R_3) R_4 + R_1 R_2 R_3}$

Coefficients of transfer function numerator	
s^0	1
s^1	0
s^2	$\frac{C_6 C_8 R_5 R_7 R_L}{R_H}$

Fig. 2. Derived coefficients of modified KHN filters section.

After transformation of the schematic elements into two-port elements, we can use the basic Ohm's law and Kirchhoff's current and voltage laws for setting up circuit equations. For each resistor or capacitor element we can use generalized Ohm's law. For each node that is a connection to at least two elements, we can setup equations using Kirchhoff's current law. For the voltage sources between two nodes we can use Kirchhoff's voltage law. The system of equations consists of 27 equations with 27 variables. The final results cannot be presented in viewable format. The transfer function coefficients are presented in Figure 2.

Assuming that amplification of the amplifiers is infinitive, A→∞, the simplified transfer function of the second-order function is obtained using limiting value of the expression when *A* approaches ∞. The simplified transfer function can be used for the filter synthesis. In Figure 3 we have used the traditional type of displaying the derived result.

$$\frac{R_3\,R_4\,(R_1+R_9)\,R_{10}\left(C_6\,C_8\,R_5\,R_7\,s^2\,R_L+R_H\right)}{R_H\,R_L\left(C_6\,C_8\,R_2\,R_3\,R_5\,R_7\,R_9\,s^2+C_8\,R_1\,R_3\,R_4\,R_7\,s+R_2\,R_4\,R_9\right)}$$

Fig. 3. The derived transfer function of KHN filter section for A→∞.

IV. Determining Design Constrains

In this section we have demonstrated how the initial schematic and the automated deriving of the transfer functions help us to choose the most preferred values for implementation.

For known transfer function and the circuit that has to implement it, we can prepare synthesis procedure. In this case the transfer function that we have to implement can be low-pass H_{LP} or low-pass-notch H_{LPN} with four symbolic parameters K_0, ω_z, ω_p, Q_p:

$$H_{LP}=K_0\frac{s^2}{s^2+\frac{\omega_p}{Q_p}++\omega_p^2} \tag{1}$$

$$H_{LPN}=K_0\frac{s^2+\omega_z^2}{s^2+\frac{\omega_p}{Q_p}++\omega_p^2} \tag{2}$$

For the 6th order filter, on disposal are 12 values that we have to determine to implement the filter section with the transfer function. Most values can be arbitrary chosen. As the first attempt, let us choose R_3, R_4, R_9, and R_H as variables that we can determine from the filter transfer function. All other values we would like to specify arbitrary. From the schematic we can determine relations between four parameters K_0, ω_z, ω_p, Q_p and four resistances. For example,

$$R_3=R_2\frac{-1+C_8Q_pR_7\omega_p}{C_8Q_pR_7\omega_p} \tag{3}$$

The value of the resistance R_3 can be negative in the case when *Q* factor is too low:

$$C_8Q_pR_7\omega_p<1 \tag{4}$$

The lowest *Q* factor for elliptic-type filters is 0.5. Therefore we can determine constrains for the resistor R_7:

$$R_7>\frac{2}{C_8\omega_p} \tag{5}$$

The design procedure is as follows assuming that R_x and C_x are from the set of the most preferred values for the implementation:

$$\begin{aligned}&R_x>\frac{2}{C_x\omega_p}\\&R_1=R_2=R_5=R_7=R_{10}=R_L=R_x\\&C_6=C_8=C_x\end{aligned} \tag{6}$$

The next step is to determine the design procedure. The design procedure means that we have to automatically derive a code for computing all filter values. Again, comparing the transfer function in terms of filter parameters (K_0, ω_z, ω_p, Q_p) and the transfer function in terms of resistances and capacitance, we can obtain all implementation values as a code in the Wolfram language [1].

V. Design of Higher-Order Filters Using Known Schematic

Once the circuit is well analyzed and design procedure is available, we can use as embedded solution in higher order systems.

Suppose we have to implement low-pass filter that satisfies the following conditions: the passband edge frequency F_p=3000 Hz, the stopband edge frequency F_s=4500 Hz, the maximal passband variation A_p=0.2 dB, the minimal stopband attenuation A_s=40 dB.

The minimal filter order of the most efficient elliptic filter is 5. Since we are planning to use the same second-order filter section, we will choose the sixth order approximation. Thera are many solutions from the design space [1]. We choose the minimal Q factor elliptic filter [1] with the selectivity factor 1.635, and the frequency normalization 1.1 (move the passband edge to the transition region). This approximation was robust as digital filters [1], and we are expecting that the analog filter can be also robust to many parasitic effects and imperfections.

The filter parameters are for three sections:

	H_3	H_2	H_1
ω_p^2	$4.8016\ 10^8$	$4.8016\ 10^8$	$4.8016\ 10^8$
Q_p	4.24831	1.13453	0.557171
ω_z^2	$9.48906\ 10^9$	$1.40637\ 10^9$	$8.29491\ 10^8$

Since all ω_p are the same, we can choose the referent capacitance C_x=10 nF. The minimal value from design constrains is $R_{x,min}$= 9127 Ω: we chose the large value R_x=10 kΩ. All three sections have the same value for resistances R_1=R_2 =R_5=R_7=R_{10}=R_L =10 kΩ and capacitances C_6= C_8=10 nF. Other resistances are computed using design procedure for known filter parameters:

	H_3	H_2	H_1
R_3 [Ω]	8930	5980	1810
R_4 [Ω]	42900	28700	8690
R_9 [Ω]	83500	14900	2210
R_H [Ω]	82900	141000	949000

In this section we have demonstrated how the initial schematic, the automated deriving of the transfer functions from the schematic description, and derived design procedure, can be used to design higher order filter as cascaded connection of initial second-order filter sections.

VI. Verification of the Magnitude Response of Designed Circuit

In the verification step, we would like to obtain the same magnitude response with computed resistances and capacitances as the taken approximation. In the derived transfer function in terms of resistances and capacitances we are replacing computed values that are within 0.1% tolerances. The magnitude responses of the theoretical and implemented filter are presented in the Figures 4 and 5.

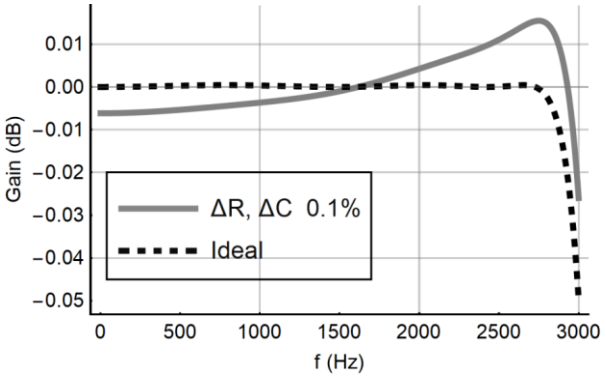


Fig. 4. Pass-band magnitude response of theoretical and implemented filter.

The pass-band variation is within required specification and significantly smaller than 0.2 dB. The minimum stopband attenuation is approximately 40 dB.

VII. Study of Imperfection

The purpose of the study of imperfection is to identify the critical components that can harm the specification and regular operation. In this section we will consider the finite gain of the amplifier as an example. The magnitude responses of the theoretical and implemented filter are presented in the Figures 6 and 7.

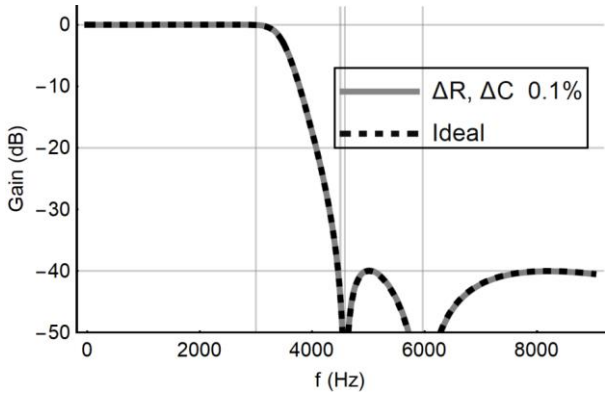


Fig. 5. Stop-band magnitude response of theoretical and implemented filter

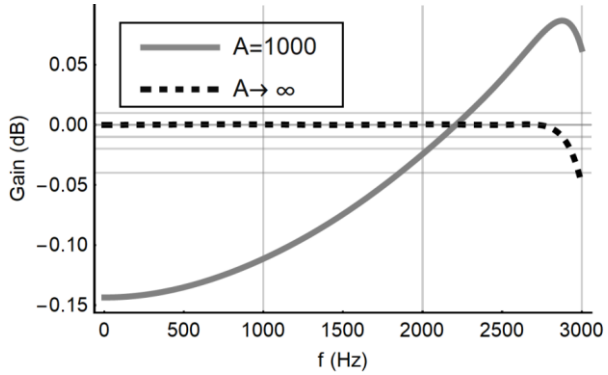


Fig. 6. Pass-band magnitude response of theoretical and implemented filter for A=1000.

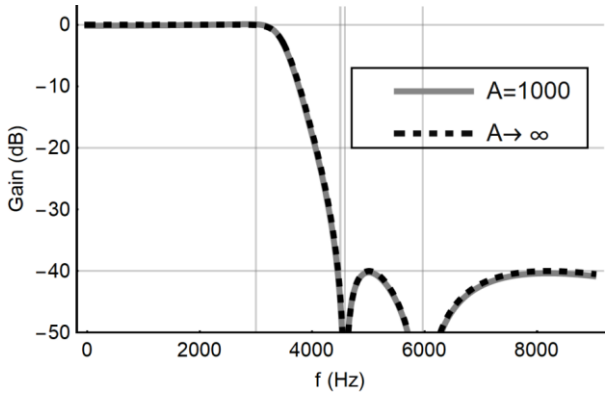


Fig. 7. Stop-band magnitude response of theoretical and implemented filter for A=1000.

The pass-band variation is slightly larger than the required specification 0.2 dB. The minimum stopband attenuation is approximately 40 dB. This implies that for implementation is required the opamp with larger gain then 1000.

VIII. Optimization and tuning procedure

The purpose of the optimization step is to optimize some values for the optimal values using different approaches. One possibility is to find referent resistance R_x for minimal gain-sensitivity product. Since all expressions are available as closed-form relations, it is simple to derive all sensitivity functions and to find the minimal value using procedures from [1]. In practice, many elements can be modeled using manufacturer instructions, and that models can be used to replace ideal values. For example, the gain of the amplifier can be modeled as single pole function that is frequency dependence. Replacing this model into derived transfer function will lead to increasing the order of the transfer function. This is not serious problem when we are using CAS, because all substitutions are in closed form and in the overall response we can identify the expression of the second-order section. From that expression we can discover the influence of R_3, R_4, R_9, and R_H as variables on four parameters K_0, ω_z, ω_p, Q_p.

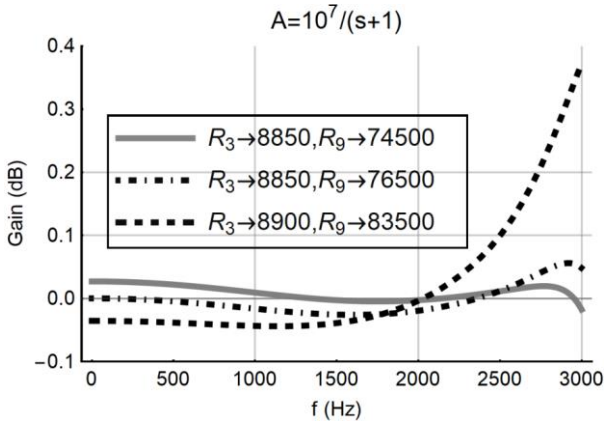


Fig. 8. Pass-band magnitude response of implemented filter for different values of R_3 and R_9.

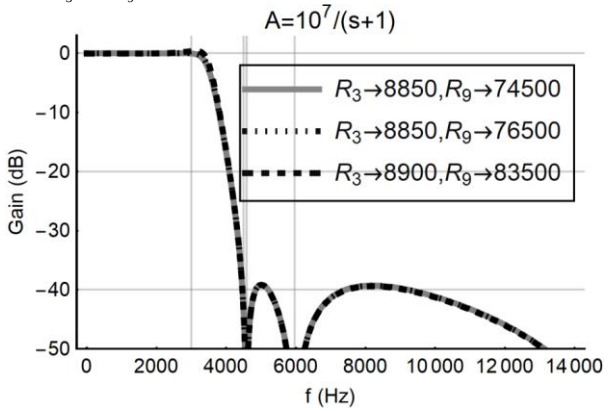


Fig. 9. Stop-band magnitude response of implemented filter for different values of R_3 and R_9.

Those resistances can be used for discovering tuning procedure. As an example, the magnitude responses of the implemented filters for several values of R_3 and R_9 are presented in the Figures 8 and 9.

After tuning R_3 and R_9, the pass-band variation is within required specification 0.2 dB. The minimum stopband attenuation is slightly smaller than 40 dB. This implies that the optimization or tuning procedure is possible by computing the new values of several resistances (for example, R_3 and R_9).

IX. Conclusion

In this paper, we have developed an environment and graphical user interface (GUI) so that the system description can be generated with the small number of tasks. The software is not like usual canned applications and it is possible to combine system description with many other specific user targets such as manipulate with symbolic expressions and analyze complex non-ideal element and system models.

Acknowledgment

This work is supported in part by the Ministry of Education and Science of Serbia under Grant TR 32023

References

[1] M. Lutovac, D. Tosic, and B. Evans, *Filter Design for Signal Processing Using MATLAB and Mathematica*, Upper Saddle River, Prentice Hall, 2001.

[2] B. Kleinjohann, G. R. Gao, H. Kopetz, L. Kleinjohann, and A. Rettberg, *Design Methods Applications Distributed Embedded Systems*, IFIP 18th World Computer Congress, Springer Publishing Company, 2013.

[3] S. Wolfram, *An Elementary Introduction to the Wolfram Language*, Champaign, Wolfram Media, IL 2015.

[4] K. Strehl, *Symbolic Methods Applied to Formal Verification and Synthesis in Embedded Systems Design*, Swiss Federal Institute of Technology Zurich, 2000.

[5] B. L. Evans and J. H. McClellan, "Symbolic Analysis of Signals and Systems," in *Symbolic and Knowledge-Based Signal Processing*, (eds. A. V. Oppenheim and S. H. Nawab), Prentice Hall, pp. 88-141, 1992.

[6] M. D. Lutovac, V. Mladenović, and M. Lutovac-Banduka , "Graphical User Interface for Electrical Engineering Systems using Wolfram Language," TELFOR, Belgrade, Serbia, pp. 1-4, 22-23 Nov. 2016.

[7] M. D. Lutovac, M. Lutovac-Banduka, and V. Mladenović, "Environment and Graphical User Interface for Design of Continuous-Time Systems," IcETRAN, Palić, Serbia, pp. 802-805, 11-14 Jun. 2018.

[8] E. Lunca, S. Ursache, and O. Neacsu, "Graphical Programming Tools for Electrical Engineering Higher Education," *International Journal of Online Engineering*, vol. 7, no. 1, pp. 19-24, iJOE, 2011.

[9] M. Lutovac-Banduka, "Robotics First – A Mobile Environment for Robotics Education," *International Journal of Engineering Education*, vol. 32, no. 2A, pp. 818–829, 2016.

Implementation of IIR Digital Filters with Variable Characteristics in GNU Octave

Darko Đ. Vračar

***Abstract*—Implementation of IIR digital filters with variable characteristics in GNU Octave software is presented. The goal is to demonstrate efficient usage of free open-source software instead of the commercial ones. The method for tuning the cut-off frequency with one parameter is based on series expansion of the low-pass-low-pass frequency transformation. The filter structure is a parallel connection of real or complex all-pass sections. The simulation results are matched with results in the references. In addition, simulation of the example elliptic filters in time domain is shown which verifies their design.**

***Index Terms*— All-pass, IIR, free software, GNU Octave implementation, recursive digital filters, variable characteristics.**

I. Introduction

MAIN function of electric filters is to adjust signal spectrum. In many cases such adjustment of signal spectrum comes down to allowing signal to pass in some frequency range (i.e. band-pass) and attenuation of signals that are outside of that range (i.e. band-stop) [1].

In many applications it is desirable to change frequency characteristics of a digital filter during its operation. This is subject of research in last several decades [2]-[6]. For example, filters with variable characteristics are used in: telecommunications, digital audio equipment, medical electronics, radar, sonar and control systems, adaptive and tracking systems, spectrum and vibration analyses, and in laboratory instruments [5].

There are many methods how to change filter frequency characteristics. A good review and systematization of such methods is given in [5]. Method used in this paper is based on Taylor series expansion of the low-pass-low-pass (LP→LP) frequency transformation [2]. The filter structure is a parallel connection of real or complex all-pass sections.

The recursive digital filter is a discrete system that is, in transformation domain, described with equation [1]

$$\mathrm{H(z)}=\frac{\mathrm{Y(z)}}{\mathrm{X(z)}}=\frac{\sum_{k=0}^{M} b_k \cdot z^{-k}}{1+\sum_{k=1}^{N} a_k \cdot z^{-k}}. \qquad (1)$$

The (1) is valid for linear time-invariant discrete systems. Note that the IIR (infinite impulse response) filter is often called a recursive one [2]-[4]. One has to keep in mind that these two terms are not synonyms although IIR filters are always realized as recursive ones [1].

In this paper, implementation of IIR digital filters with variable characteristics in free open-source software *GNU Octave* [7] is presented. In addition, for signal processing tasks, one had to use its package named *Signal* [7]. The similar results one could also get by using commercial software like *Matlab*™ and its *Signal Processing Toolbox*™. The goal of this paper is to demonstrate efficient usage of the free open-source software for such applications thus avoiding going for the commercial ones. Author's wish is to encourage audience to use more free software so the filter drawings are done by using free software as well [8].

This topic about tunable digital filters with relevant, although old, references was chosen as appropriate for such a demonstration. The task was to simulate all four transformations: low-pass-low-pass (LP→LP), low-pass-high-pass (LP→HP), low-pass-band-pass (LP→BP) and low-pass-band-stop (LP→BS) of a starting LP filter.

II. Theoretical Background

In this section only minimum theoretical background is given that was used for the software implementation. For more details the reader shall refer to [2]-[4]. One shall note that majority of the sentences in this chapter are either paraphrased or directly copied from [2]-[4]. Similar is valid for formulas as well.

If we mark transfer function of a digital filter as *H(z)*, adjustment of its frequency response can be made by using transformation of Constantinides [2], [5]-[6]

$$z^{-1} \rightarrow T(z) \qquad (2)$$

which gives us a new transfer function

$$H'(z) = H(z)\big|_{z^{-1}=T(z)} \qquad (3)$$

whose frequency response $H'(e^{j\omega})$ is appropriate transformed version of the original frequency response $H(e^{j\omega})$.

Let us have a look at the transfer function of recursive filter of N^{th} order given with (1). For LP→LP transformation, the *T(z)* function is an all-pass one of the 1^{st} order, and is given as [2]

$$T_1(z) = \frac{z^{-1}-\beta}{1-\beta \cdot z^{-1}} \qquad (4)$$

where

Author, at the time when similar results were initially obtained, was with the School of Electrical Engineering, University of Belgrade, Serbia. Author is now with Huawei Technologies Duesseldorf GmbH, Riesstr. 25, 80992 Munich, Germany; (e-mail: vracard@eunet.rs).

$$\beta = \sin\left(\frac{\Theta_P - \omega_P}{2}\right) \Big/ \sin\left(\frac{\Theta_P + \omega_P}{2}\right). \qquad (5)$$

Here Θ_P is cut-off frequency of the initial filter (i.e. prototype filter), and ω_P is the desired cut-off frequency. If $\beta=0$, then (4) reduces to delay operation. This transformation is not linear regarding ω.

Modified frequency transformation of the LP→BP type for recursive filter has form of a 2nd order all-pass function

$$T_2(z) = -z^{-1}\left(\frac{z^{-1} - \alpha}{1 - \alpha \cdot z^{-1}}\right) \qquad (6)$$

where

$$\alpha = -\cos(\Theta_C). \qquad (7)$$

Here Θ_C is the desired central frequency of the BP (BS) filter. It is calculated as a difference between upper and lower band-pass (band-stop) boundaries divided by 2.

The general form of these frequency transformations in discrete domain is [1]

$$T(z) = e^{j \cdot n \cdot \pi} \prod_{i=1}^{m} \frac{z^{-1} - a_i^*}{1 - a_i z^{-1}} \qquad (8)$$

where n and m are integers, whereas a and a^* are conjugate-complex constants. One can see that in (4) $m=1$, and in (6) $m=2$. This transformation does not have influence on stability of the transfer function.

The reader shall notice that (6) differs from *(23)* in [2] in sense that, in this paper, differences in nominator and denominator were used (i.e. not the sum as in [2]). This is what worked in software implementation and it is in accordance with (8) as well as with *(33)* in [5]. That was probably just a typo in [2].

This method uses, as basic structure, parallel connection of all-pass filter sections [2]. By introducing standard LP→LP transformation in these sections and with decomposition into Taylor series on control variable, one gets filter structure that is possible to realize. With only one parameter it is possible to control cut-off frequency.

A. Structures for Filter Realizations

The all-pass transfer function of N^{th} order with complex coefficients is given as

$$\varepsilon_N(z) = \frac{z^{-N} + a_{N-1} z^{-N+1} + \ldots + a_1 z^{-1} + a_0}{1 + a_{N-1}^* z^{-1} + \ldots + a_1^* z^{-N+1} + a_0^* z^{-N}} \qquad (9)$$

and for all-pass transfer function with real coefficients the $A_N(z)$ coefficients $\{a_i\}$ are the real ones. In [2] is shown that a big class of transfer functions with real coefficients can be realized as parallel connection of all-pass filters with real or complex coefficients as shown in Fig. 1 [2].

This type of structure has low sensitivity in band-pass and requires only N multipliers for N^{th} order filter. All standard low-pass filters of odd order (e.g. Butterworth, Chebyshev, elliptic, etc.) can be realized in this form [2].

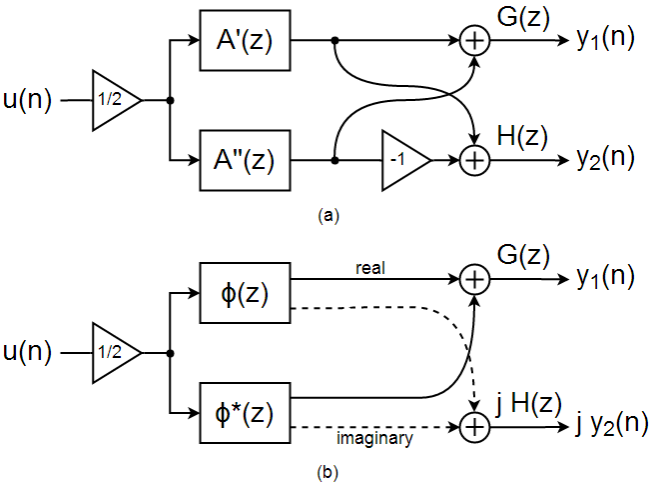


Fig. 1. Realization of a double-complementary filter pair. (a) Filters derived from LP transfer function of odd order have form of two, parallel connected, real all-pass sections. (b) Filters based on LP transfer function of even order are implemented as parallel connected complex all-pass sections.

1) Realization of Odd-Order Filters

At Fig. 1 one can see that at outputs one has complementary filter pairs. It is possible to show that, for LP-HP pair of N^{th} order, all-pass sections $A'(z)$ and $A''(z)$ are of "$<N/2$" and "$>N/2$" order, respectively; where "<" and ">" denote first lower and first bigger integer than $N/2$.

There are many structures for implementation of all-pass sections of 1st and 2nd order. At Fig. 2 two examples are shown that require only one and two multipliers per 1st and 2nd order sections, respectively [2]. They are used in this paper for software implementation of odd-order filter.

Derivation of $A'(z)$ and $A''(z)$ is given in [3] and I will repeat it here in short form for sake of paper completeness. Let us have a look at the N^{th} order transfer function of IIR filter

$$G(z) = \frac{P(z)}{D(z)} = \frac{p_0 + p_1 z^{-1} + \ldots + p_N z^{-N}}{1 + d_1 z^{-1} + \ldots + d_N z^{-N}} \qquad (10)$$

where coefficients p_i and d_i are real ones. Now let us assume that the structure is such that, despite real values of multipliers m_i, the absolute value $|G(e^{j\omega})|$ is bounded above by a fixed constant, e.g. 1:

$$|G(e^{j \cdot \omega})| \le 1 \quad \text{for all } \omega. \qquad (11)$$

The stable function $G(z)$, with real coefficients, that satisfies (11) is called bounded real function. If expression (11) is valid, with equal sign, for every ω then the $G(z)$ is called bounded real transfer function without losses. It is more known as stable all-pass function.

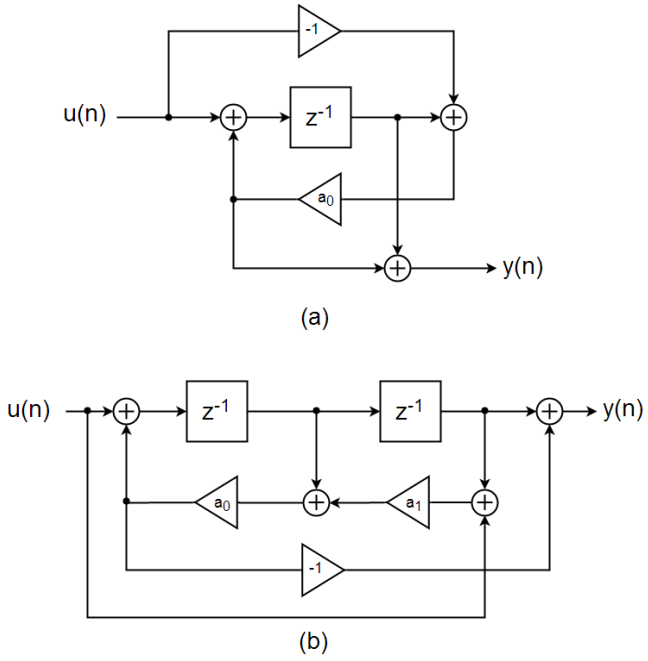


Fig. 2. All-pass structures (blocks): (a) 1st order, (b) 2nd order.

Let us consider typical bounded real LP transfer function of N^{th} order like in (11). Let us assume that $P(z)$ has linear phase. This is typical for majority of transfer functions of digital filters (because zeroes are usually located at unity circle) and, therefore, this is just mild boundedness. In order to be more precise, let $P(z)$ be symmetrical, i.e. $p_k=p_{N-k}$. Let us now look at another transfer function $H(z)$ given as

$$H(z)=\frac{Q(z)}{D(z)}=\frac{q_0+q_1z^{-1}+\ldots+q_Nz^{-N}}{1+d_1z^{-1}+\ldots+d_Nz^{-N}} \tag{12}$$

where $H(z)$ is defined so that applies

$$|H(e^{j\omega})|^2=1-|G(e^{j\omega})|^2. \tag{13}$$

In another words, the $H(z)$ is complementary to $G(z)$. From this, regarding z variable, we have following expression

$$P(z^{-1})\cdot P(z)+Q(z^{-1})\cdot Q(z)=D(z^{-1})\cdot D(z). \tag{14}$$

Let us now make a non-trivial assumption: $G(z)$ is such so that $Q(z)$, which satisfies (14), is anti-symmetric, i.e. $q_k=-q_{N-k}$. After several mathematical transformations we are coming to the expression for $A'(z)$ and $A''(z)$:

$$A'(z)=\prod_{k=r+1}^{N}\left(\frac{z^{-1}-z_k^{-1}}{1-z^{-1}\cdot z_k^{-1}}\right)$$
$$A''(z)=\prod_{k=1}^{r}\left(\frac{z^{-1}-z_k}{1-z^{-1}\cdot z_k}\right) \tag{15}$$

where r is number of zeros of the polynomial $P(z)+Q(z)$ that are located within unity circle. The calculation of $Q(z)$ is given in [3]. At the end we are coming to the expression for $G(z)$ and $H(z)$:

$$G(z)=\frac{1}{2}[A'(z)+A''(z)] \tag{16}$$

$$H(z)=\frac{1}{2}[A'(z)-A''(z)]. \tag{17}$$

This brings implementation of $G(z)$ and $H(z)$ as parallel combination of stable all-pass functions. If, e.g., $G(z)$ is a LP function then $H(z)$ is the HP one. Following is valid as well:

$$\begin{aligned}G(z)+H(z)&=A'(z)\\G(z)-H(z)&=A''(z)\end{aligned}. \tag{18}$$

If functions G(z) and H(z) are fulfilling (13) and (18) then one can say that they are double complementary.

2) Realization of Even-Order Filters

Let $G(z)$ and $H(z)$ be given as in (10) and (12), respectively, and let they form a double-complementary pair. It is shown in [2] that, when we are working with real input signals, we can use more efficient structure for realization of the even-order filter, and not the structure from Fig. 1(b). Such a structure is given on Fig. 3 [2].

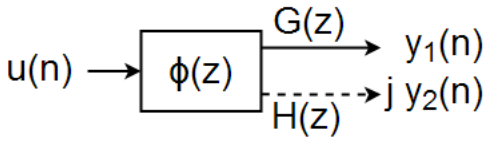


Fig. 3. A structure for implementation of even-order filter that uses one complex all-pass section.

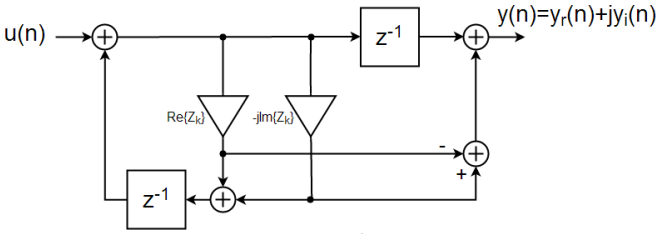


Fig. 4. The all-pass complex section of 1st order.

For the implementation of the complex all-pass section of 1st order the structure from Fig. 4 was used [4].

B. Control of Cut-Off and Central Frequency

Let us have a look at complex all-pass filter like in (9). By using LP→LP transformation of (4) type in $\varepsilon_N(z)$ we are coming to transformed all-pass section

$$\varepsilon_N'(z)=\gamma\cdot\frac{z^{-N}+\hat{a}_{N-1}z^{-N+1}+\cdots+\hat{a}_1z^{-1}+\hat{a}_0}{1+\hat{a}_{N-1}^*z^{-1}+\cdots+\hat{a}_1^*z^{-N+1}+\hat{a}_0^*z^{-N}} \tag{21}$$

where

$$\gamma=\frac{1+\sum_{i=0}^{N-1}(-1)^i a_i\beta^{N-i}}{1+\sum_{i=0}^{N-1}(-1)^i a_i^*\beta^{N-i}}. \tag{22}$$

Let us expand in Taylor series the transfer-function coefficients $\hat{a}_i$ and $\hat{a}_i^*$ *(i=0, 1… N-1)* in regard to β. If we assume that β is very small and omit expansion after linear term we are coming to

$$\begin{aligned} \hat{a}_i &\approx a_i + \beta \cdot c_i \\ \hat{a}_i^* &\approx a_i^* + \beta \cdot c_i^* \end{aligned} \tag{23}$$

where c_i and c_i^* are functions of $\varepsilon_N(z)$ coefficients in (9). If we set $\beta=0$ we will get the original transfer function $\varepsilon_N(z)$.

Let us notice that positive values of β are decreasing cut-off frequency whereas negative values are increasing it.

All-pass transfer functions of higher order, with real coefficients, can always be realized as cascade connection of real all-pass sections of 1st and 2nd order. The transfer functions of higher order, with complex coefficients, can be realized as cascade connection of complex all-pass sections of 1st order.

For the real case of 1st order we have

$$c_0 = -1 + a_0^2 \tag{24}$$

where a_0 is the coefficient of structure from Fig. 2(a). Correspondingly, for the real case of 2nd order we have

$$\begin{aligned} c_0 &= a_1(-1 + a_0) \\ c_1 &= -(2 + 2a_0 - a_1^2) \end{aligned} \tag{25}$$

where a_0 and a_1 are coefficients of structure from Fig. 2(b), respectively. The reader shall notice that (24) and (25) differ from *(16b), (18a)* and *(18b)* in [2] in sense that they are inverted and they did not work in software. The (25) is in accordance with *(39)* in [5]. That was probably a typo in [2].

From implementation point of view, equation (23) and its special cases (24)-(25) are allowing two alternatives for adjustments. We can either calculate the coefficients again by using transformation $T_1(z)$ from (4) or we can introduce braches for tuning, with coefficients $\beta \cdot c_i$, parallel to multipliers (which have nominal values a_i). In this paper the second alternative is used and is shown at Fig. 5 [2].

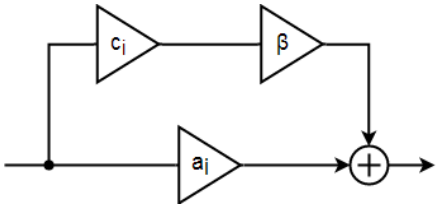


Fig. 5. The tuning branch that is connected in parallel to a nominal value of a coefficient a_i

The variable band-pass (or band-stop) filter one can get by using LP→BP transformation $T_2(z)$ from (6) to variable LP filter. This transformation does not produce loops without delays because it has a form of a delay followed by all-pass section of the 1st order. Therefore, we can control band-pass range and central frequency by using two parameters (one per every characteristic).

III. Implementation Results

As already mentioned, for implementation of such filter structures the *GNU Octave* and its package *Signal* are chosen. The developed m-files comprise around 1000 lines of software code. The *long* number format is used which means that one has 15 digits plus sign.

In [2] is stated that one can get best performance with elliptic filters. That was the reason why it is used in this paper for software implementation. The specifications of elliptic filters are freely chosen and will be stated by the results. Only results related to LP and BP filters will be shown for the sake of simplicity and due to limited number of pages. Showing characteristics of their complementary filter pairs (HP and BS) would be just simple repetition.

For all filter simulations (or processing) in time domain one can notice small signal delay after passing through recursive filter. The delay occurs because it takes some time until all auxiliary variables in 1st or 2nd order sections are initialized. In addition, it is known that IIR filter does not have linear phase characteristics.

A. Odd-Order Filters

Specification of the prototype elliptic filter: 5th order, maximum allowed attenuation in in band-pass area of 1 dB, minimum allowed damping in band-stop area of 35 dB and sampling frequency of 2 kHz.

The change of an amplitude characteristics as well as simulation in time domain of the new filter will be shown.

The LP→LP transformation of the prototype filter is given at Fig. 6. The old cut-off frequency was 400 Hz and new cut-off frequency is 300 Hz.

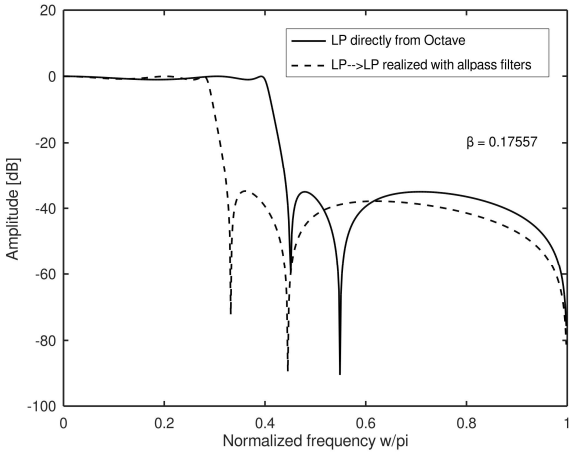


Fig. 6. The LP→LP transformation of 5th order elliptic filter. The old cut-off frequency was 400 Hz (solid line); new cut-off frequency is 300 Hz (dashed line).

The time simulation of the LP→LP transformed filter from Fig. 6 is shown at Fig. 7. Input signal is a discrete sinusoidal with 40 p.u. amplitude and 200 Hz frequency. Work in band-pass area was considered. The new filter worked as expected.

The LP→BP transformation of the prototype filter is given at Fig. 8. The old cut-off frequency was 200 Hz and the new commanded central frequency is 400 Hz.

The time simulation of the LP→BP transformed filter from Fig. 8 is shown at Fig. 9. Input signal is a discrete sinusoidal with 40 p.u. amplitude and 100 Hz frequency.

Work in band-stop area was considered. The new filter worked as expected.

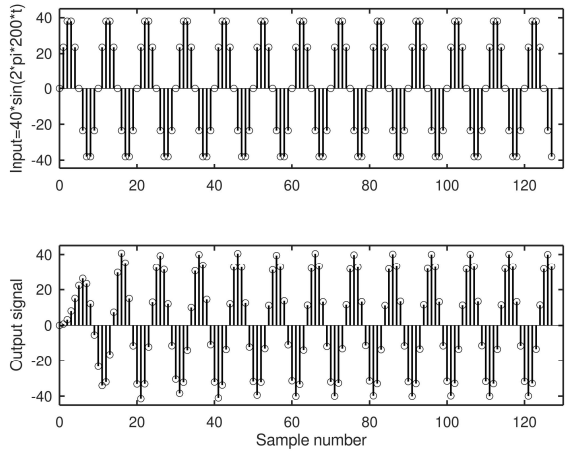


Fig. 7. Time simulation of the LP→LP transformed 5th order elliptic filter from Fig. 6. Input signal (upper trace) is a discrete sinusoidal with 40 p.u. amplitude and 200 Hz frequency. Work in band-pass area was considered.

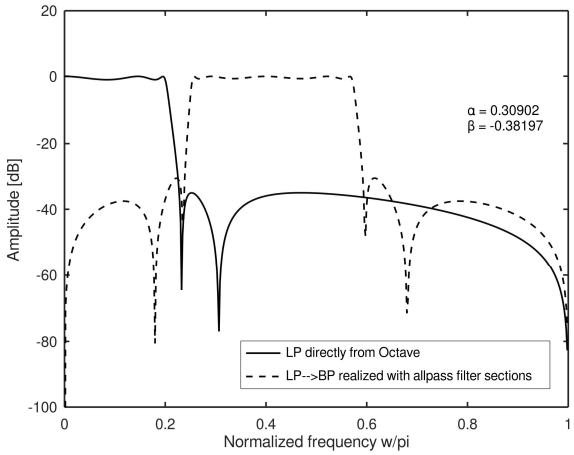


Fig. 8. The LP→BP transformation of 5th order elliptic filter. The old cut-off frequency was 200 Hz (solid line); new commanded central frequency is 400 Hz (dashed line).

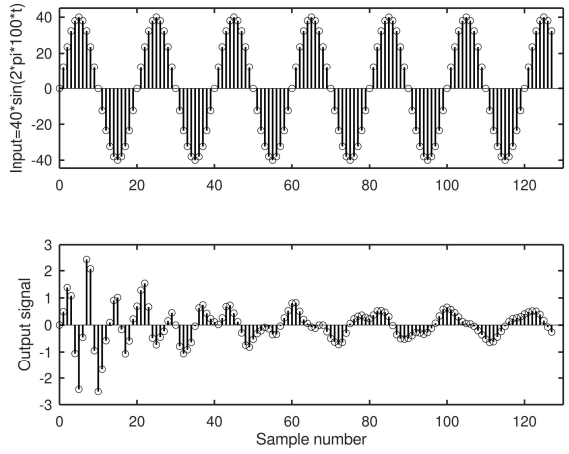


Fig. 9. Time simulation of the LP→BP transformed 5th order elliptic filter from Fig. 8. Input signal (upper trace) is a discrete sinusoidal with 40 p.u. amplitude and 100 Hz frequency. Work in band-stop area was considered.

B. *Even-Order Filters*

Specification of the prototype elliptic filter: 4th order, maximum allowed attenuation in in band-pass area of 1 dB, minimum allowed damping in band-stop area of 35 dB and sampling frequency of 2 kHz.

The change of an amplitude characteristics as well as simulation in time domain of the new filter will be shown. One shall note that structures for even-order filters have narrower allowed tuning range [2].

The LP→LP transformation of the prototype filter is given at Fig. 10. The old cut-off frequency was 200 Hz and new cut-off frequency is 300 Hz. One can notice change (i.e. error) in band-stop attenuation of around 13 dB. In [4] it is mentioned that even-order filters are not low sensitive in band-stop area and band-stop attenuation is also dependent on parameter β (i.e. the same problem is noticeable in the reference - see *Fig. 10* in [2]).

The time simulation of the LP→LP transformed filter from Fig. 10 is shown at Fig. 11. Input signal is a discrete sinusoidal with 40 p.u. amplitude and 150 Hz frequency. Work in band-pass area was considered. The new filter works as expected.

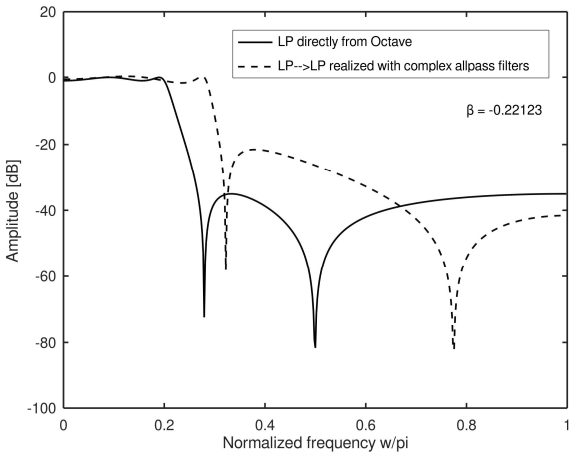


Fig. 10. The LP→LP transformation of 4th order elliptic filter. The old cut-off frequency was 200 Hz (solid line); new cut-off frequency is 300 Hz (dashed line).

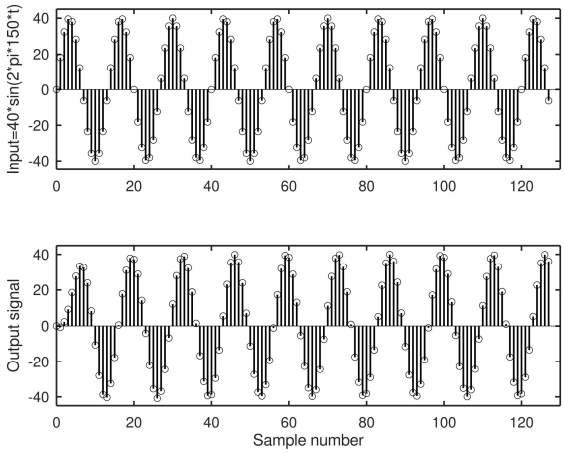


Fig. 11. Time simulation of the LP→LP transformed 4th order elliptic filter from Fig. 10. Input signal (upper trace) is a discrete sinusoidal with 40 p.u. amplitude and 150 Hz frequency. Work in band-pass area was considered.

The LP→BP transformation of the prototype filter is given at Fig. 12. The old cut-off frequency was 200 Hz and the new commanded central frequency is 400 Hz. One can notice change (i.e. error) of more than 10 dB in band-stop attenuation. In addition to comments given for Fig. 10, one can add that in [3] trade-off between errors in pass-band and band-stop areas is analyzed ("*α vs. β*", see *Fig. 10* in [3]). In this paper aim was not to find the perfect combinations of α and β, but to demonstrate that the transformation works in GNU Octave and in accordance with references.

The time simulation of the LP→BP transformed filter from Fig. 12 is shown at Fig. 13. Input signal is a discrete sinusoidal with 40 p.u. amplitude and 100 Hz frequency. Work in band-stop area was considered.

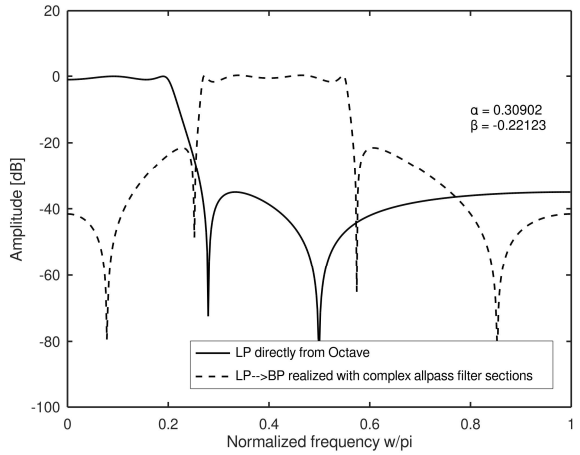


Fig. 12. The LP→BP transformation of 4th order elliptic filter. The old cut-off frequency was 200 Hz (solid line); new commanded central frequency is 400 Hz (dashed line).

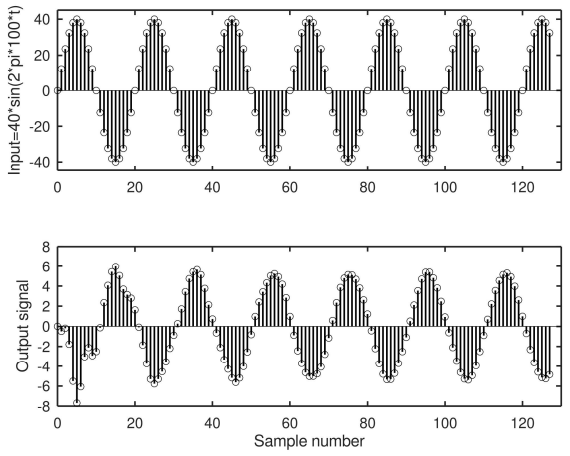


Fig. 13. Time simulation of the LP→BP transformed 4th order elliptic filter from Fig. 12. Input signal is a discrete sinusoidal with 40 p.u. amplitude and 100 Hz frequency. Work in band-stop area was considered.

IV. Conclusion

In this paper the implementation of IIR digital filters with variable characteristics in GNU Octave free open-source software is presented. The goal was to demonstrate efficient usage of such software in tunable digital filters' applications instead of expensive commercial software thus encouraging wider audience to use it.

The new filters, after transformation of their amplitude characteristics, worked as expected. In addition, some small mistakes in formulas in the reference literature [2] are noticed. This is explained and corrected in the paper as well as in the software implementation.

Besides the two aspects mentioned above there were no other new results in this paper compared to the [2]-[4]. The results presented in this paper are in accordance with relevant results in [2]-[4] including errors in band-stop attenuation of even-order filters.

It is worth mentioning that the same code was executed on Octave 4.4 (32-bit) under Windows Vista (32-bit) and Octave 5.1 (64-bit) under Windows 10 (64-bit) and no difference in the filtering results was noticed. No further work is planned on this topic.

Acknowledgment

Author would like to thank Prof. Dr. M. Popović, School of Electrical Engineering, University of Belgrade, for his supervision, advices and tips at the time when initial similar work on this topic was done. This paper is an extension of that work and is not related to author's current job, i.e. affiliation. Author thanks his family and friends for support at that time too. The filter drawings are made in free program *Draw.io* [8].

References

[1] Miodrag V. Popović, *Digitalna obrada signala (in Serbian)*. Beograd: Akademska Misao, 3rd ed, 2003.

[2] S. K. Mitra, Y. Nuevo, and H. Roivanen, "Design of recursive digital filters with variable characteristics," *International Journal of Circuit Theory and Applications*, vol.18, no. 2, pp. 107-119, Mar./Apr. 1990.

[3] S. K. Mitra, and Y. Nuevo, "A new approach to the realization of low-sensitivity IIR digital filters," P. P. Vaidyanathan, *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. ASSP-34, no. 2, pp. 350-361, Apr. 1986.

[4] P. P. Vaidyanathan, P. A. Regalia, and S. K. Mitra, "Design of doubly-complementary IIR digital filters using a single complex allpass filter, with multi-rate applications," *IEEE Trans. Circuits and Systems*, vol. CAS-34, no. 4, pp. 378-389, Apr. 1987.

[5] G. Stoyanov, and M. Kawamata, "Variable digital filters," *Journal of Signal Processing*, vol. 1, no. 4, pp. 275–290, July 1997.

[6] J. Yli-Kaakinen, and T. Saramaki, "Efficient recursive digital filters with variable magnitude characteristics," in *2006 Proc. NORSIG Symposium*, pp. 30-33.

[7] GNU Octave, version 4.4.0 (32-bit) and 5.1.0 (64-bit), and "Signal" package, versions 1.3.2 and 1.4.0; URL (accessed in July 2018 and May 2019): http://www.gnu.org/software/octave/.

[8] Draw.io, version 9.1.2; URL (accessed in August 2018): https://www.draw.io/.

Multifractal Image Forgery Using Logistic Regression

Natasa Milosavljevic, Aleksandra Pavlovic

***Abstract* – This paper presents multinomial logistic regression (MLR) using multifractal image data for forgery detection. Multifractal analysis gives a much smaller number of parameters which are sufficient to describe the image and which in this paper serve for further analysis in order to better forensics. On the basis of these parameters, a MLR method is proposed that will predict the blocks at which the change occurred in the image.**

Keywords: **multifractal, image forensic, prediction, forgery analysis, multinomial logistic regression, classification.**

I. INTRODUCTION

Supervised classification (and segmentation) of high dimensional data sets such as remotely sensed hyperspectral images is a difficult endeavor [1]. Obstacles, such as the Hughes phenomenon [2], appear as the data dimensionality increases. This is because the number of training samples used for the learning stage of the classifier is generally very limited compared with the number of available spectral bands. In order to circumvent this problem, several feature selection [3] and extraction [4] methods have been combined with machine learning techniques that are able to perform accurately in the presence of limited training sets, including support vector machines (SVMs) [5], [6] or multinomial logistic regression (MLR)-based classifiers [7], [8].

The remainder of this paper is organized as follows. Section II formulates the problem. Section III describes the proposed approach. Section IV reports results based on proposed method. Finally, Section V concludes with some remarks and hints at plausible future research lines.

II. PROBLEM FORMULATION

Digital images are used in everyday life. With development of different software, it has become very easily do change the content of digital image. So, we cannot be sure about originality of the same.

One of the most used methods for image forgery is Copy move forgery detection (CMFD). It means that part of original image is copied and pasted into another part of the same image. The goal of that type of changes is to hide some image's content, or to multiply the content of digital image. The examples of CMFD are shown on the Figures 1 and 2.

Fig. 1. Copy move forgery detection – example 1.

Fig. 2. Copy move forgery detection – example 2.

In the Figure 1, a part of forest is copied and pasted to hide the content of image (the man from the right side). Another example (Figure 2) where the cats are copied and pasted so we have four cats on the image, instead of two.

The aim of this research is to represent how we can use multinomial logistical regression for image forgery detection. For the prediction, we use the parameters obtained by multifractal processing. The benefits of using these parameters are that there are a few of them and it has been shown that it is enough to do a successful prediction forgery detection [17, 18]. There are enough and we will describe the proposed method below in more details.

III. METHODS

Multifractal Analysis parameters

A)

Fractal Dimension.

Fractals are irregular geometric objects that cannot be sufficiently specified using topological dimensions. Fractal objects are self-similar, i.e., if one zooms in or out the fractal set, its geometric shape has a similar appearance. Fractal sets have theoretical dimensions that exceed their topological dimensions and can be noninteger values. The basic idea is to cover a fractal set with measure elements (e.g. box) at different sizes and examine how the number of boxes changes with respect to the size [14, 15]. If $N(a)$ is the number of boxes that are needed to cover a fractal object with the size a, then the box-counting dimension D_B is defined as:

Natasa Milosavljevic is with State university of Novi Pazar, Vuka Karadzica, 36300 Novi Pazar, Serbia (e-mail: natasaglisovic@gmail.com).

Aleksandra Pavlovic is with School of Electrical and Computer Engineering of Applied Studies, Vojvode Stepe 283, 11010 Vozdovac, Belgrade, Serbia (e-mail: sandra.pavlo@gmail.com).

$$D_B = \lim_{a \to 0} \frac{\ln N(a)}{\ln(\frac{1}{a})} \quad (1)$$

The fractal dimension D_B characterizes the average behaviors of the image profiles via the scaling law, and profiles with different levels of roughness may have the same fractal dimension.

B) Multifractal Spectrum.

The multifractal analysis utilizes a spectrum of singularity exponents to provide a detailed local description of complex scaling behaviors. In order to quantify local densities of the fractal set, we estimate the mass probability in the box centered at x_i of the image as

$$P_i(a) = N_i(a)/N \quad (2)$$

where $N_i(a)$ is the number of mass or pixels in the *i*-th box of size a, N is the total mass of the set and x_i is the center of a box with size a. It may be noted that the scaling of mass probability $P_i(a)$ with box size a of a multifractal set also follows the power law. The multifractal spectrum is estimated as

$$f(\alpha) = \lim_{a \to 0} \frac{lnN(\alpha)}{ln\frac{1}{a}} \quad (3)$$

where $\alpha_i = \alpha(x_i) = \lim_{a \to 0^+} \frac{lnP_i(a)}{lna}$, x_i is the center of a box with size a.

Logistical regression

Let $D = \{(x^n, t^n)\}_{n=1}^{l}$ represent the training sample, where $x^n \in X \subset R^d$ is the vector of input features for the i[th] example, and $t^n \in T = \{t \mid t \in \{0,1\}^c, \|t\|_1 = 1\}$ is the corresponding vector of desired outputs, using the usual 1-of-c coding scheme. Multinomial logistic regression constructs a generalized linear model [9] with a softmax inverse link function [10], allowing the outputs to be interpreted as a-posteriori estimates of the probabilities of class membership,

$$p(t_i^n, x^n) = y_i^n = \frac{\exp\{a_i^n\}}{\sum_{j=1}^{c} \exp\{a_j^n\}} \quad (4)$$

where $a_i^n = \sum_{j=1}^{d} w_{ij} x_j^n$.

Assuming that D represents an *i*-th i *d*-th sample from a conditional multinomial distribution, then the negative log-likelihood, used as a measure of the data-misfit, can be written as,

$$E_D = \sum_{n=1}^{l} E_D^n = -\sum_{n=1}^{l} \sum_{i=1}^{c} t_i^n \log\{y_i^n\} \quad (5)$$

The parameters, ω of the multinomial logistic regression model are given by the minimizer of a penalized maximum-likelihood training criterion,

$$L = E_D + \alpha E_\omega \text{ where } E_\omega = \sum_{i=1}^{c} \sum_{j=1}^{d} |\omega_{ij}| \quad (6)$$

and α is a regularization parameter [11] controlling the bias-variance trade-off [12]. A minimum of L, the partial derivatives of L with respect to the model parameters will be uniformly zero, giving

$$\left|\frac{\partial E_D}{\partial \omega_{ij}}\right| = \alpha \text{ if } |\omega_{ij}| > 0 \text{ and } \left|\frac{\partial E_D}{\partial \omega_{ij}}\right| < \alpha \text{ if } |\omega_{ij}| = 0 \quad (7)$$

This implies that if the sensitivity of the negative log-likelihood with respect to a model parameter, ω_{ij} falls below α, then the value of that parameter will be set exactly to zero and the corresponding input feature can be pruned from the model. Details regarding the parameters can be see in [13].

IV. EXPERIMENTAL RESULTS

The implemented method is evaluated by utilizing the images from publicly available CoMoFoD dataset and Image Manipulation Dataset [16]. This research has implemented the block-based copy-move forgery detection approach. Namely, an image of interest is divided into non-overlapping blocks of different size. For each block, a feature vector is calculated, based on the multifractal spectrum and its parameters. Tested images are of 256x256 resolution, and the size of blocks vary from 8x8, 16x16 to 32x32. The tested images are shown on Figures 3-7. The left image represent original image, shown in the right image represents forgered image. Multifractal spectrum of original and modified images are shown on the Figures 8-12.

Fig. 3. The first pair of tested images (images numerated with I in Table 1).

Fig. 4. The second pair of tested images (images numerated with II in Table 1).

Fig. 5. The third pair of tested images (images numerated with III in Table 1).

Fig. 6. The fourth pair of tested images (images numerated with IV in Table 1).

Fig. 7. The fifth pair of tested images (images numerated with V in Table 1).

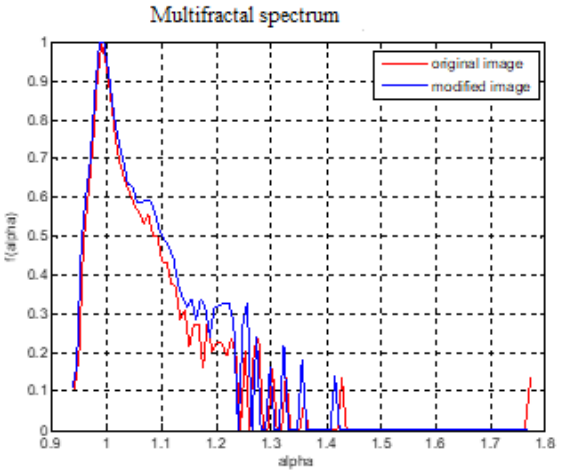


Fig. 8. Multifractal spectrums of original and modified image (image I).

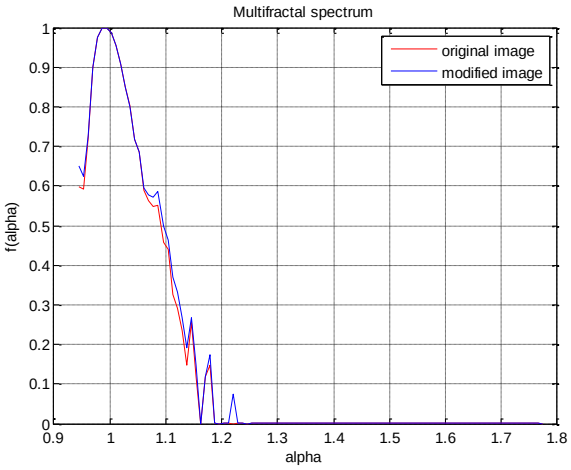


Fig. 9. Multifractal spectrums of original and modified image (image II).

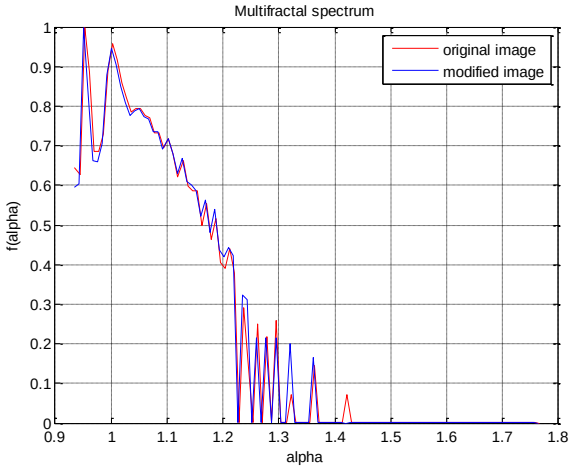


Fig. 10. Multifractal spectrums of original and modified image (image III).

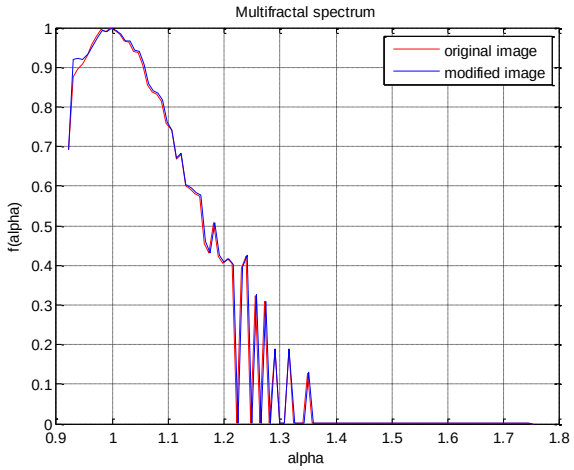


Fig. 11. Multifractal spectrums of original and modified image (image IV).

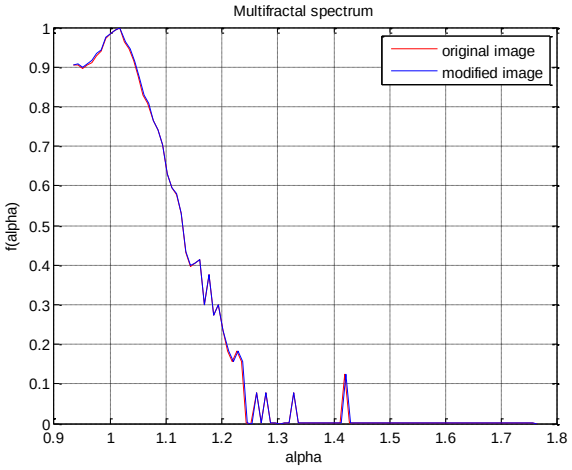


Fig. 12. Multifractal spectrums of original and modified image (image V).

From Figures 8-12, we can see that the spectrum of original and modified images are different. The main differences are in the maximum and minimum value of singularities, the minimum and maximum values of distributions of singularities, the position of the maximum and first zero in multifractal spectrum. So, these parameters are used as elements of characteristic feature vectors of non-overlapping blocks.

After applying the linear multinomial logistic regression, obtained results are shown in Table 1.

TABLE I
EVALUATION OF LINEAR MULTINOMIAL LOGISTIC REGRESSION METHODS OVER A SET OF FIVE BENCHMARK DATASETS (IMAGES)

Benchmark	Error Rate	Cross Entropy	Sparsity
I	0.0213	0.1172	0.0598
II	0.0475	0.2170	0.0691
III	0.1123	0.3661	0.3745
IV	0.2214	0.9412	0.4491
V	0.0142	0.1954	0.0433

Error rate is the error of the prediction model (the smaller it is, the prediction is better). Cross Entropy is the entropy of prediction distribution of outputs and real output. The sparsity is "the scarcity of the model", the scarcity of the model. Namely, in practice, represents how many times when we extend or decrease the test value (which is used to train the regression model). The smaller the value the model is less sensitive.

From Table 1, we can see that the best results are obtained for image I. Namely, as the values of the parameters Error Rate, Cross Entropy and Sparsity are smaller, a minor error of linear regression is achieved. Also, results obtained for other images are very good. The worst results are obtained for image IV (the lower the values of the parameters Error Rate, Cross Entropy and Sparsity, the results are better). The multifractal spectrum of original and modified image IV are very similar (Figure 11), so we get the worst results in parameters of linear regression error.

IV. CONCLUSION

In this paper we have demonstrated that the multifractal approach can be used for multinomial logistic regression. It is interesting to note that the MLR implements a strategy that is exactly the opposite of the relevance vector machine (RVM), in that it integrates over the hyper-parameters and optimizes the weights, rather than marginalizing the model parameters and optimizing the hyper-parameters. In future research we plan to compare this method to other similar method for this problem.

ACKNOWLEDGMENT

The paper is supported by the Serbian Ministry of Education and Science Project III44006.

REFERENCES

[1] D. Landgrebe, Signal Theory Methods in Multispectral Remote Sensing. Hoboken, NJ: Wiley, 2003.
[2] G. Hughes, "On the mean accuracy of statistical pattern recognizers," IEEE Trans. Inf. Theory, vol. IT-14, no. 1, pp. 55–63, Jan. 1968.
[3] S. Serpico and G. Moser, "Extraction of spectral channels from hyperspectral images for classification purposes," IEEE Trans. Geosci. Remote Sens., vol. 45, no. 2, pp. 484–495, Feb. 2007.
[4] J. Richards and X. Jia, Remote Sensing Digital Image Analysis: An Introduction. New York: Springer-Verlag, 2006.
[5] B. Scholkopf and A. Smola, Learning With Kernels-Support Vector Machines, Regularization, Optimization and Beyond. Cambridge, MA: MIT Press, 2002.
[6] G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral image classification," IEEE Trans. Geosci. Remote Sens., vol. 43, no. 6, pp. 1351–1362, Jun. 2005.
[7] D. Böhning, "Multinomial logistic regression algorithm," Ann. Inst. Stat. Math., vol. 44, no. 1, pp. 197–200, 1992.
[8] J. Li, J. Bioucas-Dias, and A. Plaza, "Semi-supervised hyperspectral image segmentation using multinomial logistic regression with active learning," IEEE Trans. Geosci. Remote Sens., vol. 48, no. 11, pp. 4085– 4098, Nov. 2010.
[9] P. McCullagh and J. A. Nelder. Generalized linear models, volume 37 of Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, second edition, 1989.
[10] J. S. Bridle. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In F. Fogelman Soulie and J. H ´ erault, editors, ´ Neurocomputing: Algorithms, architectures and applications, pages 227–236. Springer-Verlag, New York, 1990.
[11] A. N. Tikhonov and V. Y. Arsenin. Solutions of ill-posed problems. John Wiley, New York, 1977
[12] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilema. Neural Computation, 4(1):1–58, 1992.
[13] Cawley, G. C., Talbot, N. L., & Girolami, M. (2007). Sparse multinomial logistic regression via bayesian l1 regularisation. In Advances in neural information processing systems (pp. 209-216).
[14] Foroutan-Pour, K., Dutilleul, P., and Smith, D. L., 1999, "Advances in the Implementation of the Box-Counting Method of Fractal Dimension Estimation," Appl. Math. Comput., 105(2–3), pp. 195–210.
[15] Li, J., Du, Q., and Sun, C., 2009, "An Improved Box-Counting Method for Image Fractal Dimension Estimation," Pattern Recognit., 42(11), pp. 2460–2469.
[16] CoMoFoD database, available at: http://www.vcl.fer.hr/comofod. Accessed March 2019.
[17] Aleksandra Pavlović, Ana Gavrsovska, Nataša Milosavljević, „The Skyline Image Segmentation using Color and Detail Clustering", 14th Symposium on Neural Networks and Applications (NEUREL), Belgrade, Serbia, November 20-21, 2018.
[18] Aleksandra Pavlović, „Application of multifractals for copy-move forgery detection", Contemporary problems of mathematics, mechanics and informatics (CPMMI), Novi Pazar, 2018.

Low cost solution for laboratory class on fundamentals of wireless communication link design

Milutin Nesic[1], Slavica Marinkovic[2], Ivan Pavlovic[3] and Amela Zekovic[4]

***Abstract*—This paper presents inexpensive way for demonstrating concepts and options in wireless link design and deployment for the example of point to point 802.11n link. Students are first introduced to wireless link design by using Radio Mobile software. This simulation is complemented with newly introduced laboratory session that compares results obtained by simulation to results obtained by the link realization in practice with MikroTik wireless device LHG5 with integrated dual polarization grid antenna. During the laboratory sessions students are revising and expanding knowledge on Wi-Fi channels and frequencies, antenna parameters, Wi-Fi modulation and coding schemes, data rates, as well as the impact of wireless channel on the link performance.**

***Index Terms*—Teaching wireless communications; Wi-Fi point to point link; wireless communication link design.**

I. Introduction

The success of 802.11 standard is evident judging by its increasing deployment and applications. This is the widely used family of standards for wireless computer networking used by both commercial and residential users. The most common applications include in-building wireless local area networks, public access hot spots, home networking and wireless bridges. The 802.11 standard denoted by brand name Wi-Fi has evolved quickly and is still under development. The improvement in performance, in terms of data rates and range, has been achieved primarily by employing modern techniques that allow for high spectral efficiency and high data speeds such as the orthogonal frequency division multiplexing (OFDM) and the multiple input multiple output antenna technology (MIMO) [1][2]. This standard is affordable solution to explore and examine principles, concepts as well as certain aspects of practical deployment in wireless communications in general. This would be more complicated in case of licensed microwave links or cellular systems.

In this paper we consider point to point 802.11n link as a wireless link between two buildings that are 3,12 km apart. Wireless point to point link is a cost effective solution easy and fast to deploy in order to connect two locations. The purpose of implementing this link is to complement laboratory sessions on wireless link design and propagation, however the link can be used to send real world traffic between two networks. So far students have been introduced to this topic theoretically and by simulation in Radio Mobile software [3]. The newly introduced demonstration contributes to revising knowledge, expanding it as well as motivating students to learn this subject. This laboratory class is implemented within Mobile Communications course within study program Electronics and Telecommunications at the School of Electrical and Computer Engineering (VISER) in Belgrade. Here we present educational aspects of the use of this link. The choice of equipment, namely MikroTik wireless device LHG5 [4], was motivated by price-performance characteristics. The choice of locations for antennas was also constrained by the position of the VISER School building and the availability of the second site. There is line of sight, however, the environment is dense urban and highly congested.

II. Examining the Radio Link with Radio Mobile Software

In this part of the exercise students first learn how to create Wi-Fi link in the Radio Mobile software. With this software tool it is possible to predict performance of the communication system by using terrain maps (real data) and by defining real characteristics of the elements of the communication system. That is, it is possible to use this program for design and study of the systems implemented in practice. The students are introduced to terminology related to terrain data and then to fundamentals of creating a link such as: defining position of the locations of the communication nodes, defining parameters such as transmitter power, receiver sensitivity, type and height of the antenna and cable losses. They also specify system minimum and maximum operating frequencies. This is done for the example of the implemented point to point Wi-Fi link. Figure 1 shows positions of the transmitting and receiving antennas and the distance in Google maps and Figure 2 shows the same in the Radio Mobile software.

[1]Milutin Nesic is with the School of Electrical and Computer Engineering of Applied Studies Belgrade, 283 Vojvode Stepe, 11010 Belgrade, Serbia (e-mail: nesic@viser.edu.rs).

[2]Slavica Marinkovic is with the School of Electrical and Computer Engineering of Applied Studies Belgrade, 283 Vojvode Stepe, 11010 Belgrade, Serbia (e-mail: slavica.marinkovic@viser.edu.rs)..

[3]Ivan Pavlovic is with the School of Electrical and Computer Engineering of Applied Studies Belgrade, 283 Vojvode Stepe, 11010 Belgrade, Serbia (e-mail: ivanp@viser.edu.rs).

[4]Amela Zekovic is with the School of Electrical and Computer Engineering of Applied Studies Belgrade, 283 Vojvode Stepe, 11010 Belgrade, Serbia (e-mail: amelaz@viser.edu.rs).

Figure 1. Positions of the transmitting and receiving antennas and the distance in Google maps.

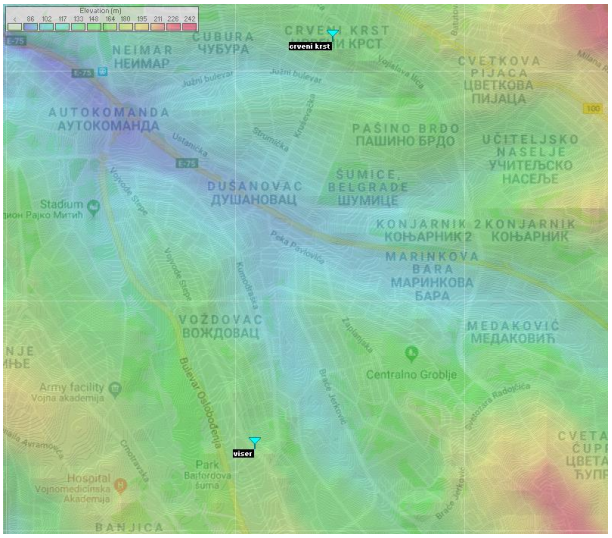


Figure 2. Positions of the transmitting and receiving antennas in Radio Mobile software.

After creating map in Radio Mobile by merging terrain and topographic data and defining positions of transmit and receive antennas, students input into the Radio Mobile program maximum and minimum operating frequencies for the link with band central frequency 5,2 GHz and 20 MHz channel width. Subsequently, they define antenna height. For the next set of parameters students have to consult the datasheet of the MikroTik LHG5 wireless device [4]. They read minimal usable signal level for the modulation and coding scheme (MCS) with binary phase shift keying and convolutional code of rate 1/2, denoted as MCS0, which is −82 dBm. Next required parameter, transmitter output power for MSC0, is taken as the minimum output power which is measured (read) from the antenna and is equal to 22 dBm. Antenna gain is obtained from the manufacturer's documentation and in this case is 24 dBi. Antenna radiation patterns are also downloaded from the manufacturer's website and processed so that the format is matched to one required by the Radio Mobile software. The standard 802.11n allows the use of multiple chains (antennas). MikroTik LHG5 wireless device uses multiple chains on a single dual polarized antenna. This is the way to realize MIMO processing which improves the system performance. The horizontal and vertical radiation pattern for chain 0 that are used in the simulation are shown in Figure 3, and 4 respectively.

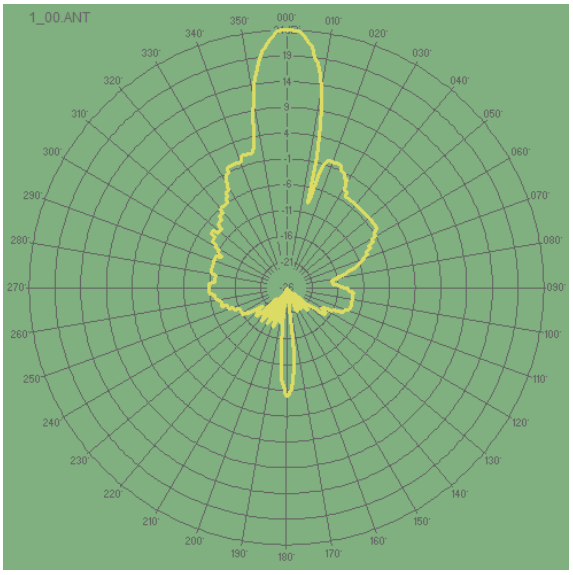


Figure 3. The horizontal radiation pattern for chain 0.

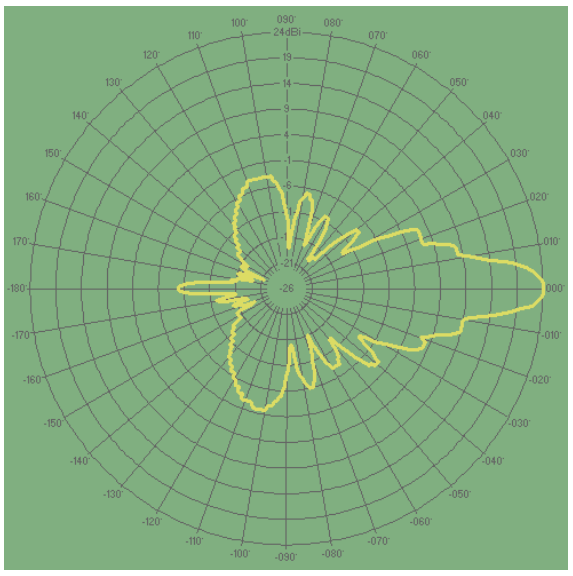


Figure 4. The vertical radiation pattern for chain 0.

The results of the simulation are shown in Figure 5. Radio Mobile uses digital terrain data for automatic extraction of

path profile between a transmitter and a receiver. Students read the predicted path loss and the received signal strength, which are 125,3 dB and −56,3 dBm, respectively. They observe Fresnel zones denoted with white color ellipses with the line of sight denoted in green color in Figure 5. In this simulation we have taken into account only terrain profile, that is only elevation of Earth surface and the default values for the clutter in urban environment and not the detail impact of the ground objects on the RF propagation.

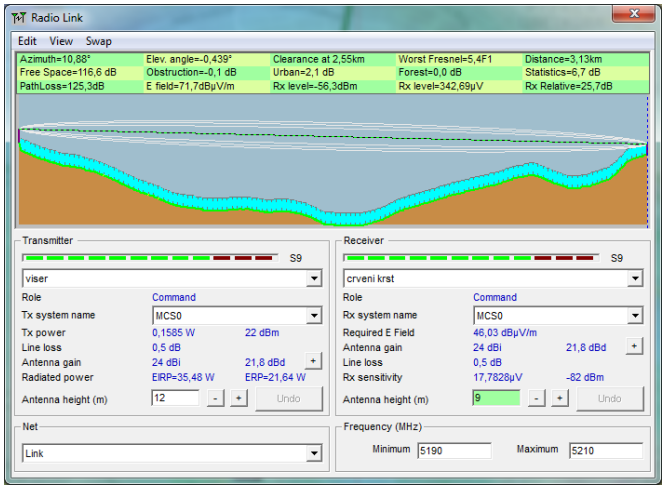


Figure 5. The results of the simulation in Radio Mobile software.

III. EXAMINING THE IMPLEMENTED RADIO LINK REALIZED USING MIKROTIK WIRELESS DEVICE LHG5

In this part of the exercise students get hands-on experience with the realized link in practice.

They observe the antenna mounted on the outer wall of the Electronics and Telecommunications laboratory (Figure 6).

The antenna is in fact compact and light 5 GHz 802.11 a/n wireless device with an integrated dual polarization 24,5 dBi grid antenna. The fact that the antenna element is tightly coupled with compact enclosure of the integrated microwave transmitter/receiver of the equipment means no loss of RF signals on cables [4].

Students further get acquainted with necessary cabling to connect the wireless device and a laptop that contains MikroTik software for configuring link parameters and reading measured parameters from the wireless device. The software tools that are used are Router OS and its GUI version WinBox. The power is supplied by Power over Ethernet (POE) technology that lets network cables carry electrical power.

The initial configuration of the point to point link means that students define central frequency and bandwidth of communication system, as well as wireless protocol and standard wireless network configuration in order to establish the link.

Here we omit initial configuration details of the point to point link and examine just the impact of changing parameters on the performance of the link.

In order to first examine spectral congestion at 5 GHz students first use MikroTik software Dude to perform spectral scan. The spectral scan for the locations of the two antennas are shown in Figure 7 and Figure 8. From the results it can be concluded that the level of interference is high. Several bands with least interference were tested, and 5,2 GHz frequency band was selected. For this part of the exercises the link was not established, because wireless device cannot scan spectrum and maintain link at the particular frequency at the same time.

Figure 6. Antenna mounted on the outer wall of the Electronics and Telecommunications Laboratory.

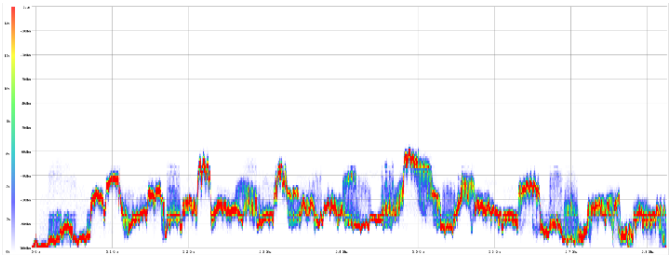


Figure 7. The spectral scan for the location VISER.

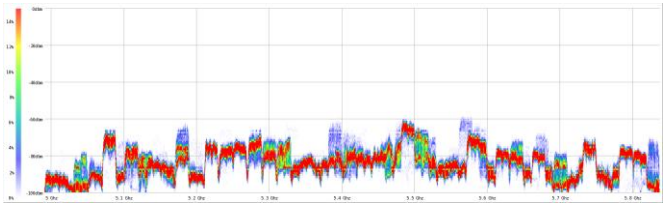


Figure 8. The spectral scan for the location Crveni Krst.

After spectral scan the link is established and students configure wireless device so that modulation and coding scheme MSC0 is active on both chains 0 and 1 and that transmitting power is 22 dBm in each chain. MCS0 is defined in table I. In this part of the exercise the same data is sent on both chains. The results are shown in Figure 9 and Figure 10 .

It can be seen that the received signal strength is −57 dBm. This is comparable with simulation results, although simulation did not consider multiple transmission chains. Students observe that signal strength on individual chains is lower than resulting signal strength, that is the resulting received signal strength is higher due to use of dual chains. They can observe bitrate which is 5,6 Mb/s. The bitrate is measured based on the bandwidth test in the wireless device which uses UDP protocol for estimating bitrate.

The students further change modulation and coding schemes and observe bitrate and received signal strength. The results that are obtained are compared to 802.11n specifications. This is shown in Table I. Table contains information on modulation and coding scheme (MCS) and measured values obtained in our experiment. It is not possible to change MCS completely independently, because wireless device we use dynamically switches to lower modulation scheme in case of change in conditions.

TABLE I
MODULATION AND CODVING SCHEMES AND BITRATE

MSC	Modulation	Coding	Data Rate [Mbps]
0	BPSK	1/2	5,6
1	QPSK	1/2	11,2
2	QPSK	3/4	16,9
3	16-QAM	1/2	22,7
8	BPSK	1/2	10,9
9	QPSK	1/2	22,8
10	QPSK	3/4	33,8
11	16-QAM	1/2	45,4

From the results shown in Table I students can observe that for modulation and coding schemes MSC8 to MSC11 different information is sent on different chains which leads to doubling the bitrate.

Second experiment consists in changing transmit signal power and observing received signal strength, as well as, observing how the modulation and coding scheme changes. The results are shown in Table II.

TABLE II
SIGNAL POWER, RECEIVED SIGNAL STRENGTH, DATA RATE AND MODULATION-CODING SCHEME

Power [dBm]	RSSI [dBm]	Data Rate [Mbps]	MCS
22	-57	22,7	3
10	-70	22,7	3
4	-75	20,7	3
2	-80	10,4	1
1	-81	5,6	0

It can be noted that, as transmit power decreases, the received signal strength decreases. Also the modulation and coding scheme is changed dynamically. It can be observed that dominant MCS is that of lower order modulation and higher rate error control coding scheme for the lower transmit signal powers.

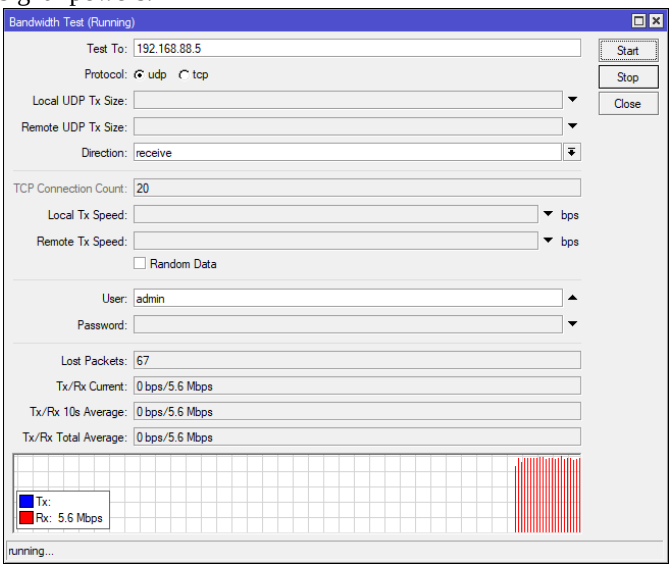


Figure 9. Bitrate obtained by experiment in the field.

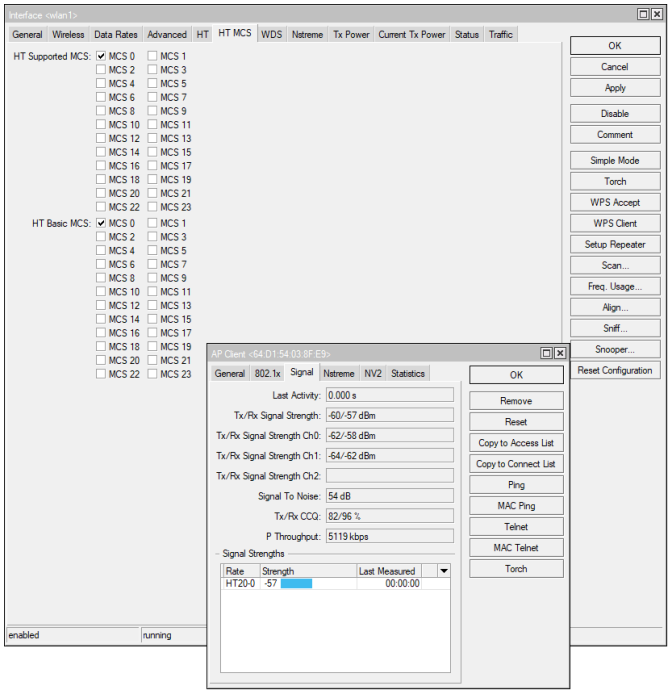


Figure 10. Received signal strength obtained by experiment in the field.

IV. CONCLUSION

The paper presents use of the simple low-cost point to point link for enhancing practical teaching of physical layer aspects of wireless communications. In particular, we have first focused on teaching students about basic elements of wireless system and their parameters such as transmitter and receiver locations, radiated power, frequency, and antenna pattern by simulation in Radio Mobile software. Then we have built on this by teaching students how to set up point to point Wi-Fi link with MikroTik wireless device LHG5. The students had opportunity to examine hardware and software necessary for setting up the link, however, the emphasis was on observing

impact of changing parameters such as modulation and coding scheme on the link performance, as well as demonstrating path loss results obtained by the simulation. The enhancement of teaching and learning process is achieved by engaging students into real world example that apply outside the lab.

REFERENCES

[1] E. Perahia, R. Stacey, *Next Generation Wireless LANs: 802.11n and 802.11ac*, Cambridge: Cambridge University Press., 2013.

[2] E. Perahia, "IEEE 802.11n Development: History, Process, and Technology," in IEEE Communications Magazine, vol. 46, no. 7, pp. 48-55, July 2008.

[3] "Radio mobile", January 2018, Available online at: http://www.ve2dbe.com/english1.html

[4] MikroTik LHG5 802.11 device manufacturer's data, April 2019: Available online at: https://mikrotik.com/product/RBLHG-5nD

Healthcare IoT Monitoring using Photoplethysmography

Milan S. Milivojević, *Student Member, IEEE,* Ana Gavrovska, *Member IEEE,* Irini Reljin, *Senior Member, IEEE* and Branimir Reljin, *SeniorMember, IEEE*

***Abstract*—The Internet of Things (IoT) has offered a variety of different possibilities for healthcare applications and systems. Information and communication technologies (ICT) and high quality micro- and nanoelectronics enable novel solutions based on body area networks for collecting valuable data describing vital signs of humans. In this paper a system for monitoring some vital parameters described by photoplethysmogram (PPG) signals is described. The system is based on low-cost Arduino platform and appropriate sensors. Heart rate is monitored and compared for telemedical and e-health purposes, showing differences between chosen PPG features in physical domain. For this task, a Python based graphical user interface has been developed. Finally, PPG has been compared in counting backwards experiment, where an increase of the selected feature values is obtained due to a cognitive load.**

***Index Terms*— photoplethysmogram, heart rate, internet of things (IoT), cognitive test, graphical user interface.**

I. Introduction

Photoplethysmography (PPG) is an optical measurement technique that can be used to detect blood volume changes in the microvascular bed of tissue [1]. It is widely used in clinical practice with commercial medical devices. The most prominent example is the pulse oximeter used for measurement of the blood saturation by oxygen (SpO2), as well as a heart rate [2]. This technique is used in vascular diagnostics, i.e. estimation of microcirculation. It is also used in digital beat-to-beat devices for blood pressure measurement [3][4].

PPG pulsed current corresponds to alternating component with the fundamental frequency related to the heartbeat of around 1Hz. Pulsating DC (DC-Direct Current) is a consequence of the tissue processes and blood volume changes. The envelope of this quasi-direct component is influenced by processes such as: respiration, vasomotor activity, thermoregulation, etc. The components can be extracted by appropriate signal processing techniques depending on the analyzed phenomenon [5].

The basic architecture of PPG device includes several optoelectronic components: light source used for the tissue (skin surface) illumination, and photodetector for detection of small variations of light rays intensities penetrated in the skin. Variation of the light intensities is a consequence of blood volume change within the tissue (fingertip, earlap/earlobe, etc.). PPG is usually recorded according to non-invasive procedure. Wavelengths used for PPG acquisition correspond to infrared (IR) spectrum part. The most prominent component of the PPG signal is the peripheral pulse synchronized with each heartbeat. Despite the simplicity of the PPG technique, the origin of all of its components is still not known. However, PPG is established as a justified technique for the cardiovascular examinations [6]. Beside commercial applications in medicine, PPG is also used in smart devices, like: fitness smart watches, fitness bracelets or headphones, for providing useful information to professional and recreational athletes during physical activities [7].

Different signals can be monitored using body area network (BAN), where heart rate (HR) is usually of interest. HR is measured by counting the number of heart beats or contractions per minute. PPG can be also used for HR estimation. It represents one of the vital signals and can be measured if appropriate sensor is available. Namely, PPG can be implemented in Internet of Things (IoT) systems for telemedical, i.e. e-health applications [8]. Data is collected from the sensors, and further processed and presented on the Cloud according to physicians' or patient's demands. Information and communication technologies (ICT) in telemedicine enable developing simple, but efficient systems for healthcare

In this paper we present simple system for acquisition of PPG signal using low-cost sensor and Arduino open source platform. We demonstrate PPG acquisition using developed graphical user interface (GUI). The PPG is processed and the HR is estimated. Moreover, we illustrate the influence of the subject cognitive activity on the estimated parameters during the acquisition.

The organization of this paper is as follows. In Section II, typical PPG signal is described. Section III is dedicated to the HR estimation based on PPG signal. Details on data acquisition and the Arduino based system are provided in Section IV. The simulation description is given in Section V, and the obtained experimental results are presented in Section VI. Finally, conclusions are summarized in Section VII.

II. PPG Signal Morphology

Typical photoplethysmogram (PPG) is proportional to the quantity of blood flowing through the blood vessels. It is used to asses blood flow in skin, muscle and bone tissue. It

Milan S. Milivojević is with the School of Electrical Engineering, University of Belgrade, Bulevar kralja Aleksandra 73, 11120 Belgrade, Serbia (e-mail: msmilance@etf.bg.ac.rs).

Ana Gavrovska is with the School of Electrical Engineering, University of Belgrade, Bulevar kralja Aleksandra 73, 11120 Belgrade, Serbia (e-mail: anaga777@etf.bg.ac.rs, anaga777@gmail.com).

Irini Reljin is with the School of Electrical Engineering, University of Belgrade, Bulevar kralja Aleksandra 73, 11120 Belgrade, Serbia (e-mail: irini@etf.bg.ac.rs, irinitms@gmail.com).

Branimir Reljin is with the School of Electrical Engineering, University of Belgrade, Bulevar kralja Aleksandra 73, 11120 Belgrade, Serbia (e-mail: reljinb@etf.bg.ac.rs).

consists of constant-level DC related to the relative tissue vascularization and pulsatile component syncronous with heart pumping. The DC variation occurring is a consequence of cardiac synchronous changes in blood volume in arterial blood vessels. Change is synchronized with heart beats and directly affects the light that falls on the tissue, or the reflected wave. The superposition of direct and reflected waves results in the characteristic morphology of the PPG signal (Fig. 1). The PPG signal has two phases: the anacrotic phase and the catacrotic phase [9].

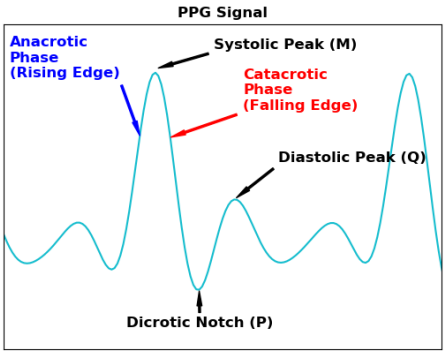


Fig. 1. The PPG signal morphology.

The anacrotic phase corresponds to an increase of the signal amplitude (rising edge) and is associated with a systole (contraction of the heart muscle). Thus, the maximum in this phase is called the systolic peak, and it is denoted by M. The time interval between M peaks defines the MM interval. This interval is strongly correlated with RR intervals in the electrocardiogram signal. Therefore, it is directly related to heart rate variability (HRV) [10][11].

The catacrotic phase is the phase of decreasing the impulse amplitude (falling edge). It is associated with the diastole (relaxation of the heart muscle). Diastolic peak is designated as the Q peak in PPG. Before the diastolic peak occurs, a minimum or dicrotic notch (P point) occurs (P point corresponging to PPG morphology). This point may be missing, and instead of the minimum, it appears as an inflection point.

III. PPG Heart Rate Estimation

PPG signal can be used for HR estimation. Each pair of successive M peaks in PPG defines the MM interval. Within each MM interval, the positions of the P and Q points corresponding to the PPG morphology, are found as the local minimum and maximum, respectively. In this way, both PP or QQ intervals are defined [12]. The HR value in bpm (beats per minute) is defined as the reciprocal value of the MM interval scaled with 60:

$$HR_{MM}\left[bpm\right]=\frac{60}{MM\left[s\right]}. \tag{1}$$

Similarly, it can also be defined as:

$$HR_{PP}\left[bpm\right]=\frac{60}{PP\left[s\right]},\ HR_{QQ}\left[bpm\right]=\frac{60}{QQ\left[s\right]} \tag{2}$$

on the basis of the other two intervals. Since the number of beats per minute is of integer type, HR is rounded up to the nearest integer value. On the basis of these values, dependency diagram of HR in time can be formed.

The values calculated by all three methods ((1)-(2)) are similar so that each approach is valid. In the literature, however, the use of the HR calculation based on the MM interval is common. The reason lies in the fact that M peak is dominant in the morphology of the PPG signal (similarly as in electrocardiogram). Thus, M peak detection is easier to perform [13]. The remaining two points can be more difficult to detect, especially in the case where P is not pointed out. Therefore, HR calculation based on M peak can be taken as a reference. Root mean square error (RMSE) for HR calculated on the PP interval is:

$$\varepsilon_{PP}=\sqrt{\frac{1}{N}\cdot\sum_{i=1}^{N}\left(HR_{PP}-HR_{MM}\right)^2} \tag{3}$$

where N is equal to the length of estimated HR sequence (number of samples). Similarly, estimation error can be performed as:

$$\varepsilon_{QQ}=\sqrt{\frac{1}{N}\cdot\sum_{i=1}^{N}\left(HR_{QQ}-HR_{MM}\right)^2} \tag{4}$$

where QQ interval is observed.

HR values are averaged over a certain number of consecutive values to give accurate estimation results. For this purpose a moving average (MA) filter can be applied:

$$\overline{HR_{MM}}\left[n\right]=\frac{1}{L}\cdot\sum_{k=0}^{L-1}HR_{MM}\left[n-k\right] \tag{5}$$

where L is a length of the window where the averaging is performed.

In the literature, there are many types of features that can be distinguished from the PPG signal. However, they seem not to be completely standardized yet, as is the case with the ECG signal. They can correspond to different domains (time, frequency, etc.). The easiest way to describe the PPG morpohology for HR estimation is to use the original, time domain [14].

IV. IOT SYSTEM AND DATA ACQUSITION

The Arduino based sytem for PPG heart rate estimation is shown in Fig. 2. For the acquisition of pulse waves we used Pulse Sensor Amped [15]. The front of the sensor is the side with the Heart logo. This side contacts the skin. Also, there is a LED (Light Emitting Diode). Under the LED ambient light sensor exists, exactly like the one used in cellphones, tablets or laptops, to adjust the screen brightness in different light conditions. The LED shines light on the skin (the fingertip or earlobe, or other capillary tissue), and sensor measures the light reflection. Since the positions of the light source and the photodetector are on the same side, the reflexive operation mode is used (otherwise transmitting mode can be used) [16].

Sensor has three ports corresponding to: positive potential of +3.3V or 5V (+Vcc), ground (GND), and signal line (S). Sensor is connected to open source Arduino Uno platform. From the sensor, signal is collected on the analog input A0. Analog part also provides power supply of the sensors. Analog-to-digital conversion is conducted with 10 bits. Arduino Uno is connected to PC using USB A/B cable, where digital data is conducted to PC over serial port. The program loaded to Arduino reads the data from the sensor and forwards it to serial port. Data signaling rate is 9600

symbols/s. The time interval between two successive samples is 10ms, and it defines the sampling frequency of 100 Hz. The sensor is positioned on the subject's forefinger of the right hand.

During the acquisition, it is important to avoid abrupt movements and changes of the pressure on the sensor's surface in order to minimize artefacts in the collected data. The room used for acquisition is soundproofed and its temperature is approximately 25°C.

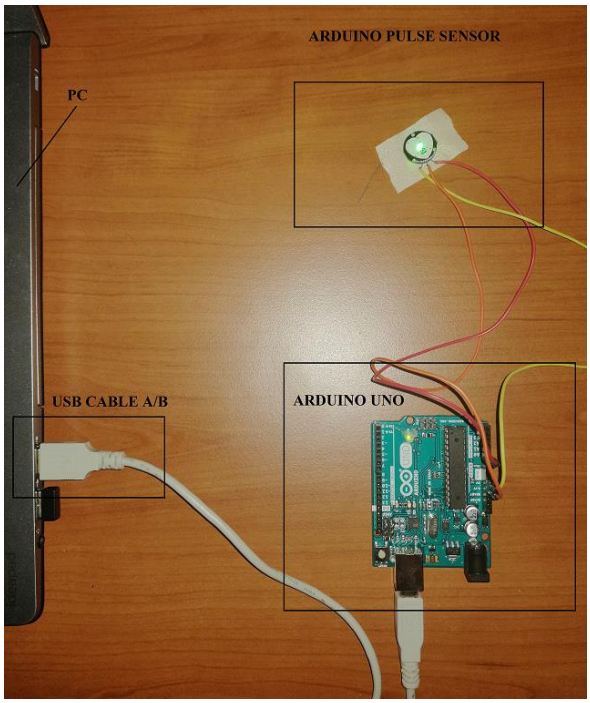


Fig. 2. The Arduino based sytem for PPG heart rate estimation.

V. Simulation

The simulation was performed using the Arduino based system for PPG heart rate estimation. Collected signals are processed within the Python environment. For this purpose, we used libraries for signal processing from the package SciPy.

Raw PPG signal may contain artefacts caused by movements during acquisition or changes of sensor pressures on the skin. In order to deal with artefacts, we used Butterworth bandpass (1-5 Hz) filter of the 4th order for preprocessing. Filtering is conducted forward and backward, therefore the delay can be neglected. This also reduces variation of the base line of the signal, 50 Hz power supply noise, as well as the artefacts caused by movements [17].

Filtered PPG signal is normalized so that the value of the samples belongs to the interval [0,1]. After normalizing the PPG signal, the detection of M peaks is made using methods from publicly available Python libraries. A stepwise form of the display is used so that the duration of each MM interval is appended to the HR value according to the expressions (1)-(5). The error values are calculated for 3 and 5 beats per minute (rounded to the nearest integer).

Continuous depiction of HR is a starting point for applications where HRV analysis is performed, similarly as in the case of ECG signal [18]. For this purpose we realized graphical user interface (GUI) for PPG signal acquisition in program package Python3.6.

An experiment was performed involving a cognitive test. A ninety seconds long recording was acquired, where the central 60s of the signal have been taken for further analysis. Particularly, recording is performed in two iterations, where the first one represents a calm state and the other state is realized during the exercise of mental activity. The subject in both cases is sitting.

Mental exercise involves counting even numbers backwards in a quiet speech manner (counting backwards from one hundred to one). Counting is performed until the recording time expires. The collected signals are filtered and characteristic M peaks are detected. Based on the detected M peaks and determined MM intervals, P and Q points are detected. Then, the amplitudes corresponding to the time positions of M, P and Q are determined. The mutual relations of these amplitudes define two types of features. The amplitude of M peak is taken as a reference for the remaining two amplitudes [19]. The relative amplitude ratio AR_{MP} is defined as:

$$AR_{MP} = \frac{A_P}{A_M} \cdot 100\% \tag{6}$$

where A_P and A_M are minimum values that correspond to the dicrotic notch and the systolic peak, respectively. A relative amplitude relation AR_{MQ} is defined as:

$$AR_{MQ} = \frac{A_Q}{A_M} \cdot 100\% \tag{7}$$

where A_Q corresponds to the diastolic maximum value. These values are calculated for each interval and define one point in the scatter diagram, which is used to present the cognitive load effect.

VI. Experimental Results

A. Heart Rate Estimation results

Fig. 3 shows raw PPG signal acquired from the sensor, as well as PPG signal after filtering. Averaging is carried out at the window of five samples. Larger window lengths lead to greater error in the case of sudden changes in pulse.

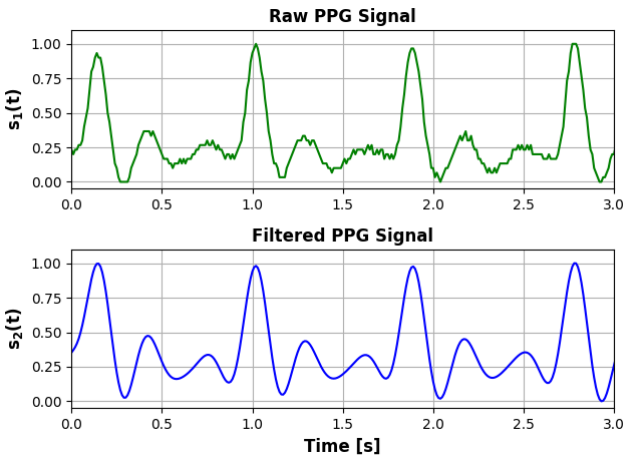


Fig. 3. Raw and preprocessed PPG signal acquired from the sensor.

Characteristic points of PPG signal are detected and presented in Fig. 4. The corresponding amplitudes are used for HR estimation.

A continuous presentation of the HR change in time is given in Fig. 5. Namely, Fig. 5 shows HR calculations based on MM, PP and QQ intervals, respectively. This leads to obtaining the instantaneous heart rate (IHR). Moreover, Fig. 5 shows the original HR and HR after filtering for the case when the assessment is performed based on the MM interval.

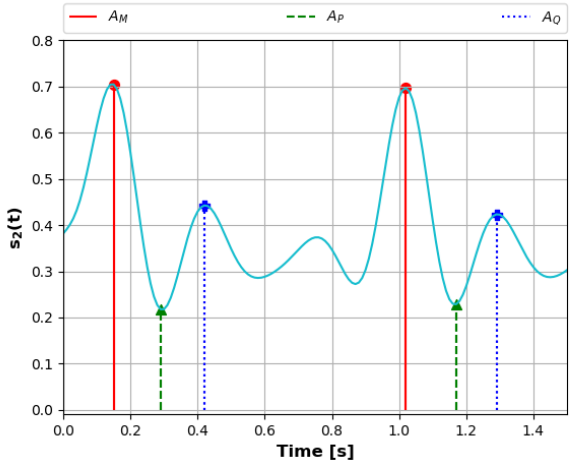


Fig. 4. Detected points and corresponding amplitudes for HR estimation.

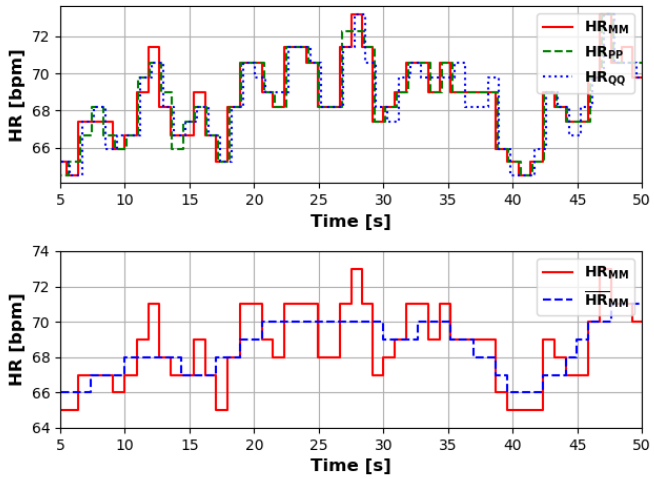


Fig. 5. Calculated HR values based on MM, PP and QQ intervals, as well as a comparative view of the original and the averaged HR based on the MM interval.

B. GUI developed for PPG analysis

In Fig. 6 developed graphical user interface for PPG signal acquisition is presented. The panel for data acquisition shows the main buttons.

Button CONNECT enables connection to Arduino microcontroller over serial port on PC. Besides the automatic serial port search, activation of this button finds the communication speed. After successful connection, the label of the used serial port is displayed in the appropriate panel. Command TEST starts the signal acquisition for 5s and the signal is afterwards shown in the graphical panel. If the signal is of adequate quality, by pressing RECORD button a user can record the signal. It also enables the user to enter the date of acquisition. User chooses the location where the record is saved. Several file extensions for records are provided: *Python* file (.npz), *Matlab* file (.mat), textual file (.txt). Command LOAD loads saved files into the program interface and displays the signal in the graphical interface. Button QUIT is for exiting the program.

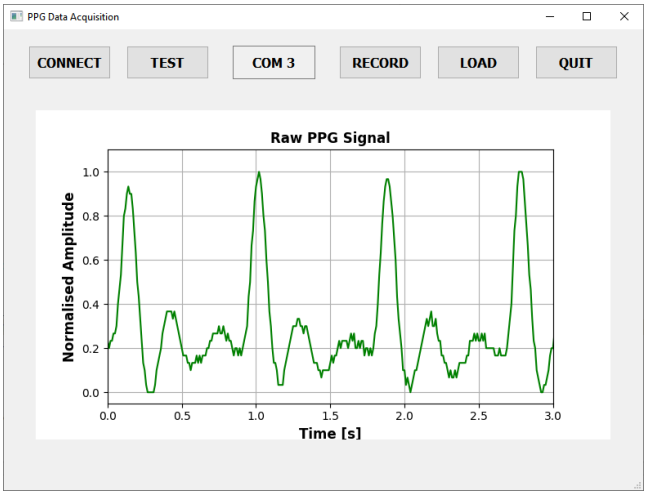


Fig. 6. Panel for acquisition of the PPG signal.

C. Cognitive influence results

The scattering in Fig. 7 presents the feature values before the cognitive experiment (blue dots in Fig. 7). The same procedure is repeated in the case of a mental exercise (red crosses in Fig. 7). A change in two phases is obvious. In the cognitive load there is a significant level of dispersion in the characteristics. The relative amplitude ratio values are greater during the mental exercise.

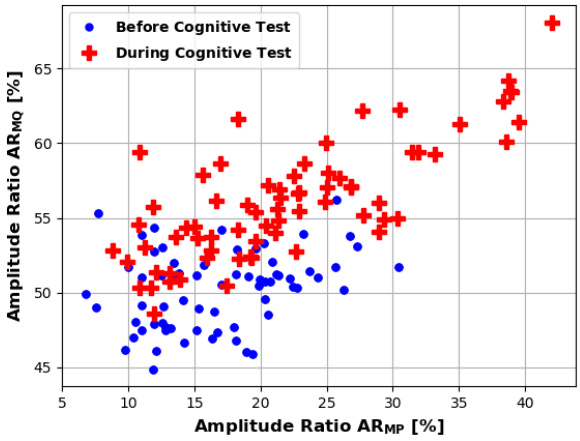


Fig. 7. Diagram of amplitude based scattering corresponding to the PPG signal before and during the cognitive load.

VII. Conclusion

This paper presents IoT system, based on the open source platforms, for telemedical (e-health) applications. The system uses PPG sensor. Graphical user interface for BAN acquisition has been developed, which can be supplemented with new features such as a PPG signal analysis block. The price of this implemented system is very low, and represents a good starting point for the use of PPG in IoT systems.

Despite the low price of realized device, the acquired signal is of good quality, and it is possible to evaluate the set of parameters that represent the condition of patients and

people who are engaged in physical activity. Amplitude features are clearly distinguished and used for calculating the adequate amplitude-based ratio at characteristic locations. The influence of cognitive load on the amplitude based values is also illustrated.

In further work, we plan to record a larger set of signals under different stress conditions, as well as to define new features in different domains that could be helpful in the healthcare systems.

ACKNOWLEDGMENT

This work was partially supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia, under the Project No.III44009, subproject 3.

REFERENCES

[1] J. E. Sinex, "Pulse oximetry: principles and limitations.," *Am. J. Emerg. Med.*, vol. 17, no. 1, pp. 59–67, Jan. 1999.

[2] S. Bagha, S. Hills, P. Bhubaneswar, and L. Shaw, "A Real Time Analysis of PPG Signal for Measurement of SpO 2 and Pulse Rate," 2011.

[3] G. Slapničar, M. Luštrek, and M. Marinko, "Continuous Blood Pressure Estimation from PPG Signal," 2018.

[4] M. Simjanoska, M. Gjoreski, M. Gams, and A. Madevska Bogdanova, "Non-Invasive Blood Pressure Estimation from ECG Using Machine Learning Techniques.," *Sensors (Basel).*, vol. 18, no. 4, Apr. 2018.

[5] M. Klum, T. Tigges, A. Pielmus, Alexandru-Gabriel Feldheiser, and R. Orglmeister, "Impedance Plethysmography for Respiratory Flow and Rate Estimation using Multilayer Perceptrons," *Int. J. Bioelectromagn.*, vol. 21, pp. 34–47, 2019.

[6] J. Allen, "Photoplethysmography and its application in clinical physiological measurement," *Physiol. Meas.*, vol. 28, no. 3, pp. R1–R39, Mar. 2007.

[7] M. J. Gregoski *et al.*, "Development and Validation of a Smartphone Heart Rate Acquisition Application for Health Promotion and Wellness Telehealth Applications," *Int. J. Telemed. Appl.*, vol. 2012, pp. 1–7, 2012.

[8] D. Castaneda, A. Esparza, M. Ghamari, C. Soltanpur, and H. Nazeran, "A review on wearable photoplethysmography sensors and their potential future applications in health care.," *Int. J. Biosens. Bioelectron.*, vol. 4, no. 4, pp. 195–202, 2018.

[9] E. Peralta, J. Lazaro, R. Bailon, V. Marozas, and E. Gil, "Optimal fiducial points for pulse rate variability analysis from forehead and finger photoplethysmographic signals," *Physiol. Meas.*, vol. 40, no. 2, p. 025007, Feb. 2019.

[10] T. Proesmans, C. Mortelmans, R. Van Haelst, F. Verbrugge, P. Vandervoort, and B. Vaes, "Mobile Phone-Based Use of the Photoplethysmography Technique to Detect Atrial Fibrillation in Primary Care: Diagnostic Accuracy Study of the FibriCheck App.," *JMIR mHealth uHealth*, vol. 7, no. 3, p. e12284, Mar. 2019.

[11] S. Das, S. Pal, and M. Mitra, "Real time heart rate detection from PPG signal in noisy environment," in *2016 International Conference on Intelligent Control Power and Instrumentation (ICICPI)*, 2016, pp. 70–73.

[12] A. R. Kavsaoğlu, K. Polat, and M. R. Bozkurt, "An innovative peak detection algorithm for photoplethysmography signals: an adaptive segmentation method," *TURKISH J. Electr. Eng. Comput. Sci.*, vol. 24, pp. 1782–1796, 2016.

[13] T. Tigges *et al.*, "Classification of morphologic changes in photoplethysmographic waveforms," *Curr. Dir. Biomed. Eng.*, vol. 2, no. 1, pp. 203–207, Jan. 2016.

[14] G. Joseph, A. Joseph, G. Titus, R. M. Thomas, and D. Jose, "Photoplethysmogram (PPG) signal analysis and wavelet de-noising," in *2014 Annual International Conference on Emerging Research Areas: Magnetics, Machines and Drives (AICERA/iCMMD)*, 2014, pp. 1–5.

[15] "Pulse Sensor - SEN-11574 - SparkFun Electronics." [Online]. Available: https://www.sparkfun.com/products/11574. [Accessed: 21-Apr-2019].

[16] M. Elgendi, "On the analysis of fingertip photoplethysmogram signals.," *Curr. Cardiol. Rev.*, vol. 8, no. 1, pp. 14–25, Feb. 2012.

[17] A. E. Awodeyi, S. R. Alty, and M. Ghavami, "On the Filtering of Photoplethysmography Signals," in *2014 IEEE International Conference on Bioinformatics and Bioengineering*, 2014, pp. 175–178.

[18] M. Milivojević, A. Gavrovska, I. Reljin, and B. Reljin, "Python Based Physiological Signal Processing for Vital Signs Monitoring," in *Proceedings of the 4th International Conference on Electrical, Electronic and Computing Engineering IcETRAN 2017*, 2017, p. EK(I)2-1-4.

[19] D. Mcduff, S. Gontarek, and R. W. Picard, "Remote Detection of Photoplethysmographic Systolic and Diastolic Peaks Using a Digital Camera," 2013.

DASH video user interface based on GPU background subtraction and OpenCL C++ framework

Katarina Popović, Ana Gavrovska, *Member, IEEE,* Irini Reljin, *Senior Member, IEEE*

***Abstract*—The media delivery over the Hypertext Transfer Protocol (HTTP) got extremely popular over the years. Dynamic Adaptive Streaming over HTTP (DASH) provided the adaptive multimedia delivery utilizing a chunk based approach. In this paper vision based user interface has been developed enabling the analysis of media presentation description files, as well as the implementation of fundamental steps for video processing such as background subtraction. This is enabled by OpenCL C++ framework. The obtained experimental results show advantages of GPU computing over CPU where 30% less processing time is needed for high resolution video.**

***Index Terms*—Video, MPEG-DASH, parallel processing, background subtraction, GPU, OpenCL.**

I. Introduction

Content delivery networks (CDNs) are oriented toward delivering video services for communication and entertainment purposes. They are changing under the internet pressure due to video consumption [1]. Broadband video delivery has experienced exponential growth over traditional broadcasting systems, where video represent a huge part of communication and IP (Internet Protocol) based traffic. VoD (Video- On-Demand) services, OTT (Over-The-Top) services, popular platforms (Netflix, Youtube, Vimeo, etc.), real-time video conferencing (Skype, Viber, etc.) contribute to the changes. Information-centric networking (ICN) has been proposed to support high quality video services and streaming [1], where a multimedia standard such as MPEG-DASH is being helpful [2]. Dynamic Adaptive Streaming over HTTP (DASH) provides the adaptive multimedia delivery utilizing a chunk based approach, standardized by MPEG (Motion Picture Experts Group) [2, 3]. Namely, the packets are sent over HTTP (HyperText Transfer Protocol) in adaptive manner according to network's conditions, where HTTP traffic has reliable transfer infrastructure [4]. DASH enables interoperability between servers and clients from different vendors and high quality multimedia streaming over the Internet delivered from conventional HTTP web servers. The MPEG-DASH Media Presentation Description (MPD) is an XML document containing information about media segments [2]. It provides information needed for decoding and streaming [4-7]. One the most relevant technologies is Software Defined Networking (SDN) making the application-centric intelligent routing strategies, providing high quality services to DASH clients [4].

With growing popularity of video content, need for real-time video processing appeared. Since video processing is very demanding, parallel programming became inevitable. The need for high speed computing popularized parallel approaches [8-10]. Parallel programming uses heterogeneous hardware, processing on CPU (*central processing unit*), GPU (*graphic processing unit*) and other processing units.

Modern GPUs have wide use not only in image processing and computer vision, but also in machine learning and big data analysis, because of their ability to execute tasks in parallel (much faster than CPU) [3, 8-12]. They are inevitable part of smart devices and are used for mobile and video services. Today we have plenty of different kinds of servers and user devices (clients) with need to effectively communicate with each other. In video processing generally, video surveillance and object tracking, the first task is usually performing background subtraction or to extract foreground from background [9-14]. Thus, the background subtraction methods are applied in this paper for experimental analysis using OpenCV and OpenCL [15-16].

In this paper graphical user interface (GUI) has been developed for streaming and processing video using DASH data. Here, two methods for background subtraction are compared (Gaussian mixture model and k-nearest neighbors) [13-15], based on computing time (clock time) and selected processing unit (CPU/GPU). The analysis is performed to test the advantages of GPU approach. For coordination of processing units OpenCL C++ framework is used.

The paper is organized as follows. After the introduction, a brief explanation regarding MPEG-DASH standard and DASH data is given in Section II. Video processing using OpenCV and OpenCL is described in Section III. The details regarding the simulation performed in this paper are explained in Section IV. The obtained experimental results followed by discussion are presented in Section V. Finally, the conclusion is given in Section VI.

Katarina Popović is with the School of Electrical Engineering, University of Belgrade, Bulevar kralja Aleksandra 73, 11020 Belgrade, Serbia (e-mail: katarina.popovic994@gmail.com).

Ana Gavrovska is with the School of Electrical Engineering, University of Belgrade, Bulevar kralja Aleksandra 73, 11020 Belgrade, Serbia (e-mails: anaga777@gmail.com; anaga777@etf.rs).

Irini Reljin is with the School of Electrical Engineering, University of Belgrade, Bulevar kralja Aleksandra 73, 11020 Belgrade, Serbia (e-mails: irinitms@gmail.com; irini@etf.rs).

II. MPEG-DASH STANDARD AND DASH DATA

DASH, known also as MPEG-DASH, represents an adaptive bitrate streaming technique over the Internet delivered from conventional HTTP web servers [2-3, 5]. It is based on breaking the content into a sequence of HTTP-based file segments (chunks). Moreover, the content is made at different bitrates enabling the adaptive approach depending on the network conditions.

Content distribution architecture based on MPEG-DASH standard consists of four main parts: media capture and encoding, media origin servers, HTTP cache server and user devices. Content is stored on servers in form of media presentation, which can be related to video or audiovisual content and/or other data described by appropriate description (metadata). Thus, besides the media segments, there is a Media Presentation Description (MPD) [2, 4-7]. After TCP connection with server is established, a client first downloads MPD file. Based on information given in MPD, network state and users requests (or device characteristics), a client selects appropriate segments. Static MPD is used when the content is stored like in VoD applications, whereas dynamic MPD is used for live streaming or PVR (Personal Video Recorder) applications.

MPD is XML document that provides information about available segments. Hierarchical structure of DASH data, i.e. MPD file, is shown in Fig.1.

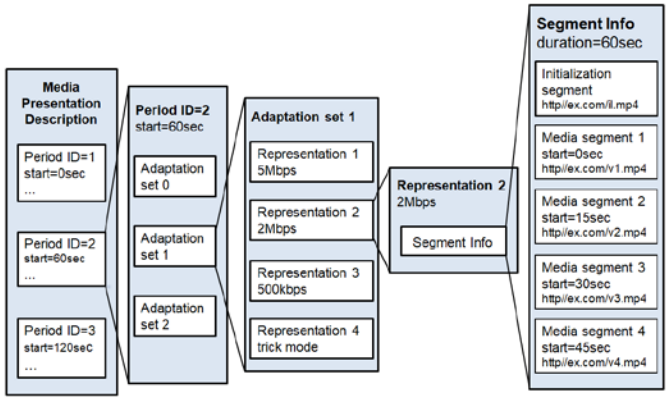


Fig. 1. Hierarchical structure of MPEG-DASH data.

MPD is divided according to time periods. One time period has multiple adaptation sets. Adaptation sets define different content (e.g. audio, video, text). One adaptation set has multiple representations. They represent the same content with different encoding or bitrates. Representation gives information about segments, their ID, location, beginning time and duration (that could be given implicitly or explicitly). First segment is usually initialization segment, and gives to the client information necessary for decoding.

For generating MPD (manifest) file in this paper MP4Box [17] is used. MP4Box is a multimedia packager that enables preparation of HTTP adaptive streaming content, and can be used for many manipulations on multimedia and ISO (International Organization for Standardization) files, like MP4.

DASH represents both media and delivery information. By parsing the manifest a DASH client should be able to reproduce the content. MPEG-DASH implementation is presented in Fig.2, where communication between Content Server and DASH Client is illustrated.

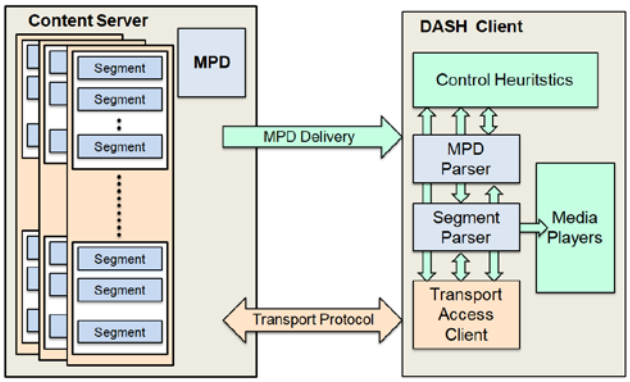


Fig.2. MPEG-DASH implementation in communication between Content Server and DASH Client.

III. VIDEO PROCESSING BASED ON OPENCV AND OPENCL

During video content analysis, motion analysis and object tracking are mostly performed. The popular cross-platform open source library for computer vision and machine learning OpenCV (Open Source Computer Vision Library) [15] contains modules providing common infrastructure for video application and service development.

One of the common tasks during video content or frame based analysis represents the background subtraction [9-12]. It represents a way to extract the moving foreground from the relatively static background and analyze motion within video sequence. Foreground modeling is usually more computationally demanding and more challenging due to the change of relevant pixel locations [11].

In this paper, for the purpose of testing two background-subtraction algorithms are used. Both of them combine statistical background image estimation and pixel-level Bayesian segmentation. One algorithm is based on GMM (Gaussian mixture model), whereas the second one represents knn (k-nearest neighbors) background subtraction [13-15]. If a pixel is adequately described by the predefined probability, the pixel can be considered as a part of background. GMM based method uses the assumption of initial state for probability description in comparison to knn based on the fixed number of frames [3, 13]. The mixture weights represent the time proportions of the colours that are more probable and stay longer within the scene. This is suitable for real time processing. Here, the methods are implemented using the default parameters [15], where their setting was not a part of this study.

A. OpenCL C++ framework

OpenCV [15] is used for image and video processing, while OpenCL (*Open Computing Language*) [16] is open source framework for hardware code accelerations. OpenCL C++ framework is used in this paper. OpenCL provides API (*application programming interface*) for coordination of code execution on heterogeneous processors [16].

OpenCV has a module dedicated to OpenCL [15-16]. In previous versions this was separate module, but from version 3.0 it is a part of core module. *Umat* (Unified matrix) class/type is introduced enabling rewriting data from one memory to another. It tells OpenCV functions to process data with OpenCL which may use OpenCL-enabled GPU or CPU otherwise. Algorithms that were implemented in *ocl* module are retained. This relatively new type enabled writing unique code that could run on both CPU and GPU or any other OpenCL device. Device/hardware selection for code executing is done by using instruction *ocl::setUseOpenCL(bool).* If it is set to false, the code will be executed on CPU, otherwise it will be executed on GPU or other OpenCL device, using parallel processing.

IV. Simulation

In the first step test video sequence is converted into several formats (five different spatial resolutions), encoded with H.264 with different bitrate. This is performed using ffmpeg [18], as in case of 160x90 video (the example of code line is shown below).

```
ffmpeg –i input.mp4 –s 160x90 –c:v libx264 –b:v 250k
–g 90 –an input_video_160x90_250k.mp4
```

In the second step MPD (manifest) file is made by MP4Box based on video sequences generated in the first step. Code line example is shown below.

```
MP4Box –dash 5000 –rap –profile
dashavc264:onDemand –mpd-title BBB –out
manifest.mpd –frag 2000 input_audio_128k.mp4
input_video_160x90_250k.mp4
input_video_320x180_500k.mp4
input_video_640x360_750k.mp4
input_video_640x360_1000k.mp4
input_video_1280x720_1500k.mp4
```

Command *-dash* represents duration of segment in milliseconds, command *-profile* sets profile to “on demand” or “live” streaming and *–out* represents the name of the MPD file. At the end of line all the sequences are listed.

User interface is prepared using Qt Creator [19] and C++ programming language. The interface allows downloading and parsing MPD file, selecting preferred video sequence for analysis, choosing processing units (CPU or GPU) and methods for background subtraction.

For MPD parsing *QdomDocument* and *QxmlReader* classes are used. After parsing, one of the video sequences is selected for testing the method (here background subtraction). Original frame of the test sequence “Big Buck Bunny” [20] is shown in Fig.3. Also, for the testing processing unit is selected (CPU or GPU). The number of tacts (clock time) is measured and used to compare the performance between CPU and GPU usage during testing. The time required for getting segments through network is taken into account. Function *clock()* from *ctime* class is used for measuring tact number per video sequence. The video sequence tact number is measured five times. This is followed by calculating mean tact number for each sequence. Absolute relative difference is calculated as:

$$D = \left| \frac{tact\_number(On) - tact\_number(Off)}{tact\_number(Off)} \right| \quad (1)$$

where On stands for computing on GPU, and Off for CPU.

Intel OpenCL platform is used for the analysis. The simulation is performed on laptop with Windows 10 (Intel Core i3-4005U CPU-1.70GHz, 4GB RAM). For the task of GPU based processing Intel HD Graphics 4400 is used.

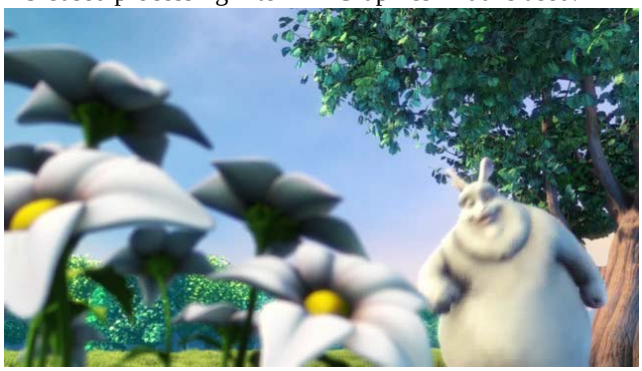


Fig. 3. “Big Buck Bunny” test video sequence (frame/scene example).

V. Experimental results

A. Manifest creating and type setting for MPD files

Results of a test sequence conversion are shown in Table I. After generating the MPD file with MP4Box and obtained results, user can get access to the manifest by using server’s public IP address (*http://[public_ip_address]/manifest.mpd*).

TABLE I
TEST VIDEO SEQUENCES

Test seq.	Resolution	Encoding	Bitrate	GOP *key frame* interval
1	160x90	H.264	250k	90
2	320x180	H.264	500k	90
3	640x360	H.264	750k	90
4	640x360	H.264	1000k	90
5	1280x720	H.264	1500k	90

In order to recognize manifest file (.mpd extension), it is necessary to set Internet Media Type (content type). This is also called MIME type (Multipurpose Internet Mail Extensions), originally created for emails helping to extend their limited capabilities. It indicates nature and format of a document. According to MPEG-DASH standard, type should be application with subtype dash+mpd [21], Fig.4.

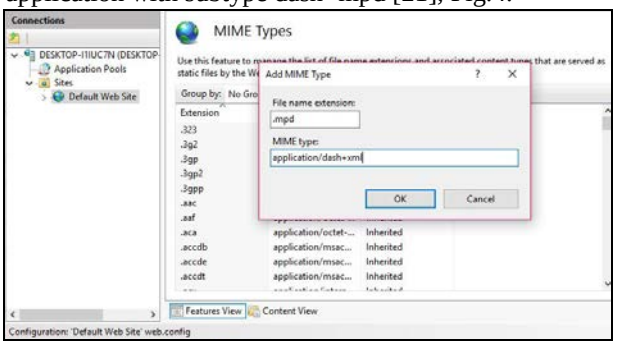


Fig. 4. Internet media type settings and MPD extension.

B. GUI for video processing

Realization of GUI using [19] is shown in Fig.5. Button MPD enables downloading and parsing MPD file. After parsing, in the right corner appears a list of all available video sequences. Streaming and processing starts by pressing Start/Refresh button. Processing unit is chosen by selecting on/off (On for GPU, Off for CPU). User can choose between GMM and knn by selecting mog2 or knn option, respectively. Also, user can set the frame number from which streaming starts.

For downloading a manifest, *QNetworkAccessManager* object is required, because *QDomDocument* cannot load xml document directly from network. *QNetworkAccessManager* object allows control over downloading content from network/internet. In this case, new class *mymanager* is written, for defining required signals and slots. For the purpose of defining and using a *QNetworkAccessManager* object, classes *QNetworkAccessManager* and *QNetworkReply* are used. In *mymanager* class, function *makeRequest* is defined. This function sends request for manifest downloading. If downloading is successful, signal *finished(QNetworkReply*)* is emitted and response is being forwarded to slot *readRead(QNetworkReply*)*. Response is written in variable *QbyteArray* type and signal *dataReadyRead(QByteArray)* is loaded. Then, within *collectData(QByteArray)* slot parsing of MPD file is done, and the first step is writing information on from variable *QByteArray* type to the variable with *QDomDocument* type.

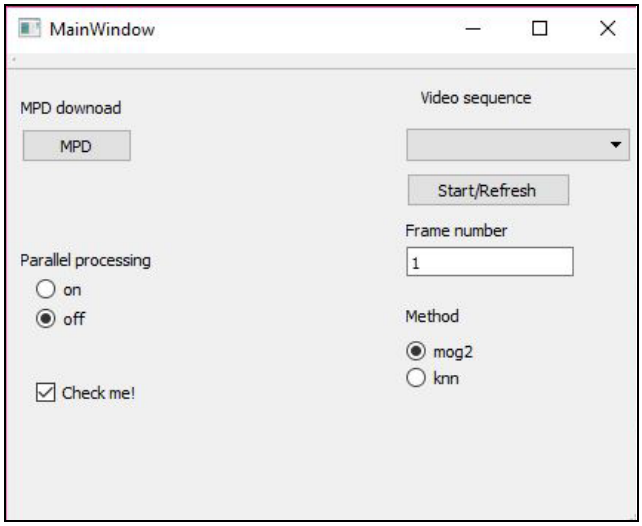


Fig. 5. MPD/DASH based graphical user interface.

C. GPU based background subtraction results

Video processing has been performed using two methods, where GMM showed better visual results in background subtraction compared to knn. GMM based video background subtraction result is presented in Fig. 6.

Five independent tests has been performed for each test sequence with different resolution and bitrate, and without or with GPU, meaning processing using CPU or processing on GPU.

Fig. 6. GMM based video background subtraction result.

Obtained results and mean values of the clock time (tact numbers) for processing using GMM are presented in Table II. In Table III CPU or GPU based background subtraction using GMM is compared to knn method.

TABLE II
VIDEO PROCESSING RESULTS USING GMM METHOD

Test sequences					
resolution, bitrate	160x90, 250k	320x180, 500k	640x360, 750k	640x360, 1000k	1280x720, 1500k
Test number	Clock time (tact numbers) without or with GPU (Off/On) – each row (1-5) is a test				
1	42201/ 47187	43611/ 43982	50510/ 43376	50676/ 42909	85074/ 58448
2	42922/ 43845	43651/ 44079	50560/ 47790	51015/ 42669	85443/ 59620
3	43360/ 43848	42717/ 43629	50922/ 41602	50757/ 43031	85089/ 58891
4	42557/ 43470	43760/ 43505	50236/ 43212	51066/ 42678	84940/ 59138
5	43421/ 43959	43494/ 43985	50411/ 42760	51364/ 42797	85099/ 59275
Mean values for five tests (rows above)					
Result	42896/ 44462	43447/ 43836	50528/ 43748	50976/ 42817	85129/ 59074

TABLE III
VIDEO PROCESSING RESULTS OF KNN COMPARED TO GMM METHOD

Test sequences (resolution, bitrate)				
160x90, 250k	320x180, 500k	640x360, 750k	640x360, 1000k	1280x720, 1500k
Clock time without or with GPU (Off/On) – mean values for GMM				
42896/ 44462	43447/ 43836	50528/ 43748	50976/ 42817	85129/ 59074
Absolute relative difference in percents (compared to CPU)				
3.7%	0.8%	13.4%	16.0%	30.6%
Clock time without or with GPU (Off/On) - mean values for knn				
42743/ 43022	47157/ 47073	73846/ 74807	73622/ 74012	180761/ 179075
Absolute relative difference in percents (compared to CPU)				
0.7%	1.8%	1.3%	0.5%	0.9%

The result of background subtraction and/or measured clock time can be shared with another user, by opening new TCP connection. Based on the results given in Table II and Table III, GMM based computing on GPU is above 30% faster compared to CPU computing for high resolution. Also, GMM based computing was 30-50% faster compared to knn for higher resolutions. Computing knn on GPU did not show much difference compared to knn computing on CPU. Around 1-2% difference was obtained.

VI. Conclusion

In this paper we performed experimental analysis based on OpenCL C++ framework platform and MPEG-DASH standard. Graphical user interface has been developed for DASH (MPD) data. It enables adaptive streaming of video content on demand and its processing. Two methods for background subtraction are compared, which are based on Gaussian mixture model and k-nearest neighbors. GMM method showed better results from the computational time aspect compared to knn regardless the processor unit. GMM method computing on GPU for higher resolutions gave good results. The results gave above 30% difference in clock time compared to CPU, and above 30% difference compared to knn regardless the processor unit.

Efficient processing is important for real-time applications. Providing new options for the user interface are expected as next steps. The future work should be oriented towards the analysis of different parallel processing and GPU based possibilities, machine and deep learning optimizations, providing efficient solutions for video processing and multimedia systems.

Acknowledgment

This research is partially funded by Serbian Ministry of Education, Science and Technological Development through the project TR32048.

References

[1] Q. Fan, H. Yin, G. Min, P. Yang, Y. Luo, Y. Lyu, H. Huang, and L. Jiao, "Video delivery networks: Challenges, solutions and future directions," *Computers & Electrical Engineering*, vol. 66, pp. 332-341, 2018.

[2] Mpeg, I. "Information technology-dynamic adaptive streaming over http (dash)-part 1: Media presentation description and segment formats," ISO/IEC MPEG, Tech. Rep. 2012.

[3] K. Popović, "Razvoj korisničkog interfejsa za manipulaciju video sadržajem zasnovan na OpenCL C++ platformi i MPEG-DASH standardu," BSc thesis (in Serbian), University of Belgrade, ETF, 2018.

[4] M. Sayit, C. Cihat, H.U. Yildiz, and B. Tavli. "DASH-QoS: A Scalable Network Layer Service Differentiation Architecture for DASH over SDN," *Computer Networks*, pp.1-43, 2019.

[5] I. Sodagar, "The mpeg-dash standard for multimedia streaming over the internet," *IEEE MultiMedia*, vol. 8, no. 4, pp. 62-67, 2011.

[6] T. Stockhammer, and I. Sodagar. "Mpeg dash: The enabler standard for video delivery over the internet," *SMPTE Motion Imaging Journal*, vol. 121, no. 5, pp. 40-46, 2012.

[7] S. Zhao, L. Zhu, D, Medhi, PL Lai, and S. Liu. "Study of user QoE improvement for dynamic adaptive streaming over HTTP (MPEG-DASH)," In Proc. of 2017 *International Conference on Computing, Networking and Communications (ICNC)*, pp. 566-570. IEEE, 2017.

[8] P. Guler, D. Emeksiz, A. Temizel, M. Teke, and T. T. Temizel, "Real-time multi-camera video analytics system on GPU," *Journal of Real-Time Image Processing*, vol. 11, no. 3, pp. 457-472, 2016.

[9] D. D. Bloisi, A. Pennisi, and L. Iocchi. "Parallel multi-modal background modeling," *Pattern Recognition Letters* 96: pp. 45-54, 2017.

[10] W. Porr, J. Easton, A. Tavakkoli, D. Loffredo, and S. Simmons, "GPU Accelerated Non-Parametric Background Subtraction," In Proc. of *International Symposium on Visual Computing*, pp. 629-639, Springer, Cham, 2018.

[11] D. Berjon, C. Cuevas, F. Moran, and N. Garcia, "GPU-based implementation of an optimized nonparametric background modeling for real-time moving object detection," *IEEE Transactions on Consumer Electronics*, vol. 59, no. 2, pp. 361-369, 2013.

[12] H. Wu, N. Liu, X. Luo, J. Su, and L. Chen, "Real-time background subtraction-based video surveillance of people by integrating local texture patterns," *Signal, Image and Video Processing*, vol. 8, no. 4, pp. 665-676, 2014.

[13] T. Bouwmans, F. El Baf, and B. Vachon, "Background modeling using mixture of gaussians for foreground detection-a survey," *Recent Patents on Computer Science*, vol. 1, no. 3, pp.219-237, 2008.

[14] Z. Zivkovic, and F. Van Der Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction," *Pattern recognition letters*, vol. 27, no. 7, pp. 773-780, 2006.

[15] OpenCV, https://opencv.org/, (*accessed* 30.06.2018).

[16] Munshi, A. "The OpenCL Specification (2012)." URL:{ https://www.khronos.org/registry/cl/specs/opencl-1.2.pdf}.

[17] GPAC-MP4Box, https://gpac.wp.imt.fr/, https://gpac.wp.imt.fr/mp4box/ (*accessed* 10.02.2019).

[18] Ffmpeg, https://ffmpeg.org/ (*accessed* 12.04.2019).

[19] QT Creator, https://doc.qt.io/ (*accessed* 30.06.2018).

[20] Big Buck Bunny, https://peach.blender.org/ (*accessed* 12.04.2019).

[21] Media Types listed by IANA (Internet Assigned Numbers Authority), http://www.iana.org/assignments/media-types/media-types.xhtml (*accessed* 12.04.2019).

Group delay equalization of discrete Butterworth tan filters in the continuous domain

Negovan Stamenković, Nikola Stojanović and Milan Savić

***Abstract*— We present new techniques to approximate the magnitude response of discrete Butterworth tan filters with the delay characteristic which is constant in a maximally flat sense. The algorithm is used to obtain stable allpass filter, which acts as a group delay equalizer, with the aim to equalize group delay of the discrete tan filter in a maximal flat sense. The proposed method relies on a set of nonlinear equations, derived directly from the flatness conditions of the group delay response of the equalized filter at the origin in the analog domain, in order to obtain the unknown coefficient values of the discrete allpass filter. The algorithm implemented in the symbolic Matlab platform returns the coefficients of allpass filter. In the given example, firstly we design an discrete tan filter with a maximally flat magnitude at origin, and secondly, we augment the system with cascade connection of non-minimum phase allpass discrete filter, in order for the group delay response of the whole filter to equalize in a maximally flat sense.**

***Keywords*— Modified-Jacobi polynomials, return loss, stepped impedance filter**

I. Introduction

In recently published papers [1] and [2], the group delay equalization of analog and discrete Butterworth sine (all-pole) filters in a maximally flat sense, respectively, is presented. The conditions for the maximally flat group delay response of a lowpass filter require that the first $2m+1$ derivatives of the group delay function in the discrete frequency variable z, evaluated at $z=1$ to be all zeros [3], where m is the equalizer degree. The proposed method derives a set of nonlinear equations that satisfy the conditions for the group delay response to be flat with certain degrees of flatness at origin. In that method the system of nonlinear equation is expressed in terms of the first $2m$ derivatives of the numerator and the first $2m$ derivatives of the denominator of group delay response.

Let us denote the constant group delay equalization of the continuous and discrete filters can be in an equal ripple or in a maximally flat sense. In point of fact, there are several techniques for designing group delay equalizers primarily for continuous filters [4], [5], [6]. Unfortunately, these techniques cannot be applied to the discrete filters design through a s-to-z transformation, such as the bilinear transformation, because the group delay characteristic is not preserved under such a transformation [7].

N. Stamenković and M. Savić, University in Pristina-Kosovska Mitrovica, Faculty of Science, L. Ribara 29, 28220 K. Mitrovica, Serbia; e-mail: (negovan.stamenkovic@pr.ac.rs).

N. Stojanović, Faculty of Electronic Engineering, A. Medvedeva 14, 18000 Niš, Serbia; e-mail: (nikola.stojanovic@elfak.ni.ac.rs).

The design the group delay equalizer, on the other hand, can be accomplished by using an iterative technique direct in the z-domain that searches the constant group delay solution in an equal ripple sense by using Fletcher's optimization algorithm [8] or in the quasi equal ripple sense by using neural networks [9]. The evolutionary algorithm, the Differential Evolution Algorithm in particular is used to obtain an equal ripple group delay frequency response in the passband of lowpass filter [10]. The mentioned procedures are for designing optimum group delay equalizers that whole filter has the group delay response in an equal ripple sense.

In this paper, a novel approach to equalization of the discrete low-pass tan filter group delay response is presented. The proposed method relies on a set of nonlinear equations, derived from the flatness conditions of the equalized group delay response at the origin in continuous plane [1], to obtain the unknown coefficient values of the discrete allpass filters. There is a trade off between creating a system of nonlinear equations without a differentiation operation and standard bilinear transformation.

II. Digital Butterworth tan filters transfer function

The magnitude squared function of the nth degree lowpass discrete Butterworth tan filter [11], [12], which is a logical equivalent of an analog Butterworth filter, is given by:

$$|H_B(e^{j\omega T})|^2 = \frac{1}{1+\varepsilon^2\left[\dfrac{\tan^2\frac{\omega}{2}}{\tan^2\frac{\omega_p}{2}}\right]^n} \tag{1}$$

where ω_p is cutoff edge passband angular frequency, ε can be calculated from the specified passband edge attenuation a_{max}, $\varepsilon^2 = 10^{a_{max}} - 1$, and $2n-1$ is the degree of the flatness of the amplitude characteristic at the origin (the first $2n-1$ derivatives with respect to ω are equal to zero).

Appealing to analytic continuation by substituting ω by $-j\log(z)$, we obtain

$$|H_B(e^{j\omega T})|^2 = \frac{\tan^{2n}\dfrac{\omega_p}{2}}{\tan^{2n}\dfrac{\omega_p}{2} + \varepsilon^2(-1)^n\left(\dfrac{z-1}{z+1}\right)^{2n}} \tag{2}$$

The poles of Butterworth tan transfer function are found by substituting

$$q = \frac{z-1}{z+1} \tag{3}$$

in (2), from which we can find that the $2n$ poles of $q = x + jy$ are uniformly spaced around a circle of radius $\rho = \tan(\omega_p/2)/\sqrt[n]{\varepsilon}$. The poles of $z = u + jv$ are then readily found by using the inverse function to (3), namely

$$z = \frac{1+q}{1-q} \tag{4}$$

From (3) and (4) the corresponding poles in the z plane are computed to be in closed form [13]. For odd values of n, the $2n$ poles of z have the u_m and jv_m coordinates

$$\begin{aligned} u_m =& \frac{1-\rho^2}{1-2\rho\cos\dfrac{m\pi}{n}+\rho^2} \\ v_m =& \frac{2\rho\sin\dfrac{m\pi}{n}}{1-2\rho\cos\dfrac{m\pi}{n}+\rho^2} \end{aligned} \tag{5}$$

for $m = 0, 1, 2, \ldots, 2n-1$. For even values of n the coordinates are

$$\begin{aligned} u_m =& \frac{1-\rho^2}{1-2\rho\cos\dfrac{(2m+1)\pi}{2n}+\rho^2} \\ v_m =& \frac{2\rho\sin\dfrac{(2m+1)\pi}{2n}}{1-2\rho\cos\dfrac{(2m+1)\pi}{2n}+\rho^2} \end{aligned} \tag{6}$$

Consider the nth degree real-coefficient discrete tan filter transfer function, where only the poles inside the unit circle are to be included in the transfer function, having the form

$$H_F(z^{-1}) = \frac{h(1+z^{-1})^n}{\sum_{i=0}^{n} \mathsf{a}_i z^{-i}}, \tag{7}$$

with $\mathsf{a}_0 = 1$, whose gain is adjusted to unity at dc. Since the numerator of the transfer function can always be chosen as a mirror-image polynomial [7, p. 534] which does not affect to the linearity of phase, this group delay equalization will be performed like as by the discrete sine transfer function.

The group delay function $\tau_F(z)$ of a discrete tan low-pass transfer function (7) is a rational function in z. In the Appendix it is shown' that the group delay function is given by

$$\begin{aligned} \tau_B(z) = -\frac{1}{2}\Bigg[& \frac{z}{H_B(z)}\frac{\mathrm{d}H_B(z)}{\mathrm{d}z} \\ & - \frac{z}{H_B(z^{-1})}\frac{\mathrm{d}H_B(z^{-1})}{\mathrm{d}z}\Bigg]_{z=e^{j\omega}} \end{aligned} \tag{8}$$

After the addition, the denominator and the numerator of the group delay function (8) are mirror-image polynomials of degree $2n$, where n is the filter degree. In other words, the group delay is a *mirror-image rational function*.

Using (8), it can be derived an expression for the group delay value of this transfer function (7) at the origin ($z = 1$) as $\tau_B(1) = -(n/2 - \sum_{i=1}^{n} i\mathsf{a}_i / \sum_{i=0}^{n} \mathsf{a}_i)$. This formula is valid for arbitrary discrete tan filters with image-mirror polynomial in the numerator, because it does not affect to the phase linearity.

III. Group delay equaliser design

The group delay of the original filter can be equalized by connecting an allpass filter (delay equalizer) in cascade with the original filter. The magnitude response of an allpass filter is unity over its entire frequency range. Let $H_E(\mathbf{c}, z)$ denote an mth degree causal real-coefficients discrete transfer function of an allpass filter, i.e.,

$$H_E(\mathbf{c}, z) = \frac{c_m z^m + c_{m-1} z^{m-1} + \cdots + c_1 z + c_0}{c_0 z^m + c_1 z^{m-1} + \cdots + c_{m-1} z + c_m}, \tag{9}$$

where $c_0 = 1$ and $\mathbf{c} = [c_1, c_2, \ldots, c_m]$ is the real vector coefficient. The numerator of the allpass filters (9) is the reciprocal polynomial of the denominator polynomial, that implies that the poles and zeros of a real coefficient allpass filter exhibit mirror image symmetry with respect to the unit circle. To guarantee the stability of allpass filter (9), the poles of its transfer function have to be inside of the unit circle.

Applying (8) to an allpass function (9), an expression for the group delay of the allpass filter can be derived as

$$\tau_E(\mathbf{c}, z) = \frac{\beta_1(\mathbf{c}) z^{2m-1} + \cdots + \beta_m(\mathbf{c}) z^m + \cdots + \beta_1(\mathbf{c}) z}{\alpha_0(\mathbf{c}) z^{2m} + \cdots + \alpha_m(\mathbf{c}) z^m + \cdots + \alpha_0(\mathbf{c})} \tag{10}$$

As expected, the numerator and denominator polynomial form a mirror-image rational function. For example, the coefficients of the corresponding mirror-image rational function of the third degree group delay allpass filter can be calculated as

$$\begin{aligned} \beta_1(\mathbf{c}) =& c_2 - c_1\, c_3 & \alpha_0(\mathbf{c}) =& c_3 \\ \beta_2(\mathbf{c}) =& 2\, c_1 - 2\, c_2\, c_3 & \alpha_1(\mathbf{c}) =& c_2 + c_1\, c_3 \\ \beta_3(\mathbf{c}) =& c_1^2 - c_2^2 - 3\, c_3^2 + 3 & \alpha_2(\mathbf{c}) =& c_1 + c_1\, c_2 + c_2\, c_3 \\ & & \alpha_3(\mathbf{c}) =& c_1^2 + c_2^2 + c_3^2 + 1 \end{aligned} \tag{11}$$

The total group delay of the cascade connection of the original filter and the equalizer is defined as a sum of the group delays on each of them, so that:

$$\tau_{BE}(\mathbf{c}, z) = \tau_B(z) + \tau_E(\mathbf{c}, z) \tag{12}$$

Let $\tau_{BE}(z) = B(z)/A(z)$. The first $2m+1$ derivatives of $\tau_{BE}(z)$ will be zero at origin, if $\tau_{BE}(z)|_{z=1} = \lambda$ and the first m even-order derivatives of $B(z)$ at $z = 1$, are equal to the first m even-order derivatives of $A(z)$ at $z = 1$ multiplied by λ, i.e.

$$\frac{\mathrm{d}^k \tau_{BE}(z)}{\mathrm{d}z^k}\Big|_{z=1} = 0, \quad \text{for } k = 1, 2, \ldots, 2m+1 \tag{13}$$

if

$$\frac{\mathrm{d}^k B(z)}{\mathrm{d}z^k}\Big|_{z=1} - \lambda \frac{\mathrm{d}^k A(z)}{\mathrm{d}z^k}\Big|_{z=1} = 0, \quad \text{for } k = 2, 4, \ldots, 2m \tag{14}$$

where m is a degree of allpass filter that controls the degree of flatness at origin. Hence, the group delay response of the tan filters can be equalised in the maximally flat sense. It should be noticed the equalization to be carried out directly in the discrete domain.

On the other hand, the mirror-image rational function in

the discrete domain is the counterpart of the even rational function in the continuous domain [14]. The well known inverse bilinear transformation

$$z = \frac{1+s}{1-s} \tag{15}$$

is used to transform both discrete group delay functions, (8) and (10), into a continuous domain. The group delay response of the Butterworth tan filter/equalizer combination in the continuous s-domain can be expressed as

$$\tilde{\tau}_{BE}(\mathbf{c},\omega) = \tilde{\tau}_B(s)|_{s=j\omega} + \tilde{\tau}_E(\mathbf{c},s)|_{s=j\omega} \tag{16}$$

In this paper, we will focus only on designing a discrete allpass filter (9) which will act as an equalizer of the discrete tan filter. If a vector coefficient c is known, then the poles and zeros of the allpass filter prototype can also be calculated. Thus, the resulting group delay function $\tau_{BE}(\mathbf{c},z)$ approximates constant group delay response in the maximally flat sense over the passband interval.

The group delay of the equalized filter $\tilde{\tau}_{BE}(\mathbf{c},\omega)$ is an even rational function in ω as

$$\tilde{\tau}_{BE}(\mathbf{c},\omega) = \frac{b_0 + \cdots + b_{2m}\omega^{2m} + \cdots + b_{2(n+m)}\omega^{2(n+m)}}{a_0 + \cdots + a_{2m}\omega^{2m} + \cdots + a_{2(n+m)}\omega^{2(n+m)}} \tag{17}$$

with the coefficients being dependent on m allpass filter coefficients $\mathbf{c}$. Since, there are only m unknown parameters of the allpass network design, only $2m+1$ derivatives of the group delay response given by (17) can be equal to zero, including all odd order derivatives which are identically zero. This results in simpler m nonlinear equations [1] and this simplifies the vector coefficient $\mathbf{c}$ calculation

$$f_k(\mathbf{c}) = \frac{b_k}{a_k} - \frac{b_0}{a_0} = 0, \text{ for } k = 2,4,\ldots,2m \tag{18}$$

The m unknown coefficients for the discrete group delay equalizer design to be solved by iterative techniques. The Matlab Symbolic Toolbox can be used to solve the system of m nonlinear equations (18) for the arbitrary degree of both the lowpass tan filter and the group delay equalizer (i.e. allpass network $H_E(\mathbf{c},z)$), by using command `[x,fval]=fsolve(f,x_o)`, whereas $\mathbf{x} = \mathbf{c}$ and $\mathbf{f}(\mathbf{c}) = [f_2(c), f_4(c), \ldots, f_{2m}(c)]^T$. The exponent T stands for matrix operation transpose.

The denominator coefficients of the mth degree Butterworth tan transfer function (7), whose group delay will be corrected, can be used as initial solution values for the coefficient vector $\mathbf{x\_o} = [c_2, c_4, \cdots, c_{2m}]$.

As is mentioned Matlab rutine, which is used to solve system of nonlinear equations $\mathbf{f}(\mathbf{c}) = 0$, employs a number of least squares algorithms that find the solution to the problem

$$\Big[\min_{c_2,\ldots,c_{2m}} \sum_{k=1}^{m} f_k^2(\mathbf{c})\Big]^{\frac{1}{2}} \le \delta \tag{19}$$

where δ is an arbitrarily small number. Since the minimum of (19) is $\delta = 0$ and this occurs when all $f_i(\mathbf{c}) = 0$, a solution $\delta > 0$ to (19) will also be a solution to the equation system (18).

Finally, the group delay of the filter/equalizer combination is constant in the maximally flat sense since it has the first $2m+1$ derivatives that are approximately equal to zero at the origin, because the zero value of derivatives cannot be achieved.

The quality of an equalizer is inversely related to the maximum variation of τ_{BE} over the frequency band of interest. A measure that can be used to assess the quality of a group delay equalizer design can, therefore, be defined as

$$Q_\tau = \frac{\max\limits_{0\le\omega\le\omega_c} \tau_{BE}(\omega) - \min\limits_{0\le\omega\le\omega_c} \tau_{BE}(\omega)}{\max\limits_{0\le\omega\le\omega_c} \tau_{BE}(\omega) + \min\limits_{0\le\omega\le\omega_c} \tau_{BE}(\omega)} 100 \tag{20}$$

We are used that the minimum value of group delay value is expected at the origin, $\min\limits_{0\le\omega\le\omega_c} \tau_{BE}(\omega) = \tau_{BE}(0)$, then the quality of group delay equalization does not depend on the passband edge attenuation. Hence,Q_τ will be referred to as the maximum group delay deviation, hereafter.

IV. Design example

In order to demonstrate the effectiveness of the present method for the group delay equalisation for discrete Butterworth tan filter a practical example has been worked out. The fifth-degree group delay equalizer for the seventh-degree discrete Butterworth tan filter with the normalized $a_{max} = 3.0103$ dB ($\varepsilon = 1$) at the cutoff point of $\omega_c = 0.30\pi$ radians is considered as a design example. The transfer function of the 7th degree Butterworth tan filter is calculated by using equations (5) and (7)

$$H_B(z^{-1}) = \frac{0.000962894(1+z^{-1})^7}{\begin{array}{r}1 - 2.78251z^{-1} + 3.96681z^{-2} - 3.40515z^{-3} \\ +1.87586z^{-4} - 0.65094z^{-5} + 0.13085z^{-6} \\ -0.01166z^{-7}\end{array}}$$

The general form of the 5th degree allpass filter transfer function is given by Eq. (10). The system of nonlinear equations (18) is solved with an initial solution `x_o= [-2.1830 2.2787 -1.3028 0.3969 -0.0508]`, calculated as the denominator coefficients of the 5th degree Butterworth tan filter. Matlab to return the values of allpass filter coefficients as follows

Listing 1. Allpass discrete filter coefficients

```
ans =

[ c1,    -2.209203171366232]
[ c2,     2.082326315189515]
[ c3,    -1.037836012889430]
[ c4,     0.271589008818471]
[ c5,    -0.029678379559256]
```

with `c0=1`.

The pole-zero locations of the 7th-degre discrete Butterworth tan lowpass filter with cutoff frequency $\omega_p = 0.3\pi$ and for the 5th-degree group delay equalizer in the maximally flat sense are shown in Fig. 1. In other word, this is pole-zero plot of the whole filter. Therefore, equalized filter is a nonminimum phase system, because allpass network introduces the zeros outside unit circle.

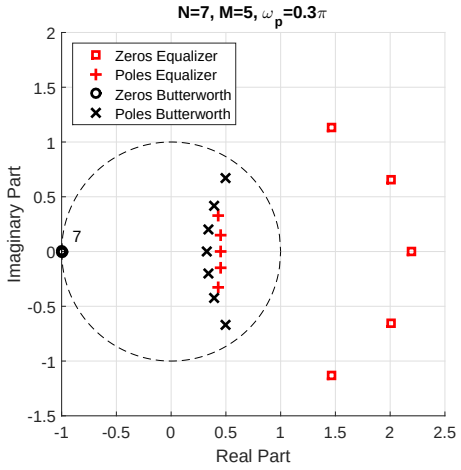


Fig. 1. The pole-zero locations of the 7th-degre discrete Butterworth tan lowpass filter with cutoff frequency $\omega_p = 0.3\pi$ and of the group delay equalizer for the maximally flat sense.

The value of the function $\mathbf{f}(\mathbf{c})$ at the solution $\mathbf{x}$ Matlab is returned in the vector `fval`, as in Listing 2 is given:

Listing 2. The accuracy of the solution of a nonlinear equation system

```
fval =

[ f1,     -0.00000000000000013603889448]
[ f2,      0.00000000000000066193322442]
[ f3,      0.00000000000000010083713782]
[ f4,      0.00000000000000206424658027]
[ f5,     -0.00000000000000233619779576]
```

Since $m = 5$ then the first eleven derivatives of group delay at $z = 1$ are approximately equal to zero, as is shown in Matlab Listing 3, where `lam`=$\lambda = 15.112039764842292$ is the group delay value of the whole filter at origin, see eq. (14).

Listing 3. Derivatives of group delay at z=1

```
ans=

[diff(B,1)-lam*diff(A,1),    0.0]
[diff(B,2)-lam*diff(A,2),    0.00000000023275]
]diff(B,3)-lam*diff(A,3),    0.00000000080888]
[diff(B,4)-lam*diff(A,4),    0.00000000203123]
[diff(B,5)-lam*diff(A,5),    0.00000000474683]
[diff(B,6)-lam*diff(A,6),    0.00000001169507]
[diff(B,7)-lam*diff(A,7),    0.00000003442634]
[diff(B,8)-lam*diff(A,8),    0.00000015204330]
[diff(B,9)-lam*diff(A,9),    0.00000218574755]
[diff(B,10)-lam*diff(A,10), 0.00000484575153]
[diff(B,11)-lam*diff(A,11),-0.00001778180469]
[diff(B,12)-lam*diff(A,12),-78.6647341133690]
```

Fig. 2 presents the comparison of a group delay response of the Butterworth tan filter with yhe group delay of the equalized system. The maxumally flat attenuation response of the Butterworth tan filter, at the same time for the whole filter, is also given in this figure. As one can notice, the compensation of constant group delay characteristics in a maximally flat sense brings that the difference between maximum and minimum values of the group delay characteristic in the passband has been lowered. The maximum group delay deviation (20) of the equalized Butterworth tan filter is $Q_{\tau_{BE}} = 0.055192597270427$, and it is about seven times smaller in comparison with the maximum group delay deviation of the Butterworth tan filter without equalizatin $Q_{\tau_B} = 0.370319660189653$.

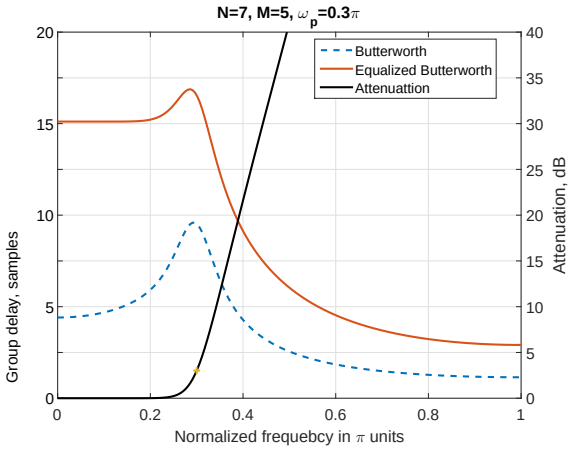


Fig. 2. Group delay equalization of a 7th-degree Butterworth tan filter with the bandwidth value of 0.30π radians; the frequency responses of the original Butterworth filter and group delay frequency responses of Butterwworth filter/equaliser with $m = 5$ combination.

On the other hand, the group delay value of group delay equalized filter (15.112039 samples) is obviously greater than in the Butterworth tan filter (4.409946 samples), however this value does not play an important role in many applications. From the engineering point of view, the important feature is constancy of the group delay in the filter pass-band.

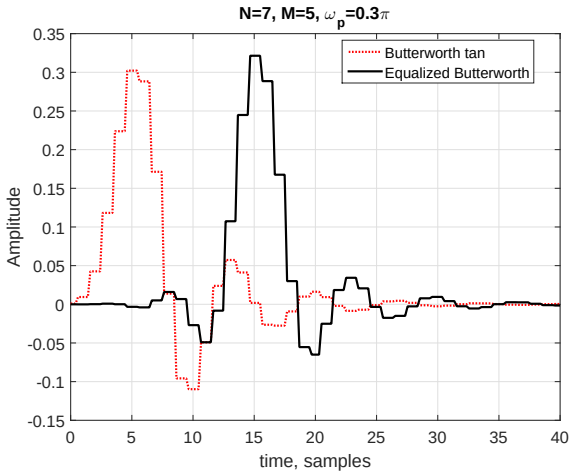


Fig. 3. Impulse response of a 7th-degree Butterworth tan filter with the bandwidth value of $\omega_p = 0.30\pi$ radians, and the impulse responses of the Butterworth tan filter/equaliser with $m = 5$ combination.

Improvement of the transient (dynamic) properties of the group delay equalized Butterworth tan filter is given in Fig. 3 which shows the impulse response of the Butterworth tan and equalized filter. It can be seen, that the introduction of a non-minimum-phase transfer function

obviously introduces ripples (or echoes) before the rise of the impulse response output from such a system. Number of pre-transition oscillation is equal to the degree of allpass equaliser.

It can be seen that side lobes of the impulse response are suppressed and man lobe is more symmetric.

V. Conclusion

A novel optimization algorithm has been applied to the design of the allpass filters for group delay equalized for a discrete Butterworth tan filters in a maximally flat sense. The proposed method relies on a set of nonlinear algebraic equations, derived directly from the flatness conditions of the group delay response at the origin in the analog domain, which is much simpler than that derived in the discrete domain, whose solution gives the unknown values of the allpass filter coefficients. Since $2m+1$ derivatives of group delay exist and they are zero at the origin, where m is the allpass filter degree, the equalized group delay response is called maximally flat.

The Matlab routine `fsolve` is used to solve this set of nonlinear equations using a quasi-Newton method. The proposed technique is illustrated by designing the 5th degree group delay equalizer for a 7th degree Butterworth tan filter with a normalized cutoff frequency of $\omega_p = 0.30\pi$ rad. The proposed method can be used for an arbitrary type discrete tan approximation, for an arbitrary filter degree, and with arbitrary filter bandwidth.

Since the numerator of the recursive rational transfer function can always be chosen as a mirror-image polynomial, which does not affect the linearity of phase, the proposed method can be extended towards the design of group delay equalizers for other types of filters with rational transfer function, for example, Elliptic and Inverse Chebyshev filters.

Appendix

To derive (8), let

$$H_F(e^{j\omega}) = |H_F(e^{j\omega})|e^{j\phi_F(e^{j\omega})}$$

Then

$$H_F(e^{-j\omega}) = |H_F(e^{j\omega})|e^{-j\phi_F(e^{j\omega})}$$

Therefore

$$\frac{H_F(e^{j\omega})}{H_F(e^{-j\omega})} = e^{2j\phi_F(e^{j\omega})}$$

Then

$$\log H_F(e^{j\omega}) - \log H_F(e^{-j\omega}) = 2j\phi_F(e^{j\omega})$$

Therefore, if we let $z = \exp(j\omega)$, we can write $\mathrm{d}z/\mathrm{d}\omega = jz$ and

$$\log H_F(z) - \log H_F(z^{-1}) = 2j\phi_F(z)$$

or

$$\phi_F(z) = \frac{1}{2j}\Big[\log H_F(z) - \log H_F(z^{-1})\Big]$$

Th group delays function of the above phase function can be derived as given below

$$\begin{aligned}\tau_F(z) &= -\frac{\mathrm{d}\phi_Z(z)}{\mathrm{d}z}\frac{\mathrm{d}z}{\mathrm{d}\omega} \\ &= -\frac{1}{2j}\Big[\frac{1}{H_F(z)}\frac{\mathrm{d}H_F(z)}{\mathrm{d}z} - \frac{1}{H_F(z^{-1})}\frac{\mathrm{d}H_F(z^{-1})}{\mathrm{d}z}\Big]jz\end{aligned}$$

The equation (8) is derived.

Acknowledgement

This paper is supported by Project TR 32009 financed by Ministry of Education and Science, Republic of Serbia. This paper is supported by the University in Pristina-Kosovska Mitrovica, Faculty of Science as a part of the Internal-Macro project. Modeling information transmition systems that use light. The authors wish to thank Dr. Vidosav Stojanović, Professor at University of Niš, Faculty of Electronic Engineering, for help and useful discussions.

References

[1] N. Stojanović, I. Krstić, N. Stamenković, and G. Perenić, "Butterworth transfer function with the equalized group delay response in the maximally flat sense," *Electronics Letters*, vol. 54, no. 25, pp. 1436–1438, Dec. 2018.

[2] N. Stamenković, N. Stojanović, and G. Perinić, "Group delay equalization of polynomial recursive digital filters in maximal flat sense," *Journal of Circuits Systems and Computers*, vol. 28, pp. 1–12, Nov. 2018, accepted manuscript to appear in JCSC.

[3] L. Bruton and P. Ramamoorthy, "Maximally flat low-pass transfer function synthesis using continuous and discrete filters," *IEEE Transactions on Circuits and Systems*, vol. 23, no. 8, pp. 510–513, Aug. 1976.

[4] X. Huang and M. Caron, "A novel type-based group delay equalization technique," in *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, Paris, France, May 30/June 02, 2010, pp. 2836–2839.

[5] P. Žiška and J. Vrbata, "Method for design of analog group delay equalizers," in *Proc. of International Symposium on Circuits and Systems*. Island of Kos, Greece: IEEE, May 21–24, 2006, pp. 445–448.

[6] K. H. K. Zaplatilek, P. Ziška, "Practice utilization of algorithms for analog filter group delay optimization," *Radioengineering*, vol. 16, no. 1, pp. 7–15, Apr. 2007.

[7] A. Antoniou, *Digital Signal Processing - Signals Systems and Filters*, 1st ed. New York: McGraw-Hill, 2006.

[8] C. Charalambous and A. Antoniou, "Equalisation of recursive digital fitters," *IEE Proceedings, G - Electronic Circuits and Systems*, vol. 127, no. 5, pp. 219–225, Oct. 1980.

[9] M. F. Quélhas and A. Petraglia, "Optimum design of group delay equalizers," *Digital Signal Processing*, vol. 21, no. 1, pp. 1 – 12, Jan. 2011.

[10] P. Žiška and M. Laipert, "Analog group delay equalizers design based on evolutionary algorithm," *Radioengineering*, vol. 15, no. 1, pp. 1–5, Apr. 2006.

[11] S. Hazra and S. D. Roy, "A comparison of digital tan and sine filters with the generating analog filter," *Journal of the Franklin Institute*, vol. 292, no. 3, pp. 225 – 230, Sept. 1971. [Online]. Available: https://doi.org/10.1016/0016-0032(71)90054-8

[12] N. Stamenković and V. Stojanović, "On the design transitional Legendre-Butterworth filters," *International Journal of Electronics Letters*, vol. 2, no. 3, pp. 186–195, 2014.

[13] C. M. Rader and B. Gold, "Digital filter design techniques in the frequency domain," *Proceedings of the IEEE*, vol. 55, no. 2, pp. 149–171, Feb. 1967.

[14] V. Ramachandran and M. Ahmadi, "Multivariable mirror-image and anti-mirror-image polynomials obtained by bilinear transformations," *IEEE Transactions on Circuits and Systems*, vol. 34, no. 9, pp. 1088–1090, Sept. 1987.

Analiza značaja DCT koeficijenata u objektivnoj proceni kvaliteta slike zasnovanoj na promeni kontrasta

Nenad Stojanović, Boban Bondžulić i Ivana Stojanović

***Apstrakt*—U radu je dat predlog metode za procenu kvaliteta slike koji se zasniva na poređenju originalne i test slike. Prilikom određivanja vrednosti kvaliteta slike koristi se diskretna kosinusna transformacija, pri čemu se na lokalnom nivou određuje kontrast koji daljim usrednjavanjem daje krajnju vrednost kvaliteta. U radu je pokazano da prilikom proračuna, na krajnju procenu kvaliteta utiče broj korišćenih koeficijenata diskretne kosinusne transformacije. Performanse predložene mere date su kroz linearnu korelaciju i korelaciju rangova sa subjektivnim ocenama kvaliteta nakon testiranja na četiri javno dostupne baze slika, pri čemu su tri baze slika sa višestrukim degradacijama.**

***Ključne reči*—Procena kvaliteta slike, očuvanje kontrasta, diskretna kosinusna transformacija, višestruke degradacije.**

I. Uvod

PROCENA kvaliteta različitih tipova digitalnih signala je postala sve značajnija u sistemima prenosa gde se kvalitet pružene usluge korisniku može prilagoditi korišćenjem složenijih, ali efikasnijih algoritama u odnosu na najčešće korišćenu verovatnoću bitske greške (*Bit Error Rate*, BER). Sa povećanjem kapaciteta telekomunikacionih kanala, kako u stacionarnim, tako i u mobilnim telekomunikacionim sistemima, povećane su i mogućnosti da se korisnicima pruže servisi većeg kapaciteta kao što su mirne slike i video zapisi visokih rezolucija ili video striming.

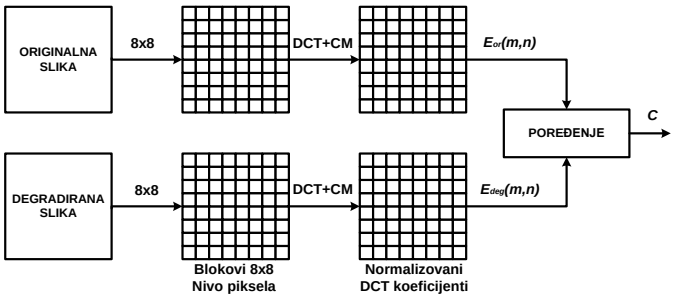
Mere za procenu kvaliteta slike sa potpunim referenciranjem (*Full-Reference*, FR), gde se prilikom procene kvaliteta slike vrši poređenje originalne (referentne, izvorne) slike i testirane (degradirane) slike, se često koriste prilikom projektovanja, testiranja i optimizacije različitih sistema. Mere bez referenciranja (*No-Reference*, NR), s obzirom da ne koriste nikakvu referencu tokom procene, već je za procenu kvaliteta slike neophodna samo degradirana slika, mogu se koristiti u sistemima koji zahtevaju rad u realnom vremenu. Tako se sistem prenosa može optimizovati odmah po detekciji određenog stepena degradacije u prenošenom sadržaju.

U radu je predstavljena mera za procenu kvaliteta slike sa potpunim referenciranjem. Prilikom procene kvaliteta, korišćena je diskretna kosinusna transformacija (*Discrete Cosine Transform*, DCT) i analiziran je značaj njenih koeficijenata prilikom određivanja kontrasta. Algoritam procene kvaliteta slike na osnovu analize promene kontrasta proizašao je iz metoda za sjedinjavanje multisenzorskih slika [1] i metoda za poboljšanje kvaliteta slike [2], kod kojih je kontrast računat u DCT domenu.

U drugom delu rada opisana je predložena mera procene kvaliteta slike. Značaj DCT koeficijenata analiziran je u trećem delu rada testiranjem mere na jednoj javno dostupnoj bazi slika. Analiza mere sa parametrima koji su pokazali najbolje performanse izvršena je u četvrtom delu rada na tri javno dostupne baze slika sa višestrukim degradacijama. Na kraju su dati osnovni zaključci i budući pravci istraživanja.

II. Procena kvaliteta slike metodom očuvanja kontrasta

Koncept procene kvaliteta slike kroz analizu očuvanja kontrasta prikazan je blok šemom na Sl. 1. Izvorna slika i slika sa degradacijom se najpre podele na blokove dimenzija 8x8 piksela. Nakon toga se odredi DCT za svaki blok posebno. DCT razlaže blok u nizove talasnih oblika, svaki sa određenom prostornom frekvencijom. Blok 8x8 piksela je izabran kao standardan i za druge primene DCT transformacije, kao što je kompresija slike.



Sl. 1. Blok šema mere očuvanja kontrasta.

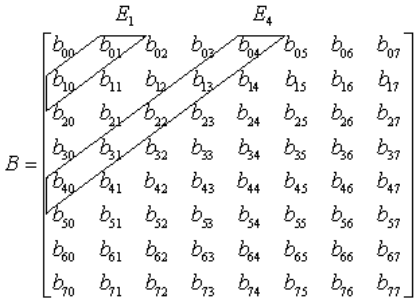
DCT koeficijenti sa uporedivom prostornom frekvencijom se grupišu i koriste za definisanje kontrasta. Radi definisanja nivoa kontrasta, potrebno je podeliti blokove u 15 različitih

Nenad Stojanović – Vojna akademija, Univerzitet odbrane u Beogradu, Generala Pavla Jurišića Šturma 33, 11000 Beograd, Srbija (e-mail: nivzvk@hotmail.com).

Boban Bondžulić – Vojna akademija, Univerzitet odbrane u Beogradu, Generala Pavla Jurišića Šturma 33, 11000 Beograd, Srbija (e-mail: bondzulici@yahoo.com).

Ivana Stojanović – Telekom Srbija a.d., Bulevar umetnosti 16a, 11000 Beograd, Srbija (e-mail: ivnvukanic@gmail.com).

frekvencijskih opsega. Opsezi su ilustrovani na Sl. 2 (prvi i četvrti opseg su naglašeni) i njih čine koeficijenti duž dijagonala matrice.



Sl. 2. Blok DCT koeficijenata.

Matrica *B* predstavlja DCT koeficijente *n*-tog bloka slike, obeležene sa $b_{i,j}$, $0 \le i,j \le 7$. Za svaki od blokova određuje se srednja vrednost amplitude *k*-tog opsega kao [1, 2]:

$$E_k = \frac{\sum_{i+j=k} |b_{i,j}|}{N}, \tag{1}$$

pri čemu *N* predstavlja broj članova unutar svakog od opsega i on se može napisati kao:

$$N = \begin{cases} k+1, k < 8 \\ 14-k+1, k \ge 8 \end{cases}. \tag{2}$$

Iz relacije (1) se vidi da *k*-ti opseg čine koeficijenti bloka DCT koeficijenata čiji je zbir jednak *k* ($i+j=k$).

Kontrast se definiše na svakoj poziciji (i,j) unutar matrice *B* kao:

$$C_{i,j} = \frac{b_{i,j}}{\sum_{k=0}^{m-1} E_k}, \tag{3}$$

gde je *m* ukupan broj opsega (za podelu na blokove 8x8 piksela $m=15$). Na ovaj način je izvršena normalizacija DCT koeficijenata i normalizovani koeficijenti $C_{i,j}$ imaju vrednosti između -1 i 1.

Ako normalizovane DCT koeficijente svih blokova izvorne slike obeležimo sa E_{or}, normalizovane koeficijente slike sa degradacijom obeležimo sa E_{deg}, sličnost dve slike na globalnom nivou (nivou slika) može se odrediti pomoću izraza:

$$C = \frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \left(\frac{2E_{or}(m,n) E_{deg}(m,n) + g}{E_{or}^2(m.n) + E_{deg}^2(m,n) + g} \right), \tag{4}$$

gde je *g* konstanta ($g=10^{-16}$), a dimenzije slika su *M*x*N* piksela.

Nakon poređenja (4) dobija se vrednost *C* koja predstavlja kvalitet slike sa degradacijom i ona može biti između -1 i 1. Slika sa većom vrednošću *C* je boljeg kvaliteta (sličnija izvornoj slici). Ako je $C=1$ slike koje se porede su identične. Konstanta *g* služi za stabilizaciju mere *C* kada su normalizovani koeficijenti jednaki nuli.

Izrazom:

$$C(m,n) = \frac{2E_{or}(m,n) E_{deg}(m,n) + g}{E_{or}^2(m.n) + E_{deg}^2(m,n) + g}, \tag{5}$$

koji predstavlja deo izraza (4), dobijaju se vrednosti očuvanja kontrasta na poziciji (m,n) – nivo piksela. Mera *C* se dobija usrednjavanjem vrednosti očuvanja kontrasta na svim pozicijama [3].

III. Analiza značaja DCT koeficijenata kod mere očuvanja kontrasta

Osobina DCT da se velika količina energije koncentriše u malom broju koeficijenata niske frekvencije koristi se u mnogim primenama, a najčešće u kompresiji podataka sa gubicima, gde se na osnovu malog broja koeficijenata može izvršiti rekonstrukcija slike zadovoljavajućeg kvaliteta [4]. Shodno tome, prilikom kvantizacije i kodovanja DCT koeficijenata vrši se skeniranje istih prema značaju, takozvano cik-cak skeniranje na način prikazan u Tabeli I. Pored najčešće korišćenog cik-cak skeniranja, postoje alternativni načini skeniranja DCT koeficijenata prema njihovom značaju, a u cilju postizanja što boljih rezultata u obradi signala. Jedan od primera alternativnog načina skeniranja je prikazan u Tabeli II [4]. Brojevi u tabelama predstavljaju redne brojeve DCT koeficijenata prilikom korišćenja (skeniranja).

TABELA I
Cik-cak način skeniranja DCT koeficijenata za blok 8x8 piksela

1	2	6	7	15	16	28	29
3	5	8	14	17	27	30	43
4	9	13	18	26	31	42	44
10	12	19	25	32	41	45	54
11	20	24	33	40	46	53	55
21	23	34	39	47	52	56	61
22	35	38	48	51	57	60	62
36	37	49	50	58	59	63	64

TABELA II
Alternativni način skeniranja DCT koeficijenata za blok 8x8 piksela

1	5	7	21	23	37	39	53
2	6	8	22	24	38	40	54
3	9	20	25	35	41	51	55
4	10	19	26	36	42	52	56
11	18	27	31	43	47	57	61
12	17	28	32	44	48	58	62
13	16	29	33	45	49	59	63
14	15	30	34	46	50	60	64

Dimenzije bloka za izračunavanje DCT koeficijenata, mogu značajno da utiču na krajnje procene kvaliteta mere očuvanja kontrasta [3]. Kako su u [3] najbolji rezultati dobijeni za blokove dimenzija 5x5 i 7x7 piksela, ove dve dimenzije bloka su izabrane i za dalju analizu, gde se analiza promene kontrasta vrši na osnovu određenog broja DCT koeficijenata prema jednom od dva navedena pravila skeniranja. Broj i značaj DCT koeficijenata koji je korišćen u proračunima, za blok dimenzije 7x7 piksela je prikazan u Tabelama III i IV, a za blok dimenzija 5x5 piksela u Tabelama V i VI.

TABELA III
Cik-cak način skeniranja DCT koeficijenata za blok 7x7 piksela

1	2	6	7	15	16	28
3	5	8	14	17	27	29
4	9	13	18	26	30	39
10	12	19	25	31	38	40
11	20	24	32	37	41	46
21	23	33	36	42	45	47
22	34	35	43	44	48	49

TABELA IV
Alternativni način skeniranja DCT koeficijenata za blok 7x7 piksela

1	5	7	19	21	33	35
2	6	8	20	22	34	36
3	9	18	23	31	37	45
4	10	17	24	32	38	46
11	16	25	28	39	42	47
12	15	26	29	40	43	48
13	14	27	30	41	44	49

TABELA V
Cik-cak način skeniranja DCT koeficijenata za blok 5x5 piksela

1	2	6	7	15
3	5	8	14	16
4	9	13	17	22
10	12	18	21	23
11	19	20	24	25

TABELA VI
Alternativni način skeniranja DCT koeficijenata za blok 5x5 piksela

1	6	8	13	18
2	7	9	14	19
3	10	15	20	23
4	11	16	21	24
5	12	17	22	25

Analiza performansi mere za tri različite dimenzije bloka i dva načina skeniranja DCT koeficijenata izvršena je testiranjem na LIVE bazi slika [5]. LIVE baza slika (*Laboratory for Image and Video Engineering*) Teksas Univerziteta (*University of Texas at Austin*) se sastoji od 982 kolor slike. Baza slika je izvedena od 29 izvornih kolor slika rezolucije uglavnom 768x512 piksela. Slike u bazi su dobijene tako što su izvorne slike degradirane korišćenjem pet tipova distorzije. Svaka od izvornih slika je degradirana sa svim tipovima distorzije i to tako da kvalitet dobijenih slika pokriva potpuni opseg kvaliteta, tj. od slika lošeg do slika dobrog kvaliteta. Kvalitet izvornih slika je narušen korišćenjem JPEG2000 kodovanja, JPEG kodovanja, dodavanjem belog šuma, Gausovim zamrljanjem – filtriranjem (*Blurring*) i simulacijom grešaka u prenosu JPEG2000 povorke bita korišćenjem Rejlijevog (*Rayleigh*) modela kanala sa brzim fedingom. Pored izvornih slika i slika sa degradacijama, dostupne su i subjektivne ocene kvaliteta koje su date u formi diferencijalnih MOS skorova (*Differential Mean Opinion Score*, DMOS), gde niže vrednosti predstavljaju bolji subjektivni kvalitet. Performanse mere su prikazane na Sl. 3 kroz stepen slaganja sa subjektivnim ocenama pomoću korelacije rangova (*Spearman Rank Order Correlation Coefficient*, SROCC).

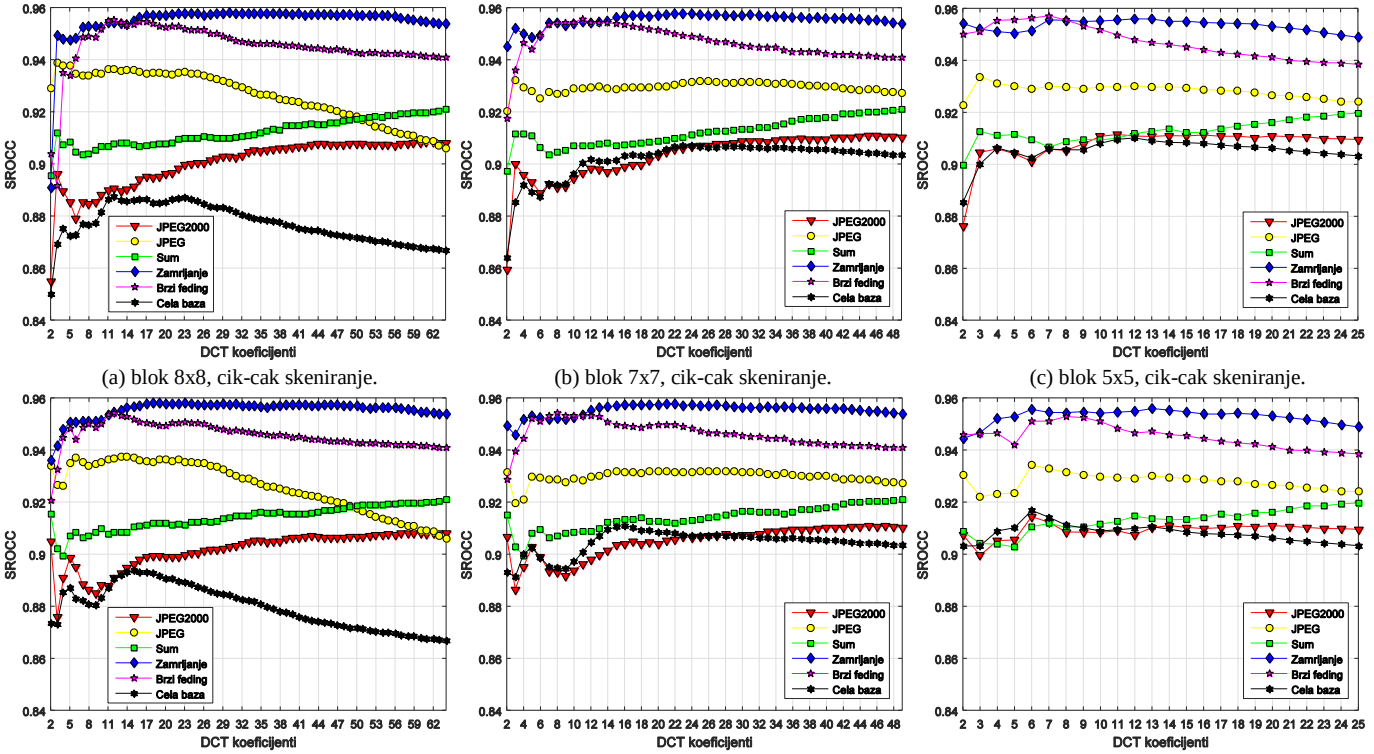
Najviši stepen slaganja sa subjektivnim ocenama mera postiže za slike sa zamrljanjem, dok se za skup slika degradiranih brzim fedingom dobijaju za nijansu lošiji rezultati. Najlošije performanse, mera pokazuje za skup slika degradiranih JPEG2000 kompresijom i na nivou kompletne baze. U zavisnosti od tipa degradacije, broj DCT koeficijenata korišćen prilikom proračuna kontrasta na lokalnom nivou različito utiče na krajnje performanse mere u korelaciji sa subjektivnim ocenama.

Kod JPEG kompresije, bolji rezultati se postižu sa malim brojem DCT koeficijenata bez obzira na dimenzije bloka i način skeniranja. Takođe, kod JPEG2000 kompresije i aditivnog belog Gausovog šuma veći broj DCT koeficijenata doprinosi boljim performansama. Kod slika sa zamrljanjem i brzim fedingom broj korišćenih DCT koeficijenata pri čemu se postiže najviši stepen slaganja sa subjektivnim ocenama je približno sličan i relativno se poklapa sa brojem DCT koeficijenata gde se postižu najbolji rezultati na kompletnoj bazi.

Korišćenjem različitog broja DCT koeficijenata prilikom određivanja vrednosti kvaliteta slike merom *C*, kod dimenzije bloka 8x8 piksela dobijeno je poboljšanje slaganja sa subjektivnim procenama od oko 2% na nivou cele baze, kod bloka 7x7 piksela oko 3-4% i za dimenziju bloka 5x5 piksela maksimalnih 5%. Na taj način je korelacija sa subjektivnim procenama sa 87% podignuta na oko 92%.

Dobijeni rezultati upoređeni su sa najčešće korišćenim merama kao što su PSNR (*Peak Signal to Noise Ratio*), SSIM_index (*Structural Similarity Index Metric*) i SSIM (vrši predobradu slike za razliku od prvobitnog SSIM_index) [6], kao i sa merama koje su postigle izuzetno visok stepen slaganja sa subjektivnim ocenama kod LIVE baze slika i to IW-SSIM (*Information Content Weighted* SSIM) [7], VIF (*Visual Information Fidelity*) [8] i MAD (*Most Apparent Distortion*) [9]. Poređenje je vršeno pomoću linearne korelacije (*Pearson Linear Correlation Coefficient*, PLCC) - Tabela VII i korelacije rangova - Tabela VIII. Pored početne verzije mere *C* (opisana u poglavlju II), za dalje poređenje je izabrana ista mera gde je prilikom proračuna korišćena

dimenzija bloka 5x5 piksela i šest DCT koeficijenata po alternativnom načinu skeniranja, što je i naznačeno u Tabeli VI. Navedeni parametri su izabrani, jer je njima postignut najveći stepen korelacije sa subjektivnim procenama na nivou cele baze, a pritom je taj broj DCT koeficijenata doprineo maksimalnim vrednostima korelacije na skoro svim skupovima slika po tipovima distorzije. Mera *C* sa navedenim parametrima u daljem radu je označena sa *C_5x5_a6*. U odnosu na mere PSNR i SSIM_index, mera *C_5x5_a6* postigla je bolje rezultate, za razliku od mere *C* sa početnim parametrima koja je bila u rangu sa merom PSNR i nešto lošija u odnosu na SSIM_index na nivou kompletne baze.



(a) blok 8x8, cik-cak skeniranje. (b) blok 7x7, cik-cak skeniranje. (c) blok 5x5, cik-cak skeniranje.

(d) blok 8x8, alternativno skeniranje. (e) blok 7x7, alternativno skeniranje. (f) blok 5x5, alternativno skeniranje.

Sl. 3. Korelacija rangova subjektivnih i objektivnih procena kvaliteta mere očuvanja kontrasta na LIVE bazi slika za različit broj DCT koeficijenata.

TABELA VII
Koeficijenti linearne korelacije na LIVE bazi slika

PLCC	JPEG2000	JPEG	Gausov šum	Zamrljanje	Brzi feding	Cela baza
PSNR	0.900	0.888	0.986	0.783	0.889	0.870
SSIM_index	0.941	0.950	0.969	0.874	0.943	0.901
SSIM	0.966	0.979	0.970	0.945	0.949	0.938
IW-SSIM	0.971	0.981	0.969	0.963	0.930	0.942
VIF	**0.977**	**0.986**	0.984	**0.974**	0.961	0.960
MAD	0.975	0.981	**0.990**	0.947	0.956	**0.967**
C	0.917	0.899	0.944	0.951	0.947	0.871
C_5x5_a6	0.920	0.931	0.936	0.933	**0.962**	0.922

TABELA VIII
Koeficijenti korelacije rangova na LIVE bazi slika

SROCC	JPEG2000	JPEG	Gausov šum	Zamrljanje	Brzi feding	Cela baza
PSNR	0.895	0.881	0.985	0.782	0.891	0.876
SSIM_index	0.935	0.945	0.963	0.894	0.941	0.910
SSIM	0.961	0.976	0.969	0.952	0.956	0.948
IW-SSIM	0.965	0.981	0.967	0.972	0.944	0.957
VIF	**0.969**	**0.985**	**0.986**	**0.973**	**0.965**	0.964
MAD	0.968	0.976	0.984	0.946	0.957	**0.967**
C	0.908	0.906	0.921	0.954	0.941	0.867
C_5x5_a6	0.914	0.934	0.910	0.956	0.951	0.917

U odnosu na ostale testirane mere, mera *C_5x5_a6* postiže nešto slabije rezultate. Na nivou cele baze taj zaostatak je od 3-4%, s obzirom da mere MAD i VIF dostižu korelaciju sa subjektivnim ocenama blizu 97%. Slična, pa i nešto veća, je razlika između ovih mera kada se posmatraju pojedinačni tipovi degradacije, osim u slučaju brzog fedinga gde je za linearnu korelaciju mera *C_5x5_a6* postigla najbolji rezultat.

IV. TESTIRANJE MERE OČUVANJA KONTRASTA NA BAZAMA SLIKA SA VIŠESTRUKIM DEGRADACIJAMA

Dalje testiranje performansi mere označene sa *C_5x5_a6* izvršeno je na bazama slika sa višestrukim degradacijama. U tu svrhu korišćene su LIVE MD (LIVE *Multiply Distorted*) [10], IVL (*Imaging and Vision Laboratory, University of Milan-Bicocca*) [11, 12] i MDID (*Multiply Distorted Image*

Database) [13] baze slika.

LIVE MD baza slika je kreirana od 15 izvornih slika rezolucije 1280x720 piksela. Baza je organizovana kroz dva scenarija. Prvi scenario je predstavljen sa 225 slika (90 sa jednom i 135 slika sa dve degradacije) degradiranih sa zamrljanjem i JPEG kompresijom, a drugi scenario je takođe predstavljen sa 225 slika (90 sa jednom i 135 slika sa dve degradacije) degradiranih zamrljanjem i Gausovim šumom. Korišćeno je tri nivoa degradacija.

IVL baza slika se sastoji od ukupno 1130 degradiranih slika od čega je 380 slika sa jednostrukom degradacijom (200 slika sa Gausovim šumom i 180 slika sa JPEG kompresijom) i 750 slika sa dvostrukom degradacijom (350 slika sa zamrljaniem + JPEG kompresijom i 400 slika sa Gausovim šumom + JPEG kompresijom). Degradirane slike su dobijene od 20 izvornih slika, rezolucije 886x591 piksel, pri čemu je 10 od tih slika korišćeno za generisanje dela slika sa dve distorzije. Korišćeno je do 10 različitih nivoa degradacija prilikom kreiranja test slika.

Baza slika MDID se sastoji od 1600 degradiranih slika kreiranih od 20 izvornih slika. Sa jednom degradacijom ima 275 slika, sa dve 415 slika, sa tri 444 slike i 466 slika sa četiri degradacije. Slike su dimenzija 512x384 piksela. Za generisanje test slika u bazi korišćeno je pet različitih tipova distorzija i to Gausov šum, zamrljanje, promena kontrasta, JPEG kompresija i JPEG2000 kompresija. Na slikama se javljaju od jedne do četiri različite distorzije tako da se dva tipa kompresije nikad ne pojavljuju zajedno. Korišćeno je pet različitih nivoa distorzija.

Tabela IX
Koeficijenti linearne korelacije na bazama slika sa višestrukim degradacijama

Redni broj	PLCC	LIVE MD			IVL			MDID		
		Jednostruke degradacije	Višestruke degradacije	Kompletna baza	Jednostruke degradacije	Višestruke degradacije	Kompletna baza	Jednostruke degradacije	Višestruke degradacije	Kompletna baza
1	PSNR	0.7937	0.4691	0.7398	0.8738	0.6803	0.7515	0.4438	0.5292	0.6120
2	SSIM_index	0.7499	0.4280	0.7333	0.7738	0.6778	0.7250	0.6788	0.6223	0.7066
3	SSIM	0.9076	0.7416	0.8914	0.8465	0.8610	0.8390	0.8490	0.8004	0.8454
4	IW-SSIM	**0.9224**	**0.7899**	**0.9105**	0.8936	**0.9041**	0.8642	0.8948	0.8723	0.8983
5	VIF	0.8955	0.7836	0.8985	**0.9561**	0.8731	0.8394	**0.9418**	**0.9188**	**0.9366**
6	MAD	0.9023	0.7559	0.8944	0.9070	0.8985	**0.8716**	0.6996	0.6884	0.7527
7	*C*	0.7181	0.7187	0.8078	0.6906	0.6908	0.6859	0.8783	0.8520	0.8807
8	*C_5x5_a6*	0.8913	0.7376	0.8843	0.7823	0.7915	0.7867	0.7477	0.8472	0.8825

Tabela X
Koeficijenti korelacije rangova na bazama slika sa višestrukim degradacijama

Redni broj	SROCC	LIVE MD			IVL			MDID		
		Jednostruke degradacije	Višestruke degradacije	Kompletna baza	Jednostruke degradacije	Višestruke degradacije	Kompletna baza	Jednostruke degradacije	Višestruke degradacije	Kompletna baza
1	PSNR	0.8039	0.4469	0.6771	0.8675	0.6136	0.7218	0.5113	0.4858	0.5784
2	SSIM_index	0.7607	0.4185	0.6459	0.7872	0.6037	0.6911	0.7478	0.6081	0.6928
3	SSIM	0.9102	0.7381	0.8604	0.8457	0.7966	0.8176	0.8592	0.7832	0.8328
4	IW-SSIM	**0.9220**	**0.7804**	**0.8836**	0.8846	0.8588	0.8493	0.8321	0.8621	0.8910
5	VIF	0.8963	0.7793	0.8823	**0.9367**	0.8381	0.8254	**0.8715**	**0.9102**	**0.9306**
6	MAD	0.9096	0.7618	0.8646	0.8998	**0.8644**	**0.8587**	0.6623	0.6529	0.7249
7	*C*	0.6708	0.7089	0.8029	0.7049	0.7126	0.6946	0.7275	0.8457	0.8767
8	*C_5x5_a6*	0.8870	0.7291	0.8488	0.7734	0.7795	0.7726	0.7578	0.8407	0.8774

U Tabeli IX su prikazani rezultati linearne korelacije na tri opisane baze koje sadrže slike sa višestrukim degradacijama. Generalno, sve mere imaju lošije performanse nego što je to slučaj kod LIVE baze slika sa jednostrukim degradacijama. Mera *C_5x5_a6* je pokazala značajno poboljšanje performansi u odnosu na početnu meru očuvanja kontrasta iz [3]. Kod LIVE MD baze slika to poboljšanje iznosi 8% na nivou cele baze, a za slike sa jednostrukom degradacijom čak 18%.

Poboljšanje stepena slaganja sa subjektivnim ocenama je postignuto i kod IVL baze za oko 10% na svim podskupovima slika. Kod baze MDID, rezultati su slični kao i kod prvobitne verzije mere *C*, osim kod slika sa jednostrukom degradacijom gde je evidentiran značajniji pad performansi, za oko 7%. U odnosu na ostale testirane mere, mera *C_5x5_a6* je pokazala bolje performanse u odnosu na PSNR i SSIM_index gotovo na svim podskupovima slika na sve tri testirane baze, dok je koeficijent korelacije sa subjektivnim ocenama kvaliteta značajno približen merama SSIM, IW-SSIM, VIF i MAD.

Rezultati dobijeni kroz korelaciju rangova su prikazani u Tabeli X i neznatno se razlikuju u odnosu na rezultate dobijene kod PLCC. Osobina koju je pokazala mera *C* [14], kada se posmatra SROCC, da su bolji rezultati dobijeni na slikama sa višestrukim degradacijama, mera *C_5x5_a6* nije uspela da pokaže. Poboljšanja koje je mera *C_5x5_a6* donela su značajno veća kada se posmatraju slike sa jednostrukim degradacijama i kompletne baze, dok su poboljšanja kod podskupova slika sa višestrukim degradacijama, iako postoje, znatno manja i to je praktično dovelo do toga da kao i sve ostale testirane mere, mera *C_5x5_a6* ima bolje performanse kod slika sa jednom u odnosu na slike sa više degradacija.

V. ZAKLJUČAK

U radu je predložena objektivna mera za procenu kvaliteta slike sa potpunim referenciranjem koja se zasniva na primeni diskretne kosinusne transformacije. Pored osnovne mere, predložene su i njene modifikacije u smislu korišćenja određenog broja DCT koeficijenata na tri različite dimenzije bloka prilikom proračuna krajnje vrednosti kvaliteta. Pokazano je na jednoj javno dostupnoj bazi slika da broj korišćenih DCT koeficijenata u krajnjem proračunu kvaliteta može značajno da poboljša performanse mere i da je dodatno približi rezultatima mera koje se često koriste i već imaju široku primenu.

Parametri mere očuvanja kontrasta koji su pokazali najbolji rezultat na bazi slika sa jednostrukom degradacijom, iskorišćeni su za dalje testiranje na tri javno dostupne baze slika sa višestrukim degradacijama. Na dve od tri baze slika postignuta su značajna poboljšanja u korelaciji sa subjektivnim ocenama, bilo posmatrajući kompletne baze, bilo posmatrajući određene podskupove slika u okviru baza na kojima je mera testirana.

U daljem radu, planira se unapređenje mere očuvanja kontrasta, kako kombinovanjem koeficijenata diskretne kosinusne transformacije na različite načine, tako i uvođenjem predfiltracije i skaliranja slika čiji se kvalitet procenjuje. Dalje testiranje objektivne mere procene kvaliteta slike vršiće se na većem broju baza slika sa što više različitih tipova degradacija.

LITERATURA

[1] J. Tang, "A contrast based image fusion technique in the DCT domain," *Digital Signal Processing 14* (2004) pp. *218-226*.

[2] J. Tang, E. Peli, S. Acton, "Image enhancement using a contrast measure in the Compressed Domain," *IEEE Signal Processing Letters*, vol. *10*, no. *10*, pp. *289-292*, October 2003.

[3] N. Stojanović, B. Bondžulić, D. Mikluc, "Procena kvaliteta slike analizom promene kontrasta," *XXI naučna i biznis konferencija YUINFO 2015, Zbornik radova*, str. 200-205, Kopaonik, Srbija, 2015.

[4] M. Popović, *Digitalna obrada slike*, Beograd, Srbija, Akademska misao, 2006.

[5] H. R. Sheikh, Z. Wang, L. Cormack, A. C. Bovik, "LIVE image quality assessment database release2", http://live.ece.utexas.edu/research/quality.

[6] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. *13*, no. *4*, pp. *600-612*, 2004.

[7] Z. Wang, Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Transactions on Image Processing*, vol. *20*, no. *5*, pp. *1185-1198*, 2011.

[8] H. R. Sheikh, A. C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. *15*, no. *2*, pp. *430-444*, 2006.

[9] E. C. Larson, D. M. Chandler, "Most apparent distortion: A dual strategy for full-reference image quality assessment," *International Society for Optics and Photonics*, vol. *7242*, pp. *72420S*, January, 2009.

[10] D. Jayaraman, A. Mittal, A. K. Moorthy, A. C. Bovik, "Objective quality assessment of multiply distorted images," *Conference Record of the 46th Asilomar Conference on Signals, Systems and Computers*, pp. *1693-1697*, November, 2012.

[11] S. Corchs, F. Gasparini, R. Schettini, "Noisy images-JPEG compressed: subjective and objective image quality evaluation," *Image Quality and System Performance XI*, vol. *9016*, pp. *90160V*, International Society for Optics and Photonics. February, 2014.

[12] S. Corchs, F. Gasparini, "A multidistortion database for image quality," *International Workshop on Computational Color Imaging*, Springer, Cham, pp. *95-104*, March, 2017.

[13] W. Sun, F. Zhou, Q. Liao, "MDID: A multiply distorted image database for image quality assessment," *Pattern Recognition 61*, pp. *153-168*, 2017.

[14] D. Mikluc, N. Stojanović, B. Bondžulić, V. Petrović, "Uticaj višestrukih distorzija na objektivnu procenu kvaliteta slike," *XXV naučna i biznis konferencija YUINFO 2019*, Kopaonik, 2019. (prihvaćen u zbornik radova).

ABSTRACT

The paper proposes a method for image quality assessment, which is based on original and test image comparison. During determination of image quality, the measure use discrete cosine transform, whereby local values of contrast is determined which with further averaging is obtaining final qualiy value. In the paper is shown that during calculation, number of used discrete cosine transform's coefficients has impact in final quality assessment. The performance of the proposed measure is given through linear correlation and rank correlation with subjective quality scores after its testing on four public available image datasets, whereby three of four datasets are with multiply degradation.

DCT Coefficients Significance Analysis in Contrast-Based Objective Image Quality Assessment

Nenad Stojanović, Boban Bondžulić, Ivana Stojanović

Poredjenje zavisnosti verovatnoće lažnog alarma i faktora skaliranja CA-CFAR i OS-CFAR detektora za različite tipove klatera

Dušan Ristić, Slobodan Simić

***Apstrakt*—U ovom radu je napravljena komparativna analiza CA-CFAR i OS-CFAR detektora. Uporedjivali smo odzive CA-CFAR i OS-CFAR detektora pod identičnim uslovima u različitim tipovima klatera. Pokazali smo koji detektor ima manju verovatnoću lažnog alarma u zavisnosti od vrednosti faktora skaliranja. Poredjenje detektora je prikazano na graficima. Za potrebe ovog rada kreirana je GUI aplikacija u programskom paketu MATLAB.**

***Ključne reči*—CA-CFAR, OS-CFAR, klater, verovatnoća lažnog alarma, faktor skaliranja.**

I. Uvod

Zahvaljujući novim generacijama računara, danas je moguće potvrditi mnoge statističke teorije koje ranije nismo mogli. U radarstvu, radarska okruženja, radarski ciljevi i ostale pojave se mogu opisati različitim statističkim raspodelama. Uz pomoć softverskih paketa koje imamo danas, moguće je opisati različita radarska okruženja. Statističke raspodele koje su razmatrane u ovom radu za opisivanje radarskih okruženja su Rejlijeva, Rajsova, lognormalna i eksponencijalna raspodela. Cilj ovog rada je da u laboratorijskim uslovima i bez eksperimentalnog merenja, pokažemo kako bi se OS-CFAR i CA-CFAR približno ponašali u realnim uslovima i da pored toga utvrdimo zavisnost verovatnoće lažnog alarma i faktora skaliranja za oba detektora. U tu svrhu je kreirana GUI aplikacija u softverskom paketu MATLAB.

II. Teorijski modeli klatera

Klater je neželjeni eho signal koji potiče od svih objekata koji nisu predmet osmatranja radara. Najčešće je to prirodno okruženje koje ometa radar i smanjuje verovatnoću ispravne detekcije cilja. Pod klaterom se podrazumevaju eho signali zemlje, mora, atmosferskih pojava i životinja. Klater je prostorno rasporedjen i veći je u fizičkom smilsu od rezolucione ćelije radara. Objekti koje su napravili ljudi (npr. zgrade) su „tačke“, ili diskretni klaterski odrazi. Veliki klaterski odrazi mogu da maskiraju eho signale ciljeva. Klater može biti površinski i zapreminski. Eho signali zemlje ili mora su primeri površinskog klatera. Eho signal kiše je primer zapreminskog klatera. Amplituda površinski raspodeljenog klatera je proporcionalna površini osvetljenoj radarskim snopom [1]. Za modelovanje radarskog okruženja, koristili smo Rejlijevu, Rajsovu, lognormalnu i eksponencijalnu raspodelu.

Rejlijeva raspodela je bazirana na pretpostavci da se u rezolucionoj ćeliji radara nalazi veliki broj slučajno raspodeljenih uniformnih reflektora koji su medjusobno nezavisni. Funkcija gustine verovatnoće naponske anvelope za Rejlijevu raspodelu klatera je

$$p(v) = 2 * \frac{v}{m_2} * e^{\left[-\frac{v^2}{m_2}\right]} \qquad (1.)$$

gde je m_2 srednja kvadratna vrednost (drugi momenat raspodele) anvelopoe v, a m_1 je srednja vrednost (prvi momenat raspodele). Srednja vrednost Rejlijeve raspodele je proporcionalna standardnoj devijaciji [1]

$$StDev = \sqrt{\frac{4}{\pi} - 1} * \mu = 0.523 * \mu \qquad (2.)$$

gde je μ srednja vrednost Rejlijeve raspodele.

Rejlijev model klatera se primenjuje kada je rezoluciona ćelija velika i sadrži mnogo reflektora, medju kojima nema dominantnih reflektora. Koristi se za opisivanje relativno uniformnog klatera. Medjutim, ova raspodela ne predstavlja dobro klater kada su rezoluciona ćelija i upadni ugao zraka mali. Pod ovim uslovima, postoji veća verovatnoća dostizanja velikih vrednosti klatera nego što je to opisano Rejlijevom raspodelom. Jedna od prvih raspodela koja treba bolje da opiše klater od Rejlijeve raspodele je lognormalna raspodela. Lognormalna raspodela ima veću verovatnoću dostizanja velikih vrednosti od Rejlijeve raspodele (raspodela sa dugim repom, „long tail distribution“). Funkcija gustine verovatnoće lognormalne raspodele je

$$p(P) = \frac{1}{\sqrt{2\pi} s P} e^{\left[-\frac{1}{2s^2}(\ln(\frac{P}{P_m}))^2\right]}, P \geq 0 \qquad (3.)$$

Dušan Ristić – Vojna akademija, Univerzitet odbrane, Veljka Lukića Kurijaka 33, 11000 Beograd, Srbija (e-mail: dusanristic152@gmail.com).

Slobodan Simić – Vojna akademija, Univerzitet odbrane, Veljka Lukića Kurijaka 33, 11000 Beograd, Srbija (e-mail: simasimic01@gmail.com).

gde je s standarda devijacija od ln(P), i Pm medijana vrednosti P. Odnos srednje vrednosti i medijane je $e^{\frac{s^2}{2}}$.

Lognormalnu raspodelu karakterišu dva parametra (standardna devijacija i srednja vrednost), dok Rejlijevu raspodelu karakteriše samo jedan parametar (srednja kvadratna vrednost). Klater lognormalne raspodele je često opisan preko odnosa njegove srednje vrednosti i medijane. Treba očekivati da, zbog dva parametra, funkcija gustine verovatnoće lognormalne raspodele može da da bolje eksperimentalne rezultate od Rejlijeve raspodele [1].

Eksponencijalna funkcija gustine verovatnoće data je formulom

$$f(x) = \begin{cases} \lambda * e^{-\lambda x}, x \geq 0 \\ 0, drugde \end{cases} \quad (4.)$$

gde je λ konstanta [2].

III. Modelovanje CFAR detektora

CFAR (Consat False Alarm Rate) detektor se sastoji od 2*n ćelija koje okružuju ćeliju koja se testira. Svaka ćelija sadrži vrednosti signala iz odgovarajuće rezolucione ćelije i ove vrednosti se pomeraju u desno kada dodju vrednosti signala iz nove rezolucione ćelije. Oko ćelije koja se testira se nalazi 2*m ćelija čuvara koje služe da se izbegnu problemi sa interferencijom prilikom procene šuma. Rastojanje izmedju ćelija odgovara rezolucionoj ćeliji radara [3]. CA-CFAR i OS-CFAR su adaptivni detektori koji, na dva različita načina, prilagodjavaju vrednost praga detekcije u zavisnosti od nivoa klatera u rezolucionoj ćeliji. CA-CFAR (Cell-Averaging Consat False Alarm Rate) to radi usrednjavanjem rezolucionih ćelija pre i posle ćelije koja se testira, dok OS-CFAR (Order Statistic Consat False Alarm Rate) odredjuje vrednost praga sortiranjem vrednosti iz CFAR prozora. Detekcija CA-CFAR detektora je optimalna u slučajevima kada imamo homogeno okruženje i kada referentne ćelije sadrže medjusobno nezavisne i slučajno rasporedjene reflektore. Verzije CA-CFAR detektora (kao što su CAGO-CFAR i CASO-CFAR) su napravljene u ciljlu poboljšanja orginalnog CA-CFAR detektora za prelaze u regionima osmatranja i za situacije u kojima imamo više ciljeva. Ove verzije računaju odvojeno srednje vrednosti ćelija levo i desno od ćelije koja se testira, pa onda od ova dva rezultata biraju manji ili veći za nivo okruženja.

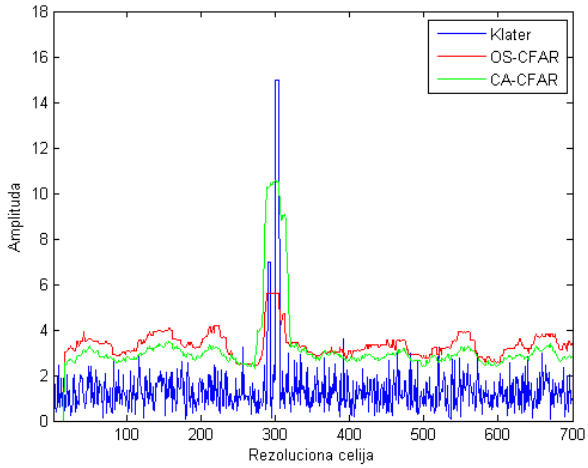
Da bi se unapredila detekcija u ovakvim situacijama, osmišljeni su OS-CFAR detektori. Detektori zasnovani na statistici, OS-CFAR detektori, imaju dobre karakteristike za nehomogena okruženja. Kod ovih detektora se sortiraju svi odbirci u rastućem redosledu, pa se potom bira referentni odbirak na osnovu kojeg se odredjuje nivo interferencije u ćeliji koja se testira. Kao rezultat, OS-CFAR pokazuje mnogo bolje rezultate u okruženju sa više ciljeva i u heterogenom klaterskom okruženju [4].

IV. Eksperimentalni rezultati

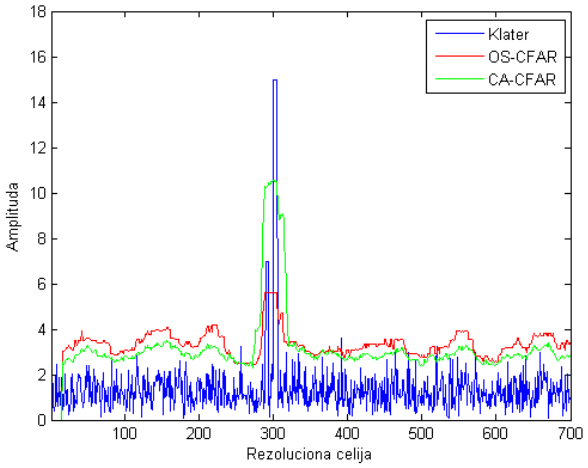
U eksperimentu su generisani odzivi CA-CFAR i OS-CFAR detektora za radarska okruženja modelovana različitim statističkim raspodelama. Odzivi su generisani pod identičnim uslovima. Na osnovu generisanih signala klatera i odziva CFAR detektora, proračunat je broj lažnih alarma za svaki detektor u svim radarskim okruženjima zasebno.

A. Poredjenje zavisnosti verovatnoće lažnog alarma i faktora skaliranja za Rejlijevu i Rajsovu raspodelu

Na graficima koje generiše GUI aplikacija, možemo videti odzive CA-CFAR i OS-CFAR detektora. Već na osnovu samih odziva detektora možemo uočiti da OS-CFAR detektor ima manje detektovanih lažnih alarma od CA-CFAR detektora, pod istim zadatim uslovima. Na slikama 1. i 2. su prikazani odzivi CFAR detektora u okruženjima modelovanim Rejlijevom i Rajsovom raspodelom.



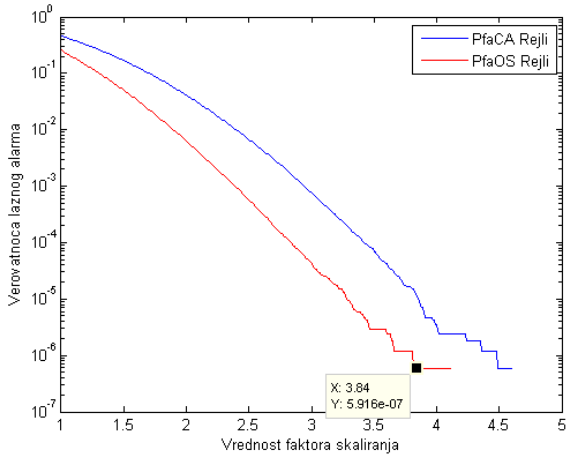
Sl. 1. Odzivi CA-CFAR i OS-CFAR detektora u okruženju modelovanom Rejlijevom raspodelom.



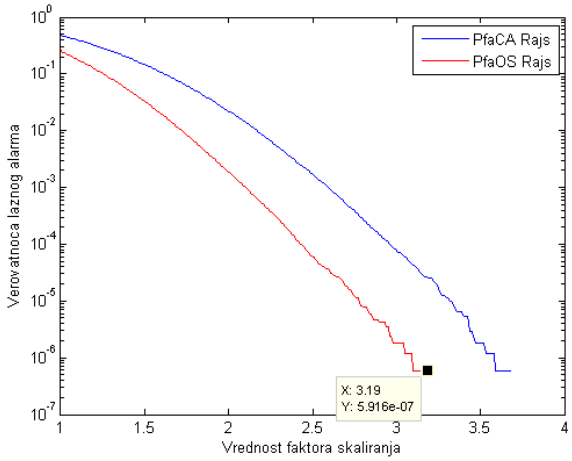
Sl. 2. Odzivi CA-CFAR i OS-CFAR detektora u okruženju modelovanom Rajsovom raspodelom.

Ovu pojavu, smo na osnovu rezultata generisanih u aplikaciji, prikazali na grafiku gde je iscrtana zavisnost

verovatnoće lažnog alarma od faktora skaliranja. Na slikama 3. i 4. su prikazane ove zavisnosti.



Sl. 3. Zavisnost verovatnoće lažnog alarma i faktora skaliranja za Rejlijevu raspodelu.

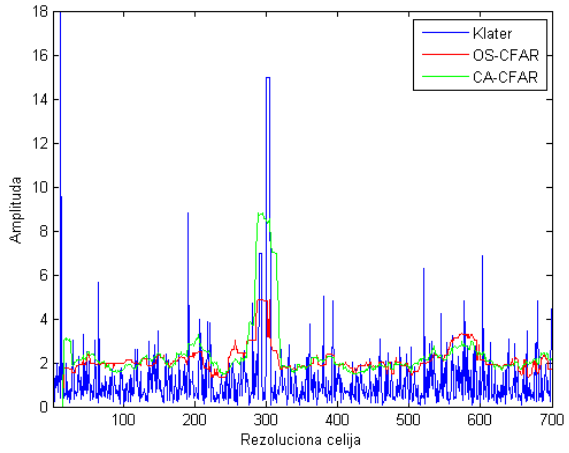


Sl. 4. Zavisnost verovatnoće lažnog alarma i faktora skaliranja za Rajsovu raspodelu.

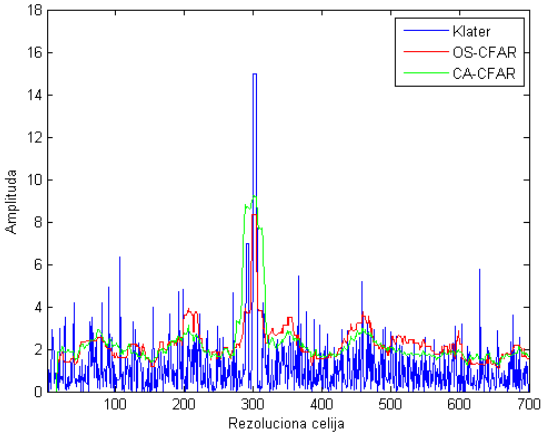
Sa grafika možemo uočiti da OS-CFAR ima mnogo manju verovatnoću lažnog alarma za istu vrednost faktora skaliranja u odnosu na CA-CFAR. Rejlijeva i Rajsova raspodela su zajedno poredjene zato što imaju zajedničku karakteristiku da imaju malo vrednosti klatera koje su približno jednake vrednostima cilja.

B. Poredjenje zavisnosti verovatnoće lažnog alarma i faktora skaliranja za lognormalnu i eksponencijalnu raspodelu

Za razliku od Rejlijeve i Rajsove raspodele, u kojima skoro da nema većih skokova signala klatera, eksponencijalna i lognormalna raspodela generišu veliki broj skokova koji po svojoj amplitudi mogu čak biti i veći od eho signala cilja. Ove raspodele se nazivaju raspodelama sa dugačkim repovima („long tail distributions“) zbog velike verovatnoće pojavljivanja vrednosti koje su znatno veće od srednje vrednosti signala. Posledica ovih raspodela je veliki broj lažnih alarma u odzivima CFAR detektora, koje možemo videti na graficima generisanim u GUI aplikaciji. Odzivi CFAR detektora za lognormalnu i eksponencijalnu raspodelu su prikazani na slikama 5. i 6.

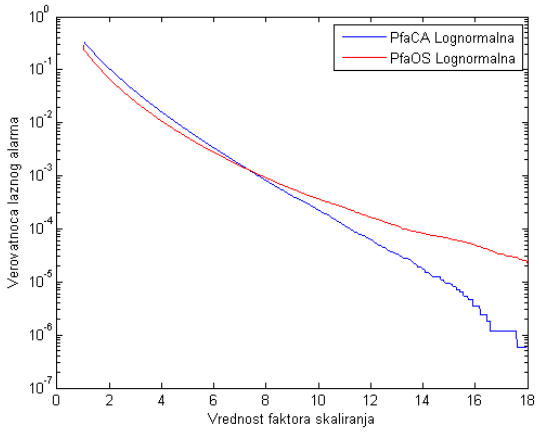


Sl. 5. Odzivi CA-CFAR i OS-CFAR detektora u okruženju modelovanom Lognormalnom raspodelom.

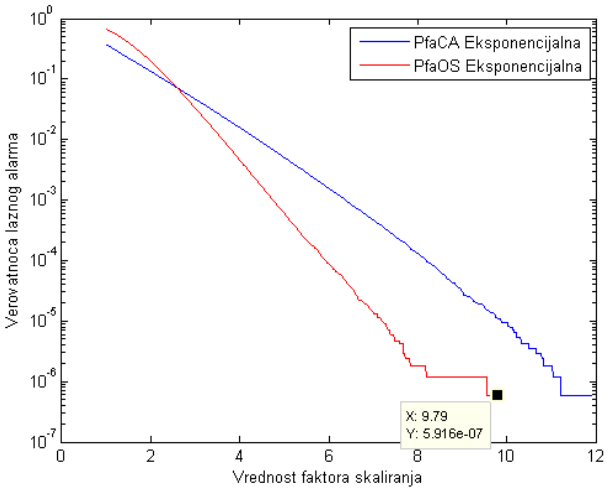


Sl. 6. Odzivi CA-CFAR i OS-CFAR detektora u okruženju modelovanom Eksponencijalnom raspodelom.

Ovu pojavu smo prikazali i na graficima zavisnosti verovatnoće lažnog alarma i faktora skaliranja. Zavisnosti su prikazane na slikama 7. i 8.



Sl. 7. Zavisnost verovatnoće lažnog alarma i faktora skaliranja za Lognormalnu raspodelu.

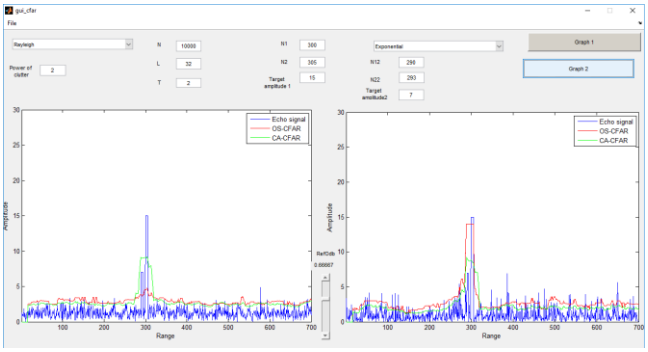


Sl. 8. Zavisnost verovatnoće lažnog alarma i faktora skaliranja za Eksponencijalnu raspodelu.

U Rohlingovom radu [5], data je tabela zavisnosti faktora skaliranja i rednog broja odbirka u odnosu na koji se radi sortiranje. Tabela je napravljena za verovatnoću lažnog alarma od 10^{-6} u okruženju modelovanom eksponencijalnom raspodelom. U skladu sa tom tabelom, možemo zaključiti da je simulacija u ovom radu dala približno isti rezultat za verovatnoću lažnog alarma 10^{-6}.

V. Realizacija eksperimenta

Za potrebe komparativne analize CFAR detektora kreirana je GUI aplikacija u softverskom paketu MATLAB. U aplikaciiji su uporedjivana oba detektora pod identičnim uslovima. Kao ulazne podatke unosili smo snagu klatera (drugi moment raspodele), broj odbiraka sa kojima će aplikacija raditi, dužinu CFAR prozora, faktor skaliranja, redni broj rezolucionih ćelija u kojima se javljaju ciljevi i amplitude eho signala. U padajućim menijima smo birali tip raspodele po kojoj će biti generisano radarsko okruženje. Kao ulazni parametar se unosi i referentni odbirak u odnosu na koji se radi procena nivoa interferncije kod OS-CFAR detektora. Kao izlazne rezultate smo dobili odzive CFAR detektora za odredjeni tip klatera i zavisnost verovatnoće lažnog alarma u odnosu na vrednost faktora skaliranja. Prikaz okruženja GUI aplikacije nalazi se na slici 9.



Sl. 8. Prikaz okruženja u GUI aplikaciji.

Za svaku statističku raspodelu kreirana je zasebna funkcija za generisanje klatera i odziva CFAR detektora. Sve funkcije su na kraju uvezene u glavnu funkciju koja pokreće aplikaciju i generiše grafike. Kodovi ovih funkcija su takodje iskorišćeni za iscrtavanje grafika zavisnosti verovatnoće lažnog alarma i faktora skaliranja.

VI. Pravci daljeg rada

Ovaj rad može poslužiti za dalje ispitivanje uticaja klatera na odzive CFAR detektora. Kao predlog za neka dalja istraživanja može biti ispitivanje uticaja referentnog odbirka i faktora skaliranja na željenu vrednost verovatnoće lažnog alarma. Ovakav rad bi bio sličan poput rada [], s tim što bi u našem slučaju, zbog savremenih softverskih paketa, bilo moguće raditi ispitivanja za sve tipove raspodela. Predlog za dalje unapredjenje rada može biti i dodavanje neke od raspodela koje nisu pomenute u ovom radu.

Literatura

[1] Blasch, Erik P., and Mike Hensel. “Fusion of distributions for radar clutter modeling”. AIR FORCE RESEARCH LAB WRIGHT-PATTERSON AFB OH, 2004.

[2] Dolecek, Gordana Jovanovic. “Random signals and processes primer with MATLAB”. Springer Science & Business Media, 2012.

[3] Simić, Slobodan, Milenko Andrić, and Bojan Zrnić. "An FPGA based implementation of a CFAR processor applied to a pulse-compression radar system." *Radioengineering* 23.1 (2014): 73.

[4] Magaz, Boualem, and Adel Belouchrani. "A new adaptive linear combined CFAR detector in presence of interfering targets." *Progress In Electromagnetics Research* 34 (2011): 367-387.

[5] Rohling, Hermann. "Radar CFAR thresholding in clutter and multiple target situations." *IEEE transactions on aerospace and electronic systems* 4 (1983): 608-621.

Abstract

In this paper comparative analysis of CA-CFAR and OS-CFAR has been made. The responses of the detectors CA-CFAR and OS-CFAR with the same parameters were compared in different clutter types. It was shown which detector has less probability of false alarm in the dependence of value of scaling factor. Comparison of detectors was shown on figures. For this paper, the GUI application in the software package MATLAB was created.

Comparison of dependence of probability of false alarm on scaling factor for CA-CFAR and OS-CFAR in differnet types of clutter

Dušan Ristić – Vojna akademija, Univerzitet odbrane, Veljka Lukića Kurijaka 33, 11000 Beograd, Srbija (e-mail: dusanristic152@gmail.com).

Slobodan Simić – Vojna akademija, Univerzitet odbrane, Veljka Lukića Kurijaka 33, 11000 Beograd, Srbija (e-mail: simasimic01@gmail.com).

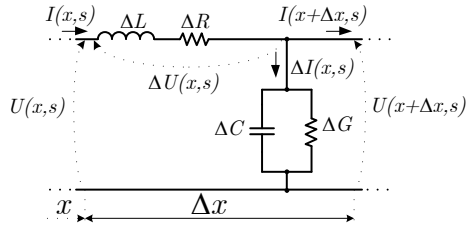
Фреквенцијске карактеристике два тополошка уопштења фракционе једначине телеграфичара

Стеван М. Цветићанин, Душан Зорица, Милан Р. Рапаић

Сажетак—Једначина телеграфичара уопштена је модификацијом Хевисајдовог елементарног кола тако да се узму у обзир мемориски ефекти феномена поларизације и магнетизације медијума, што је постигнуто фракционим калемом и фракционим кондензатором у елеменентарном колу, уместо класичних. Такође, разматрана су и два тополошка уопштења једначине телеграфичара, која су последица додавања кондензатора у редну грану елементарног кола, а чиме је узет у обзир ефекат нагомилавања наелектрисања дуж вода. Урађена је фреквенцијска анализа модула и аргумента функције преноса, добијене из уопштених фракционих једначина телеграфичара које су последица поменутог тополошког уопштења, укључивањем паралелно и редно додатног кондензатора у редну грану елементарног кола вода.

I. Увод

Класична једначина телеграфичара се изводи тако што се електрични вод издели на једнаке секције, а затим се свака секција вода моделира Хевисајдовим елементарним електричним колом приказаним на сл. 1. Овај класичан приступ моделрања вода занемарује



Слика 1. Класична топологија елементарног кола.

одређене физичке феномене, као што су: утицај струја у оточним гранама на стварање магнетског поља, као и акумулацију наелектрисања дуж вода. Такође, класичан модел претпоставља да промене електричног поља тренутно изазову промене у поларизацији диелектрика и да промене магнетског поља тренутно изазову промене у магнетизацији материјала.

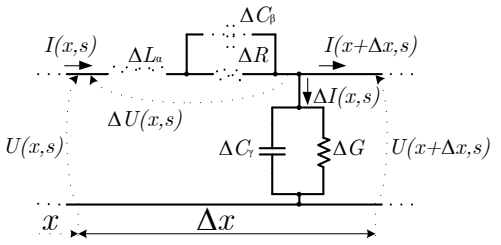
Наведена занемарења су потуно оправдана у великој већини практичних случајева, а поготово када се разматрају веома ниске учесталости, као што је случај код електроенергетских водова. Међутим, у пракси постоје случајеви када класичан модел није адекватан, нпр. THz-line (водови за пренос сигнала учесталости реда терахерца), као и CRLHT-line (composite right/left-handed transmission line) где су се фракциони модели показали много погоднији од класичних модела, види [1]. Наравно, фракциони модели нису једини начин да се опишу фреквенцијске карактеристике феномена на електричним водовима. Нпр. у раду [2] нису коришћени фракциони изводи, а постигнута је веома добра фреквенцијска зависност величина од интереса, и то на учесталостима до неколико GHz.

Моделирање операторима нецелог реда веома је практично за моделирање различитих физичких феномена, па и процеса поларизације и магнетизације материјала, када се жели узети у обзир меморијски ефекат. На пример, конеднзатори и калемови моделирани изводима нецелог реда су разматрани у [3]–[8].

У претходном раду [9] аутори су извели и анализирали уопштену фракциону једначину телеграфичара (1) засновану на паралелном елементарном колу са сл. 2, које представља уопштење класичног елементарног кола, сл. 1.

$$K^2\left(\tau_\alpha\tau_\beta\tau_\gamma\,{}_0\mathrm{D}_t^{\alpha+\beta+\gamma}+\tau_\alpha\tau_\beta\,{}_0\mathrm{D}_t^{\alpha+\beta}+\tau_\alpha\tau_\gamma\,{}_0\mathrm{D}_t^{\alpha+\gamma}+\tau_\alpha\,{}_0\mathrm{D}_t^{\alpha}+\tau_\gamma\,{}_0\mathrm{D}_t^{\gamma}+1\right)u(x,t)=\left(\tau_\beta\,{}_0\mathrm{D}_t^{\beta}+1\right)\frac{\partial^2}{\partial x^2}u(x,t)\,. \quad (1)$$

Уопштење се огледа у томе што је у редну грану

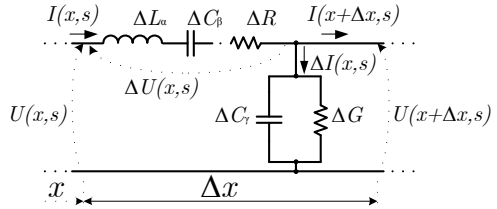


Слика 2. Паралелна топологија елементарног кола.

С. М. Цветићанин (stevan.cveticanin@uns.ac.rs), М. Р. Рапаић (rapaja@uns.ac.rs) – Универзитет у Новом Саду, Факултет техничких наука, Трг Доситеја Обрадовића 6, 21000 Нови Сад, Србија

Д. Зорица (dusan_zorica@mi.sanu.ac.rs) – Математички институт Српске академије наука и уметности, Кнеза Михаила 36, 11001 Београд, Србија – Универзитет у Новом Саду, Природно-математички факултет, Департман за физику, Трг Доситеја Обрадовића 3, 21000 Нови Сад, Србија

додат кондензатор, паралелно отпорнику, чиме је узет у обзир ефекат нагомилавања наелектрисања дуж вода. У (1) оператор ${}_0\mathrm{D}_t^{\xi}$, $\xi \in \mathbb{R}_+$, представља Риман-Лиувилов фракциони извод, за $\alpha, \beta, \gamma \in [0,1]$, где је $K > 0$ статички коефицијент слабљења, а $\tau_\alpha, \tau_\beta, \tau_\gamma > 0$ фракционе временске константе. Фреквенцијском анализом (1), приказаном у [10] и [11], показано је да (1) описује дифузне процесе за $\alpha + \gamma \in (0,1)$, а за $\alpha + \gamma \in (1,2)$, узима у обзир и таласне процесе.



Слика 3. Редна топологија елементарног кола.

$$K^2\left(\tau_\alpha\tau_\beta\tau_\gamma\,{}_0\mathrm{D}_t^{\alpha+\beta+\gamma} + \tau_\alpha\tau_\beta\,{}_0\mathrm{D}_t^{\alpha+\beta} + \tau_\beta\tau_\gamma\,{}_0\mathrm{D}_t^{\beta+\gamma} + \tau_\beta\,{}_0\mathrm{D}_t^{\beta} + \tau_\gamma\,{}_0\mathrm{D}_t^{\gamma} + 1\right)u(x,t) = \tau_\beta\,{}_0\mathrm{D}_t^{\beta}\frac{\partial^2}{\partial x^2}u(x,t)\,. \quad (2)$$

У наставку је разматрана уопштена фракциона једначина телеграфичара (2) заснована на елементарном колу са сл. 3, које је такође добијено додавањем кондензатора у редну грану класичног елементарног кола, сл. 1, али сада редно са калемом и отпорником. Слична тополошка уопштења елементарног кола коришћена су у [12] за моделирање водова за пренос сигнала веома високих учесталости.

Фракциона и тополошка уопштења модела електроенергетског вода у практичним случајевима углавном нису од интереса. Међутим, у неким другим примерима електроенергетике тополошка уопштавања елементарног кола, како би се добио општији и бољи модел, односно уважио већи скуп физичких феномена, и те како се користе, види нпр. [13].

Овај рад је проширење раније објављених радова [9]–[11] у којима је разматран фракциони модел вода (1), док је у овом раду извршена и анализа модела (2), као и поређење модела у фреквенцијском домену.

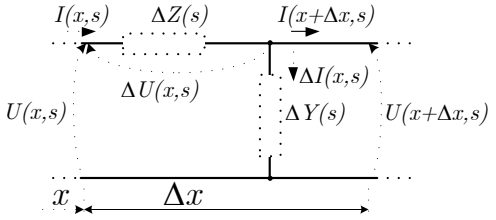
II. Формулација модела

Импедансе фракционог калема и фракционог кондензатора дефинисане су следећим изразима

$$Z_{L_\xi} = s^\xi L_\xi \quad \text{и} \quad Z_{C_\xi} = \frac{1}{s^\xi C_\xi}, \quad (3)$$

који су добијене применом Лапласове трансформације $F(s) = \mathcal{L}[f(t)](s) = \int_0^\infty f(t)\mathrm{e}^{-st}\,\mathrm{d}t$ на њихове конститутивне релације (за нулте почетне услове), користећи Лапласову трансформацију Риман-Лиувиловог фракционог извода $\mathcal{L}\left[{}_0\mathrm{D}_t^{\xi} f(t)\right](s) = s^\xi F(s)$.

Математички модели, односно уопштене фракционе једначине телеграфичара (1) и (2), који одговарају елементарним колима са сл. 2 и 3, могу се извести у комплексном домену (Лапласова трансформација) разматрајући еквивалентну импедансу редне гране ΔZ и еквивалентну адмитансу оточне гране ΔY елементарног кола, види сл. 4.



Слика 4. Елементарно коло у комплексном домену.

Применом Кирхофових закона на елементарно коло са сл. 4 добија се систем једначина

$$U(x+\Delta x,s) - U(x,s) = -\Delta Z(s) I(x,s), \quad (4)$$

$$I(x+\Delta x,s) - I(x,s) = -\Delta Y(s) U(x+\Delta x,s). \quad (5)$$

Преласком на континуум, $\Delta x \to 0$, једначине (4) и (5) прелазе у

$$\frac{\partial U(x,s)}{\partial x} = -Z(s) I(x,s) \text{ и } \frac{\partial I(x,s)}{\partial x} = -Y(s) U(x,s), \quad (6)$$

где је $Z(s) = \lim\limits_{\Delta x \to 0} \frac{\Delta Z(s)}{\Delta x}$ подужна редна импеданса, а $Y(s) = \lim\limits_{\Delta x \to 0} \frac{\Delta Y(s)}{\Delta x}$ подужна оточна адмитанса.

Једначина телеграфичара у комплексном домену

$$\frac{\partial^2 U(x,s)}{\partial x^2} - k^2(s)\, U(x,s) = 0, \quad (7)$$

где је $k(s) = \sqrt{\psi(s)} = \sqrt{Z(s)Y(s)}$ тзв. коефицијент простирања, добија се увршавањем $(6)_1$, претходно диференциране по x, у $(6)_2$.

Када је у питању класична топологија, сл. 1, редна подужна импеданса и оточна подужна адмитанса дефинишу се као

$$Z_k(s) = \lim_{\Delta x \to 0} \frac{\Delta Z_L(s) + \Delta R}{\Delta x} = sL + R \text{ и}$$

$$Y_k(s) = \lim_{\Delta x \to 0} \frac{\Delta Y_C(s) + \Delta G}{\Delta x} = sC + G,$$

где су L, R, C, и G подужни параметри вода: индуктивност, отпорност, капацитивност и проводност. Редна подужна импеданса, у случају паралелне и редне топлогије елементарног кола са сл. 2 и 3, јесу

$$Z_p(s) = \frac{RL_\alpha C_\beta s^{\alpha+\beta} + L_\alpha s^\alpha + R}{RC_\beta s^\beta + 1} \text{ и}$$

$$Z_r(s) = \frac{L_\alpha C_\beta s^{\alpha+\beta} + RC_\beta s^\beta + 1}{C_\beta s^\beta},$$

док је оточна подужна адмитанса иста за обе топологије $Y(s) = C_\gamma s^\gamma + G$. Коефицијенти простирања у случају класичне, паралелне и редне топологије су

$$k_k(s) = K\sqrt{(\tau_L\, s+1)(\tau_C\, s+1)},$$
$$k_p(s) = K\sqrt{\frac{(\tau_\alpha\tau_\beta\, s^{\alpha+\beta} + \tau_\alpha\, s^\alpha + 1)(\tau_\gamma\, s^\gamma + 1)}{\tau_\beta\, s^\beta + 1}}, \quad (8)$$
$$k_r(s) = K\sqrt{\frac{(\tau_\alpha\tau_\beta\, s^{\alpha+\beta} + \tau_\beta\, s^\beta + 1)(\tau_\gamma\, s^\gamma + 1)}{\tau_\beta\, s^\beta}}. \quad (9)$$

Уопштене једначине телеграфичара (1) и (2), које одговарају паралелној и редној топологији, добијене су применом инверзне Лапласове трансформације на (7) користећи одговарајуће коефицијенте простирања k_p и k_r, (8) и (9). Константе које фигуришу у једначинама и коефицијентима простирања дефинисане су као: $K = \sqrt{RG}\,\left[m^{-1}\right]$, $\tau_L = L/R\,[s]$, $\tau_C = C/G\,[s]$, $\tau_\alpha = L_\alpha/R\,[s^\alpha]$, $\tau_\beta = RC_\beta\,\left[s^\beta\right]$ и $\tau_\gamma = C_\gamma/G\,[s^\gamma]$.

Једначине (1) и (2) могу се извести и директно у временском домену, као што је урађено у [9] и [14].

III. Функција преноса и фреквенцијска анализа

Једначинама (1) и (2) моделираће се полубесконачан вод, $x \geq 0$, из чега произлазе одговарајући гранични услови $u\,(0,t) = u_0\,(t)$ и $\lim_{x\to\infty} u\,(x,t) = 0$, односно њихови комплексни ликови $U(0,s) = U_0(s)$ и $\lim_{x\to\infty} U(x,s) \to 0$, где $u_0\,(t)$ представља напон извора на који је вод прикључен.

A. Функција преноса

Опште решење једначине (7) је облика $U(x,s) = A(s)\mathrm{e}^{xk(s)} + B(s)\mathrm{e}^{-xk(s)}$, где се константе A и B одређују из наведених граничних услова. Према томе, израз за функцију преноса (импулсни одзив), која је дефинисана као количник комплексних ликова напона вода на произвољној позицији x и напона извора, јесте

$$W(x,s) = \frac{U(x,s)}{U_0(s)} = \mathrm{e}^{-xk(s)}, \quad (10)$$

где је $k\,(s)$ коефицијент простирања одговарајуће топологије, (8) и (9). Због квадратног корена и поткорених израза у $k\,(s)$, функција преноса, осим $s = 0$, може да има и додатне тачке гранања које одређују тип прелазних процеса на воду. У зависности од вредности параметара модела (α, β, τ_α и τ_β), могућа су три случаја: да нема додатних тачака гранања, тада је одзив апериодичан; да постоји једна реална негативна тачка гранања, тада је одзив критично апериодичан; и да постоји пар конјуговано комплексних тачака гранања са негативним реалним делом, тада је одзив пригушено осцилаторан. За детаље видети [9] и [14].

B. Фреквенцијска анализа

Изрази за фреквенцијске зависности модула и аргумента функције преноса (10)

$$|W\,(x,\omega)|_{\mathrm{dB}} = -20\,x\,\mathrm{Re}\,k\,(\omega)\log\mathrm{e} \quad \text{и} \quad (11)$$
$$\arg W\,(x,\omega) = -x\,\mathrm{Im}\,k\,(\omega)\,, \quad (12)$$

добијени су након смене $s = \mathrm{i}\omega$ у израз за $k\,(s)$, затим раздвајањем његовог реалног и имагинарног дела $k\,(\omega)$, па уврштавањем у (10). Зависност модула и аргумента (11) и (12) функције преноса од позиције дуж вода је линеарна, док је фреквенцијска зависност одређена топологијом елементарног кола преко фреквенцијске зависности реалног и имагинарног дела одговарајућег коефицијената простирања k_p и k_r, односно (8) и (9).

За одређивање реалног и имагинарног дела коефицијента простирања коришћене су следеће математичке релације

$$\mathrm{Re}\,k\,(\omega) = \sqrt{\frac{|k^2(\omega)| + \mathrm{Re}\,k^2\,(\omega)}{2}} \quad \text{и} \quad (13)$$
$$\mathrm{Im}\,k\,(\omega) = \mathrm{sgn}\left(\mathrm{Im}\,k^2(\omega)\right)\sqrt{\frac{|k^2(\omega)| - \mathrm{Re}\,k^2\,(\omega)}{2}}. \quad (14)$$

Фреквенцијске зависности реалног и имагинарног дела коефицијента простирања, према (13) и (14), у случају паралелне топологије су

$$\begin{matrix}\mathrm{Re}\,k_p^2(\omega)\\ \mathrm{Im}\,k_p^2(\omega)\end{matrix} = \frac{K^2}{|\tau_\beta(\mathrm{i}\omega)^\beta + 1|^2}\Big(\tau_\alpha\tau_\beta^2\tau_\gamma\,\omega^{\alpha+2\beta+\gamma}\begin{matrix}\cos\\ \sin\end{matrix}\frac{(\alpha+\gamma)\pi}{2}$$
$$+ \tau_\alpha\tau_\beta^2\,\omega^{\alpha+2\beta}\begin{matrix}\cos\\ \sin\end{matrix}\frac{\alpha\pi}{2}$$
$$+ 2\tau_\alpha\tau_\beta\tau_\gamma\,\omega^{\alpha+\beta+\gamma}\cos\frac{\beta\pi}{2}\begin{matrix}\cos\\ \sin\end{matrix}\frac{(\alpha+\gamma)\pi}{2}$$
$$+ 2\tau_\alpha\tau_\beta\,\omega^{\alpha+\beta}\begin{matrix}\cos\\ \sin\end{matrix}\frac{\alpha\pi}{2}\cos\frac{\beta\pi}{2}$$
$$+ \tau_\alpha\tau_\gamma\,\omega^{\alpha+\gamma}\begin{matrix}\cos\\ \sin\end{matrix}\frac{(\alpha+\gamma)\pi}{2}$$
$$\begin{matrix}+\\ -\end{matrix}\tau_\beta\tau_\gamma\,\omega^{\beta+\gamma}\begin{matrix}\cos\\ \sin\end{matrix}\frac{(\beta-\gamma)\pi}{2} + \tau_\alpha\,\omega^\alpha\begin{matrix}\cos\\ \sin\end{matrix}\frac{\alpha\pi}{2}$$
$$\begin{matrix}+\\ -\end{matrix}\tau_\beta\,\omega^\beta\begin{matrix}\cos\\ \sin\end{matrix}\frac{\beta\pi}{2} + \tau_\gamma\,\omega^\gamma\begin{matrix}\cos\\ \sin\end{matrix}\frac{\gamma\pi}{2} + \begin{matrix}1\\ 0\end{matrix}\Big),$$

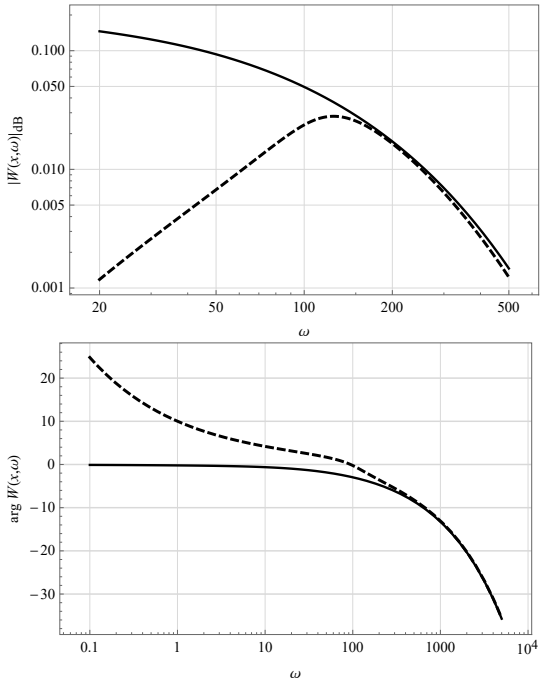
док су у случају редне топологије

$$\begin{matrix}\mathrm{Re}\,k_r^2(\omega)\\ \mathrm{Im}\,k_r^2(\omega)\end{matrix} = K^2\Big(\tau_\alpha\tau_\gamma\,\omega^{\alpha+\gamma}\begin{matrix}\cos\\ \sin\end{matrix}\frac{(\alpha+\gamma)\pi}{2}$$
$$+ \tau_\alpha\,\omega^\alpha\begin{matrix}\cos\\ \sin\end{matrix}\frac{\alpha\pi}{2} + \tau_\gamma\,\omega^\gamma\begin{matrix}\cos\\ \sin\end{matrix}\frac{\gamma\pi}{2} + \begin{matrix}1\\ 0\end{matrix}$$
$$+ \frac{\tau_\gamma}{\tau_\beta}\,\omega^{\gamma-\beta}\begin{matrix}\cos\\ \sin\end{matrix}\frac{(\gamma-\beta)\pi}{2} + \begin{matrix}+\\ -\end{matrix}\frac{1}{\tau_\beta\omega^\beta}\begin{matrix}\cos\\ \sin\end{matrix}\frac{\beta\pi}{2}\Big).$$

За детаљније извођење израза фреквенцијске анализе, два разматрана тополошка уопштења фарккционе једначине телеграфичара, погледати [14], а начелно за фреквенцијску анализу електричних кола [15].

IV. Нумерички примери

Приказане су фреквенцијске карактеристике модула и аргумента функције преноса полубесконачног вода, која произлази из редног и паралелног елементарног кола и њима одговарајућих коефицијената простирања, тј. произлази из једначина (1) и (2) и



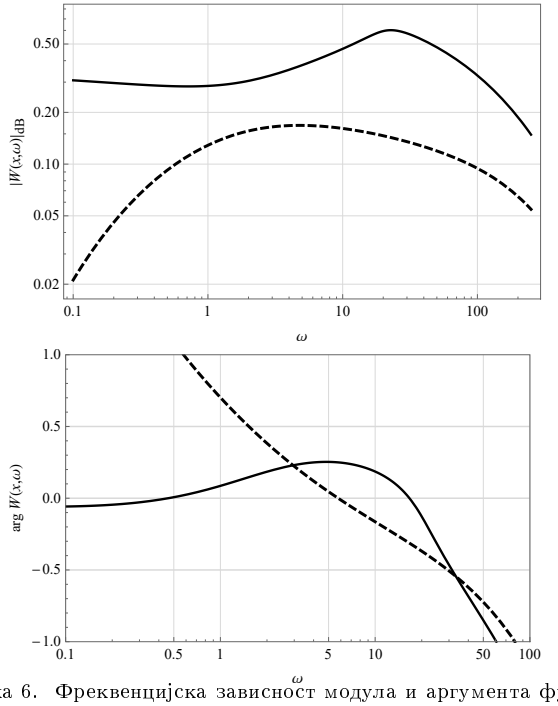
Слика 5. Фреквенцијска зависност модула и аргумента функције преноса, према паралелној (пуна линија) и редној (испрекидана линија) топологији, на позицији $x = 1$, за $K = 1$, $\alpha = 0.9$, $\beta = 0.85$, $\gamma = \frac{1}{3}$, $\tau_\alpha = 0.05$, $\tau_\beta = 0.005$ и $\tau_\gamma = 1$.

граничних услова. Урађено је поређење фреквенцијских карактеристика двеју топологија за идентичне параметре, али различите случајеве броја и позиције тачака гранања коефицијента простирања, сл. 5 - 7.

На сл. 5 приказане су карактеристике када коефицијент простирања паралелне топологије k_p нема тачака гранања, док коефицијент простирања редне топлогије k_r има пар конјуговано комплексних тачака гранања. Сл.6 резервисана је за случај када k_p има пар конјуговано комплексних тачака гранања, док их k_r нема уопште; а на сл. 7 приказане су фреквенцијске карактеристике када коефицијенти простирања обе топологије имају пар конјуговано комплексних тачака гранања.

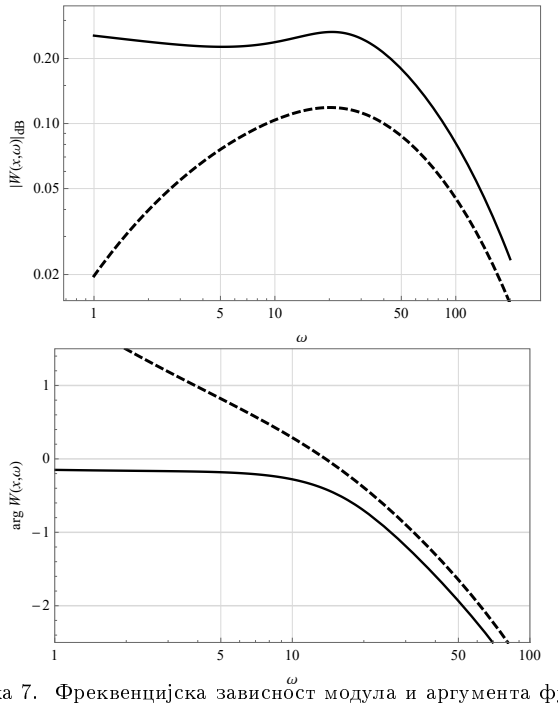
Са сл. 5 - 7 може се уочити да је код паралелне топологије фреквенцијска карактеристика модула монотоно опадајућа (нископропусник) када не постоје тачке гранања, а немонотона са израженим минимумом и максимумом када постоји пар конјуговано комплексних тачака гранања, док се код редне топологије увек јавља само један максимум (пропусник опсега) без обзира на број и природу тачака гранања.

Што се тиче фреквенцијских карактеристика аргумента функције преноса, у случају паралелне топологије она не мора бити немонотона и када постоји пар



Слика 6. Фреквенцијска зависност модула и аргумента функције преноса, према паралелној (пуна линија) и редној (испрекидана линија) топологији, на позицији $x = 1$, за $K = 1$, $\alpha = 0.9$, $\beta = 0.85$, $\gamma = \frac{1}{3}$, $\tau_\alpha = 0.01$, $\tau_\beta = 0.5$ и $\tau_\gamma = 1$.

конјуговано комплексних тачака гранања (упоредити графике са сл. 6 и 7), док је у случају редне топологије фреквенцијска карактеристика аргумента функције преноса мнотоно опадајућа, без обзира на број и позицију тачака гранања коефицијента простирања.



Слика 7. Фреквенцијска зависност модула и аргумента функције преноса, према паралелној (пуна линија) и редној (испрекидана линија) топологији, на позицији $x = 1$, за $K = 1$, $\alpha = 0.9$, $\beta = 0.85$, $\gamma = \frac{1}{3}$, $\tau_\alpha = 0.05$, $\tau_\beta = 0.1$ и $\tau_\gamma = 1$.

V. Закључак

При извођењу уопштених фракционих једначина телеграфичара, као модела електричног вода, коришћен је исти приступ као и при извођењу класичне једначине телеграфичара. Наиме, вод је издељен на сесегменте, а затим је сваки сегмент моделиран паралелним или редним елементарним колом, приказаним на сл. 2 и 3, респективно. Применом Кирхофових закона и преласком на континуум добијене су једначине (1) и (2), као уопштени фракциони модели електричног вода.

Класичне конститутивне релације, односно импедансе, калема и кондензатора замењене су фракционим, чиме је узет у обзир меморијски ефекат при процесу поларизације и магнетизације медијума. Додатним кондензатором у редној грани уважен је и ефекат нагомилавања наелектрисања дуж вода, а како је и тај додатни кондензатор фракциони узета је у обзир и историја ефекта нагомилавања наелектрисања.

Дата је функција преноса за полубесконачан вод (10), која се за паралелну и редну топологију разликује избором одговарајућег коефицијента простирања, (8) и (9), респективно. Урађена је фреквенцијска анализа и дат упоредни приказ модула и аргумента функције преноса за паралелну и редну топологију, као и класификација према броју и природи тачака гранања коефицијената простирања (8) и (9).

Поменути резултат може имати практичну примену при избору одговарајућег модела електричног вода (топологије елементарног кола и/или једначине) за који су познате фреквенцијске карактеристике модула и аргумента.

Захвалница

Аутори се захваљују на подршци пројектима ИИИ42004 (СМЦ), 174005 (ДЗ) и ТР32018, ТР33013 (МРР) Министарства просвете, науке и технолошког развоја Републике Србије, као и пројекту 114-451-2098 Аутономне покрајине Војводине.

Литература

[1] Y. Shang, W. Fei, and H. Yu, "A fractional-order RLGC model for terahertz transmission line," in IEEE MTT-S International Microwave Symposium Digest (IMS), Seattle, WA, 2013, pp. 1–3.

[2] D. B. Kandić, B. D. Reljin, and I. S. Reljin, "On modelling of two-wire transmission lines with uniform passive ladders," Mathematical Problems in Engineering, vol. 2012, p. 42, 2012.

[3] R. Süsse, A. Domhardt, and M. Reinhard, "Calculation of electrical circuits with fractional characteristics of construction elements," Forsch Ingenieurwes, vol. 69, pp. 230–235, 2005.

[4] R. Martin, J. J. Quintana, A. Ramos, and I. Nuez, "Modeling electrochemical double layer capacitor, from classical to fractional impedance," in Electrotechnical conference, MELECON 2008, The 14th IEEE Mediterranean, Ajaccio, Corsica, France, 2008, pp. 61–66.

[5] J. A. T. Machado and A. M. S. F. Galhano, "Fractional order inductive phenomena based on the skin effect," Nonlinear Dynamics, vol. 68, pp. 107–115, 2012.

[6] J. J. Quintana, A. Ramos, and I. Nuez, "Modeling of an EDLC with fractional transfer functions using Mittag-Leffler equations," Mathematical Problems in Engineering, vol. 2013, pp. 807 034–1–7, 2013.

[7] A. G. Radwan and K. N. Salama, "Fractional-order RC and RL circuits," Circuits, Systems and Signal Processing, vol. 31, pp. 1901–1915, 2012.

[8] I. Schäfer and K. Krüger, "Modelling of coils using fractional derivatives," Journal of Magnetism and Magnetic Materials, vol. 307, pp. 91–98, 2006.

[9] S. M. Cvetićanin, D. Zorica, and M. R. Rapaić, "Generalized time-fractional telegrapherâ€™s equation in transmission line modeling," Nonlinear Dynamics, vol. 88, pp. 1453–1472, 2017.

[10] S. M. Cvetićanin, M. R. Rapaić, and D. Zorica, "Frequency analysis of generalized time-fractional telegrapherâ€™s equation," European Conference on Circuit Theory and Design, Catania, Italy, Septembar 4-6, 2017.

[11] S. M. Cvetićanin, D. Zorica, and M. R. Rapaić, "Frekvencijska analiza frakcionog modela elektriičnog voda," ETRAN, Kladovo, Srbija, Jun 5-8, 2017.

[12] C. Yang-Yang and S.-H. Yu, "A compact fractional-order model for terahertz composite right/left handed transmission line," in General Assembly and Scientific Symposium (URSI GASS), 2014 XXXIth URSI, Beijing, 2014, pp. 1–4.

[13] J. Č. Mikulović and T. B. Šekara, "The numerical method of inverse laplace transform for calculation of overvoltages in power transformers and test results," SERBIAN JOURNAL OF ELECTRICAL ENGINEERING, vol. 11, pp. 243–256, 2014.

[14] S. M. Cvetićanin, "Frakciono i topološko uopštenje jednačine telegrafičara kao model električnog voda," Ph.D. dissertation, Fakultet tehničkih nauka, Univerzitet u Novom Sadu, 2017.

[15] B. D. Reljin, Teorija električnih kola II. Beograd: Akademska misao, 2009.

Abstract

The classical telegrapher's equation is generalized by modification of the Heaviside's elementary circuit such that the memory effects of polarization and magnetization phenomena of the medium are taken into account, which is achieved by using the fractional inductor and the fractional capacitor in the elementary circuit, instead of the classical ones. Also, two topological generalizations of the telegrapher's equation were considered, which are the consequence of the addition fractional capacitor in the serial branch of the elementary circuit, which takes into account the effect and history of charge accumulation along the line. The frequency analysis of the transfer function modulus and the transfer function argument, which is derived from the general telegrapher's equations, i.e. consequences of the above-mentioned topological generalizations of elementary circuit, has been performed.

Frequency Characteristics of two Topological Generalization of the Fractional Telegrapher's Equation

S. M. Cvetićanin, D. Zorica, M. R. Rapaić

Design Space Exploration in Advanced CMOS Process: IIR filter case study

Dejan Mirkovic, *Member, IEEE* and Milena Stanojlovic Mirkovic

***Abstract*—This paper deals with ever increasing complexity of digital integrated circuits design in advanced CMOS process. Concretely, hardware synthesis of IIR filter in 45nm CMOS process is considered. Filter synthesis procedure, starting form specs all the way to the generation of HDL code, is exemplified though the design of simple, third order, all-pole, selective IIR filter. Two options for standard cells under worst case conditions are examined with the help of custom written tcl scrip for early design space exploration. Obtained results gives valuable information crucial for searching of optimal synthesis solution in target CMOS process.**

***Index Terms*—Integrated circuit; Digital Synthesis; IIR Filters; CMOS process.**

I. Introduction

THERE is no doubt that CMOS technology prevails as the main implementation technology for electronics circuitry. Supersonic progress of CMOS technology in past several decades certainly ensured significant improvement in circuits performance and power consumption meaning more functionality can be implemented in the same area [1]. This inevitably led to increased complexity of the implemented systems especially when digital sub-systems are concerned. SoC (System on Chip) become main-stream concept in the world of integrated circuits (IC) design. Techniques like multiple frequency and supply/threshold voltage are common practice in today's ICs. All this is made possible with introduction of the advanced sub-hundred-micron process nodes. Besides complexity, initial (fixed) costs for the design in advanced process nodes also increases due to more masks per chip and more sophisticated photolithography techniques and testing equipment employed in the manufacturing process. This means that finding the optimal solution early in the design process is crucial. This paper tries to give insight in such process at the hardware synthesis level. Practically, this is the first stage in the digital design flow where, if appropriate models are available, designer can try to search for optimal solution in target CMOS process.

Fortunately, leading CAD/EDA tools vendors, like Cadence, Mentor Graphics and Synopsis, recognized this need and implemented routines for design space exploration (DSE) in their products. This means that digital designer needs to add another skill into his/her toolbox. Unfortunately, even those new routines are available in CAD/EDA tools, they are usually buried in piled documentation. Therefore, authors of this paper tried to do their best to highlighting and explain some of these options available in CAD/EDA tools.

Since tcl is widely adopted as the EDA language in CAD/EDA tools, DSE options are invoked by appropriate commands in conjunction with standard tcl branching statements [2]. For this purpose, custom tcl script is written. Before coping with DSE, synthesizable HDL description of the designed system has to be available. Simple third order, selective, high-pass filter will serve as a case study.

Following section of the paper is dedicated to the brief overview of the IIR filter design. In the third section process of digital synthesis and DSE are explained. Summarized simulation results are presented in the fourth section. Finally, paper concludes with suggestions and thoughts regarding presented DSE mythology.

II. IIR Filter Design

Generally, there are two approaches when synthesizing IIR filters. First approach is the traditional one where filter function is synthesized in analog domain (s domain) and digital filter is obtained by proper transform into discrete domain (z domain) [3]. Second approach is based on direct synthesis in z domain [4]. Here, first approach is adopted.

For analog prototype Least-Square Monotonic (LSM) low-pass filter function is utilized. LSM filter function belongs to the class of the filter functions with Critical Monotonic Amplitude Characteristics (CMAC) developed by Rakovic and Litovski [5]. This class of the filters offers good tradeoff between selectivity, step response and group delay characteristics. Since detailed discussion about CMAC class of filters is beyond the scope of this paper interested reader is advised to dive into [6].

For a sake of example simple high-pass, third order selective filter with f_c=5MHz cutoff frequency is designed. Sampling frequency of f_s=61.44MHz is used targeting Multi-Standard Radio (MSR) application.

Design of the filter is done with the help of the toolbox developed by the author [7]. Toolbox covers filter design flow from specs to the VHDL code generation at Register Transfer Level (RTL) and simulation.

For concrete example, poles of the analog LSM low-pass prototype filter are $s_{p1,2}= -0.4076505822 \pm j0.8728824407$ and

Dejan Mirkovic is with the University of Nis, Faculty of Electronic Engineering, 14 Aleksandra Medvedeva, 18106 Niš, Serbia (e-mail: dejan.d.mirkovic@elfak.ni.ac.rs).

Milena Stanojlovic Mirkovic is with the Innovation Centre of Advanced Technologies Ltd. Niš-Crveni krst, Bulevar Nikole Tesle 61, loc. 5, 18000 Nis, Serbia (e-mail: milena.stanojlovic@icnt.rs).

s_{p3}= −0.7958988354. Poles of the real, translated, high-pass analog filter are easily obtained by scaling with cutoff, angular frequency, ω_c=2π(5MHz). Since high-pass filter is designed, three zeros at zero are present in the *s* domain transfer function of the resulting filter.

Applying the bilinear transform to the *s* domain transfer function, digital counterpart is obtained [4]. To minimize degradation of the filter characteristics due to final world length, transfer function needs to be decomposed as the sum or product of first and/or second order transfer functions. These sub-systems described with first and/or second transfer functions are often called sections (or cells). Decomposing filter function as a product of the sections results with the cascade (serial) realization. For cascode (parallel) realization filter function is decomposed as a sum of sections. Both of them has its own strengths and weaknesses [7]. Here, cascade realization is adopted since it is the most popular one. Final transfer function of the digital filter is given in (1) and corresponding architecture is shown in Fig. 1.

$$H(z)=\frac{c_{0,I}+c_{1,I}z^{-1}}{1+d_{1,I}z^{-1}}\cdot\frac{c_{0,II}+c_{1,II}z^{-1}+c_{2,II}z^{-2}}{1+d_{1,II}z^{-1}+d_{2,II}z^{-2}} \qquad (1)$$

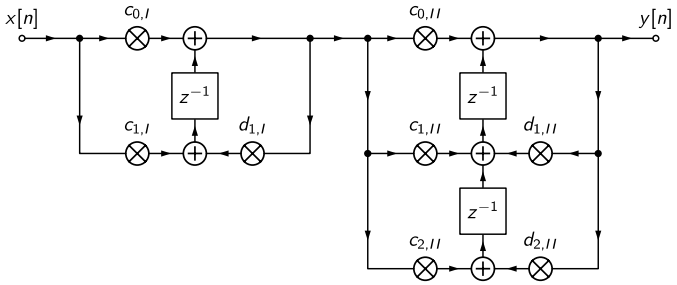


Fig. 1 Architecture of the designed IIR filter.

One can note that Transpose Direct Form II (TDF-II) form is used. Since this form is canonical it is always desirable when hardware realization is intended [3, 4]. Being the third order, filter transfer function is decomposed into cascade of the first and second order sections. Sixteen-bit word length proved to be safe enough in order to preserve good filter characteristics in discrete domain. Filter coefficients are quantized in format Q[16 14], where *W*=16 is the word length and *F*=14 is the number of bits reserved for fractional part. Hexa-decimal representation of the filter coefficients is given in Table I.

TABLE I.A
QUANTIZED NUMERATOR COEFFICIENTS OF THE THIRD-ORDER LSM HIGH-PASS IIR FILTER SECTIONS

Section	Numerator coefficients			Denominator coefficients	
	c_0	c_1	c_2	d_1	d_2
I	33C9	CC37	0000	2A5C	0000
II	33C9	986E	33C9	5EBD	D506

III. DIGITAL SYNTHESIS

After HDL description of the design filter is available, next stage in the design flow is hardware synthesis. Generally, hardware synthesis represents sophisticated optimization process. Therefore, industry standard CAD/EDA tools supported with appropriate technology vendor database must be exploited for this task. These tools provide reliable way of mapping behavioral HDL description of the digital system into gate-level netlist for target CMOS process node. Since this process is similar to the compilation of the software source code, these tools are usually called RTL compilers. There are three main-stream tools for digital synthesis namely Design Compiler (DC), Genus Synthesis Solution (GSS), and Oasys-RTL by Synopsis, Cadence, and Mentor Graphics, respectively. All of them support powerful environment with GUI and tcl script-based design flows. In this work Oasys-RTL tool is used.

A. The Flow

Although from different CAD/EDA tools vendors, RTL compilers share some common stages in the digital synthesis which are illustrated in Fig. 2.

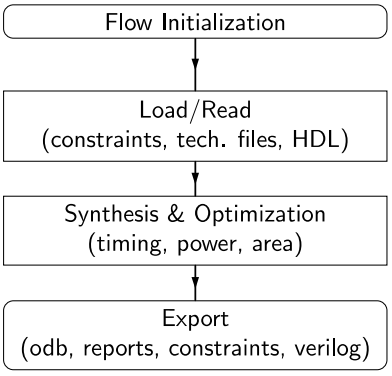


Fig. 2 Simplified flow of digital hardware synthesis process.

In flow initialization stage user's and tool specific variables are defined (e.g. paths to constraints/technology/HDL files). Technology specific information is captured in Liberty Timing (LIB) and Library Exchange Format (LEF) files. These files are prepared and delivered by the technology vendor and digital designer takes them as is. Technology vendor often provides several versions of these files covering important Process Voltage Temperature (PVT) corners of the process as well as various options for standard cells. Traditionally, at least timing constraints in a form of Synopsys Delay Constraints (SDC) file must be assigned [7].

Keeping track of the rate in the development of CMOS technology, RTL compilers now offer early estimate of the power and/or area consumption. This is done by defining desired power consumption intent (power domains) and rough floorplan through files in Common/Unified Power Format (C/UPF) and Design Exchange Format (DEF), respectively. Practically, flow initialization phase is the phase where the

most designer-flow interaction happens. The rest of the flow is almost grossly automated.

Second stage is where the tool specific tcl commands for loading/reading/compiling and/or elaborating previously assigned files.

Digital synthesis & optimization takes place in the third phase of the flow. Generally, synthesis is two stage process. First tool translate (compiles) the behavioral HDL description into structural, gate-level, netlist. Next, gate-level netlist is mapped (linked) to the concrete standard cells available in the target CMOS process. Further, tool automatically invokes a set of analyses with the goal of finding optimal synthesis solution for given timing/power/area constraints. At the end of the synthesis process tool generates new constraints (new SDC, UPF, DEF files) which contain more realistic prediction of circuit performance.

Finally, in the fourth stage synthesized design is exported in one of the standard formats (Verilog, DEF, etc.).

Technology Verilog netlist and constraints obtained in digital synthesis stage further drives the process of digital implementation. In digital implementation stage final, physical, abstraction of the design is generated (GDSII files).

B. Design Space Exploration

Bearing in mind that chip fabrication is very expensive, it is crucial to explore various options for standard cells for target design. In this example Nangate OpneCell 45nm technology with nominal power supply voltage of 1.25V is used [8]. LIB files for three standard cells options are available:

- High Threshold Voltage (HVT), slow but low-power;
- Low Threshold Voltage (LVT), fast but high-power;
- Standard Threshold Voltage (SVT), typical.

All three options are characterized under the wort case conditions i.e. at −40°C with two 0.85V and 0.95V power supply voltages.

Today's CAD/EDA tools support DSE early in the design flow i.e. at the synthesis level. The way of how DSE analysis can be performed is exemplified through the custom written tcl script for Oasys-RTL compiler.

Code of the DSE tcl script is shown in Fig. 3. Design variables are defined from line one to sixteen. It is important to note that default time scale is in ns. Accordingly, each numerical variable related to time units has to be converted into ns (lines two and three).

For the sweep of clock period three options are created namely, *Tclk_m20per*, *Tclk*, *Tclk_p20per* denoting 0%, +20% and −20% variation (lines eight to ten).

Two worst case process options, HVT and LVT, are to be examined. These options are defined as *hvt* an *lvt* (lines eleven to thirteen).

For the voltage variations two cases are considered. These options are defined as *VDD_0p95v* and *VDD_0p85v* representing 0.95V and 0.85V power supply voltage (lines fourteen to sixteen).

```
1 # Define clock frequency
2 set fclk 61.44e6
3 set tclk [ expr floor( ( 1.0 / $fclk ) * 1e9 ) ]
4
5 # Set up explore variables
6 set_explore_setup -job_limit 2
7
8 set_explore -variable clock_period -option {
9   Tclk_m20per Tclk Tclk_p20per
10 }
11 set_explore -variable lib_vt -option {
12   hvt lvt
13 }
14 set_explore -variable voltage -option {
15   VDD_0p95v VDD_0p85v
16 }
17
18 # Initialize script parameters
19 source scripts/init_design.tcl
20
21 # Read logical libraries
22 foreach lib $high_vt_libs {
23   read_library $lib -target_library high_vt
24 }
25 foreach lib $low_vt_libs {
26   read_library $lib -target_library low_vt
27 }
28
29 read_lef $tech_file
30
31 foreach lef $lef_files {
32   read_lef $lef
33 }
34
35 # Sweep Liberty Timing Library
36 explore lib_vt {
37   hvt { set_target_library high_vt }
38   lvt { set_target_library low_vt }
39 }
40 set_dont_use [get_lib_cell * ] false
41
42 # Sweep voltage
43 explore voltage {
44   VDD_0p95v {
45     load_upf $work_dir/constraints/upf95.tcl
46   }
47   VDD_0p85v {
48     load_upf $work_dir/constraints/upf85.tcl
49   }
50 }
51
52 # Read the VHDL file
53 read_vhdl $hdl_files
54
55 # Synthesize the top module
56 synthesize -module ${top_module}
57
58 # Create clock
59 create_clock -period $tclk -name GCLK clk
60
61 # Sweep clock
62 explore clock_period {
63   Tclk_m20per { scale_clock -all -20 }
64   Tclk_p20per { scale_clock -all  20 }
65 }
66
67 # Optimize
68 optimize -virtual
69 report_clocks
70 report_power
71 report_design_metrics
```

Fig. 3 Example of the tcl script used for early design space exploration.

Altogether, tool is setup to perform twelve design synthesis/analysis & optimization runs (three for clock × two for process × two for supply voltages runs).

After initialization of the design (line nineteen), appropriate LIB and LEF files are read into work memory (lines twenty-two to thirty-three). In lines thirty-six to fifty tool is setup to perform multiple optimizations combining low and high threshold/supply voltages. One should note that for each supply voltage option appropriate UPF file has to be prepared (upf95.tcl and upf85.tcl). Design compilation and gate-level synthesis is performed in lines fifty-three and fifty-six after which clock driver is created (line fifty-nine) and tool is setup for exploring defined options for clock period (lines sixty-two to sixty-five).

Optimization process is issued in line sixty-eight. Here, command switch *virtual* is used to relax optimization process. Specifying this option ensures that the optimization process converges by ignoring potential physical level problems that can lead to unsuccessful timing closure. With *virtual* switch tool performs only coarse (virtual) placement of the cells. Ones all timing problems can be fixed with *virtual* option, full optimization can be invoked. In early DSE, where multiple analyses are to be performed, using *virtual* option offers significant time savings. Finally, obtained results of DSE flow are reported (lines sixty-nine to seventy-one). More in-depth explanation of the Oasys-RTL commands can be found in [9].

IV. Simulation Results

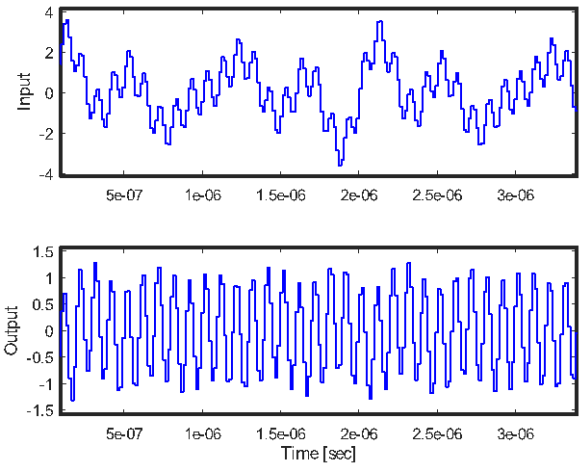
In order to verify HDL description of the designed IIR filter, logic simulation is performed. Filter is excited with composite signal given in (2).

$$x[n] = \sum_{k=0}^{3} \sin\left(2\pi \frac{f_k}{f_s} n\right), \qquad (2)$$

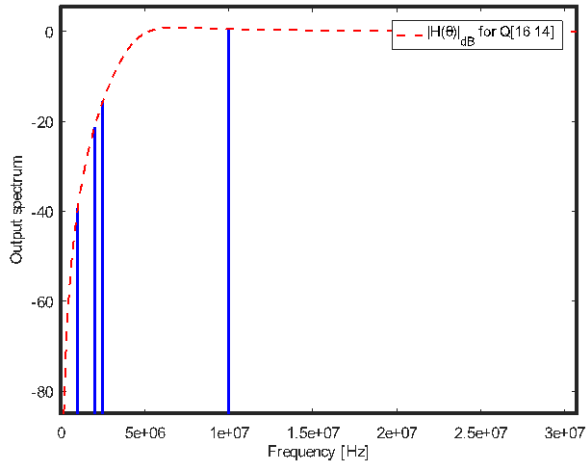
where $f_k \in \{1\text{MHz}, 2\text{MHz}, 2.5\text{MHz}, 10\text{MHz}\}$ and f_s=61.44MHz. These test frequencies are chosen in such a way to easy check filter selectivity. Time response and output spectrum of the filter are shown in Fig. 4. Output spectrum is estimated with 2^{18}-point FFT.

Based on the time response given in Fig. 4a one can clearly see that the high frequency (10MHz) component prevails. Comparing values of the magnitude of the frequency response, output spectrum shown in Fig. 4b, at 1MHz and 2MHz 18dB/oct slope is observed. Similar can be concluded when looking output spectrum at 2.5MHz and 5MHz. This response clearly corresponds to the third-order transfer function which is plotted with dashed line in the output spectrum.

Summarized DSE results are given in Table III. When looking last two columns of the table it can be seen that there is no significant variation in area or number of gates. On the other hand, Worst Negative Slack (WSN) and total power consumption experience significant changes over examined cases. Generally, optimization of WNS parameter is the prime goal in the digital design. This parameter is the most important results of Static Timing Analysis (STA) [10].



a) Time response



b) Output spectrum

Fig. 4 Time response (a) and output spectrum (b) of the RTL model of the third-order, selective, LSM IIR filter.

Practically, its value determines whether design is feasible in target CMOS process under given constraints. WNS is defined as difference between arrival time (AT) and required time (RT).

AT is calculated for each standard cell based on timing information available in LIB files and it represents time needed for the signal to arrive at certain point. RT is the least time delay allowed for the signal which guarantees that clock period will not need to be increased (or equivalently clock frequency lowed).

Therefore, for positive value of the WNS there is enough margin in timing constraints. This means that, not only that the designed circuitry meets timing requirements, but that it can even run at higher clock rate then it was intended. On the other hand, negative value of the WNS indicates that timing requirements are not met and re-design or/and relaxation of the timing constraints is needed.

When the time closure of the design is achieved, one can try to optimize power consumption.

TABLE III
SUMMARIZED DSE RESULTS

Supply voltage [V]	V_t library	Clock period variation [%]	Clock frequency [MHz]	WSN [ns]	Total power [µW]	Area [µm^2]	Number of gates
0.95	hvt	0	62.500	7.6671000	264.531769	3030	1305
		+20	52.083	10.867100	220.446198	3030	1305
		−20	78.125	4.467100	330.659851	3030	1305
	lvt	0	62.500	15.436200	320.37915	3032	1307
		+20	52.083	18.636200	269.296906	3032	1307
		−20	78.125	12.236200	397.002777	3032	1307
0.85	hvt	0	62.500	2.688800	218.100967	3030	1305
		+20	52.0833	5.888800	175.045609	3030	1305
		−20	78.125	0.796200	282.736511	3061	1382
	lvt	0	62.500	12.806600	266.098267	3032	1307
		+20	52.083	16.006599	225.735413	3032	1307
		−20	78.125	9.606600	326.642456	3032	1307

Changes in WNS and total power are illustrated in Fig. 5. Both parameters of interest are scaled with its maximum values in order to produce comparable representation.

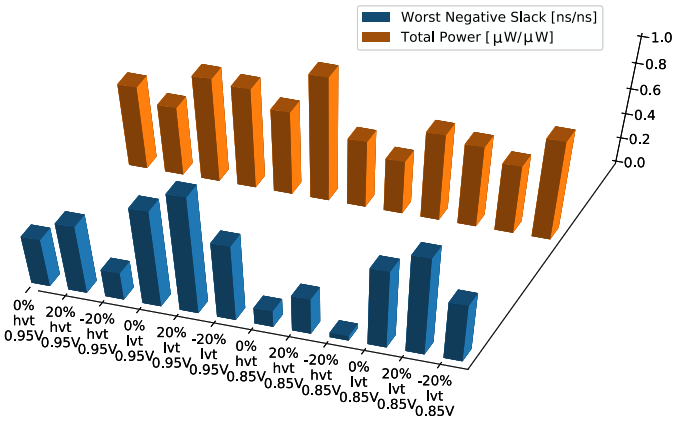


Fig. 5 Comparative graphical representation of DSE results for Worst Negative Slack (WNS) and Total power consumption.

Graphical representation given in Fig. 5 can be used as a guide for tradeoff between WNS and total power consumption. For example, if main goal is WNS maximization regardless of the power consumed case with 20% increase in clock period, 0.95V supply voltage and cells with low threshold voltage should be used. By contrast, if the power consumption is the prime goal then option with −20% reduction of clock period, 0.85V power supply voltage and cells with high threshold voltage can be used.

If the optimal design is sought it can be found with the help of Fig. 6. Since WNS needs to be maximized and total power minimized, ratio of the two will unveil the optimal synthesis solution.

According to Fig. 6, case with 20% increase in clock period, 0.85V power supply voltage and cells with low voltage threshold gives optimal synthesis solution. Of course, here equal weight is assumed for both parameters of interest.

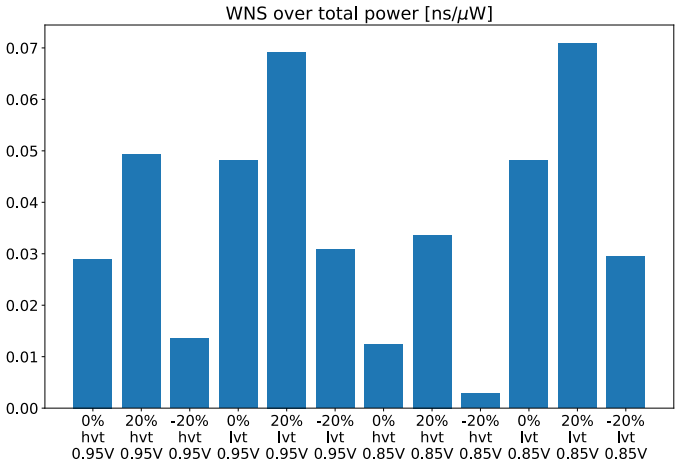


Fig. 6 WNS over total power estimation

Depending on the target application one can favor one parameter over the other. This can be easily done by assigning desired weights to the WSN and total power and recreating plots shown in Fig. 5 and 6 with the data given in Table III.

V. CONCLUSION

The main reasons for early design space exploration in the design of digital ICs targeting advanced CMOS process is explained and elaborated.

Brief overview of the IIR filter synthesis and hardware realization is given.

One example of the design space exploration in advanced 45nm CMOS process is presented. DSE flow is applied to the design of the simple, third-order, selective LSM IIR filter. For performing this task custom tcl script is written. Execution of the script is discussed in detail.

Twelve cases covering two values for power supply voltages, three for the clock period and two options for standard cells are examined.

HDL description of the RTL model of the filter is verified in time and frequency domain.

Based on DSE results important conclusions are drawn regarding optimal synthesis solution.

Obtained results proved that early DSE can give valuable information before coping with more complex implementation stages of the digital ICs design flow. This way, design time can be significantly reduced.

Important feature of this work is that it can be easily extended and modified to suite similar CAD/EDA tools for digital synthesis.

ACKNOWLEDGMENT

This work is funded by Serbian Ministry of Education Science and Technological development under contract no. TR32004.

REFERENCES

[1] H. Iwai, "Technology roadmap for 22nm and beyond," in 2009 2nd International Workshop on Electron Devices and Semiconductor Technology, June 2009, pp. 1–4.

[2] E. Todorovich and O. Cadenas, "Tcl/tk for eda tools," in 2007 3rd Southern Conference on Programmable Logic. IEEE, 2007, pp. 107–112.

[3] L. D. Thede, Practical analog and digital filter design. Artech House Norwood, Mass, USA, 2005.

[4] A. V. Oppenheim and R. W. Schafer, Discrete-time signal processing. Pearson Education, 2014.

[5] B. D. Rakovich and V. B. Litovski, "Least-squares monotonic lowpass filters with sharp cutoff," Electronics Letters, vol. 9, no. 4, pp. 75–76, February 1973.

[6] D. Topisirovic, V. Litovski, and M. Andrejevic Stosovic, "Unified theory and state-variable implementation of critical-monotonic all-pole filters," International Journal of Circuit Theory and Applications, vol. 43, no. 4, pp. 502–515, 2015.

[7] D. D. Mirkovic, "Design of selective iir digital filters with linear phase utilizing analog prototypes," Ph.D. dissertation, University of Niš, Faculty of Electronic Engineering, 2018.

[8] "NanGate FreePDK45 Generic Open Cell Library," https://projects.si2.org/openeda.si2.org/projects/nangatelib, accessed: 2019.

[9] Oasys-RTL User's Guide, Software version 2018.1 ed., Mentr Graphics Corporation, 8005 S.W. Boeckman Road, Wilsonville, Oregon 97070-77777, 2018, support.mentor.com.

[10] J. Bhasker and R. Chadha, Static timing analysis for nanometer designs: A practical approach. Springer Science & Business Media, 2009.

Classification of Nonlinear Loads using Current Spectrum

Marko Dimitrijević, *Member, IEEE*, Miona Andrejević Stošović *Member, IEEE* and Dejan Stevanović

***Abstract*—One of the most prominent characteristics of nonlinear loads is existence of the higher harmonics in current spectrum. They cause losses and disturbance in power grid; thus, the power factor must be generalized to a total or true power factor where the apparent power, involved in its calculations, includes all harmonic components. Nevertheless, harmonic components of the current spectrum can be regarded as specific "signature" of some nonlinear load, therefore providing the means for identifying classes of nonlinear loads connected to the power grid. In this paper, we will present the method for identifying and classification of nonlinear loads using harmonics' amplitudes as inputs for the artificial neural network.**

***Index Terms*— artificial neural network; harmonics; nonlinear loads.**

I. Introduction

Transition to the new millennium also marked transition in our use of electric energy. The significant growth in electronic industry had immense impact on world economy and led to significant changes in our lifestyle. The electronic devices and appliances which contain electronics are everywhere. In addition, growing awareness of global warming influenced automotive industry to change from fossil fuel engines to electric motors. As a consequence, electronic devices take a significantly bigger share in overall power consumption.

Electronic devices are nonlinear loads by nature. They are usually very complex circuits, consisting active semiconductor components that require direct current (DC) for polarisation. However, electricity is delivered to end-users using an alternating current (AC), suitable for the transmission of electricity, which cannot be directly applied to electronic circuits. The means of conversion, AC/DC (power) converters, can be analyzed as two-port networks with reactive and nonlinear impedance seen from the power grid. The characterization of the power converter, and hence the electronic device that contains it, can be performed by determining the current spectrum, i.e. calculating current harmonics.

Miona Andrejević-Stošović is with Faculty of Electronic Engineering, University of Niš, Aleksandra Medvedeva 14, 18000 Niš, Serbia (e-mail: miona.andrejevic@elfak.ni.ac.rs).

Marko Dimitrijević is with Faculty of Electronic Engineering, University of Niš, Aleksandra Medvedeva 14, 18000 Niš, Serbia (e-mail: marko.dimitrijevic@elfak.ni.ac.rs).

Dejan Stevanović is with ICNT, Bulevar Nikole Tesle 61, 18000 Niš, Serbia (e-mail: dejan.stevanovic@icnt.rs).

There are many approaches that are used in identifying and monitoring electronic devices (nonlinear loads) connected to the power grid. They are divided into two categories, depending on the means of collecting data: Intrusive Load Monitoring (ILM) and Non-Intrusive Load Monitoring (NILM). ILM methods are implemented with a number of voltage and current probes placed on each monitored device. NILM methods require only one measuring device for the group of devices, for instance entire household. Although ILM methods are more accurate for obvious reasons, NILM methods are commonly used, due to the implementation costs.

In our previous proceedings, we have presented a method for classification of nonlinear loads using artificial neural networks (ANNs), based on active, reactive and distortion power [1,2].

In this paper, we will use amplitudes of current harmonics as inputs for classification of nonlinear loads. The presented method fits into NILM category – we have used one measuring apparatus for various combinations of devices. The method consists of three phases: signal acquisition, current harmonic extraction and device identification/classification using ANNs. The acquisition and harmonic extraction are performed using system for nonlinear load analysis [3,4]. The extracted parameters, i.e. amplitudes of current harmonics are used for ANN training. Finally, the trained ANN is employed for identification of similar nonlinear loads and unknown combinations of loads connected to the power grid.

The paper is organized as follows: in the second section we will present signal acquisition system and algorithms for harmonic analysis, third sections describes ANN topology, training process and obtained results. Fourth section concludes the paper.

II. Data acquisition and harmonic analysis

The signal acquisition is performed using acquisition system consisting of acquisition modules for A/D conversion [5], and computer interface. A/D resolution is 16-bit, with 100 kSa/s sampling rate and ±50 A dynamic range. This system is described in great detail in [4]. Although this system is capable for real-time operation, due to the offline nature of ANN training, this capability is not used. However, the authors are aware that when the ANN training is finished, the trained ANN can be integrated into the system, enabling identification of nonlinear loads in real time.

In the presence of nonlinear loads, the circuit no longer operates in sinusoidal condition and use of circuit's fundamental frequency analysis could not be applied any

more. Traditional power system quantities such as effective value, power (active, reactive, apparent), and power factor need to be numerically calculated from sampled voltage and current sequences by performing DFT or FFT algorithm.

Having in mind that voltage waveform almost insignificantly deviates from sinusoidal (typically $THD_V < 3\%$), we will focus only on current waveform.

We will review some basic definitions briefly. The RMS value of current is calculated according to the well-known formula:

$$\tilde{I} = \sqrt{\frac{1}{\mathrm{T}} \int_{t_0}^{t_0+\mathrm{T}} \left(i(t)\right)^2 dt} \tag{1}$$

where $i(t)$ represents instantaneous current, T is the period and t_0 is some arbitrary time. We can develop Fourier representation:

$$\begin{aligned} i(t) &= a_0 + \sum_{k=1}^{+\infty} \left(a_k \cdot \cos\left(k\omega t\right) + b_k \cdot \sin\left(k\omega t\right)\right) \\ i(t) &= i_0 + \sum_{k=1}^{+\infty} i_k \cdot \cos\left(k\omega t + \psi_k\right) \end{aligned} \tag{2}$$

where $i_0 = a_0$ represents DC component, $i_k = \sqrt{a_k^2 + b_k^2}$ amplitude and $\psi_k = \arctan\frac{b_k}{a_k}$ phase of the k^{th} harmonic. $\omega = \frac{2\pi}{\mathrm{T}}$ represents angular frequency.

The Fourier coefficients a_k, b_k can be calculated as:

$$a_0 = \frac{1}{\mathrm{T}} \int_{-\mathrm{T}/2}^{+\mathrm{T}/2} i(t)dt, \quad a_k = \frac{2}{\mathrm{T}} \int_{-\mathrm{T}/2}^{+\mathrm{T}/2} i(t) \cdot \cos\left(\frac{2k\pi t}{\mathrm{T}}\right) dt \tag{3}$$

and

$$b_k = \frac{2}{T} \int_{-\mathrm{T}/2}^{+\mathrm{T}/2} i(t) \cdot \sin\left(\frac{2k\pi t}{\mathrm{T}}\right) dt. \tag{4}$$

The RMS value of the k^{th} harmonic, further used for ANN training, as well as for identification, is:

$$\tilde{I}_k = \sqrt{\frac{a_k^2 + b_k^2}{2}}. \tag{5}$$

As we are dealing with discrete sets of values obtained by sampling, harmonic analysis is achieved by software implementation of some FFT algorithm.

III. ANN TRAINING AND LOAD IDENTIFICATION

For the purpose of this paper, we used current harmonics up to 40$^{\text{th}}$ harmonic. Since even harmonics are not present in the spectrum of most electronic devices (as well DC component), only odd harmonics are considered. In this analysis, only amplitudes are taken into consideration.

The measurement is performed on 12 different combinations of various devices, shown in the Table I. We took into account devices that are typical for an office, for example. In fact, we have one PC, one server, one monitor, one kettle, one LED bulb, and an air-condition that can be in two operation modes: cooling and heating. Both the PC and the server operated with idle CPU utilization in all cases, so they can be treated as stationary nonlinear loads with constant power consumptions and current waveforms. The codes are given in the first column of the Table I, and also in the first row of the Table II. First column of the Table II gives the order of the harmonic.

TABLE I
CHARACTERISTIC COMBINATIONS OF THE DEVICES

Code	Device
1	Led Bulb
2	Kettle
3	Server
4	Air-condition heating
5	Air-condition cooling
6	PC + Monitor
7	Server + PC + Monitor
8	Server + PC + Monitor + Kettle
9	Server + PC + Monitor + Led Bulb
10	Server + PC + Monitor + Air-condition heating
11	Server + PC + Monitor + Air-condition cooling
12	Server + PC + Monitor + Air-condition heating + Led Bulb

Artificial neural network needs to be trained for modeling the look-up table. It is a feed-forward neural network with one hidden layer. The measured values from the Table II are inputs to the network, and the Code is network output to be learned. It means that the neural network has twenty input neurons and one output neuron. After training was completed, the number of hidden neurons in the resulting ANN was two, what was found by trial and error after several iterations starting with an estimation based on [6] and [7]. The training error was 3.16916e-017.

This number of hidden neurons in the neural network is very small. This opened a question if twenty input neurons were really needed, or we can train a neural network with smaller set of data. So, we tried to reduce training set to ten inputs. After training was completed, the number of hidden neurons in the resulting ANN was also two, and the error was very small, 2.68838e-019. Next step was to reduce training set to five inputs. We first tried to train an ANN with two hidden neurons, but it was not possible because training error was unacceptable. Then we enlarged the number of hidden neurons to three, and training results were good, the training error was 1.9106e-015. We concluded that we could use five input neurons, but in that case, there were three hidden neurons needed.

So, when we compare the obtained networks, the first one has 20 input and 2 hidden neurons, the second has 10 input and 2 hidden neurons, and the last has 5 input and 3 hidden neurons. The simplest realization would be to realize the third network. It also requires the smallest set of measured data.

TABLE II

VALUES OF HARMONICS' AMPLITUDES ($\tilde{I}_k\,[\mathrm{mA}]$) FOR THE COMBINATIONS GIVEN IN TABLE II

code → harmonic (k)↓	1	2	3	4	5	6	7	8	9	10	11	12
1	43.24	7450	280.47	3460	3420	235.57	742.57	7990	609.35	3230	2980	3710
3	32.18	22.44	38.09	123.5	215.32	104.12	77.65	27.86	110.79	161.32	211.11	178.84
5	21.22	175.98	33.66	569.31	582.51	51.27	93.59	182.57	44.16	585.62	516.69	553.6
7	20.75	143.29	17.19	266.65	234.98	29.13	34.95	177.49	72	245.08	251.75	231.13
9	20.49	11.72	7.53	39.89	43.92	2.76	8.18	9.86	35.35	35.06	40.04	75.64
11	16.66	6	7.27	28.5	31.05	4.26	8.72	9.69	10.41	15.68	10	35.75
13	14.59	42.15	6.08	25.35	26.21	16.65	28.23	45.23	19.5	25.21	25.75	15.84
15	13.53	6.4	6.23	7.43	8.27	10.28	15.89	10.66	15.74	11.47	12.02	26.5
17	10.71	5.66	5.75	10.03	10.5	2.51	9.66	5.83	13.98	6.88	10.34	5.52
19	8.84	4.86	3.47	2.71	3.1	3.61	1.8	4.41	7.39	8.62	9.95	6.04
21	8.49	7.59	5.19	5.53	3.66	8.39	16.43	11.31	16.39	3.75	7.79	6.47
23	7.18	2.45	2.83	1.7	0.93	3.77	3.96	5.78	5.65	3.09	3.9	7.79
25	6.16	3.14	4.61	1.5	1.22	2.64	6.94	9.15	2.38	5.7	7.31	6.37
27	5.98	3.28	4.38	1.13	0.85	2.3	3.08	11.36	3.43	7.05	8.05	0.94
29	4.98	1.63	2.98	0.8	0.34	2.58	5.74	5.65	8.32	4.91	5.7	10.59
31	4.26	1.13	2.79	0.56	0.4	2.71	8.96	7.06	11.22	8.41	8.32	10.88
33	4.12	0.78	1.38	0.48	0.46	1.13	5.46	5.18	3.56	7.82	7.42	2.19
35	3.45	0.94	0.4	0.31	0.17	1.81	5.33	3.67	6.03	3.73	4.97	7.82
37	3.07	0.49	0.87	0.27	0.37	2.39	3.93	2.16	7.59	4.07	3.36	4.08
39	3.23	0.21	0.77	0.24	0.15	0.73	3.42	1.21	4.19	1.78	2.35	4.05

The structure and the parameters of both obtained ANNs are verified by exciting the ANN with the given inputs. Responses of the ANN show that there were no errors in identifying the codes what is presented in the Table III. Negligible discrepancies may be observed.

TABLE III

ANN OUTPUT CODES

Expected Code	ANN Output (20 hidden neurons)	ANN Output (5 hidden neurons)
1	1.00002	1.00005
2	2.00003	2.00007
3	3.00002	3.00007
4	4.00003	4.00023
5	5.00003	5.00009
6	6.00002	6.00008
7	7.00001	7.00008
8	8.00002	8.00008
9	9.00001	9.00006
10	10	10.0001
11	11	11.0001
12	12	12.0001

Also, we used one more way to verify this neural network. Namely, we measured all the harmonics when all devices were operating at the same time (Table IV). This in unknown situation for the network, because this data were not used in the training process.

So, this was used as an excitation to the network with 20 inputs. The response of that network was 11,71. This value is closest to 12, giving us information that a combination similar to combination 12 is working. When we check the Table I, we notice that this is correct, while all devices except Kettle are included in combination 12, what is the closest to situation when all devices are operating at the same time. Of course, we cannot have an exact information, but it can be useful in the diagnosis process.

TABLE IV

VALUES OF HARMONICS FOR ALL DEVICES OPERATING AT THE SAME TIME

k	1	3	5	7	9	11	13
$\tilde{I}_k\,[\mathrm{mA}]$	11400	124.58	692.57	299	65.53	22.38	46.06

k	15	17	19	21	23	25
$\tilde{I}_k\,[\mathrm{mA}]$	30.16	4.19	11.79	15.75	14.58	14.28

k	27	29	31	33	35	37	39
$\tilde{I}_k\,[\mathrm{mA}]$	4.81	9.84	10.69	3.04	3.47	4.41	3.68

IV. Conclusion

Most of the known NILM methods for identification devices use artificial neural networks. We have shown one implementation of such method, with current harmonic amplitudes used as ANNs inputs. In our previous work, different definitions of power were used – active, reactive and distortion power. Having in mind that the powers are calculated from voltage and current harmonics, we have concluded that information that characterizes nonlinear loads encoded in those parameters also exists in harmonics.

Although we concluded that we needed only 5 measured harmonics (1^{st}, 3^{rd}, 5^{th}, 7^{th}, 9^{th}), but only for this specific case. This means that in some other situation, when measured values of harmonics have less distinguished values, we would need more input values in order to better distinguish which consumer is on the network.

In this stage of our work, we have deliberately neglected harmonics' phases, which also contains some information of nonlinear load. The development of the method for identification and classification that takes harmonics' phases into account is the next logical step.

Acknowledgment

Results described in this paper are obtained within the project TR32004 founded by Serbian Ministry of Science and Technology Development.

References

[1] Andrejević Stošović, M., Stevanović, D., Dimitrijević, M., "Monitoring and Classification of Nonlinear Loads Based on Artificial Neural Networks", *13th International Conference on Advanced Technologies, Systems and Services in Telecommunications* (*TELSIKS*), Niš, October 2017, pp. 443-446, 978-1-5386-1799-1, doi: 10.1109/TELSKS.2017.8246320.

[2] Andrejević Stošović, M., Stevanović, D., Dimitrijević, M., "Classification of Nonlinear Loads Based on Artificial Neural Networks", *IEEE 30th International Conference on Microelectronics* (*MIEL*), Niš, October 2017, pp. 221-224, 978-1-5386-2561-3, doi:10.1109/MIEL.2017.8190107.

[3] M. Dimitrijević, *Elektronski sistem za analizu polifaznih opterećenja baziran na FPGA*, PhD Thesis, Faculty of Electronic Engineering, University of Niš, Niš, Serbia, 2012.

[4] M. Dimitrijević, Stevanović, D., Andrejević Stošović, M., "Real-time System for Nonlinear Load Analysis in 50A Current Range", Proc. Of the 8^{th} Small Systems Simulation Simposium, pp. 83-88, February 12-14, 2018, Niš, Serbia

[5] National Instruments: NI 9225 Operating Instructions and Specifications.

[6] T. Masters, "Practical Neural Network Recipes in C++", Academic Press, San Diego, 1993.

[7] E. B. Baum, and D. Haussler, "What size net gives valid generalization", Neural Computing, Vol. 1, 1989, pp. 151-60.

A Flexible FPGA-Based Data Acquisition System with Integrated ADCs and 32-bit RISC-V Softcore

Nikola Ž. Petrović, *Student Member, IEEE,* and Vladimir M. Milovanović, *Senior Member, IEEE*

***Abstract*—A general purpose low-cost data acquisition system which includes integrated 12-bit analog-to-digital data converters (ADCs) and a 32-bit RISC-V soft microprocessor implemented on an FPGA-fabric is presented. Major part of the proposed system is not conceived as an instance, but rather as a design generator and is hence described in Chisel hardware construction language. The use of such a modern generator-oriented language enables extreme flexibility through comprehensive set of parameters and rapid system customization approach. The open instruction set architecture core that can be exploited for custom data processing is also obtained from the free Rocket System-on-Chip generator together with a standard set of peripherals thus supporting various connectivity interfaces and protocols. A prototype data acquisitioner is built around commercially-available Arty development platform which features internal 1 MSPS dual-channel ADC and its operation as a low-end oscilloscope is demonstrated.**

***Index Terms*—ADC, Chisel, RISC-V, Rocket, System-on-Chip, TileLink.**

I. Introduction

THE RISC-V specification is a royalty-free Instruction Set Architecture (ISA) standard for low-cost processors aimed to support architecture research and education [1]. The specification defines 32-bit, 32-bit embedded, 64-bit, and 128-bit base ISAs and optional extensions for compressed, multiplication, atomic, single, double, and quad-precision floating point instructions. RISC-V ISA is easily portable between different development environments and highly flexible to match the requirements of different applications. These features made RISC-V natural choice for FPGA softcore processor and that led to its adoption in academia and industry alike and an overwhelming proliferation of projects. For example, UC Berkeley provides open-source Rocket-chip System on a Chip (SoC) generator, which has been successfully taped-out over a dozen times in multiple different modern technologies by multiple groups [2], [3]. Also, Craft2-chip from UC Berkeley is freely available as a github repository. Company SiliconFive offers both open-source SoC generator for FPGA and commercial RISC-V IPs. Finally, in [4] and [5] processors are reported capable of running Linux using RISC-V instruction set.

We are presenting the implementation of a generator-based flexible data acquisition system implemented on FPGA with integrated Successive Approximation Register (SAR) ADCs and a 32-bit RISC-V softcore. Features of the system are: TileLink bus for interfacing between processor and the peripherals attached to it, a 16KiB instruction 2-way cache, 16KiB data SRAM, 8KiB read-only memory, Serial Peripheral Interface slave for output, dedicated Quad-SPI flash interface for holding code and data, UART for serial communication, General Purpose I/O (GPIO) complex, 16-bit Pulse Width Modulation (PWM) periphery, JTAG and debug unit with two hardware breakpoints and 12-bit SAR Analog-to-Digital converters. Most of the peripherals were written in hardware construction language (HCL) Chisel which is developed at UC Berkeley and it is based on Scala programming language. The design was implemented and tested on Xilinx Arty A7 FPGA.

N. Ž. Petrović is with the Department of Electronics, School of Electrical Engineering, University of Belgrade, Bulevar kralja Aleksandra 73, 11120 Belgrade, Serbia, and is also with NovelIC Microsystems d.o.o., Veljka Dugoševića 54/B5, 11060 Belgrade, Serbia (e-mail: p.z.nikola@etf.bg.ac.rs).

V. M. Milovanović is with the Faculty of Engineering, University of Kragujevac, Sestre Janjić 6, 34000 Kragujevac, Serbia (e-mail: vlada@kg.ac.rs).

II. The Rocket core

Data acquisition system was based around Rocket Chip SoC generator. The Rocket Chip is an open-source SoC design generator developed at UC Berkeley suitable for research and industrial purposes [6]. It generates general-purpose processor cores that use the open RISC-V ISA, and provides an in-order core generator. Instead of being a single instance of an SoC design, Rocket chip is highly parametrized and customized design generator that emits synthesizable Register Transfer Level (RTL) Verilog code. This extensive parametrization enables SoC designers to generate multiple SoCs from single high-level source thus enabling easy customization for specific application. For example, by changing few parameters a SoC designer can generate SoCs ranging from embedded microcontrollers to multi-core server chips. In order to provide generators for cores, caches, and interconnects, Rocket Chip is written in Chisel hardware construction language that supports advanced hardware design using highly parameterized generators and layered domain-specific hardware languages [7]. Chisel can generate a high-speed C++ based cycle-accurate software simulator, or low-level Verilog designed to map to either FPGAs or to a standard ASIC flow for synthesis.

Rocket chip deals with new modules by using Scala's object trait cake pattern which basically involves placing code inside traits. In the Rocket core there are two kinds of traits: a LazyModule trait and a module implementation trait. The LazyModule trait runs setup code that must execute before all the hardware gets elaborated. For a simple memory-mapped peripheral, this just involves connecting the peripheral's TileLink node to the Memory-Mapped I/O (MMIO) crossbar.

III. Architecture

In this work we are using the TileLink protocol developed by UC Berkeley for communication process. TileLink is a protocol designed to be a substrate for cache coherence transactions implementing a particular cache coherence policy within an on-chip memory hierarchy [8]. Its purpose is to orthogonalize the design of the on-chip network and the implementation of the cache controllers from the design of the coherence protocol itself. Any cache coherence protocol that conforms to TileLink's transaction structure can be used interchangeably with the physical networks and cache controllers from the Rocket Chip generator.

Used Rocket core supports local (including software and timer) interrupts and global interrupts. Local interrupts are signaled directly to an individual hart with a dedicated interrupt value which reduces interrupt latency as there is no arbitration required to determine which hart will service a given request, nor additional memory accesses required to determine the cause of the interrupt. Software and timer interrupts are local interrupts generated by the Core Local Interruptor (CLINT). Global interrupts are routed through a Platform Level Interrupt Controller (PLIC), which can direct interrupt to any hart in the system via the external interrupt.

The architecture of the microcontroller is shown in Fig. 1. All peripheral interfaces are interconnected to the TileLink bus. All peripherals are written as parametrized LazyModule and module implementation traits in order to enable easy integration with the Rocket core. Verilog code for Xilinx XADC IP is instantiated inside Chisel BlackBox module which is then connected to the TileLink bus by using TLRegisterRouter. TLRegisterRouter is used to abstracts away the details of handling the TileLink protocol and provides a convenient interface for specifying memory-mapped registers. The peripheral memory maps were designed to only require naturally aligned 32-bit memory accesses.

IV. Peripherals

In the this section the short overview of used peripherals is presented.

A. General Purpose I/O

The GPIO controller is a peripheral device mapped in the internal memory map. The GPIO complex manages the connection of digital I/O pads to digital peripherals, including SPI, UART, and PWM controller, as well as for regular programmed I/O operations. It is responsible for low-level configuration of the actual GPIO pads on the device (direction, pullup-enable, and drive value), as well as selecting between various sources of the controls for these signals. The GPIO controller allows separate configuration of each of *N* GPIO bits. GPIO module has parametrized number of GPIO bits and by changing the *width* parameter given bellow, different number of GPIO bits can be achieved.

```
case PeripheryGPIOKey=>List(
 GPIOParams(address=0x10012000,width=32))
```

B. Quad-SPI

A dedicated QSPI flash interface is provided to hold code and data for the system. The QSPI interface supports burst reads of 32 bytes over TileLink to accelerate instruction cache refills. The QSPI interface also supports single-word data reads over the primary TileLink interface, as well as programming operations using memory-mapped control registers.

The SPI controller supports master-only operation over the single-lane, dual-lane, and quad-lane protocols. The baseline controller provides a FIFO-based interface for performing programmed I/O. Software initiates a transfer by en-queuing a frame in the transmit FIFO; when the transfer completes, the slave response is placed in the receive FIFO. In addition, the dedicated QSPI_0 controller implements a SPI flash read sequencer, which exposes the external SPI flash contents as a read/execute-only memory-mapped device. The QSPI_0 controller is reset to a state which allows memory-mapped reads. QSPI_0 has Execute in Place mode enabled by default and one chip select signal. QSPI_1 has 4 chip select signals enabled.

In order to archive better performance, sequential accesses are automatically combined into one long read command. Switching between the programmed-I/O and memory-mapped modes is enabled by the *fctrl* control register. While in programmed-I/O mode, memory-mapped reads do not access the external SPI flash device and instead return 0 immediately. Hardware interlocks ensure that the current transfer completes before mode transitions and control register updates take effect.

Adding additional SPI to the system can be easily done just by adding new set of parameters inside PeripherySPIKey list. For example, in our data acquisition system we had QSPI_1 with four chip select signals at address 0x10024000. If we want to add another QSPI with only one chip select signal, we could write:

```
case PeripherySPIKey=>List(
 SPIParams(csWidth=4,rAddress=0x10024000),
 SPIParams(csWidth=1,rAddress=0x10034000))
```

By changing one line of code, new QSPI periphery with one chip select signal would be added at address 0x10034000.

C. UART

The UART peripheral supports the 8-N-1 and 8-N-2 formats: 8 data bits, no parity bit, 1 start bit, 1 or 2 stop bits. It supports 8-entry transmit and receive FIFO buffers with programmable watermark interrupts and 16×Rx oversampling with 2/3 majority voting per bit. Baud rate of the UART can be changed by changing the value of UART *div* register. Baud rate can be calculated as:

$$f_{baud} = \frac{f_{in}}{1 + div} \tag{1}$$

The UART peripheral does not support hardware flow control or other modem control signals, or synchronous serial data transfers.

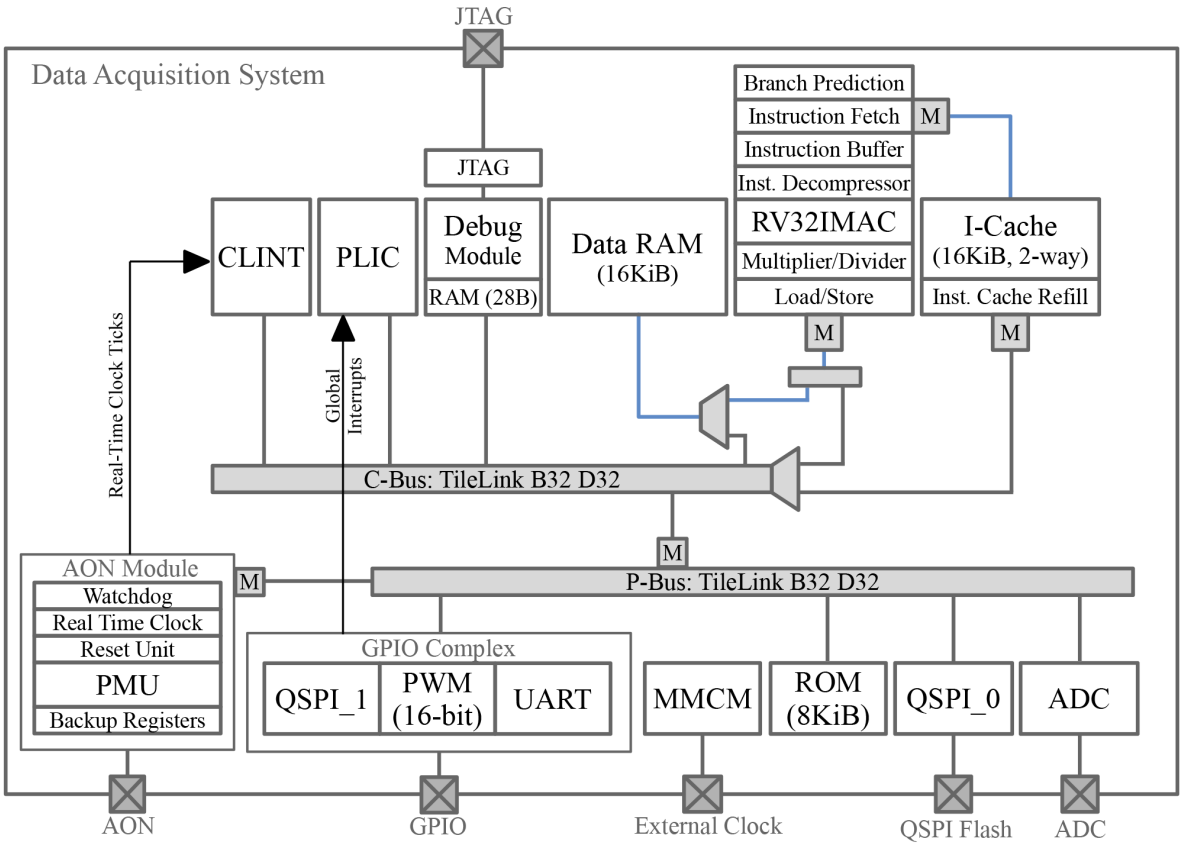


Fig. 1. Data acquisition system architecture.

Same as with SPI periphery, additional UART peripheries could be added by adding new set of parameters inside PeripheryUARTKey list:

```
case PeripheryUARTKey=>List(
 UARTParams(address=0x10013000),
 UARTParams(address=0x10023000))
```

D. PWM

The default PWM configuration contains four independent PWM comparators (*pwmcmp_0–pwmcmp_3*), but custom configurations with different number of comparators can be generated. The pulse width modulation (PWM) peripheral can generate multiple types of waveforms on GPIO output pins, and can also be used to generate several forms of internal timer interrupt.

The comparator results are captured in the *pwmcmp_X_ip* flops (where $X \in [0, 1, 2, 3]$) and then fed to the PLIC as potential interrupt sources. The *pwmcmp_X_ip* outputs are further processed by an output ganging stage before being fed to the GPIOs. The PWM unit can be provided in different comparator precisions up to 16 bits. In order to support clock scaling, the PWM count register is 15 bits wider than the comparator precision.

Different number of PWM peripherals can be added by adding new sets of parameters inside PeripheryPWMKey list. By changing the value of *cmpWidth* parameter, comparator precision can be set. For example, to set three different PWM modules with comparator precisions of 16, 12 and 8 bits respectively, PeripheryPWMKey list should contain:

```
case PeripheryPWMKey=>List(
 PWMParams(address=0x10015000,cmpWidth=16),
 PWMParams(address=0x10025000,cmpWidth=12),
 PWMParams(address=0x10035000,cmpWidth=8))
```

E. Always-on (AON) block

The AON block contains the reset logic for the chip, a watch-dog timer, connections for an off-chip low-frequency oscillator, the real-time clock, a programmable power-management unit, and 16×32-bit backup registers that retain their state while the rest of the chip is in a low-power mode. The AON can be instructed to put the system to sleep and the Always-on block can be programmed to exit sleep mode on a real-time clock interrupt or when the external digital wakeup pin is pulled low. An AON reset can be triggered by an external active-low reset pin or expiration of the watchdog timer. AON domain is continuously powered from an off-chip power source, either a regulated power supply or a battery.

F. Debug module

The debug module is accessed through JTAG, and has support for two programmable hardware break points. The

debug RAM has 28 bytes of storage. A JTAG connection is used to connect the external debugger to the internal debug module.

G. Analog-to-Digital converter

In order to perform analog to digital conversion, Arty A7 on-chip SAR Analog-to-Digital converters are used. The Xilinx IP XADC includes a dual 12-bit, 1 Megasample per second (MSPS) ADCs and on-chip sensors. The ADCs provides a general-purpose, high-precision analog interface for a range of applications. These ADCs support a range of operating modes, such as externally triggered and simultaneous sampling on both ADCs and various analog input signal types, for example, unipolar and differential. The ADCs can access up to 17 external analog input channels and also includes several on-chip sensors that support measurement of the on-chip power supply voltages and die temperature. The ADC conversion data is stored in dedicated registers called status registers. These registers are accessible through the FPGA interconnect using a 16-bit synchronous read and write port called the dynamic reconfiguration port (DRP). ADCs were used in continuous mode and four different external analog inputs were configured.

In order to connect Xilinx IP XADC to the TileLink, BlackBox wrapper for the Verilog ADC module is written. TLRegisterRouter and BlackBox ADC module were added to the module implementation trait, and the LazyModule trait was written for the ADC. By doing that, ADC module could be added to data acquisition system by extending our data acquisition class with ADC traits:

```
class dataAcquisitionSystem
  extends RocketSubsystem
    with HasPeripheryADCTL{
      override lazy val module
         = new dataAcquisitionModule(this)
    }

class dataAcquisitionModule
  [+L <: dataAcquisitionSystem](_outer: L)
  extends RocketSubsystemModuleImp(_outer)
    with HasPeripheryADCTLModuleImp
```

V. Results

The proposed data acquisition system was implemented on Xilinx Arty A7 FPGA. Mixed-Mode Clock Manager (MMCM) was used to generate 90MHz clock for Rocket core, 30MHz clock for JTAG and 10MHz clock for the AON module from the external 100MHz oscillator. Verification of the peripherals has been performed from the perspective of the Rocket core and the UART where UART was used for serial communication between FPGA and PC. The dedicated QSPI flash controller was used to fetch code from the off-chip SPI flash.

A simple C code was written to enable the ADC conversion and to send the sampled data via the UART. On the FPGA board, buttons were connected to the design and the interrupt routine was written for those buttons. By pressing the button, the ADC sampling flag was set. Inside the main loop 10k ADC samples were sent via the UART if ADC sampling flag was set. After sending sampled data, ADC sampling flag was unset. In order to send sampled data from the ADC with width of 16 bits, two UART transfers were needed considering UART symbol width of 8 bits. First 8 bits to be sent via UART were the upper 8 bits of the ADC data register and then the lower 8 bits were sent. Before sending ADC samples, 4 bytes used for data aligning were sent via UART and after sending ADC samples, two bytes were send to denote the end of UART transaction. Python script was written to collect data sent from microcontroller via UART and to plot received data.

In order to test functionality of the used integrated ADCs, signal generator was connected to the ADC input pins of the FPGA. Waveforms used for testing were sinusoidal, triangular and rectangular waveforms between 0V and 3V with frequencies 0.25kHz and 0.5kHz. Obtained diagrams for 0.25kHz waveforms are given in Fig. 2 and 0.5kHz waveforms are given in Fig. 3.

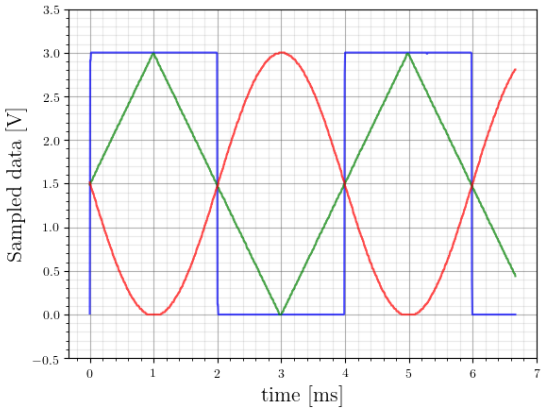


Fig. 2. Data sampled with waveform frequency of 0.25kHz.

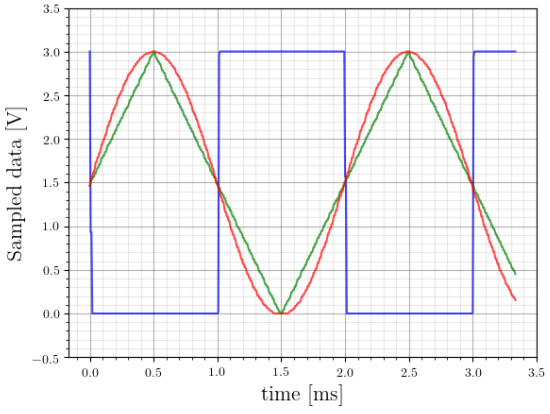


Fig. 3. Data sampled with waveform frequency of 0.5kHz.

In Table 1 post-implementation resource utilization is given and power estimation for the design is 0.308W where 42% is from MMCM.

TABLE I
POST-IMPLEMENTATION RESOURCE UTILIZATION

	Utilization	Available	Utilization %
LUT	14285	20800	68.7 %
LUTRAM	695	9600	7.24 %
FF	8159	41600	19.6 %
BRAM	5.5	50	11.0 %
DSP	2	90	2.23 %
MMCM	1	5	20.0 %

Layout of the implemented design on the Arty A7 FPGA board is presented in Fig. 4.

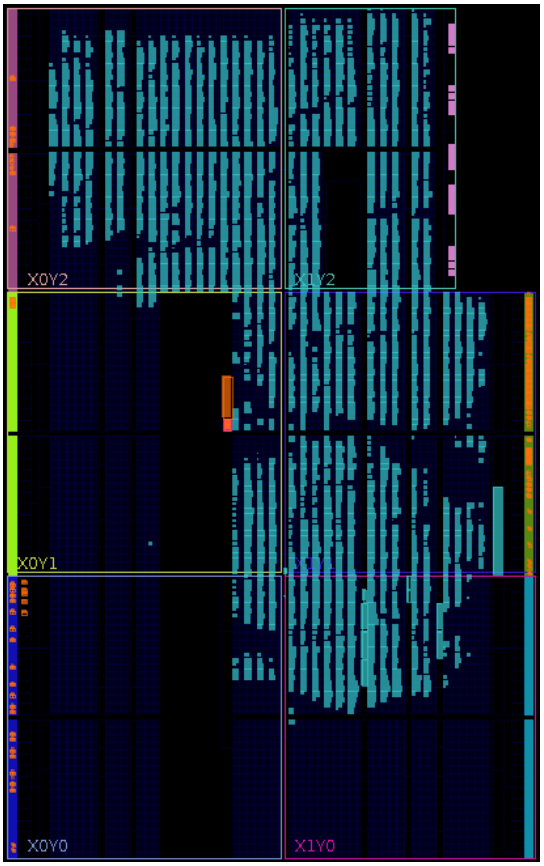


Fig. 4. Layout of the implemented design on Arty A7 FPGA board.

VI. CONCLUSIONS

In this work, a general purpose low-cost data acquisition system with integrated 12-bit SAR ADCs was presented and implemented on Xilinx Arty A7 FPGA. Data acquisition system was built around Rocket System-on-Chip design generator which is based on RISC-V ISA. By using extensive parametrization of the Rocket core and flexibility of peripherals written in Chisel hardware construction language, data acquisition system with extreme flexibility through comprehensive set of parameters was implemented and its operation as a low-end oscilloscope was demonstrated.

REFERENCES

[1] A. Waterman, Y. Lee, D. A. Patterson, and K. Asanovi, "The risc-v instruction set manual. volume 1: User-level isa, version 2.0," EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2014-54, May 2014.

[2] S. Bailey, J. Han, P. Rigge, R. Lin, E. Chang, H. Mao, Z. Wang, C. Markley, A. Izraelevitz, A. Wang, *et al.*, "A generated multirate signal analysis risc-v soc in 16nm finfet," in *2018 IEEE Asian Solid-State Circuits Conference (A-SSCC)*, pp. 285–288, IEEE, 2018.

[3] A. Wang, W. Bae, J. Han, S. Bailey, P. Rigge, O. Ocal, Z. Wang, K. Ramchandran, E. Alon, and B. Nikolić, "A real-time, analog/digital co-designed 1.89-ghz bandwidth, 175-khz resolution sparse spectral analysis risc-v soc in 16-nm finfet," in *ESSCIRC 2018-IEEE 44th European Solid State Circuits Conference (ESSCIRC)*, pp. 322–325, IEEE, 2018.

[4] B. Zimmer, Y. Lee, A. Puggelli, J. Kwak, R. Jevtic, B. Keller, S. Bailey, M. Blagojevic, P.-F. Chiu, H.-P. Le, *et al.*, "A risc-v vector processor with tightly-integrated switched-capacitor dc-dc converters in 28nm fdsoi," in *2015 Symposium on VLSI Circuits (VLSI Circuits)*, pp. C316–C317, IEEE, 2015.

[5] Y. Lee, A. Waterman, R. Avizienis, H. Cook, C. Sun, V. Stojanović, and K. Asanović, "A 45nm 1.3 ghz 16.7 double-precision gflops/w risc-v processor with vector accelerators," in *ESSCIRC 2014-40th European Solid State Circuits Conference (ESSCIRC)*, pp. 199–202, IEEE, 2014.

[6] K. Asanovic, R. Avizienis, J. Bachrach, S. Beamer, D. Biancolin, C. Celio, H. Cook, D. Dabbelt, J. Hauser, A. Izraelevitz, *et al.*, "The rocket chip generator," *EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2016-17*, 2016.

[7] J. Bachrach, H. Vo, B. Richards, Y. Lee, A. Waterman, R. Avižienis, J. Wawrzynek, and K. Asanović, "Chisel: constructing hardware in a scala embedded language," in *DAC Design Automation Conference 2012*, pp. 1212–1221, IEEE, 2012.

[8] H. Cook, W. Terpstra, and Y. Lee, "Diplomatic design patterns: A tilelink case study," in *1st Workshop on Computer Architecture Research with RISC-V*, 2017.

Stressing Issue of a Piezoceramic Cylinder with Radial Polarization

Igor Jovanović, Ljubiša Perić, Uglješa Jovanović and Dragan Mančić

***Abstract*—This paper presents a general case of stressing a circular-ring cross-sectional cylinder with a radial type of polarization. It is assumed that the cylinder is infinitely long in direction of the *z*-axis and that the componential strain in that direction is equal to zero.**

By applying the equations of electroelasticity in polar-cylindrical coordinates as well as satisfying electrical boundary conditions for the electric potential and mechanical boundary conditions for the mechanical stress of a cylinder with a radial type of polarization, componential displacements in radial direction are determined for the piezoceramic cylinder made from PZT4 piezoceramic material.

***Index Terms*—componential displacements, radial polarization, PZT4 piezoceramic materials.**

I. Introduction

Piezoceramics are one of the most common active materials used for ultrasonic transducers, actuators, resonators, wave filters delay lines, transformers, pressure sensing devices, energy harvesting devices, etc [1]. When used as piezoelectric actuators and sensors, their applications cover various fields of technology from the aerospace, military and industry, to instrumentation and medicine [2], [3].

Piezoelectric cylinders are basic actuating elements in various ultrasonic applications. The main subject of this study is investigation of free vibration of a circular-ring cross-sectional cylinder with a radial type of polarization. Knowing resonant frequencies of piezoelectric cylinders is an initial condition when designing sensors and transducers [4]. For the precise analysis of the oscillations of piezoelectric cylinders, it is ideal to use models that take into account the coupling between different oscillations (thickness, radial, flexion, etc.), which leads to a complex system of nonlinear equations that are difficult to solve [5].

In this paper, in general, the method of solving the problems of oscillating circular-ring cross-sectional cylinder with radial polarization and electrode coatings on cylindrical surfaces, on which an AC voltage is applied, is shown.

II. Fundamental Equations

At first, problem of radial polarization of hollow cylinder with electrode coatings on cylindric surfaces, shown in Fig. 1, is considered.

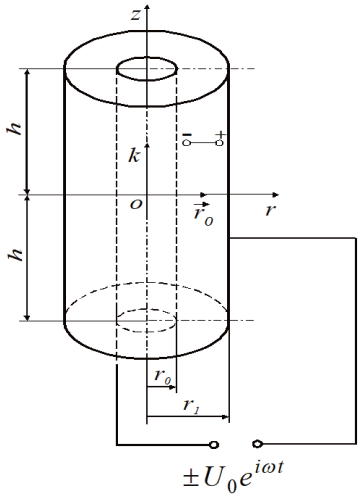


Fig. 1. Cylinder with radial polarization and electrodes on cylindric surfaces.

It is assumed that cylindrical surfaces of the annular cylinder $r=r_0$ and $r=r_1$ are completely covered with electrode coatings, on which electric voltage $\pm U_0 e^{i\omega t}$ is applied. So, oscillations of the cylinder are excited by alternate electric potential difference on electrodes, which is changing harmonically in time by circular frequency ω, obtained from electric voltage generator.

Electric boundary conditions, in case when electric voltage is applied onto the electrodes on cylindric surfaces, are [6]:

$$\psi\Big|_{r=\begin{Bmatrix} r=r_0 \\ r=r_1 \end{Bmatrix}} = \pm U_0 e^{i\omega t}, \tag{1}$$

$$D_z\Big|_{z=\pm h} = \left[-d_{11}^{\varepsilon} \frac{\partial \psi}{\partial z} + e_{15} \left(\frac{\partial u}{\partial z} + \frac{\partial w}{\partial r} \right) \right]\Bigg|_{z=\pm h} = 0, \tag{2}$$

This way of excitation of piezoceramic circular-ring cylinder causes occurrence of only radial oscillations, i.e., vector of componential displacements is $\vec{s} = u(r, t)\,\vec{r}_0$, and it has only radial component. Electric potential ψ is also function of only radial coordinate and time, that is $\psi=\psi(r,t)$. For this case of oscillations of kineto-electricity state parameters depend only of coordinate *r*, that is:

Igor Jovanović, Uglješa Jovanović and Dragan Mančić are with the University of Niš, Faculty of Electronic Engineering, 14 Aleksandra Medvedeva, 18000 Niš, Serbia (e-mail: igor.jovanovic@elfak.ni.ac.rs, ugljesa.jovanovic@elfak.ni.ac.rs and dragan.mancic@elfak.ni.ac.rs).

Ljubiša Perić is with the Regional Chamber of Economy Niš, Dobrička 2, 18000 Niš, Serbia (e-mail: ljubisa.peric@rpknis.rs).

$$u = u(r,\ t),$$
$$\psi = \psi(r,\ t), \qquad (3)$$
$$v = 0,\ \ w = 0,$$

By substituting coordinates (3) into Cauchy's kinematic equations:

$$\varepsilon_r = \frac{\partial u}{\partial r},\ \ \gamma_{rz} = 2\varepsilon_{rz} = \left(\frac{\partial u}{\partial z} + \frac{\partial w}{\partial r}\right),$$
$$\varepsilon_\phi = \frac{1}{r}\frac{\partial v}{\partial r} + \frac{u}{r},\ \ \gamma_{r\phi} = 2\varepsilon_{r\phi} = \left(\frac{\partial v}{\partial r} - \frac{v}{r} + \frac{1}{r}\frac{\partial u}{\partial \phi}\right), \qquad (4)$$
$$\varepsilon_z = \frac{\partial w}{\partial z},\ \ \gamma_{z\varphi} = 2\varepsilon_{z\varphi} = \left(\frac{\partial v}{\partial z} + \frac{1}{r}\frac{\partial w}{\partial \varphi}\right).$$

and into the equations of electroelasticity [7]:

$$\vec{E} = -grad\psi, \qquad (5)$$

one gets:

$$\varepsilon_r = \frac{\partial u}{\partial r},\ \ \varepsilon_\phi = \frac{u}{r},\ \ E_r = -\frac{\partial \psi}{\partial r},$$
$$E_r = -\frac{\partial \psi}{\partial r},\ E_r = E_\varphi = 0 \qquad (6)$$
$$\varepsilon_z = 0,\ \ \gamma_{\phi z} = \gamma_{rz} = \gamma_{r\phi} = 0.$$

In deriving of the coupled equations of electroelasticity for piezoceramic cylinders with radial polarization, one must start from equations of piezoelectric effect:

$$\sigma_r = c_{33}^E\varepsilon_r + c_{13}^E\left(\varepsilon_\phi + \varepsilon_z\right) - e_{33}E_r,$$
$$\sigma_\phi = c_{13}^E\varepsilon_r + c_{11}^E\varepsilon_\phi + c_{12}^E\varepsilon_\phi - e_{31}E_r,$$
$$\sigma_z = c_{13}^E\varepsilon_r + c_{12}^E\varepsilon_\phi + c_{11}^E\varepsilon_z - e_{31}E_r,$$
$$\tau_{r\phi} = c_{44}^E\gamma_{r\phi} - e_{15}E_\phi,$$
$$\tau_{rz} = c_{44}^E\gamma_{rz} - e_{15}E_z, \qquad (7)$$
$$\tau_{\phi z} = \frac{1}{2}\left(c_{11}^E - c_{12}^E\right)\gamma_{\phi z},$$
$$D_r = d_{33}^\varepsilon E_r + e_{31}\left(\varepsilon_\phi + \varepsilon_z\right) + e_{33}\varepsilon_r,$$
$$D_\phi = d_{11}^\varepsilon E_\phi + e_{15}\gamma_{r\phi},$$
$$D_z = d_{11}^\varepsilon E_z + e_{15}\gamma_{rz}.$$

By inserting the obtained expressions (6) into the state equations for radial type of cylinder polarization (7), one gets expressions for componential mechanical stresses (σ) and piezoelectric displacements (D):

$$\sigma_r = c_{33}^E\frac{\partial u}{\partial r} + c_{13}^E\frac{u}{r} + e_{33}\frac{\partial \psi}{\partial r},$$
$$\sigma_\phi = c_{13}^E\frac{\partial u}{\partial r} + c_{11}^E\frac{u}{r} + e_{31}\frac{\partial \psi}{\partial r},$$
$$\sigma_z = c_{13}^E\frac{\partial u}{\partial r} + c_{12}^E\frac{u}{r} + e_{31}\frac{\partial \psi}{\partial r}, \qquad (8)$$
$$D_r = -d_{33}^\varepsilon\frac{\partial \psi}{\partial r} + e_{31}\frac{u}{r} + e_{33}\frac{\partial u}{\partial r},$$
$$\tau_{\phi z} = \tau_{rz} = \tau_{r\phi} = 0,$$
$$D_\phi = D_z = 0.$$

By substituting expression (8) into Navier's equations of motion:

$$\frac{\partial \sigma_r}{\partial r} + \frac{1}{r}\frac{\partial \tau_{r\phi}}{\partial \phi} + \frac{\partial \tau_{rz}}{\partial z} + \frac{\sigma_r - \sigma_\phi}{r} = \rho\frac{\partial^2 u}{\partial t^2},$$
$$\frac{\partial \tau_{r\phi}}{\partial r} + \frac{1}{r}\frac{\partial \sigma_\phi}{\partial \phi} + \frac{\partial \tau_{z\phi}}{\partial z} + \frac{2\tau_{r\phi}}{r} = \rho\frac{\partial^2 v}{\partial t^2}, \qquad (9)$$
$$\frac{\partial \tau_{rz}}{\partial r} + \frac{1}{r}\frac{\partial \tau_{z\phi}}{\partial \phi}\frac{\partial \sigma_z}{\partial z} + \frac{2\tau_{rz}}{r} = \rho\frac{\partial^2 w}{\partial t^2},$$

and into the equation of electroelasticity [7]:

$$\vec{E} = -\frac{\partial \psi}{\partial r}\vec{r}_0 - \frac{1}{r}\frac{\partial \psi}{\partial \phi}\vec{c}_0 - \frac{\partial \psi}{\partial z}\vec{k}, \qquad (10)$$

one gets system of two partial differential equations in form of [6]:

$$c_{33}^E\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial u}{\partial r}\right) - \frac{c_{11}^E}{r^2}u = \rho\frac{\partial^2 u}{\partial t^2} - e_{33}\frac{\partial^2 \psi}{\partial r^2} - \frac{e_{33} - e_{31}}{r}\frac{\partial \psi}{\partial r}, \qquad (11)$$
$$d_{33}^\varepsilon\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial \psi}{\partial r}\right) = e_{33}\frac{\partial^2 u}{\partial r^2} + \frac{e_{33} + e_{31}}{r}\frac{\partial u}{\partial r}. \qquad (12)$$

Further, it is assumed that on cylindric surfaces $r=r_0$ and $r=r_1$ external mechanical loads are absent, so the general expression for the vector of total stress on contour surfaces is equal to zero. Conditions for absence of shear stresses are identically satisfied because of absence of displacements v and w, and condition for absence of normal stress σ_r leads to the following equation (mechanical boundary conditions):

$$\sigma_r\Big|_{r=\begin{Bmatrix} r_0 \\ r_1 \end{Bmatrix}} = \left(c_{33}^E\frac{\partial u}{\partial r} + c_{13}^E\frac{u}{r} + e_{33}\frac{\partial \psi}{\partial r}\right)\Bigg|_{r=\begin{Bmatrix} r_0 \\ r_1 \end{Bmatrix}} = 0. \qquad (13)$$

Solutions of partial differential equations (11) and (12) are searched for in the following form:

$$u\left(r,\ t\right) = \hat{u}\left(r\right)e^{i\omega t}, \qquad (14)$$

$$\psi(r,t)=\hat{\psi}(r)\,e^{i\omega t}. \tag{15}$$

By substituting assumed solutions (14) and (15) into the system of differential equations (11) and (12), one gets the following system of common differential equations for determination of amplitude eigenfunctions:

$$c_{33}^{E}\frac{1}{r}\frac{d}{dr}\left(r\frac{d\hat{u}}{dr}\right)-\frac{c_{11}^{E}}{r^2}\hat{u}+\rho\omega^2\hat{u}=-e_{33}\frac{d^2\hat{\psi}}{dr^2}-\frac{e_{33}-e_{31}}{r}\frac{d\hat{\psi}}{dr}, \tag{16}$$

$$d_{33}^{\varepsilon}\frac{1}{r}\frac{d}{dr}\left(r\frac{d\hat{\psi}}{dr}\right)=e_{33}\frac{d^2\hat{u}}{dr^2}+\frac{e_{33}+e_{31}}{r}\frac{d\hat{u}}{dr}. \tag{17}$$

The equation (17) is integrated with respect to r, and as a result one gets:

$$d_{33}^{\varepsilon}\frac{d\hat{\psi}}{dr}=e_{33}\frac{d\hat{u}}{dr}+e_{31}\frac{\hat{u}}{r}+\frac{C}{r},$$

$$\int r\frac{d^2\hat{u}}{dr^2}\,dr=r\frac{d\hat{u}}{dr}-\hat{u}, \tag{18}$$

$$\frac{d^2\hat{\psi}}{dr^2}=\frac{e_{33}}{d_{33}^{\varepsilon}}\frac{d^2\hat{u}}{dr^2}-\frac{e_{31}}{d_{33}^{\varepsilon}}\frac{\hat{u}}{r^2}+\frac{e_{31}}{d_{33}^{\varepsilon}}\frac{1}{r}\frac{d\hat{u}}{dr}-\frac{C}{d_{33}^{\varepsilon}}\frac{1}{r^2},$$

where C is an arbitrary constant.

Now the expression (18) is substituted in the equation (16), amplitude eigenfunction of the electric potential $\hat{\psi}$ is excluded, and one gets:

$$r^2\frac{d^2\hat{u}}{dr^2}+r\frac{d\hat{u}}{dr}+\left(\lambda^2r^2-\nu_1^2\right)\hat{u}=\frac{e_{31}C}{c_{33}^{E}d_{33}^{\varepsilon}+e_{33}^2}, \tag{19}$$

standard form of differential equation of the second order, for amplitude eigenfunction of the componential displacement $\hat{u}(r)$, which contains on the right side an addend with arbitrary constant C.

Characteristic numbers λ and ν_1 are determined according to expressions:

$$\lambda=\omega\sqrt{\frac{\rho\,d_{33}^{\varepsilon}}{c_{33}^{E}\,d_{33}^{\varepsilon}+e_{33}^2}},\quad \nu_1=\sqrt{\frac{c_{11}^{E}d_{33}^{\varepsilon}+e_{31}^2}{c_{33}^{E}d_{33}^{\varepsilon}+e_{33}^2}}. \tag{20}$$

General solution of the common differential equation (19) can be expressed through Bessel's functions of first and second rank of J_ν and Y_ν real argument, that is:

$$\hat{u}(r)=AJ_\nu(\lambda r)+BY_\nu(\lambda r)+\frac{\pi}{2}\frac{e_{31}C}{c_{33}^{E}d_{33}^{\varepsilon}+e_{33}^2}\cdot$$

$$\cdot\left\{Y_\nu(\lambda r)\int_{r_0}^{r}J_\nu(\lambda r)\frac{dr}{r}+J_\nu(\lambda r)\int_{r}^{r_1}Y_\nu(\lambda r)\frac{dr}{r}\right\}, \tag{21}$$

one may see that new arbitrary constants A and B are introduced. By introducing the solution (21) into the first equation (18) and integrating, one gets an expression for amplitude eigenfunction of the electric potential:

$$\hat{\psi}(r)=\frac{e_{33}}{d_{33}^{\varepsilon}}\left[\hat{u}(r)-\hat{u}(r_0)\right]+\frac{C}{d_{33}^{\varepsilon}}\ln r+\frac{e_{31}}{d_{33}^{\varepsilon}}\int_{r_0}^{r}\hat{u}(r)\frac{dr}{r}+C_1. \tag{22}$$

In the obtained expressions for amplitude eigenfunction $\hat{u}(r)$ and $\hat{\psi}(r)$ one may see that four arbitrary constants exist. They are determined from the condition of satisfying electric and mechanical boundary conditions on the surfaces of the piezoceramic annular specimen (1) and (13). By introducing the expression (21) and (22) into mentioned boundary conditions one gets system of four algebraic equations with four unknowns – A, B, C, and C_1, which are very complex.

III. Numerical Analysis and Discussion

Subject of observation in this paper is stressing of circular-ring cross-sectional PZT4 [8] cylinder with a circular type of polarization, with the following dimensions: r_0=7.5 mm, r_1=19 mm and density ρ=7500 kg/m^3 (Fig. 1).

Numerical analysis was performed using Matlab software, and radial displacement distributions of the points between cylindrical surfaces of piezoceramic cylinder were obtained for the first and the second radial resonant modes.

Figures 2 and 3 present radial displacements of points between the inner and outer cylindrical contour surfaces in the function of radius, r, and frequency, f. One may notice that in case of both modes, internal and external cylindrical surfaces move in the same directions (Fig. 2 and Fig. 3), while for the second mode (Fig. 3), the cylinder acts as a full-wave resonator. This means that for this case at r=10.37 mm and r=16.13 mm exists a circular line that represents the wave node, for which the displacement is u_r=0.

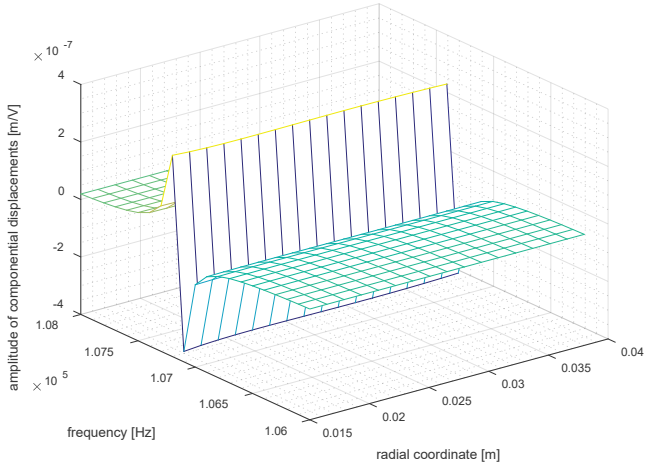


Fig. 2. Amplitude of componential displacement $u(r,f)$ in the vicinity of the first resonant mode

With the proposed method it is possible to perform a

dependence analysis of the amplitude of componential displacement on the ratio of the inner and outer radius, and in this way to predict the behavior of the cylinder of known piezoceramic characteristics at certain frequencies prior to production. Based on this, it is possible to optimize the dimensions of the piezoceramic cylinder (Fig. 4).

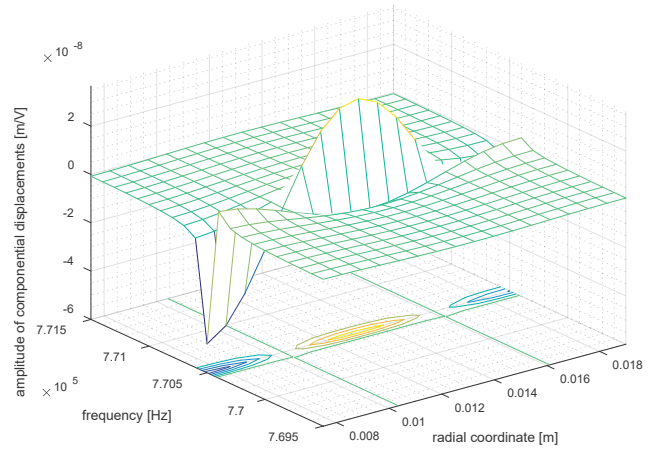


Fig. 3. Amplitude of componential displacement $u(r,f)$ in the vicinity of the second resonant mode

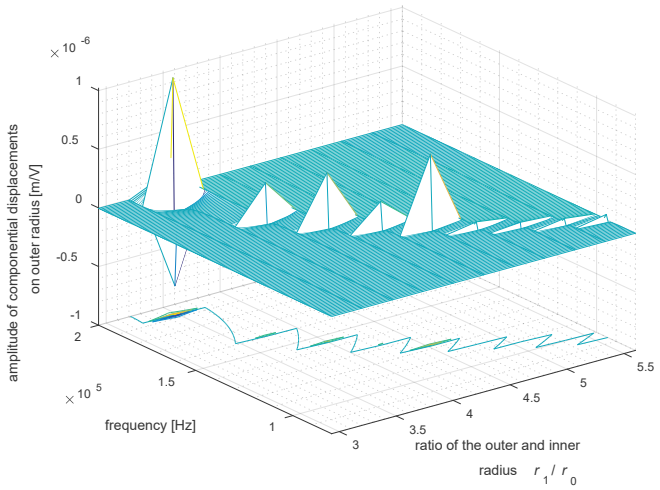


Fig. 4. Theoretical relationship between the amplitude of componential displacement on outer radius $u(r_1,f)$ and the ratio of the outer and inner radius r_1/r_0 for first resonant mode.

Fig. 4 shows that the optimal ratio of the external and internal radius is r_1/r_0=3.467, and the maximum amplitude of componential displacement are obtained for the external radius r_0=26 mm, for the given inner radius r_0=7.5 mm. In this case, the maximum amplitude of the componential displacement on the outer radius is 1.02 μm/V at the frequency of 193 kHz.

Fig. 4 also shows the influence of cylinder dimensions on the first resonant mode. The higher the radial dimensions of the cylinder, the first resonant mode slides to lower frequencies.

IV. Conclusion

In addition to the existence of the unknown coefficients A, B, C, and C_1 in (21) and (22), algebraic equations also include Bessel's functions of fractional index and their integrals, which are expressed by Lamé's functions. However, calculation using Matlab software significantly extenuate this problem.

This paper presents a method of studying elastic wave propagation in circular-ring piezoceramic cylinder with a radial type of polarization. When electric voltage is applied to the observed cylinder, two resonant modes occur at frequencies up to 800 kHz, at the following frequencies: 107.1 kHz and 770.5 kHz. In the second mode, the cylinder acts as a full-wave resonator (Fig. 3). It is also justified movement of radial resonant modes to lower frequencies in the case of an increase in the radial dimensions of the cylinder (Fig. 4).

Measurement of displacement of the cylinder will be the subject of further research.

Acknowledgment

The research presented in this paper is financed by the Ministry of Education, Science and Technological Development of the Republic of Serbia under the project TR33035.

References

[1] T. Stevenson, D. Martin, P. Cowin, A. Blumfield, A. Bell, T. Comyn, P. Weaver, "Piezoelectric materials for high temperature transducers and actuators", *Journal of Materials Science: Materials in Electronics*, vol. 26, pp. 9256-9267, 2015.

[2] H. Irschik, "A review on static and dynamic shape control of structures by piezoelectric actuation", *Eng. Struct.*, vol 24, no. 1, pp. 5-11, 2002.

[3] P. B. Petrović, V. B. Pavlović, B. Vlahović, V. Mijailović, "A high-sensitive current-mode pressure/force detector based on piezoelectric polymer PVDF", *Sensors and Actuators A: Physical*, vol. 276, pp. 165-175, 2018.

[4] D. D. Ebenezer, P. Abraham, "Eigenfunction analysis of radially polarized piezoelectric cylindrical shells of finite length", *Journal of the Acoustical Society of America*, vol. 102, pp. 1549-1558, 1997.

[5] D. Mančić, M. Radmanović, K. Hedrih, Lj. Perić, "2-D model tankog piezokeramičkog diska sa debljinskom polarizacijom", (in serbian), Proc. XLV Konferencija za ETRAN, Bukovička banja, Serbia, pp. 318-321, 4-7. Jun, 2001.

[6] Lj. Perić, *Coupled Tensors of Piezoelectric Materials State and Applications*, Le Locle, Switzerland, 2005.

[7] D. Rašković, *Teorija elastičnosti (in serbian)*, Beograd: Naučna knjiga, 1985.

[8] *Five piezoelectric ceramics*, Bulletin 66011/F, Vernitron Ltd., 1976.

Improving the Production Efficiency by Using the InfinityQS - a Real-time SPC Software

Miljana Milić, Zoran Milić, and Alex Crittenden

***Abstract*—Statistical process control (SPC) is a method of production process quality control which employs statistical calculations to monitor the manufacturing process, and keep it under control. InfinityQS is a software solution which enables implementation of a real-time SPC. The idea is to collect a real-time data from sensors and statistically process it in order to obtain a real-time image of the process stability and quality. This software is specialized in this field allowing for and easy data connectivity, and a real-time reporting/charting. Among numerous features, the most useful are basic SPC control charts. Control limits, specification limits, target values, mean values, and process capability indicators are measured/calculated and displayed on the control chart. This allows for a very effective visual recognition of situations where process is out of control, and immediate detection and action can be initiated to maintain the quality and minimize the scrap. The software also facilitates other actions which will eliminate special i.e. systematic causes of variations in parameters of the process, and make it stable and predictive, i.e. under control.**

***Index Terms*—Statistical process control; reliability; sensor networks; production lines.**

I. Introduction

Improving the quality of a product is the deciding aspect of any production or service, and it leads to enhancement, growth, and success in business and consequently better competitiveness [1]. The quality of a product is one of the most important factors that a consumer (regardless of whether it is an individua, an industry or etc.) takes into consideration when choosing among many competing products and services. The return on investment is more certain when high quality of the product is considered as the targeting production parameter in any planning of business [1].

To quantify the amount of quality of a certain product, one has to answer the following questions, and determine the level of corresponding issues:

- Will the product perform specific functions and how well it performs them? - Product performance,
- How often does the product fail when operating under a stated operating condition? – Product reliability,
- How long is the service life of the product? – Product durability,
- How difficult and financially affordable are maintenance and repair procedures for the product? – Product serviceability,
- How attractive is the visual appeal of the product? Product aesthetics,
- What are the features added to the basic performance of the product? – Product added features,
- What is the past reputation of the company concerning quality of their products? – Perceived quality of the product
- How exactly does the manufactured product parts meet the customer's/designer's requirements? – Product's conformance to Standards [2].

In order to fulfil the quality standards of the product or a production [2], [3] and achieve high reputation among customers, it is highly recomended to implement a specific control over the production process on the plant floor. Statistical process control (SPC) is defined as a group of software tools that use statistical measures to maintain the stability of the product manufacturing process, improve its capabilities, and minimize process variabilities [4].

The implementation of the SPC is highly corelated and dependent on the parameters of the manufacturing process, obtained i.e. measured from sensor networks. Sensors for this purpose represent sensing and measuring devices which can communicate through the network and whose activities can be controlled, since these nodes may have limited capacities [5]. Process measurements are obtained in real-time during manufacturing. If the right kinds of data are being collected from the right points of the production lines, it is easier to determine the cause of quality issues [6]. Sensing devices should be positioned in a way to enable maximal information extraction from the field. Collected data are later analysed and often support decision-making at the higher-level of management [5].

This paper describes some basic features and benefits of a particular SPC environment implementation. The InfinityQS represents a leading SPC solutions applied by a number of leading corporations. Some basic definitions and equations from the theory of the SPC will be given next. Then, some possibilities and features of the InfinityQS will be listed. A practical case study implementation of the InfinityQS for the particular production line will be shown at the end, as well as some concluding remarks.

Miljana Milić is with the Faculty of Electronic Engineering, University of Niš, Aleksandra Medvedeva 14, 18000 Niš, Serbia (e-mail: miljana.milic@elfak.ni.ac.rs).

Zoran Milić, bulevar Nikole Tesle 21, 18000 Niš, Serbia (e-mail: zoran.m.milic@gmail.com).

Alex Crittenden, 6133 Timberwood Lane, Texarkana, AR 71854, USA, (e-mail: zacrittenden@gmail.com).

II. About Statistical Process Control

There is a very limited amount of variations in the process that mass-production can tolerate. When variations do not exceed certain limits, it is sad that the manufacturing process is under control in each of its phases. In general, there are five groups of causes of unwanted variations that can be identified in the process. They are: raw materials, equipment, human actions, environment conditions, and methodology [7]. Variations are categorized in two groups:

- usual – tolerable variations, which are normal in the mass-production and have random nature, but are predictable, and
- specific – unwanted variation, whose causes are beyond the usual random changes of the process parameters, and cannot be predicted.

The key task of the SPC is to determine the cause of the specific process variations, by monitoring all relevant process parameters in time. Statistically predictable and tolerable variation do not require any actions, but for processes with some specific causes of variations, some actions need to be conducted to diminish them and they should be defined with a production Control plan [8].

With a SPC, process measurements' data obtained during manufacturing is then plotted in real-time on a graph referred to as a Control chart. Each process as well as the chart has two types of pre-defined limits: control limits are determined by the capability of the process (Voice of a process – VOP), and specification limits that are defined by the client (Voice of a customer – VOC). The parameter distribution as well as the corresponding limits are shown in Fig. 1. It should be emphasized that the VOC and the VOP are independent.

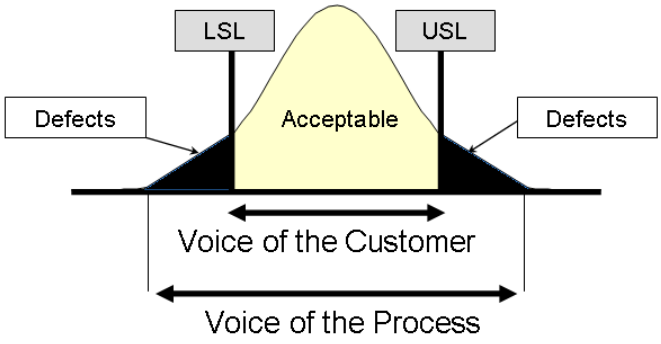


Fig. 1. Types of process limits.

When a process forms a stable distribution over time it is said to be under control. These distributions express the VOP. Process capability on the other hand expresses the goodness of a process, and is strongly related to customer specifications. These specification ranges i.e. the tolerances express the VOC here. This is illustrated in Fig. 2 [9].

Besides the usual statistical measures of the statistical process, such as the mean μ of the parameter and its standard deviation σ, some additional indicators need to be defined in order to implement SPC. They are: Cp - Process Capability, Cpk - Process Capability Index, Pp - Process Performance, and Ppk - Process Performance Index. Each of them will be defined and briefly explained next.

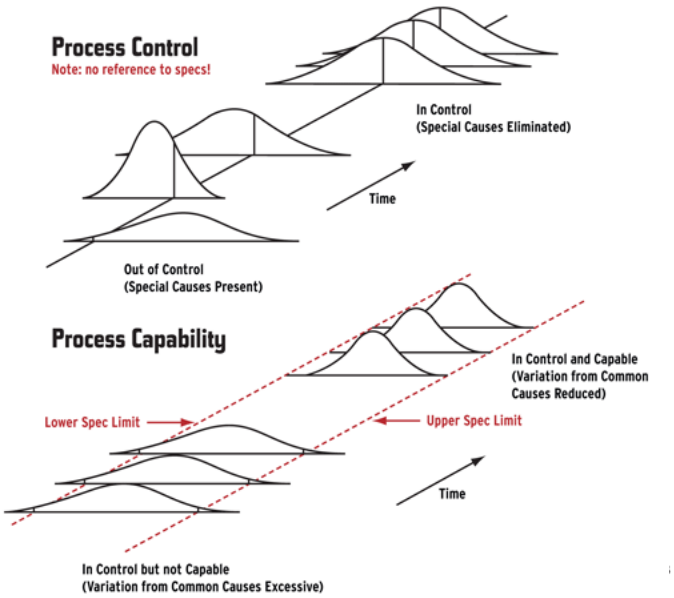


Fig. 2. Process control versus process capability.

Process capability can be calculated using as in (1):

$$Cp = \frac{USL\text{-}LSL}{6\sigma} \tag{1}$$

where USL and LSL represent upper and lower specification limits, respectively [1], while σ stands for the standard deviation of observed process parameter. It is the measure of how well the process meets customer specifications. At the other hand Process capability index Cpk, takes into consideration process centring and is calculated using (2):

$$Cpk = \min(Cpu, Cpl) = \min\left(\frac{USL\text{-}\mu}{3\sigma}, \frac{\mu\text{-}LSL}{3\sigma}\right). \tag{2}$$

In the simplified explanation, Cpk represents the one-sided Cp for the specification limit nearest to the process average μ [1]. Process Performance Pp, and Process Performance Index Ppk, can be calculated using the same equations (1), and (2), respectively. Differences between these two groups of measures are as follows. Process Capability indices Cp, and Cpk are calculated on samples; they are short term indices, and describe how well the process will perform in the future. At the other hand, Process performance indices Pp, and Ppk are calculated on the population; they are long term indices, and define how well the process has performed in the past. If the process is in statistical control, Cpk and Ppk are essentially the same [10].

Control charts are the basic graphical representation of the process quality that enable displaying and predicting of the process performances. By using them one can simply identify the existence of special causes of variations. With an SPC, control charts are used to provide information about weather a process is in statistical control or out of it, which is

determined by the nature of the process variations, as stated earlier in the text. One example of a control chart is shown in Fig. 3 [11]. The x axis represents the time, while the y axis represents the observed parameter values. Components of the control chart are the following lines: USL and LSL, target value of the parameter, average value, upper (+3σ) and lower (-3σ) control limits. Some areas of the control chart are also important to monitor. They are: a region of the expected variations - placed between upper and lower control limits, and a region of the unwanted variations – placed out of the control limits. Usually, the statistical control limits are set to +/- 3σ. In such cases, 99.73% of the charted variable fall within those limits. When a point appears outside of the control limits of the control chart, this should be alarmed, and an operator should perform actions defined by the Control plan or the plan of the reactions.

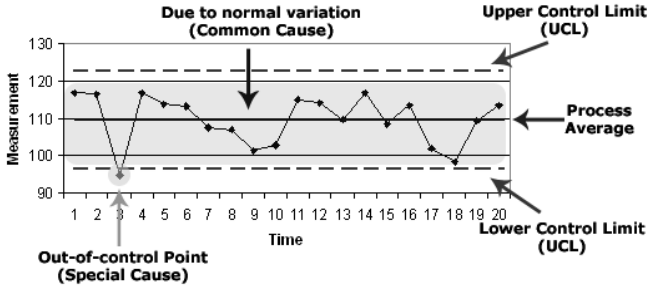


Fig. 3. The control chart.

In order to keep the process under control it is necessary to fulfil the following: the mean of the observed parameter should be as close as possible to the target value, and the Ppk should be greater than 1.3.

One implementation of a SPC using the InfinityQS solution for a simple example of the manufacturing process will be given next.

III. Implementing the InfinityQS on a Simple Production Process

ProFicient is the most important application of the InfinityQS solution that provides a customizable data-collection interface that supports multiple types of data input, manual and automatic [12]. Different charts are available for detailed insight into both real-time data, and historical trends of the process: control charts, Pareto charts, Cpk reports, box & whisker plots, capability analysis, scatter plots, trend charts, SPC monitor, etc. It enables detection of the cause codes, and corrective action codes, to quickly identify and prioritize the potential issues. There are also some very useful traceability features, such as lot genealogy, descriptors and serial numbers.

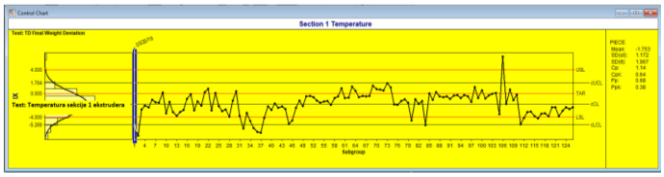


Fig. 4. The InfinityQS Xbar and Range control chart.

A case study of the simplified example of a part of the manufacturing process will be given next. The studied production system part is a line for extrusion of plastic. The extrusion is often part of the mass-production process. The plastic material is melted with the application of heat and extruded through a die into a desired shape. A cylindrical rotating screw is placed inside the barrel which forces out molten plastic material through the die. The extruded material takes shape according to the cross-section of the die. There are five important process parameters to be monitored during the extrusion process using a network of sensors: melting temperature of plastic, speed of the screw, the extrusion pressure, shape of the material that exits the die, and the cooling mechanism.

The Xbar and range control charts for the width of the extruded material is shown in Fig. 4. The Xbar gives the real-time histogram of the material width variations, while the Range control chart displays width variation trend with the respect to USL and LSL, and a target value.

A process under control for the described extrusion of plastic, monitored with the InfinityQS control chart is shown in Fig. 4. The parameter that is observed is the temperature of the melting plastic.

If the process exceeds some of the limits, the chart alarms it with the change of the color, and turns red. This is shown in Fig. 5. In this figure, an alarm for high temperature of the plastic is activated.

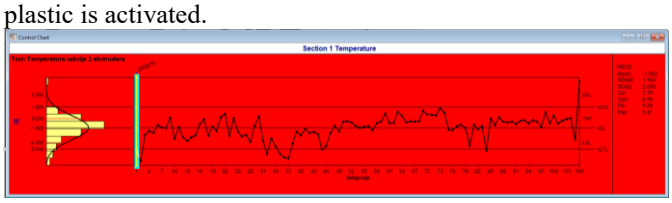


Fig. 5. Alarming an issue with the InfinityQS – process is out of control (high melting temperature of plastic).

In this case, the machine operator will easily notice the color change of the control chart at the machine Control panel. The corresponding extrusion section of the plastic should then be discarded as scrap, or will be properly marked for later disposal.

Also, the operator should then inform the subject matter expert about the event, who will analyze the occurrence of the event and perform steps in order to eliminate its causes.

Control charts are well known in mass production. But the obtained data are either manually entered, or the analysis are being performed on the historical data. The described software solution enables the real-time display of the quality parameters for the particular manufacturing process. With it, an operator at the machine (instead of a specialist) can easily and timely detect different quality issues, and the situations when the process is out of specified tolerances and control. The timed reaction of the operator significantly improves the process quality and reduces scrap.

IV. Conclusion

In this paper, one highly reputable, commercial SPC solution have been presented. It is greatly recemented in the mass-production for implementing the production quality

standards on the plant floor. The most important benefits of the ProFicient as the most important application of the InfinityQS SPC solutions are: significant reduction of the process variability and scrap, significant reduction of the production expenditures, scientific improvement of the process productivity, instant alarming and reactions to unwanted process variations, enabling the corresponding real-time decisions at the plant floor. The software also enables improved data collection from the network sensors, their analysis, processing, reports, storage, in order to compare qualities across multiple products, production lines and production sites. The complex analysis of data can be performed in real-time or using historical data within one chart or report.

ACKNOWLEDGMENT

This research is partially funded by The Ministry of Education and Science of Republic of Serbia under a contract no. TR32004.

REFERENCES

[1] D. C. Montgomery, *Introduction to Statistical Quality Control, 6*th ed. Arizona State University, John Wiley & Sons, Inc., 2009.

[2] *ISO/TS* 16949, June 1999.

[3] L. D Pop, and N. Elod, "Improving product quality by implementing ISO/TS 16949," *Procedia Technology*, *vol. 19,* pp.1004-1011, 2015.

[4] G. Škulj, R. Vrbič, P, Butala, and A. Sluga, "Statistical process control as a service: An industrial case study," *Procedia CIRP, 7,* pp.401-406, 2013.

[5] W. Li, and C. G. Cassandras, "Distributed cooperative coverage control of sensor networks," In *Proc. of the 44th IEEE Conference on Decision and Control,* pp. 2542-2547, December 2005.

[6] *, InfinityQS: SPC Software, accessed on https://www.infinityqs.com/

[7] B. Robinson, "Five Sources of Process Variation in Manufacturing," Management, 2017., retrieved from https://bizfluent.com/info-8505404-five-sources-process-variation-manufacturing.html

[8] R. Hills, "Statistical Process Control Basic Control Charts," online lecture retrieved from https://www.youtube.com/watch?v=WdqSm0DiYtY

[9] D. V. Neubauer, "Understanding Process Capability," ASTM Standardization News, June 2011, retrieved from https://www.astm.org/standardization-news/?q=data-points/understanding-process-capability-mj11.html

[10] B. McNeese, "Cpk vs Ppk: Who Wins?," Process capability reviews, May 2014, retrieved from https://www.spcforexcel.com/knowledge/process-capability/cpk-vs-ppk-who-wins.

[11] D. Miyake, "Balanced Scorecard Measurement + Control Charting Theory," blog on Strategic Impact, 2010, retrieved from https://www.ascendantsmg.com/blog/index.cfm/2010/11.

[12] *, ProFicient: On-Premises Quality Management and SPC Software, retrieved from https://www.youtube.com/watch?v=XPKHIDEpWNk.

Sensor node architecture for network control applications

Ivan Popović, Aleksandar Rakić, *Member, IEEE,*
Wenjun Zhang, Minrui Fei, Chen Peng and Dajun Du

***Abstract* – A novel architecture for sensor nodes is presented for specific application in networked control systems, where basic sensing and data acquisition functionalities are tied to several other and equally important functionalities regarding time-related services and middleware support for data transfer over the network. The main benefits of the proposed service-oriented architecture can be found in the domain of simplified design of the sensor node as well as its integration within the networked control systems.**

Key words: Network control systems; sensor node architecture; time-aware processing; service-oriented

I. Introduction

Growing interest in applying Internet of Thing (IoT) technologies in the field of Industry 4.0, smart cities, smart grid and connected infrastructures is derived from intensive deployment of Wireless Sensor Networks (WSN) in last decades. The extensive development of WSN is supported through the technology advances in the area of computer networking and communication. More generally, these advances enable integration of various end-devices in complex distributed systems and applications. One class of distributed systems are Network Control Systems (NCS), where process control loop is closed through the communication network. Regarding the design of NCS there are several, system operation and implementation related challenges that give rise to many important research topics. NCS architecture, process control algorithms, model of system components' implementation, information security, communication protocols and data formats, reliability of system operation, resilience to network communication problems, fault detection and monitoring, are some of the related topics studied in depth [1-5]. However, implementation of NCS sensing functionality according to the unified service-oriented model was not addressed in particular.

Adoption of service-oriented architecture (SOA) [6-9] and agents [10-14] in the industrial control system design is recognized as a promising opportunity to integrate the network capable heterogeneous devices and systems. The network environment enables dislocation of the key control system elements, i.e. sensors, actuators and control algorithms, to the separate network nodes. Additionally, smart transducer concept, defined through the IEEE 1451 family of standards, provides solution for service-oriented implementation of sensor and actuator components.

The family of standards is imposed in the form of upgrade of the basic IEEE 1451.0 services and functionalities [15], providing support for SOA concept. SOA is implemented using the Web service technology, as it uses SOAP/XML message format and standard protocols such as HTTP. Although SOA can be implemented in a different, application specific way [16-20], SOAP/XML Web services are preferred implementation solution as it does not conflict with firewalls and HTTP proxies. In [21], Web services are defined to reflect the IEEE 1451.0 smart transducer operations leading to concept known as Smart Transducer Web Services (STWS).

Although IEEE 1451 together with service-oriented architectural style defines framework for the implementation of basic sensing functionality, we are looking for other top-level architectural assets found in NCS operation.

Toward the goal design engineers need to understand and analyze the typical application properties given in the form of the context and context-aware behavior related to the NCS execution. An understanding of how context can be used will help the design engineers to agree-on what context-aware behavior is to be supported at the different levels of the system architecture. In the past various categorizations of the common application context properties were proposed. Nevertheless, the time property is found to be one of the most important application contexts, resulting in the time-aware model of the information processing. From the NCS point of view, time-awareness or time-sensitivity defines the critical aspect for the reliable execution of the data acquisition, processing, communication and control operation. Regarding the operation of sensors and actuators, time property is related to the continuous sampling functionality, while the time-aware controller operation is addressed in the domain of the control theory.

On the other hand, time related communication properties, introduced in the networked environment, such as varying network latency, packet losses are influencing the overall NCS operation. Therefore, special attention needs to be paid

Ivan Popović is associate professor at School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail: popovici@etf.rs).

Aleksandar Rakić is associate professor at School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail: rakic@etf.rs).

Wenjun Zhang is with School of Communication and Information Engineering, Shanghai University, 200444, China (e-mail: wjzhang@shu.edu.cn).

Minrui Fei is with Shanghai Key Laboratory of Power Station Automation Technology, School of Mechatronic Eng. and Automation, Shanghai University, 200444, China (e-mail: mrfei@staff.shu.edu.cn).

Chen Peng is with Shanghai Key Laboratory of Power Station Automation Technology, School of Mechatronic Eng. and Automation, Shanghai University, 200444, China (e-mail: c.peng@shu.edu.cn).

Dajun Du is with Shanghai Key Laboratory of Power Station Automation Technology, School of Mechatronic Eng. and Automation, Shanghai University, 200444, China (e-mail: ddj@shu.edu.cn).

to the design of the middleware as the middle-layer in the software architecture that enables communication and data management in the networked environment. The middleware provides common programming abstraction and infrastructure for the integration of different heterogeneous computing and communication devices. Since in the case of the NCS all system components share the common temporal property, as a model for the middleware implementation, we selected the time-aware approach. As a result, we take on the service-oriented concept as a uniform framework for the implementation of the process control and transducer nodes in the form of a group of network accessible services. On the other hand, the automated data transport is managed by a collection of service agents as configurable middleware components as shown in [22,23]. Thus, sensor node operation integrates basic sampling functionality, but also the group of middleware-related and time-related services. Therefore, these properties are supported through an introduced top-level component-based architectural style.

The immediate benefits of the proposed architecture are:

- layered hierarchy enables seamless integration of system application-layer components, service collaboration, implementation of security and discovery procedures and physical migration of services by the introduced service access layer,
- regarding data access model, architecture supports both passive sensor operation, where sensor node acts a data server, and active sensor operation, where sensor node acts as a configurable data publisher,
- the architecture itself it agnostic to physical deployment of system components, network topology and utilized application/transport layer communication protocols, and specific data acquisition and processing requirements.

The following sections of the paper give the overview of the corresponding sensor node architecture appropriate for network control applications and details regarding the sensor node operation, respectively.

II. Sensor Node Architecture

The proposed sensor node architecture is presented on Fig. 1. The architecture supports identified, NCS-specific, and traditional sensor side operations in the form of network accessible services, as configurable system components. Sensor node architecture assumes that each physical network node can contain single or a group of system components.

At the application (e.g. process) layer, standard sensor operations regarding A/D conversion, sampling control, raw sensor data processing, cross-correction, data aggregation, etc. are supported. These operations are specific to particular sensor type, its operation principle, application-level requirements and hardware platform. On the other hand, application-related processing regarding data protection, communication security, data storage, sensor node deployment and discovery, system integration, etc., is given in the form of collaborative execution utilizing services from other architectural layers. Service access is governed through service manager component at the service access layer.

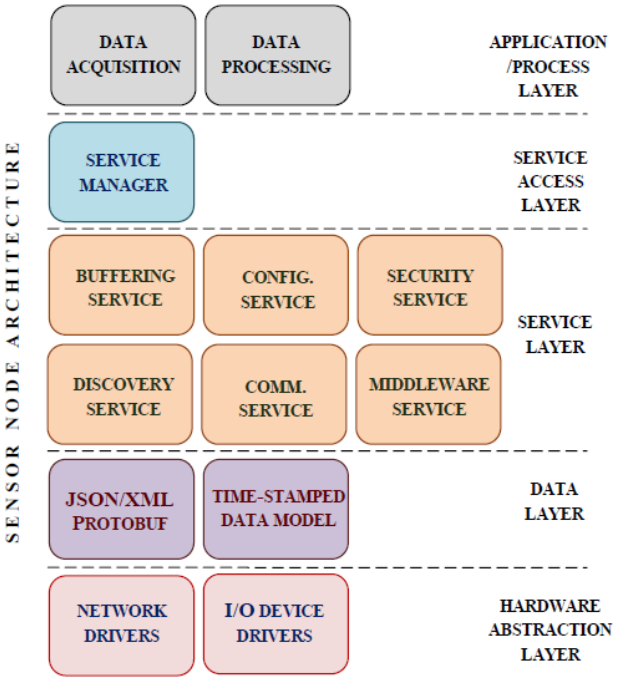


Fig. 1. Sensor node architecture for NCS

This component enables unified data-agnostic service access model, support for local or remote inter-node data communication and access control. It acts as a dispatcher of service request initiated from local application processes and local or remote services. The common system services and the support for different data representations, including time-stamped data model is given at service and data layers. NCS specific services are related to data buffering, configuration of individual services and sensor integration. The access to platform-specific or board-specific hardware resources, regarding network communication, system clock and RTC, data storage and I/O operations, is provided through dedicated device drivers found at hardware abstraction layer.

The component view of sensor node architecture is presented in Fig. 2. Service based implementation model, support for the configuration of network accessible components and channel-based data access model are adopted from the underlying IEEE 1451 group of standards. Inter-node data communication within the service-oriented NCS is performed over standard communication client-server architectural model. Different connectivity options and communication standards are available, including Smart transducer web service (STWS) interface, implemented as an additional layer at the top of the IEEE 1451.0 stack. From the information point of view, communication with passive system components is performed over two different data paths supporting data exchange (IDP) and service configuration (CDP). The information transferred over IDP is given according to time-referenced data model. Together with inter-node time synchronization, this data model enables time-aware execution of sensor services. The IDP data flow is managed by service agents (SAs), unique middleware components capable for inter-channel data communication [22-24].

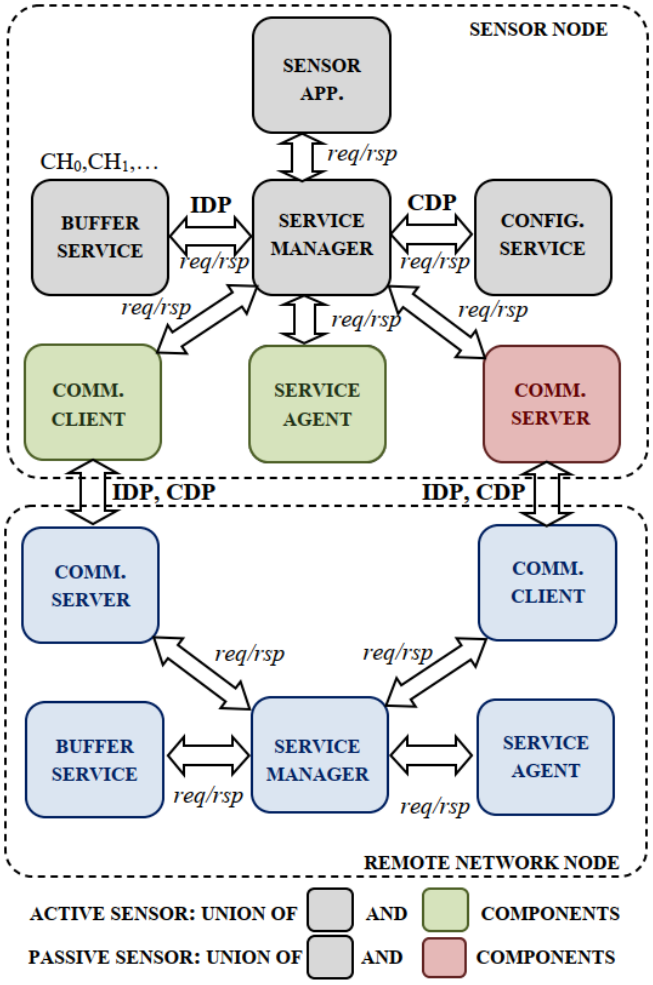


Fig 2. Different sensor node configurations from component-oriented view: (a) passive configuration and (b) active sensor node configuration

In the case of sensing functionality, channel (CH_i) represents the logical abstraction of physical sensor's buffering data structure, e.g. IDP communication endpoint [25]. While the buffer write operations are related to the storing of time-stamped data samples, the buffer read operations are associated to the execution of remote or local service request initiated by dedicated SA. The details of SA operation are explained in [24].

Since the service agent's operation is independent of its physical location, differently customized sensor node architectures are available, as shown in Fig. 2. Passive sensor node operation resembles pull data-processing model, since the service agent, accessing the sensor data, is located at remote network node (Fig 2a). On the other hand, active sensor operation is also supported in a push-like data access model, where transport of sensor data is triggered and managed by dedicated service agent, located at the sensor node (Fig. 2b). The model for inter-component data exchange is given in the form of request-response messaging pattern (*req/rsp*) through the service manager component, which acts as a service proxy and an access and security control gateway [26].

III. SENSOR NODE OPERATION

The detailed thread-level sequence diagram of all system components relevant for passive sensor node operation is given in Fig. 3. Practically all threads, according to the event-based processing model, are waiting for the notifications on particular events, which will trigger thread specific actions.

During the standard sensor operation, data acquisition and processing triggers local service request, resulting in buffer write operation *buff_wr()* according to the time-stamped data model. For the sake of simplicity, execution of time-related services is omitted from the presented diagram, although is part of data processing operations at the application layer.

As a part of NCS control loop operation, the sensor data access over the IDP is also illustrated in Fig. 3.

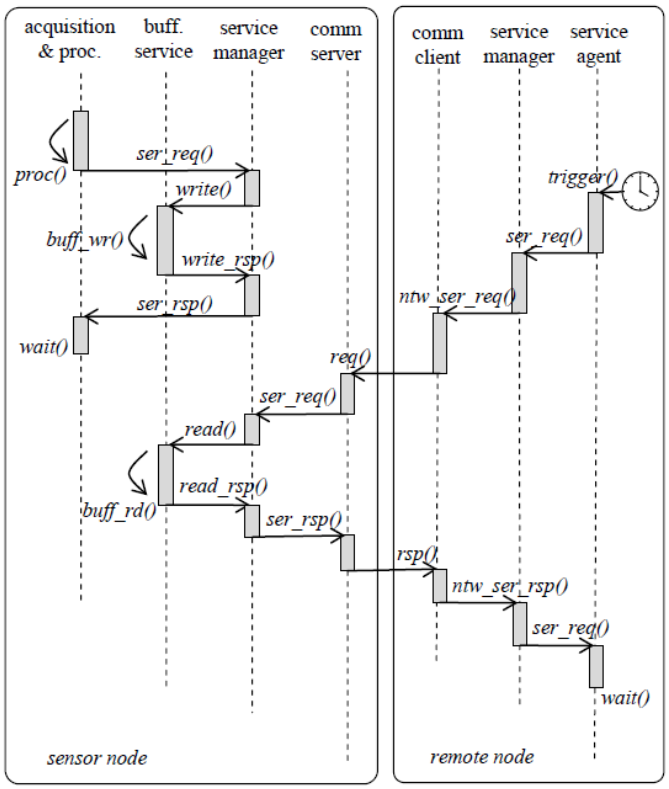


Fig.3 Diagram of the standard passive sensor node execution sequence during NCS operation

After the timer event notification triggers SA, it performs the sequence of read and write operations directed toward particular service provider and consumer endpoints specified in IDP configuration. Since the passive sensor node architecture, SA component is deployed at remote network node. Endpoint read operation, invoked by SA, in the form of local service request, e.g. *ntw_ser_req()* over the service manager effectively invokes buffer *buffer_rd()* operation at sensor node side. This service request is processed by sensor thread providing the read service response with the previously undelivered time-stamped sensor data from local sensor buffer. The read service response is passed by sensor threads to SA over *ntw_ser_rsp()*. Upon the reception of sensor data, SA continues with local or remote write operations to transfer the data to the destination endpoint buffer. The thread

execution sequence associated to write operations is not included in the presented diagrams, since is not related to the group of sensor side operations. One should have in mind that the sensor operation analyzed in the previous discussion assumes the thread-level actions of individual components is driven by the regular sequence of the events.

IV. CONCLUSION

In this paper, a novel architecture of the NCS-suited sensor node is presented. While maintaining the benefits of service-oriented architecture, introduction of the service access layer over the pool of services bring additional flexibility for the NCS integration regarding implementation of the security and discovery mechanisms. Additionally, network transfer of the sensed data is supported both in passive mode, where sensor node effectively acts as a classical data server, and in active mode, where sensor node acts as a configurable data publisher in publish-subscribe data exchange model. The architecture is agnostic to utilized network topology and protocols, as well as specific data acquisition and processing requirements. Since these functionalities are placed in different architectural layers, their implementation is decoupled from service management.

Development of the complementary architectures for the actuator nodes and control-performing nodes in NCS, as well as investigation of the optimal data transfer scheduling within the SOA-based NCS in regular and security threatened operation are planned as a part of the future work.

ACKNOWLEDGMENT

The authors gratefully acknowledge financial support from the Ministry of Education, Science and Technological Development of Republic of Serbia under contracts no. TR-32043, TR-33024, and the project of bilateral cooperation between the People's Republic of China and Republic of Serbia "Distributed Network Control and Its Application in Smart Grid".

REFERENCES

[1] Z. Lixian, G. Huijun, O. Kaynak, Network-Induced Constraints in Networked Control Systems—A Survey, IEEE Transactions on Industrial Informatics, Vol. 9 (2013) pp. 403-416.

[2] A. Doulamis, N. Matsatsinis, Visual understanding industrial workflows under uncertainty on distributed service oriented architectures, Future Generation Computer Systems, Vol. 28 (2012) pp. 605–617.

[3] R.A. Gupta, C. Mo-Yuen, Networked Control System: Overview and Research Trends, IEEE Transactions on Industrial Electronics, Vol. 57 (2010) pp. 2527-2535.

[4] J. Baillieul, P.J. Antsaklis, Control and Communication Challenges in Networked Real-Time Systems, in: Proceedings of the IEEE, Vol. 95, 2007, pp. 9-28.

[5] O. Esquivel-Flores, H. Benítez-Pérez, J. Ortega-Arjona, Issues on Communication Network Control System Based Upon Scheduling Strategy Using Numerical Simulations, in: D.M. Andriychuk (Ed.) Numerical Simulation - From Theory to Industry, InTech 2012, pp. 49-66.

[6] L.D. Xu, W. He, S. Li, Internet of things in industries: A survey, IEEE Transactions on Industrial Informatics, Vol. 10 (2014) pp. 2233-2243.

[7] N. Komoda, Service Oriented Architecture (SOA) in Industrial Systems, in: IEEE Int. Conf. on Industrial Informatics, 2006, pp. 1-5.

[8] T. Cucinotta, A. Mancina, G.F. Anastasi, G. Lipari, L. Mangeruca, R. Checcozzo, F. Rusina, A Real-Time Service-Oriented Architecture for Industrial Automation, IEEE Transactions on Industrial Informatics, Vol. 5 (2009) pp. 267-277.

[9] S. Karnouskos, A.W. Colombo, F. Jammes, J. Delsing, T. Bangemann, Towards an architecture for service-oriented process monitoring and control, in: IECON 2010 - 36th Annual Conference on IEEE Industrial Electronics Society, 2010, pp. 1385-1391.

[10] A.W. Colombo, S. Karnouskos, J.M. Mendes, P. Leitão, Industrial Agents in the Era of Service-Oriented Architectures and Cloud-Based Industrial Infrastructures, Industrial Agents, Emerging Applications of Software Agents in Industry, Elsevier Science 2015, pp. 67-87.

[11] T. Strasser, A. Zoitl, Distributed Real-Time Automation and Control - Reactive Control Layer for Industrial Agents, Industrial Agents, Emerging Applications of Software Agents in Industry, Elsevier Science 2015, pp. 89-107.

[12] R.S. Alonso, D.I. Tapia, J. Bajo, Ó. García, J.F.d. Paz, J.M. Corchado, Implementing a hardware-embedded reactive agents platform based on a service-oriented architecture over heterogeneous wireless sensor networks, Ad Hoc Networks, Vol. 11 (2013) pp. 151-166.

[13] K. Nagorny, A.W. Colombo, U. Schmidtmann, A service- and multi-agent-oriented manufacturing automation architecture: An IEC 62264 level 2 compliant implementation, Computers in Industry, Special Issue on Sustainable Interoperability: The Future of Internet Based Industrial Enterprises, Vol. 63 (2012) pp. 813-823.

[14] P. Leitao, V. Marik, P. Vrba, Past, Present, and Future of Industrial Agent Applications, in: IEEE Transactions on Industrial Informatics, Vol. 9, 2013, pp. 2360-2372.

[15] IEEE 1451.0, Standard for a Smart Transducer Interface for Sensors and Actuators — Common Functions, Communication Protocols, and Transducer Electronic Data Sheet (TEDS) Formats, IEEE Instrumentation and Measurement Society 2007, pp. 1-335.

[16] D. Wenbin, J. Peltola, V. Vyatkin, P. Cheng, Service-oriented distributed control software design for process automation systems, in: IEEE International Conference on Systems, Man and Cybernetics (SMC), San Diego, CA, 2014, pp. 3637 – 3642.

[17] F. Jammes, Real time device level Service-Oriented Architectures, in: IEEE International Symposium on Industrial Electronics (ISIE), Gdansk, Poland, 2011, pp. 1722-1726.

[18] G. Veiga, J.N. Pires, K. Nilsson, Experiments with service-oriented architectures for industrial robotic cells programming, Robotics and Computer-Integrated Manufacturing, Vol. 25 (2009) pp. 746–755.

[19] Siew Poh Lee, Lai Peng Chan, Eng Wah Lee, Web Services Implementation Methodology for SOA Application, in: IEEE Int. Conf. on Industrial Informatics, Singapore, 2006, pp. 335-340.

[20] Shangguang Wang, Yan Gong, Guangxiao Chen, Qibo Sun, Fangchun Yang, Service vulnerability scanning based on service-oriented architecture in Web service environments, Journal of Systems Architecture, Vol. 59 (2013) pp. 731-739.

[21] E. Y. Song, K. B. Lee, Smart Transducer Web Services Based on the IEEE 1451.0 Standard, in: IEEE Instrumentation and Measurement Technology Conference (I2MTC), Warsaw, 2007, pp. 1-6.

[22] I. Popović, A. Rakić, Architectural Approach to Cope with Network-Induced Problems in Network Control Systems Design, Journal of Electrical Engineering, vol. 69, No. 4, pp. 270-278, doi: 10.2478/jee-2018–0037, 2018, ISSN 1339-309X.

[23] A. Rakić, N. Bežanić, I. Popović, Novel Architecture for Networked Control Systems, in 2016 International Symposium on Industrial Electronics, Banja Luka, Bosnia and Herzegovina, Nov. 3–5, 2016, pp. 1-6. doi: 10.1109/INDEL.2016.7797806.

[24] I. T. Popović and A. Ž. Rakić, "The Fog-Based Framework for Design of Real-Time Control Systems in Internet of Things Environment," 2018 International Symposium on Industrial Electronics (INDEL), Banja Luka, Bosnia and Herzegovina, 2018, pp. 1-6. doi: 10.1109/INDEL.2018.8637639

[25] N. Bežanić, I. Popović, Service-oriented Implementation Model for Smart Transducers Network, Computer Standards & Interfaces, Elsevier, 38, C, pp. 78 - 83, ISSN: 0920-5489, doi: 10.1016/j.csi.2014.10.004, Feb. 2015.

[26] Y. Shen, M. Fei, D. Du, W. Zhang, S. Stanković, A. Rakić, "Cyber Security Against Denial of Service of Attacks on Load Frequency Control of Multi-Area Power Systems," in Int. Conf. on Intelligent Computing for Sustainable Energy and Environment, Nanjing, China, Sep. 22-24, 2017, pp. 439-449. doi:10.1007/978-981-10-6364-0_44

Exploring the limits of hardware/software co-design

Haris Turkmanović, Filip Mijušković, Ivan Popović

***Abstract* – Various software and hardware techniques have been developed to improve operational properties of an embedded system. In this paper we investigate the potential of hardware/software co-design in the domain of system execution performance. As a target application domain, we selected a class of multimedia applications characterize with extensive number of memory operations. In order to analyze the effectiveness of hardware/software co-design we introduced embedded system metric parameters to quantify system execution performance and its hardware complexity. The profiling of system performance was conducted for different hardware configurations build on FPGA platform while executing several characteristic control flow structures. The analysis has revealed the significant potential of hardware/software co-design in the domain of embedded system optimization.**

Key words: embedded system; performance; FPGA platform; co-design; optimization techniques, control flow

I. INTRODUCTION

Embedded system optimization domain is commonly related with the system execution performance, resource utilization, system energy and power consumption, cost and size reduction, etc. System operational properties including its execution performance and can be improved through the utilization of different software and hardware techniques. Some basic software approaches are: selection and efficient implementation of data processing algorithms, organization of data structures, code-level optimization, etc. On the other hand, hardware techniques are based on concurrent hardware processing, parallel computing, custom logic design, CPU core design, hierarchical organization of memory subsystem, etc. The effectiveness of the particular software-hardware optimization techniques depends on the specific target application domain of the embedded system. This target application domains can be classified into three main classes: multimedia applications, front-end telecommunication applications and network component applications [1]. Focus of our paper is in multimedia application domain with the diversity of image and video processing applications. The common attribute of this application domain is related to the processing of the huge amount of data, preferably organized as loop-based repetitive control flow structure accessing multidimensional array of data structures.

There are numeral research studies and scientific papers which evaluate different software and hardware optimization techniques based on embedded system execution performance optimization. One of the software optimization approach found in [2] analyzes the possibilities and outcomes related to the reduction of processing overhead derived from looped program execution. An additional research presented in [3] analyzes the influence of *inner for loop* and *no inner loop* construct on execution performance. The analysis was carried out in various programming languages and it was shown that in all used program languages *inner for-loops* have better execution performance than *no inner loops*. As a general remark it was given that two equivalent programs written in the same language and running on the same platform, have different performance depending on the coding style. The investigation given in [4] explores the effects of *loop unrolling* techniques on execution performance. Similar analysis found in [5] is investigating applicability and effectiveness of code optimization techniques based on function inline expansion methods. It has been shown that method can improve spatial locality of data access resulting in more efficient use of caching mechanism.

In addition to software-based approaches, different hardware techniques were investigated in the context of embedded system performance optimization. The most commonly investigated hardware design approach is related to a form of memory sub system organization. Especially, accessing data structures located in slower off-chip memories introduces significant CPU stall cycles, decreasing its execution performance. Research published in [6] analyses properties of hieratically organized memory subsystem and cache optimization techniques. Similar results presented in [7] quantify the effectiveness of caching mechanism in the multimedia application domain. As a general observation it was found that multimedia application has a lower data cache miss rate.

The aim of our research is to investigate the limits which can be achieved by a combination of software and hardware techniques. Consequently, already known techniques, which may influence the achievement of these limits, are considered in our analysis. On the other hand, in order to investigate the impact of various hardware configurations, an FPGA platform has been selected. This platform type provides the ability to create an arbitrary hardware and software environment for exploiting different optimization techniques. In particular, we analyze sequential and looped control flow structures with no

Haris Turkmanović, MSEE student and teaching assistant at School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail: haris.turkmanovic@gmail.com).

Filip Mijušković, MSEE student at School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail: filipmijuskovi7@gmail.com).

Ivan Popović is associate professor at School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail: popovici@etf.rs).

inner loop constructs, with and without inline function expansions. From the pool of hardware-related techniques, we analyze the effects of advanced CPU core features on execution performance, as pipelined execution and branch prediction, as well as the effects of different program and data memory organizations.

After this introductory part, the following section introduces the metrics parameters which quantify contribution of hardware and software techniques to the embedded systems execution performance. Section III presents realization details of the system hardware configurations and describe program control flow structures used in our analysis. The results of the applied performance optimization and as well, hardware requirements of utilized hardware and software techniques, with the associated discussions, are given in section IV. Concluding remarks and plans for future work are given in the final section.

II. System Design and performance metrics

This chapter introduces embedded system metric parameters in order to quantify effectiveness of different software and hardware techniques in the context of system execution performance and its complexity. Introduced parameters enable numerical comparison and characterization of different hardware/software configurations.

On one hand, the complexity of an overall embedded system given as a collection of data processing and data storage elements is counted through hardware related parameters. On the other hand, software related parameters quantify the performance properties of the program execution related to the control flow structure, properties of instruction set architecture, etc.

In the analysis presented in this paper, we selected the subset of system metrics parameters containing Hardware Complexity and Execution Performance parameters.

System Hardware Complexity (HC) is introduced to uniquely describe the complexity of each hardware configuration. The complexity of the hardware configuration *n* (described as a parameter HC_n) is given as a function of required number of logical elements (LE_n) and the number of occupied memory bits (M_n), in the form:

$$HC_n = f(LE_n, M_n)$$

The expression $HC_{n,m}$ is used to describe the Normalized Hardware Complexity of hardware configuration *n* in respect to the selected reference hardware configuration *m*.

Execution performance parameter given as T_n^i represents the number of system clock cycles required to execute the program flow *i* on the hardware platform *n*.

Similarly with normalized hardware complexity parameter, the symbol $T_{n,m}^i$ is used to express a normalized execution performance of a program flow *i* on a hardware platform *n* (given as T_n^i) in relation to T_m^i, e.g. $T_{n,m}^i = T_n^i / T_m^i$.

Normalized execution performance is used as a Measure of system effectiveness from the execution performance viewpoint. This quantitative measure enables ranking of comparable systems in the domain of applied optimization techniques. [8]

III. System overview

This section describes the hardware and software related details of an embedded system used in our analysis. Hardware configurations, realized on FPGA platform, are selected in order to investigate the influence of the processing unit and memory subsystem properties on execution performance [2][11]. Similarly, to examine the effects of software optimization techniques, we consider several control flow structures, as given below.

A. Hardware configurations

All presented hardware configurations are given as uniprocessor systems with the NiosII processor core [9]. Two soft-core CPU implementations, *NiosII/f* fast core and *NiosII/e* economy core, are considered as boundary choice, in terms of the most complex and the simplest core design, respectively. The NiosII fast core, in addition to features of the economy core, supports *pipelined program execution*, *dynamic branch prediction*, *instruction and data caching*, *tightly coupled memory interfaces* as well as enhanced instruction execution time.

The details of selected hardware configurations are summarized in Table 1. Configurations 3 and 4 are chosen in order to examine the effects of memory subsystem organization on execution performance, where configuration 3 supports L1 instruction caching while configuration 4 has L1 instruction and L1 data cache. Configuration 5 and 6 are built with fast On-Chip SRAM memory directly accessible by processor core through the dedicated tightly coupled interface, where configuration 5 utilizes this memory only for the instructions while configuration 6 utilizes memory for data and instruction access.

TABLE I
OVERVIEW OF HARDWARE CONFIGURATIONS

Configuration NO	Processor Core	Data Memory	Data Cache	Instruction Memory	Instruction Cache
1	e	*SDRAM*	*NO*	*SDRAM*	*NO*
2	f	*SDRAM*	*NO*	*SDRAM*	*NO*
3	f	*SDRAM*	*NO*	*SDRAM*	*YES*
4	f	*SDRAM*	*YES*	*SDRAM*	*YES*
5	f	*SDRAM*	*NO*	*On-chip RAM*	*NO*
6	f	*On-chip RAM*	*NO*	*On-chip RAM*	*NO*

All hardware configurations presented in Table 1. are built on *EP4CE115F29C7* FPGA chip from *Altera Cyclone IV* device family [10][11]. As a hardware environment for our analysis we selected *Terasic DE2-115* development board [12] with external SDRAM memory [13] available through *Altera SDRAM controller* [14]. The *Cyclone EP4CE115* device equipped on the *DE2-115* features 114480 logical elements (LEs) and up to 3.9 Mbits of RAM [11].

The hardware requirements for each of the hardware configurations in the context of utilized LE and memory bits, quantified through calculated Hardware Complexity, parameter is given in Table 2.

TABLE II
HARDWARE CONFIGURATIONS IMPLEMENTATION DETAILS

Configuration No i	No Of Logic elements	No Of bits	HC_i
1	3332 (*2.91%*)	11264 (*0.28%*)	0.61
2	4536 (*3.96%*)	11776 (*0.3%*)	0.82
3	4642 (*4.05%*)	16320 (*0.41%*)	0.84
4	5165(*4.51%*)	51776 (*1.3%*)	0.95
5	4622(*4.04%*)	273920 (*6.88%*)	0.9
6	4833(*4.22%*)	1322496 (*33.22%*)	1.18

It should be noted that for profiling NiosII system, *Performance Counter* [11] unit is embedded in each hardware configuration. The unit is used to measure a number of clock cycles which is required for the execution of the program section defined through BEGIN/END statements [15].

B. Control flow structures

In order to analyze the influence of software optimization techniques on the execution performances of applications with intensive memory accesses [1], we selected the initial simple control flow structure 1 in the form of a sequence of memory write operations to successive memory locations. The similar structure, but with non-sequential memory access is given as control flow 2. Control flow structures 3, 4 and 5 are realized in the form of a repetitive control flow with sequential, non-sequential memory access, and with nested function call encapsulating sequential memory access, respectively. Characteristics of each control flow structures are listed in the Table 3.

TABLE III
PROPERTIES OF PROGRAM FLOW STRUCTURES

Control Flow Structure No	Code size	Memory access	Instruction memory access	Data memory access
1	*2KB*	700	500 (71%)	200 (29%)
2	*2.4KB*	800	600 (75%)	200 (25%)
3	*32B*	2200	1600 (72%)	600 (28%)
4	*32B*	2200	1600 (72%)	600 (28%)
5	*136B*	4600	3400 (74%) (1600+1800)[1]	1200 (26%)

Properties given in Table 3 are derived from assembly codes which are generated from corresponding control flow structures written in C programming language.

IV. HARDWARE/SOFTWARE CO-DESIGN ANALYSIS

The results obtained by measuring execution performance, after the application of different hardware and software configurations, are analyzed in this section. Some of the effects, at first glance, give unexpected performance since the spatial and temporal properties of memory access resulting in dynamic system behavior. The details of the analysis related to the hardware and software design issues are given in the following paragraphs.

A. Software related effects

System hardware configuration 1 is used to quantify the control flow structure effects on the execution performance. Since this hardware configuration does not include any of advanced hardware features, it is ideal found to quantify these effects. Obtained results are shown in Figure 1.

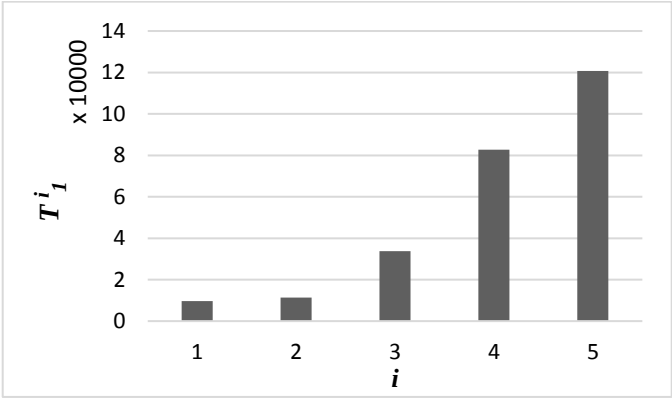


Figure 1 – Execution time of different control flow structures, measured on hardware configuration No.1

Based on the execution time properties shown in Fig. 1, it is noticeable that the control flow structure significantly affects the execution performance. In relation with control flow properties from Table 3, we can conclude that these effects are directly related to the total number of memory accesses. In other words, the program flow structure 1 achieves the best performance on a simple processor core because it requires the least memory accesses. However, control flow 1 exhibits low code density and poor maintainability since it has the highest number of code lines.

TABLE IV
RATIO OF EXECUTION PERFORMANCE WITHOUT ($T^i_{wo,4}$) AND WITH ($T^i_{w,4}$) CACHE FLUSHING MEASURED ON CONFIGURATION NO.4

Control flow structure	1	2	3	4	5
$T^i_{wo,4}/T^i_{w,4}$	2.66	5.52	1.20	1.66	1.37

The execution performance properties related to compulsory cache misses, found in initial code execution or after the context switching, are analyzed as well. The profiling of execution performance is performed with and without flushing cache content prior to code execution. As these effects are associated with cache-based systems, our analysis included hardware configuration 3 and 4 with L1 cache memory. The results obtained by the measurement program execution properties, for hardware configuration 4, are given in Table 4.

As noticeable from Table 4, these effects are the most observable for the control flow structure 2, since the spatial

[1] 1600 memory accesses are realized in main program while 1800 memory accesses are obtained in function

and temporal properties of memory access.

B. *Hardware related effects*

In order to analyze the CPU design related effects on execution performance, hardware configuration 1 and 2 are selected. Both configurations utilize SDRAM-only based design of the memory subsystem. The results shown in Figure 2 represent execution performance of defined control flow structures executed on hardware configuration 2. The results in Figure 2 are normalized with respect to execution performance of same control flow structures executed on hardware configuration 1.

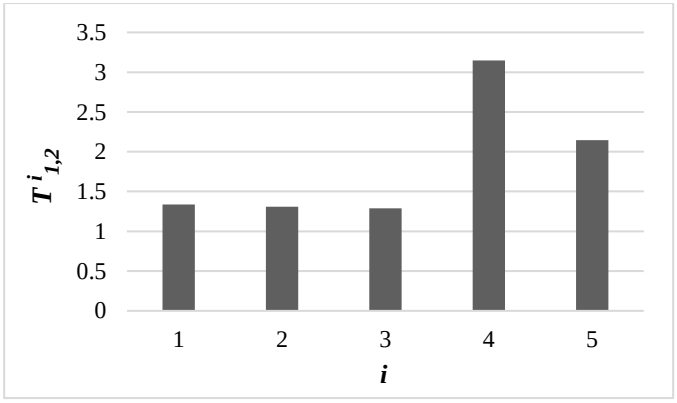


Figure 2 – Normalized execution performance for SDRAM-only hardware configurations, T_1^i/T_2^i

The results given on Fig. 2, illustrate the potential of performance-cost tradeoff in *NiosII* core design. As noticeable advanced CPU related features found in *NiosII/f* core bring execution performance improvements in all control flow structures. The improvement is most prominent for control flow 4 organized as repetitive structure with non-sequential memory access since *NiosII/f* processor core supports dynamic branch prediction and pipelined code execution. During results analysis one should have in mind that hardware configuration 2 requires a significantly higher number of hardware resources compared to configuration 1, as given in Table 2.

The following discussion analyzes the effects of memory subsystem organization on execution performance. First approach explores the benefits of hierarchically organized memory system which is exploited in hardware configuration 3 and 4. The second approach considers utilization of fast on-chip memories. These memories have low memory access time for storing code and data sections and they are used in hardware configuration 5 and 6. In order to quantify the contribution of each hardware configuration, configuration 2 is selected as the reference system configuration since it has the same core design as configurations 3, 4, 5 and 6.

Normalized execution performance, observed for hardware configuration 3 with L1 instruction cache, is presented in Figure 3.

From the Fig. 3 we can conclude that the instruction cache brings improvement in all control flows, especially in those given in the form of repetitive structures due to higher code spatial locality. The support for data caching, found in configuration 4, brings additional improvements in the execution performance of flows 3 and 5 because of higher spatial locality in consequence of sequential data memory access. Obtained results confirmed that assumption, since the obtained execution performance was improved about 9 times compared to the system configuration without instruction and data caches (see Table 6).

TABLE V
OBTAINED EXECUTION SPEEDUP AFTER ADDING DATA CACHE IN SYSTEM WITH INSTRUCTION CACHE

Control flow structure	1	2	3	4	5
$T^i_{3,4} = T^i_3/T^i_4$	0.89	**0.37**	2.40	1.68	1.93

During the analysis it was noticed that adding data cache in the system configuration with instruction cache, can lead to significant decrease of execution performance as can be viewed from Table 5. This effect is manifested in the case of control flow 2, where performance was decreased almost 3 times. This significant reduction in execution performance comes from the fact that this control flow structure has low temporal and spatial locality which increases data miss rate. One should have in mind that single data miss producing additional CPU stall cycles.

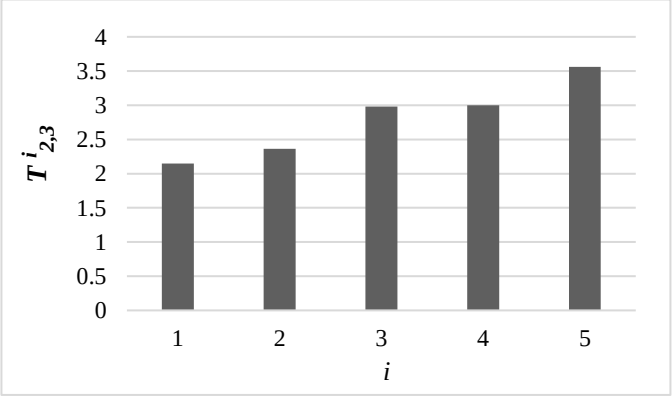


Figure 3 –Normalized execution performance of different control flow structures measured for hardware configuration No. 2 in respect to No. 3.

Hardware configuration 5 and 6 are used to explore the system performance limits since utilizing fast on-chip memories directly accessible through dedicated CPU tightly coupled interface. The use of fast memory for program code sections can significantly contribute to the execution performance, especially in the case of program flows 1 and 2. The performance improvement is expected since most memory accesses are related to instruction fetching. Although the execution performance is improved extensively (around 15 times), one should have in mind that hardware complexity of configuration 5 and 6 is significantly increased, as given in Table 2.

In comparison to configuration 3 with L1 instruction cache, configuration 5 introduces performance speedup, especially noticeable for control flows 1 and 2 with sequential structure of program flow.

Hardware Configuration 6 was introduced to examine the extreme limits of system performance, where both - data and instructions - are directly accessible from fast on-chip memory without misses. The results obtained in this analysis show that the individual performance increase is up to 60% for all control flow structures.

Obtained limits of applied hardware and software techniques and their combinations are summarized in Table 6.

TABLE VI
EXPLORED LIMITS OF HARDWARE/SOFTWARE CO-DESIGN IN PARTICULAR PROBLEM DOMAIN

No	Technique	Parameter	Max Speedup
1	Loop unrolling	T_5^3/T_5^1	9.54
2	Sequential memory access	T_4^2/T_4^1	2.61
3	Inline function	T_6^5/T_6^1	13.32
4	Hierarchical memory organization - caching	T_2^5/T_4^5	9.43
5	Core design	T_2^4/T_1^4	3.15
6	Directly coupled fast memory access	T_2^1/T_6^1	14.28
7	Inline functions and caching	T_1^5/T_4^1	32.43
8	Inline functions and directly coupled fast memory access	T_1^5/T_6^1	239.15
9	Inline functions and core design	T_1^5/T_2^1	16.81
10	Loop unrolling and caching	T_1^3/T_4^1	8.97
11	Loop unrolling and directly coupled fast memory access	T_1^3/T_6^1	66.98
12	Loop unrolling and core design	T_1^3/T_2^1	4.30

V. CONCLUSION

We analyzed the effectiveness of most commonly used hardware and software techniques in embedded system performance optimization. The limits of embedded system execution performance are revealed through the combination of applied techniques. Our analysis has shown that in the domain of software-based techniques, the significant advances can be achieved by applying adequate control flow structures. Based on the control flow properties, such as the number of memory accesses, spatial and temporal locality, the efficiency of particular software techniques is directly linked to the selected system hardware configuration, providing the background for hardware/software co-design. Furthermore, introduction of embedded system metrics parameters enables additional analysis of different complexity-performance tradeoffs in hardware/software co-design, as a part of future work.

VI. ACKNOWLEDGMENT

The authors gratefully acknowledge financial support from the Ministry of Education and Science, Government of the Republic of Serbia through the Project No. 32043: "Development and modeling of energy efficient, adaptive, multi-processor and multi sensor low power electronic systems".

VII. REFERENCES

[1] H. Falk, P. Marwedel, "*Source Code Optimization Techniques for Data Flow Dominated Embedded Software,*" Springer science business Media, New York 2004.

[2] Wikipedia, "Loop optimization techniques" available online on https://en.wikipedia.org/wiki/Loop_optimization Automation of Electronic Systems, Vol. 2, No. 4, October 1997.

[3] T.P. Adewumi, "Inner loop program construct: A faster way for program execution," Open Comput. Sci 2018.

[4] J.W. Davidson, J. Sanjay, "An Aggressive Approach to loop unrolling," Department of Computer Science, Thornton Hall, University of Virginia, Charlottesville, VA 22903 U.S.A.

[5] William Y. Chen, Pohua P. Chung, Thomas M. Conte, Member, IEEE, and Wen-mei W. Hwu, Member, IEEE," The Effect of Code Expanding Optimizations on Instruction Cache Design", IEEE Transaction on computers, VOL. 42, NO. 9, September 1993.

[6] M. Kowarschik, C. Weiß, "An Overview of Cache Optimization Techniques and Cache-Aware Numerical algorithms," In: Meyer U., Sanders P., Sibeyn J. (eds) Algorithms for Memory Hierarchies. Lecture Notes in Computer Science, vol 2625. Springer, Berlin, Heidelberg.

[7] B. Zhang, G. Wang, J.J. Sedano, "*Analysis and Improvement of Cache performance for Multimedia Applications*," IEEE Transactions on Computers, February 2004.

[8] N. Smith, T. Clark, "A Framework to Model and Measure System Effectiveness", *ICCRTS Cambridge*, 2006.

[9] Intel, "*NiosII Processor Reference Guide*", NII-PRG,2018.04.18

[10] Altera, "*Cyclone IV Device Datasheet*," Volume 3, March 2016

[11] Intel, "*Embedded Peripherals IP User Guide*," September 2018.

[12] Altera, "*DE2-115 User manual*", 2012 Terasic Technologies

[13] IS42/45R86400D/16320D/32160D, IS42/45S86400D/16320D/32160D SDRAM Device Datasheet, May 2015

[14] Altera, "*SDRAM Controller Core*", November 2009, v9.1.0

[15] Altera, "*Profiling Nios II Systems*," Version 3.0, July 2010

Analyzing the Thermal Imaging Histogram using FPGA

Igor Beracka, Petar Marin, Nikola Latinović, *Member, IEEE*, Ilija Popadić, *Member, IEEE,* Miroslav Perić, *Member, IEEE*

***Abstract*—This paper describes a histogram calculation method for thermal images implemented in FPGA technology. Thermal images display the infrared radiation of objects, which renders them particularly useful in situations where objects of interest are to be separated regardless of the state of illumination in the image. This is achieved by altering the brightness and contrast of the raw image. For that purpose, image histogram is beneficial as it provides information about the values of these correction parameters. Additionally, histogram estimation is essential for picture compression, given that it indicates the areas of image where most information is stored. Laboratory tests results are displayed in this paper and conclusions are drawn.**

***Index Terms*—histogram, FPGA, contrast, brightness, compression.**

I. Introduction

A thermal image (thermogram) is a digital representation of a scene and a measure of the thermal radiation emitted by objects in the picture [1]. Every object hotter than absolute zero (-273.15℃) emits infrared radiation as a function of its temperature - the level of radiation raises with increase of object temperature. Thermal images are captured via thermographic cameras, which are devices capable of sensing this radiation in the form of infrared light. Thermal images are usually presented as levels of grey ranging from black (cold) to white (hot).

The histogram of digital image with grey levels in the range [0, L - 1] is a discrete function

$$h(r_k) = n_k, \tag{1}$$

where r_k is the kth grey level (called bin) and n_k is the number of pixels in the image having grey level r_k [2]. It is common practice to normalize a histogram by dividing each of its values by the total number of pixels in the image. Histogram manipulation can be used effectively for image enhancement because they are simple to calculate in software and economical for hardware implementations. Histograms are also very useful in image compression.

The main objective of this work is to make experimental environment for creating the processing function for dynamic range compression of luminance signal from N to M (N>M) bits without loosing important details from Fig 1.

Heaving in mind good performance results for usage FPGA technology for image processing reported in [3] and [4] we have decided to use this technology, rather than multi-core microprocessor and graphic processing unit that would be used for further processing steps. The main contribution of this paper is that with our design we have proved efficiency of FPGA implementation of histogram and its analysis usefulness in thermal imaging. The paper is structured as follows: in section II we described overall video signal processing that we have implemented, in section III we have focused on FPGA part of implementation, in section IV we give remarks of practical implementation and laboratory setup. Finally, in section V we gave test results followed by conclusion given in section IV.

II. System Description

This system contains thermal camera, CL (Camera Link™) [5] to parallel converter board, frame grabber implemented on FPGA module, parallel to HDMI (High-Definition Multimedia Interface) [6] conversion board and PC (Personal computer).

Camera sends raw thermal image through CL to parallel conversion board to frame grabber, which calculates histogram of raw frame, processes frame and calculates histogram of corrected frame. Video is streamed from frame grabber through ADV7511 integrated circuit to monitor.

Frame grabber is also connected with PC through serial port and histograms are plotted in MATLAB® GUIDE (Graphical User Interface Design Environment) [7].

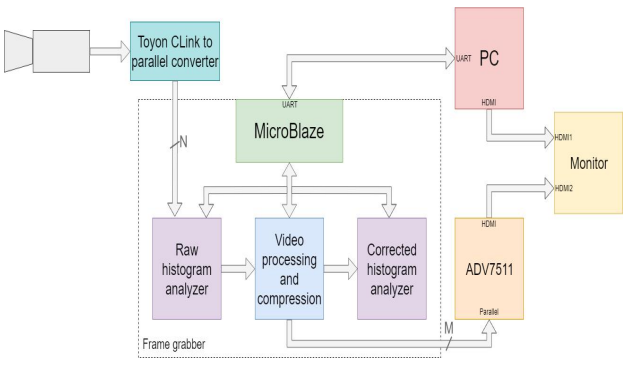


Fig.1. System block diagram.

Igor Beracka is with the Institute Vlatacom, 5 Bulevar Milutina Milankovića, 11020 Belgrade, Serbia (e-mail: igor.beracka@vlatacom.com).

Petar Marin is with the Institute Vlatacom, 5 Bulevar Milutina Milankovića, 11020 Belgrade, Serbia (e-mail: petar.marin@vlatacom.com).

Nikola Latinović is with the Institute Vlatacom, 5 Bulevar Milutina Milankovića, 11020 Belgrade, Serbia (e-mail: nikola.latinovic@vlatacom.com).

Ilija Popadić is with the Institute Vlatacom, 5 Bulevar Milutina Milankovića, 11020 Belgrade, Serbia (e-mail: ilija.popadic@vlatacom.com).

Miroslav Perić is with the Institute Vlatacom, 5 Bulevar Milutina Milankovića, 11020 Belgrade, Serbia (e-mail: miroslav.peric@vlatacom.com).

III. FPGA IMPLEMENTATION

FPGA system (Frame grabber) contains two identical histogram calculator modules (one for raw frame histogram calculation and another for calculating histogram of corrected frame after processing and compression), module for correction of brightness and contrast and MicroBlaze™ micro-controller for controlling the system and connecting it with PC through UART (Universal Asynchronous Receiver/Transmitter).

Histogram calculation module have 3 generic parameters:
g_NUM_OF_BINS - number of histogram bins;
g_FRAMES_TO_ACQUIRE - Number of frames to acquire for histogram calculation;
g_HIST_BIN_BIT_DEPTH - Depth of histogram bin counters and output registers.

With these generics user can define how many MSB (Most significant bits) to take from input signal for calculation, how many frames to take for results averaging and depth of counters results registers in bits respectively.

FSM (Finite State Machine) of this module contains 4 states and it is shown on Fig 3. In IDLE state FSM waits for start command from MicroBlaze™, in WAIT_NEW_FRAME state it waits for new frame to start calculating histogram from first pixel of the picture, in CALC_HIST state it calculates histogram for specified number of frames and in COPY_RESULTS state it stores the results into registers dedicated for that.

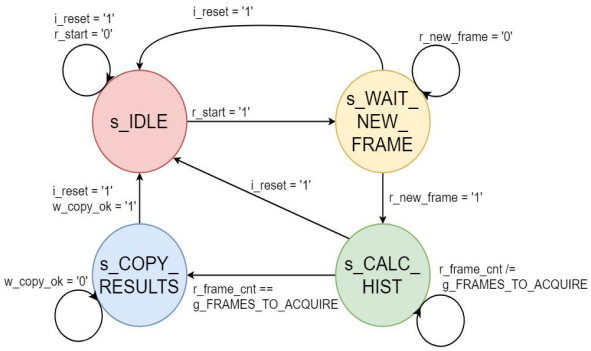


Fig. 2. Histogram calculator FSM state diagram .

FPGA module logic cell usage is measured for different number of histogram bins (16, 32, 64, 128 and 256) used for histogram calculation module and results are given in graph below.

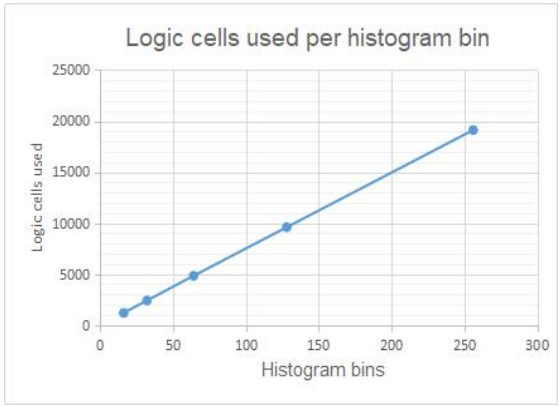


Fig. 3. Logic cells used per histogram bins.

Brightness and contrast correction module is implemented with the use of DSP (Digital Signal Processing) IP (Intellectual Property) core and registers for storing values of user defined brightness and contrast. The function for correction is defined with equation (2):

$$Y = C \cdot X + B, \tag{2}$$

where X is input data (16 bits long), C is contrast (range 0 to 31.984375), B is brightness (range -8192 to 8191) and Y is result truncated to 8 MSB. At this stage we have assumed that linear function (2) is good enough to show important parts of histogram. Better option for correction of brightness and contrast would be non-linear function, but that's out of scope of this paper.

MicroBlaze™ is instantiated and connected with these modules, so user can send commands with MATLAB® GUI (graphical user interface) application through UART.

IV. PRACTICAL IMPLEMENTATION

In this chapter we will show and describe used components and test the frame grabber module implemented on FPGA board.

Thermal camera used in this work is Xenics XTM-640™ [8] with Ophir lens, which sends 50 frames per second of VGA (640x480 pixels) signal with 16b pixel depth. For Camera Link™ to parallel signal conversion we are using Toyon Boccacio™ [9] converter board. Frame grabber is implemented on Mercury PE1 development board with Mercury+ KX2 FPGA module with Xilinx Kintex 7™ [10] on it. For parallel to HDMI conversion we are using Toshiba ADV7511 parallel to HDMI integrated circuit [11].

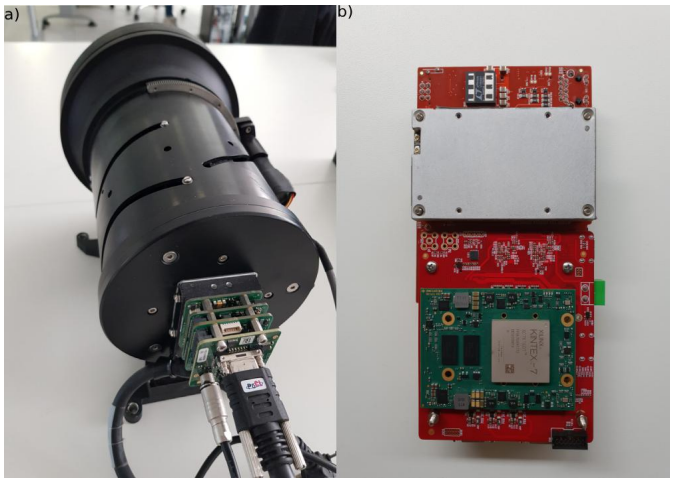


Fig. 4. a) Xenics XTM-640™ camera with Ophir lens, b) Video signal processing module with Mercury PE1 development board and Mercury+ KX2 FPGA module.

As we mentioned in the system description part, software used for creating graphical user interface is Matlab™ GUIDE. On Fig 6 you can see implemented GUI, where are shown controls for serial port configuration, raw and corrected frame histograms calculation, brightness and contrast adjustment, grab threshold points (for grabbing low and high threshold and automatically calculation of brightness and contrast values) and others which are out of scope of this

paper. There is also shown information about Camera Link interface (input) and HDMI (output) signals.

Big black surface on the left side of GUI is place where we put grabbed frame from input signal, which user can save in different picture formats.

Histograms of raw and corrected frame are plotted in new figures and they both have colour bar below to make it intuitive to the user which bin is corresponding colour representing. Number of pixels in frame is normalized to 1, so on the y axis is shown the probability of occurrence of grey level r_k.

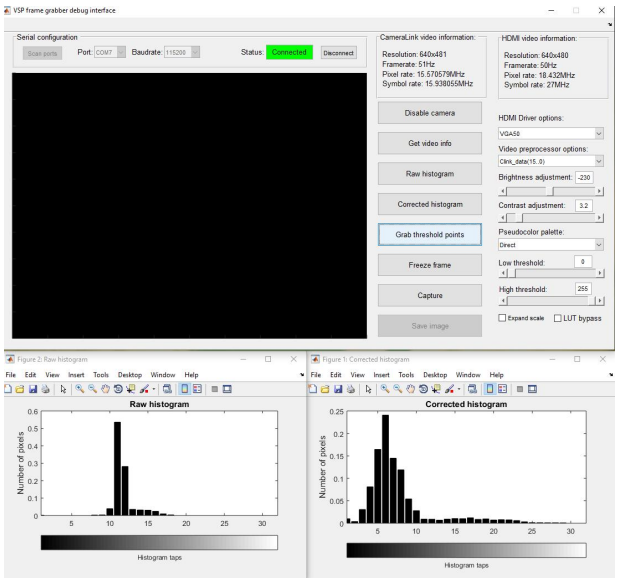


Fig. 5. Graphical user interface.

V. Results

For presentation of the results we have created an environment with a man holding a cups of hot and cold water in his hands. This set up is chosen because it's good for separating of cold and hold objects from picture. On figure below you can see thermal camera, video signal processing module (connected to PC) and person with hot water in black cup (in his right hand) and cold water in white cup (in his left hand).

Fig. 6. Experimentation environment.

The first thing that we have tested is brightness and contrast correction module, where we have captured raw frame from thermal camera (Fig 8a), calculated the histogram of that frame (Fig 8b) and from that histogram extracted the region of interest. Red vertical lines with arrows are added for better description of what's happening with histogram after correction (which part of histogram is expanded).

In the first test we have chosen to expand region from 6th to 15th bin, because there are most pixels from image stored. Resulting image grabbed after correction is shown on Fig 8c, and its histogram is shown on Fig 8d. The main differences between figures 8a and 8c can be seen on cold and hot cups, but other details are also improved on the corrected image (subject face, hands and watch).

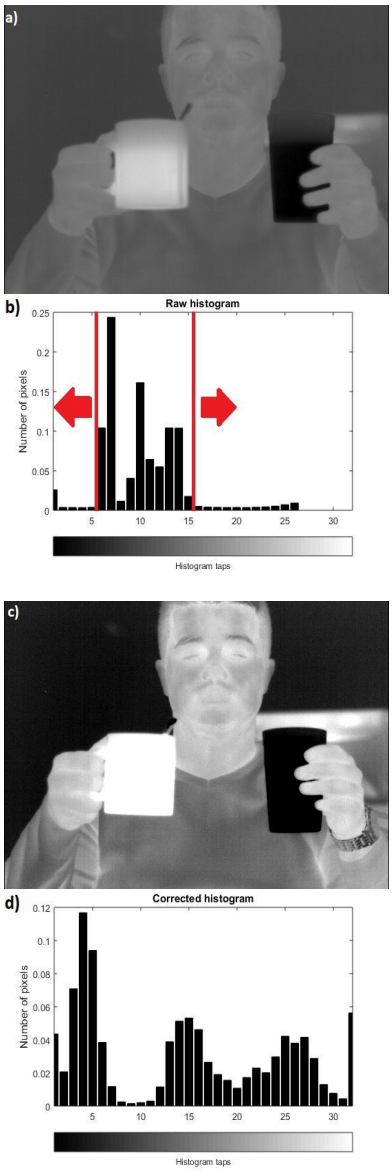


Fig. 7. a) Raw thermal image, b) Histogram of raw thermal image with two vertical lines and arrows which shows what part of histogram is expanded, c) Corrected image, d) Histogram of corrected image.

On the next two figures (9 and 10) we will show cold and hot object detection from raw thermal image histogram. The environment is same for all experiments, but some small details can change in time (subject movement, background changes...), so there must be a little changes between raw histograms shown on figures 8, 9 and 10.

In order to detect cold objects on the picture we have expanded 1st to 6th bin from raw histogram shown on Fig 9b. Resulting image after histogram correction and corrected

histogram are shown on Fig 9a and 9c respectively. On Fig 9a we can clearly see that cold water is detected, while we can also still see the edges of subject because he is hotter than background.

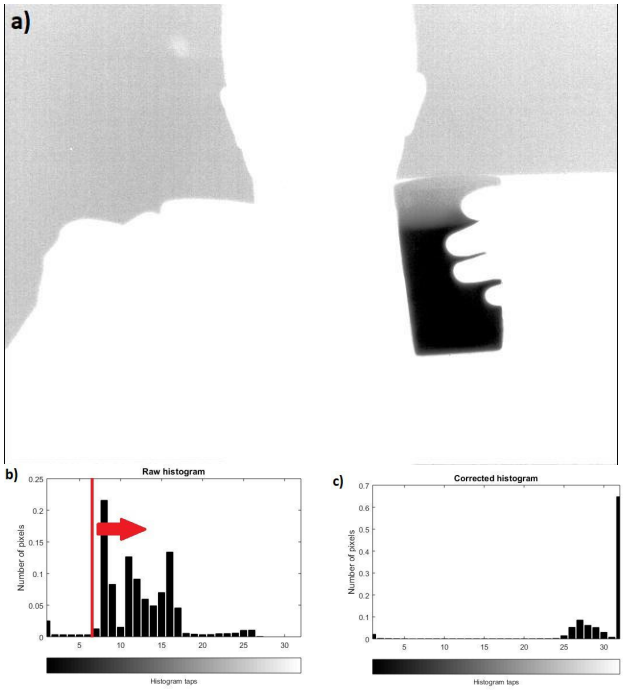


Fig. 8. a) Image with separated cold water, b) Histogram of raw thermal image with vertical line and arrow which shows what part of histogram is expanded, c) Histogram of image with separated cold (dark) pixels.

The last thing that we wanted to detect is hot water, and in order to do that we have expanded 19th to 32nd bin. Raw and corrected thermal image histograms are shown on figures 10b and 10c respectively. Resulting image is shown on Fig 10a, where we can see a cup of hot water, part of monitor from background and contour of hottest parts of subject face.

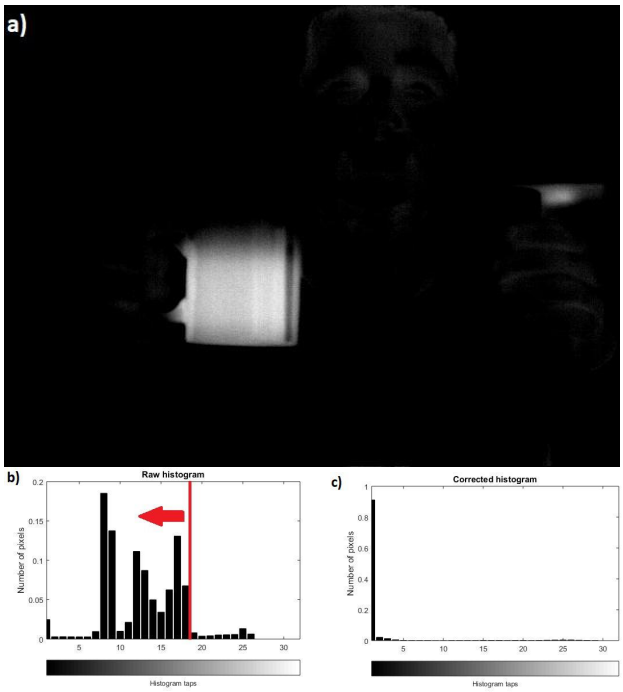


Fig. 9. a) Image with separated hot water, b) Histogram of raw thermal image with vertical line and arrow which shows what part of histogram is expanded, c) Histogram of image with separated hot (light) pixels.

Values chosen for B and C parameters from function (2) for these three different corrections are stored in the table below.

TABLE I
BRIGHTNESS AND CONTRAST PARAMETERS FOR DIFFERENT FIGURES

Brightness	Contrast	Figure
-128	3.2	8a
0	3.555	9a
-435	3.3	10a

VI. CONCLUSION

In this paper, histogram calculation method for thermal images was implemented in FPGA technology. Proposed method reveals superiority of FPGA technology in terms of processing speed when compared to microprocessor or GPU realization. Furthermore, tests carried out in laboratory environment demonstrate how efficient histogram evaluation is for specified linear compression technique.

Future work will focus on evaluating non-linear processing techniques for object detection in low-contrast imagery scenarios.

REFERENCES

[1] M. Vollmer, K. P. Mollmann, "Fundamentals of Infrared Thermal Imaging", in *Infrared Thermal Imaging,* New York, USA: Wiley-VCH, 2010, ch. *1*, sec. *1*, pp. *1*–6.

[2] R. C. Gonzalez, R. E. Woods, "Image Enhancement in Spatial Domain", in *Digital Image Processing, Upper Saddle River,* New Jersey, USA: Abbrev. of Publisher, 1992, ch. *3*, sec. *3*, pp. *88*–90.

[3] G. Bieszcard, T. Sosnowski, H. Madura, M. Kastek, J. Barela, "Adaptable infrared image processing module implemented in FPGA", SPIE Defense, Security and Sensing, Orlando, FL, USA, 2010.

[4] H. Amrouch, T. Ebi, J. Schneider, S. Parameswaran, J. Henkel, "Analyzing the thermal hotspots in FPGA-based embedded systems", 23rd International Conference on Field programmable Logic and Applications, Porto, Portugal, 2-4 Sept. 2013.

[5] "http://www.imagelabs.com/wp-content/uploads/2010/10/CameraLink5.pdf", "Specifications of the Camera Link Interface Standard for Digital Cameras and Frame Grabbers", PULNiX America, Inc. 2000. [Online].

[6] "https://www.hdmi.org/manufacturer/specification.aspx", "HDMI Specification Version 1.4a", HDMI Forum, 2010 [Online].

[7] "http://www.apmath.spbu.ru/ru/staff/smirnovmn/files/buildgui.pdf", "MATLAB® Creating Graphical User Interfaces", The MathWorks Inc. 2015. [Online].

[8] "http://ftp.elvitec.fr/Xenics/FICHES-PRODUITS/xb-050_06_xtm-640_modulescomponents_lowres.pdf", "XTM-640 High resolution ucooled thermal OEM module", Xenics Inc. 2008. [Online].

[9] "https://www.visiononline.org/userassets/aiauploads/file/Toyon_FMC_CameraLink.pdf", "Boccaccio - FMC Camera Link", Toyon Inc. [Online].

[10] "https://www.xilinx.com/support/documentation/data_sheets/ds180_Series_Overview.pdf", "7 Series FPGAs Data Sheet: Overview", Xilinx 2018. [Online].

[11] "https://www.analog.com/media/en/technical-documentation/data-sheets/ADV7511W.pdf", "165 MHz, High Performance HDMI Transmitter", Analog Devices Inc. 2011. [Online].

Design and realization of a Class EF_2 Power Amplifier with GaN FET

Zoran Živanović, *Member, IEEE* and Vladimir Smiljaković

***Abstract*—This paper presents the design of a high frequency class EF_2 switch-mode power amplifier optimized for a 50 ohm load at 13.56MHz. The amplifier basics are briefly explained, including simplified schematics. The prototype has been built and experimental results are presented to support theoretical analysis and to demonstrate the amplifier performance. This practical realization utilizes the newest GaN FET in a low inductance package with small parasitic capacitances capable of switch mode operation up to 50MHz. Previously, the GaN transistors have become widely used in high frequency switching power converters.**

***Index Terms*—Class EF_2, duty cycle, efficiency, GaN, switch-mode power amplifier, zero voltage switching.**

I. Introduction

A large variety of industrial, scientific and medical processes require reliable, low cost high frequency regulated power. Applications include RF plasma processing, dielectric heating, RF welding, magnetic resonance imaging and wireless power systems. Previously, high frequency power was provided by linear power amplifiers with transistors operated in the linear mode with efficiency around 70%. Opposite to them, transistors in switch-mode power amplifiers are operating as a saturated switch with efficiency up to 95%. Unlike standard 50 ohm applications, the load in ISM applications is usually extremely dynamic, ranging from a few ohms to a several hundred ohms including reactive parts. The problem is further complicated by the fact that the process requires the amplifier to operate with large mismatch.

II. Class EF_2 Power Amplifier Basics

The Class E power amplifier is a type of switch-mode RF power amplifier, introduced by Sokal and Sokal in the 1970's [1]. The topology consists of a MOSFET transistor, shunt capacitor, RF choke and a series resonant circuit. The transistor is driven hard enough to saturate at 50% duty cycle. Due to its operation at zero voltage switching, they can deliver RF power up to tens of MHz with efficiencies up to 95%. Single ground referenced power switch experiences a higher voltage stress compared to other topologies ($3.6V_{in}$) [2]. Also, deviations of the load can cause the efficiency drop and permanent damage to the switch. In 2002, Kee presented an overview of the Class EF inverters and a generalized frequency domain based analysis method to determine the voltage and current waveforms of LC resonant networks [3]. Later it was shown that the voltage and current stresses of the Class E amplifier can be reduced by adding resonant network either in parallel or series to its load network [4-6]. The method of adding resonant networks is already used in Class F and Class F^{-1} inverters. Applying this method to the Class E inverters, results in hybrid inverters, which has been referred to as the Class EF_n or Class E/F_n inverters. The subscript n refers to the ratio of the resonant frequency of the added resonant network to the inverters switching frequency and is an integer number greater or equal than 2. If n is an even integer the EF_n term is used and if *n* is an odd integer the E/F_n term is used.

Class EF_2 power amplifier shown in Fig. 1 overcomes some limitations of the class E amplifier. Like the Class E, the resonant components (C1, C3, L3) are tuned to provide zero voltage switching in order to lower the switching losses.

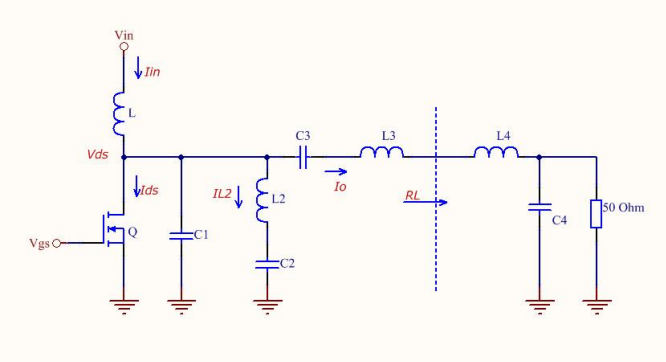


Fig. 1. Class EF_2 power amplifier

Components L2 and C2 are tuned to have low impedance at the 2nd harmonic at the terminals of switching transistor. This leads to a trapezoidal waveform of drain voltage with a peak up to $2.5V_{in}$. Choke inductance L must be high enough such that the input current I_{in} is DC current. The capacitance of shunt capacitor C1 includes the output parasitic capacitance of the power switch. This parasitic capacitance is voltage dependent and it can be difficult to choose an exact value. Loaded Q factor L3C3 branch must be high enough such that the current I_o is sinusoidal. An output matching network is needed (L4, C4) to transform the 50 Ω impedance into the required load resistance R. Class EF_2 amplifiers have higher power-output capability then class E amplifiers. Also its DC input voltage is higher than that of a Class E amplifiers,

Zoran Živanović is with the IMTEL KOMUNIKACIJE AD, Bul. Mihajla Pupina 165b, 11070 Belgrade, Serbia (e-mail: zoki@insimtel.com).

Vladimir Smiljaković is with the IMTEL KOMUNIKACIJE AD, Bul. Mihajla Pupina 165b, 11070 Belgrade, Serbia, (e-mail: smiljac@insimtel.com).

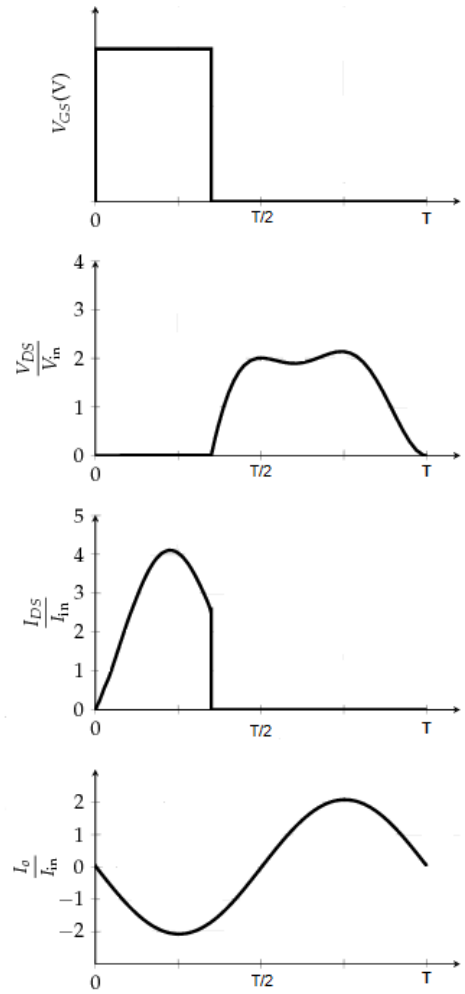


Fig. 2. Class EF_2 voltage and current waveforms

because the Class EF2 reflects a higher DC load to the DC power supply. Practical duty cycle can vary from 30 to 40% opposed to 50% in a class E.

Fig. 2 shows power switch's normalized drain voltage and current waveforms and normalized output current at 35% duty cycle.

III. Design and Analysis

To design class EF_2 RF power amplifier [7] we will start from the design specification given in Table I.

TABLE I
DESIGN SPECIFICATIONS

	Parameter	Value	Units
DC voltage	V_{CC}	28	V
Output power	P_O	40	W
Load	Z	50	Ω
Drain efficiency	η	90	%
Frequency	f	13.56	MHz

Inductor L2 and capacitor C2 must be tuned to a frequency in between the switching frequency and the second harmonic of the switching frequency. The ratio of the resonant frequency of L2C2 to the switching frequency is represented by parameter q_1 and is given by

$$q_1 = \frac{1}{\omega\sqrt{L_2 C_2}} \tag{1}$$

The parameters k is defined as

$$k = \frac{C_1}{C_2} \tag{2}$$

Also the parameter p is given by

$$p = \frac{1}{k+1}\frac{I_m}{I_{in}} \tag{3}$$

The minimum choke inductance required is

$$L_{\min} = 2\pi D \frac{V_{in}}{\omega \Delta i_{L\max}} \tag{4}$$

Series inductance L3 is defined as

$$L_3 = \frac{QR_L}{\omega} \tag{5}$$

Design begins with choosing the value for q_1. Tuning the resonant frequency of L2C2 network to around 1.5 times the switching frequency allow load independent operation to be achieved. We will adopt the value of 1.66 in order to have maximum power output capability. Duty cycle will be set at 30%. The large value for p could result in small value of shunt capacitance C1, lower then parasitic capacitance of available power switches. Low value for p will result in lower efficiency. Finally, we adopt value of 2. Also we will set the maximum input ripple current to be 10%. Knowing that, we have calculated the values of all elements which are shown in the Table II. Based on the value of the input DC voltage and output power we have selected 100V eGaN FET EPC2016C for the power switch.

TABLE II
CALCULATED VALUES

Component	Parameter	Value	Units
Load resistance	R_L	2.50	Ω
RF choke	L	6	µH
Shunt capacitance	C_1	832	pF
Added capacitance	C_2	655	pF
Added inductance	L_2	76	nH
Series capacitance	C_3	783	pF
Series inductance	L_3	232	nH
Matching inductance	L_4	128	nH
Matching capacitance	C_4	1023	pF

It has ON resistance of 16mΩ and the output capacitance of approximately 350pF, which is lower than calculated value for C1. The result is that we have to add the capacitors in the value of 482pF to the drain node. Input capacitance is 360pF.

IV. Realization

The class EF_2 RF power amplifier has been realized on a two layer FR-4 substrate with a relative dielectric constant of 4.3, thickness of 1.6mm and 50µm copper. The oscillator at 13.56MHz drives low side gate driver UCC27611. The RF choke inductor L is made on the toroidal core. The series inductor L3 and matching inductor L4 are integrated in one air core inductor. All the RF capacitors used are low ESR capacitors from ATC Corporation. The load is made as a combination of 50Ω/100W thick film resistors from Diconex. The gate drive signal with 30% duty cycle is given in Fig. 3. From the drain waveforms (Fig. 4 and 5.), it can be seen that ZVS is maintained with the load change.

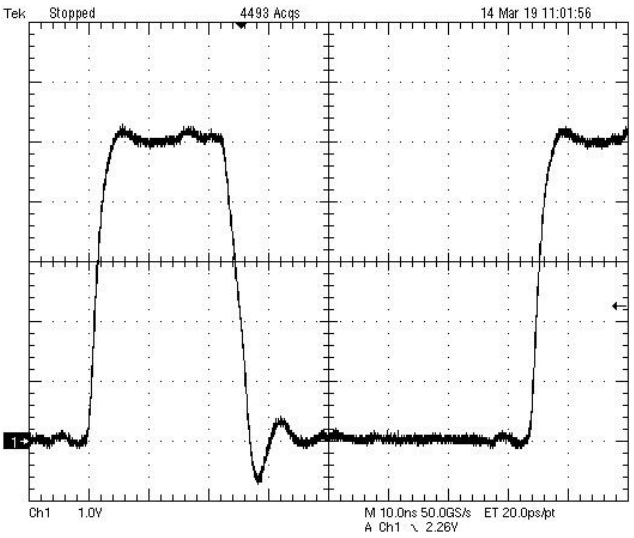


Fig. 3. Gate voltage

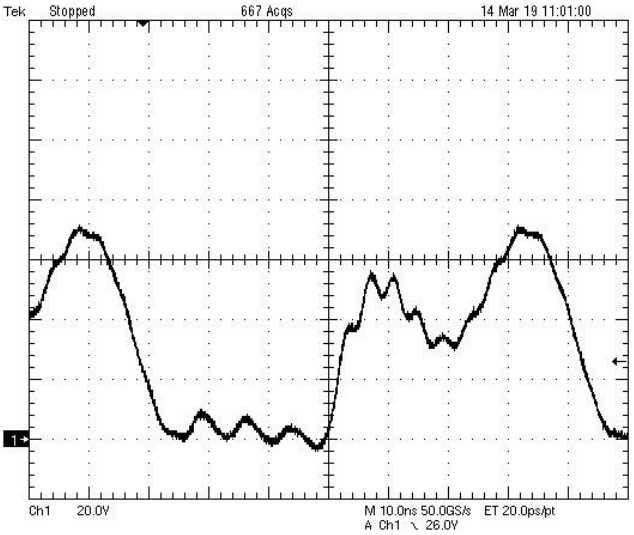


Fig. 4. Drain voltage at 50 ohm load

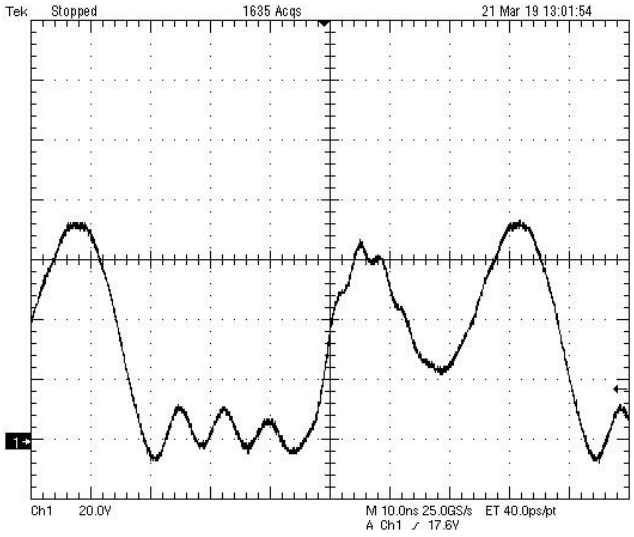


Fig. 5. Drain voltage at 100 ohm load

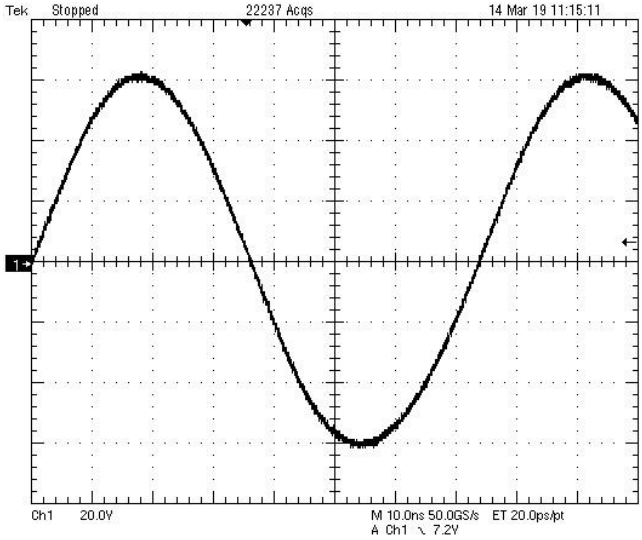


Fig. 6. Output voltage

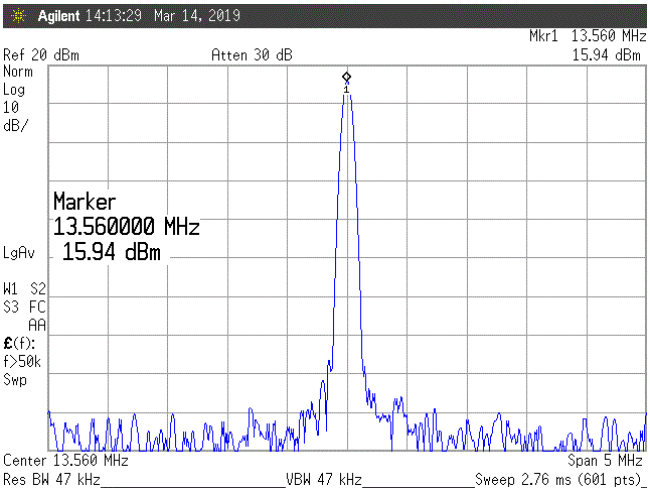


Fig. 7. Output power at fundamental frequency

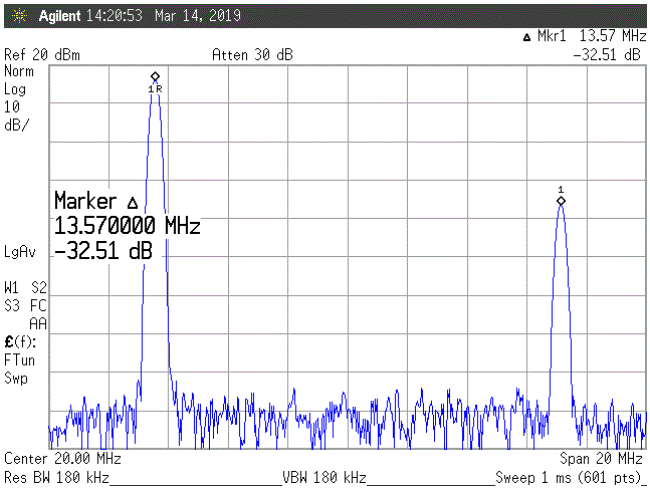


Fig. 8. Second harmonic power relative to fundamental

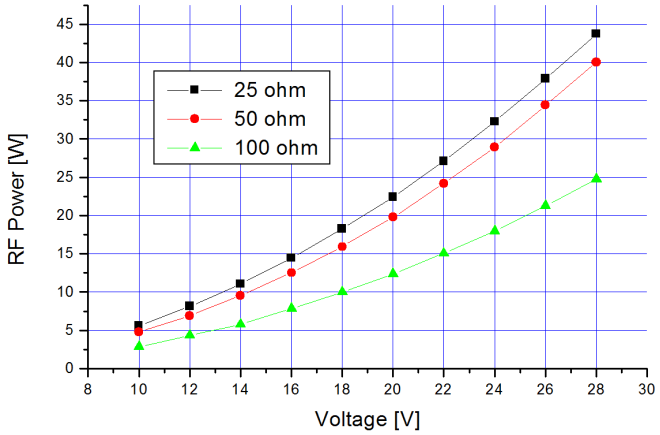


Fig. 9. Output power vs. power supply voltage

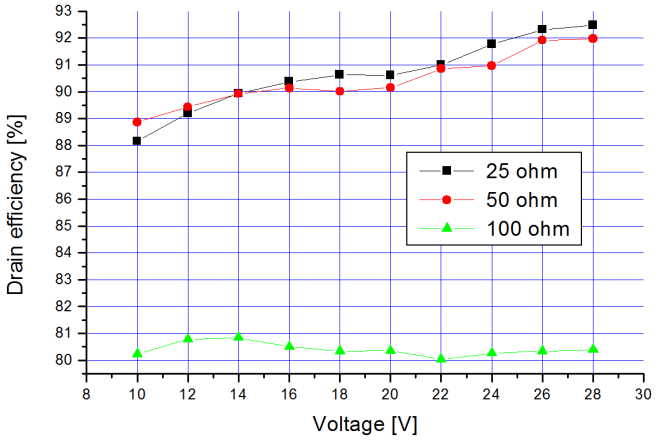


Fig. 10. Efficiency vs. power supply voltage

The peak drain voltage is slightly higher than expected because of FETs non-linear output capacitance. The output in time domain is given in Fig. 6. The output at the fundamental frequency is given in Fig. 7 (recorded with 30 dB attenuator). The second harmonic power is 32.5 dB lower relative to fundamental (Fig. 8). The output power vs. power supply voltage for various loads is given in Fig. 9.

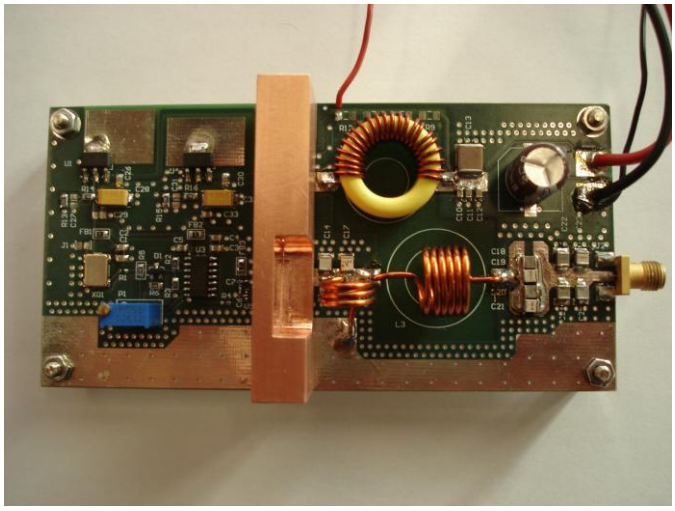


Fig. 11. Class EF_2 power amplifier prototype

Efficiency vs. power supply voltage for various loads is given in Fig. 10. It is always higher than 80% for the load change from 25 to 100 ohms. Class EF_2 power amplifier prototype is given in Fig. 11.

V. Conclusion

The class EF_2 power amplifier is designed, built and tested in the laboratory. The result verified that this amplifier can deliver safely up to 40W of RF power at 13.56 MHz into load ranging from 25 to 100 ohms. Zero voltage switching is maintained over a wide load change with efficiency from minimum 80% up to maximum over 92% unlike conventional Class E and Class EF2 who can only ensure high efficiency at fixed load.

Acknowledgment

The work is partially supported by the Serbian Ministry of Education and Science (Project III-44009). The authors would like to thank Texas Instruments for providing samples of GaN driver integrated circuits.

References

[1] N. O. Sokal, A. D. Sokal A.D. “Class E – A new Class of High-Efficiency Tuned Single-Ended Switching Power Amplifiers, ” IEEE Journal of solid-state circuits. Vol. SC-10, pp. 168-176, June 1975.

[2] Z. Zivanovic, V. Smiljakovic, “Design and realization of a Switch-mode Power amplifier with GaN FET,” IcETRAN Proceedings, pp. 883-886, June 2018

[3] S. D. Kee, I. Aoki, A. Hajimiri and D. Rutledge, ‘‘The Class E/F family of ZVS switching amplifiers,’’ IEEE Trans. Microw. Theory Techn., vol.51, no. 6, pp. 1677-1690, Jun. 2003

[4] A. Mediano and N. Sokal, “A Class-E RF power amplifier with flattop transistor-voltage waveform,” IEEE Trans. Power Electron., vol. 28, no. 11, pp. 5215–5221, Nov. 2013.

[5] A. Grebennikov, “High-efficiency Class E/F lumped and transmissionline power amplifiers,” IEEE Trans. Microw. Theory Techn., vol. 59, no. 6, pp. 1579–1588, Jun. 2011.

[6] Z. Kaczmarczyk, “High-efficiency Class E, EF2, and E/F3 inverters,” IEEE Tran Ind. Electron., vol. 53, no. 5, pp. 1584–1593, Oct. 2006.

[7] Samer Aldhaher, Paul D. Mitcheson and David C. Yates,“Load-Independent Class EF Inverters for Inductive Wireless Power Transfer, ” IEEE Wireless Power Transfer Conference (WPTC), May 2016

Measuring EMG signal with EMG click and Arduino UNO

Nemanja Peruničić, Đorđe Novaković

***Abstract*—This paper gives an insight into the basics of gathering and processing of bioelectric signals generated by our skeletal muscles. Two-part system is used as a technical solution, and the end result is shown on a computer. Signal acquisition is done with a specially designed EMG click board, and its digitization using Arduino UNO's ATmega328P microcontroller. When programming Arduino it is vital to pay attention to the hardware elements and theoretical principles on which the measurement is based; otherwise, false results will be presented – so a part of the paper is dedicated to overcoming them. The third chapter includes the complete sampling code.**

***Index Terms*—EMG signal; biceps; EMG click; Arduino.**

I. Introduction

Bioelectric recordings are of the small electrical potentials produced by the living cells and organs [1-2]. The frequencies of these signals range from 0 to 10 kHz, and their characteristics can be highly dependent on the degree of activity of the cells [3]. From a hardware standpoint EEG is the most difficult measurement to acquire [2]. Bioelectric signals are non-stationary, but their power spectrum analysis of not only provides a summary of the signal in a convenient graphic form, but also facilitates statistical analysis of signal changes that may not be evident on simple inspection of the records [1]. Besides standard methods for bioelectric signal measurement, development of methods for stochastic measurements of bioelectric signals is reported in [5-6].

Electromyography (EMG) is an electro diagnostic medicine technique for evaluating and recording the electrical activity produced by skeletal muscles. Intensity of muscle contraction is proportional to intensity of the measured signal, and it depends on the number of impulses brought by the nerves. Different muscles will generate various signals, but when it comes to humans EMG amplitude range is from 0.05 to 5 mV, and frequency from 10 to 400 Hz. As shown in [4] the vast majority of spectrum components are around 100 Hz.

We are doing non-invasive measurement, set to give valid results for signals coming from right arm biceps. Electrodes used here are disposable, made from Ag/AgCl, and designed for gel-free measurement. Each electrode is built into a hydrogel-coated adhesive foam pad that goes directly onto the skin. They connect to an ECG/EMG cable like a button – electrode's stainless steel extension is the part that lays in. Single-channel bipolar measurement is conducted, i.e. we are measuring the voltage between the pair of active electrodes located on the tendons, whilst the third one, Driven Right Leg (DRL), can be positioned anywhere on the body.

Fig. 1 shows the principle block scheme of bioelectric signal gathering system. Signal gathered by the electrodes must be amplified at least a thousand times, plus we have unwanted signals coming from other parts of the body and noises from external electro-magnetic fields. Therefore, a circuit containing amplifiers and filters is needed; in our case – EMG click. The DRL part of that circuit eliminates the common-mode voltage (utility frequency coming from the power lines; amongst others) from the input of the first amplifier to which the signal is driven (preamplifier), by inverting it and sending it back to the body.

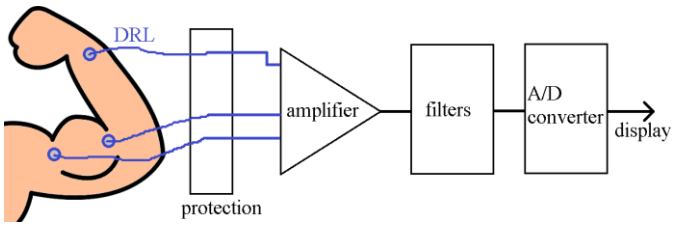


Fig. 1. Simplified block scheme of EMG acquisition device.

All of the analog filters contained within this system are RC based. Since we didn't use armored wires to connect the click board to the microcontroller there is a need for one more filter. It is a digital filter, implemented within the Arduino software.

II. Click Board

EMG click (shown on Fig. 2) serves for single-channel collection of electrical signals from motor neurons of skeletal muscles. The 3.5 mm audio jack is used to connect electrode cables to the board. The board carries two chips: MCP609 operational amplifier and MAX6106 voltage reference. Sixteen pins are located within the MCP609, out of whom three analog are used here: for the power supply (5V), ground (GND) and analog EMG signal (AN).

Nemanja Peruničić is with the Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia (e-mail: palatarion@yandex.com).

Đorđe Novaković is with the Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia (e-mail: djordjenovakovic@uns.ac.rs)

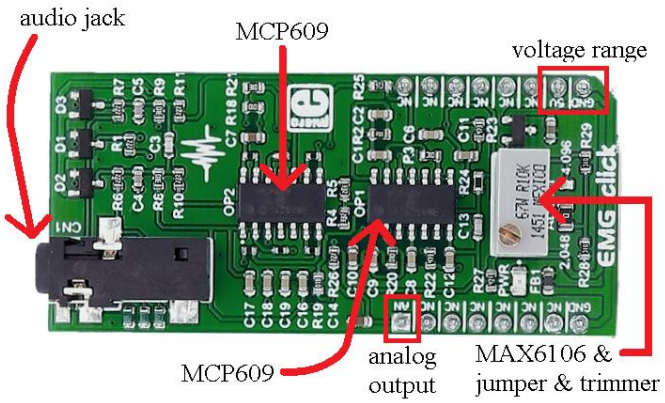


Fig. 2. EMG click board.

The differential electrode signal is run to, in order: preamplifier (instrumentation amplifier), high-pass filter, non-inverting amplifier, high-pass filter, low-pass filter, AN pin of the Arduino UNO. High-pass filters are passive first order, and separated because both need to come right after the amplifiers in order to eliminate possible signal offset. Low-pass filter is of the third order – the second order active is serial connected to the first order passive.

The DRL circuit, also a part of the MCP609, takes the signal from the input of the preamplifier and drives it through two inverters, the latter one also being a high-pass filter. Two-way zener diodes and ESD diode pairs are used as a protection for the measuring channel and the DRL electrode. The MCP609 input also contains passive low-pass filters that prevent radio waves from entering the preamplifier.

MAX6106 moves the circuit global reference mass to 2.048 V. This moves the baseline of the EMG signal up from zero – otherwise the negative values would be cut off. In case the range of the amplified EMG needs to be adjusted to another type of microcontroller, with a different input range of his A/D converter, a jumper and trimmer potentiometer are used to move the reference mass (see Fig. 3). Additionally, the non-inverting amplifier (who provides the bulk of the amplification) can be adjusted with another trimmer to accommodate new input range.

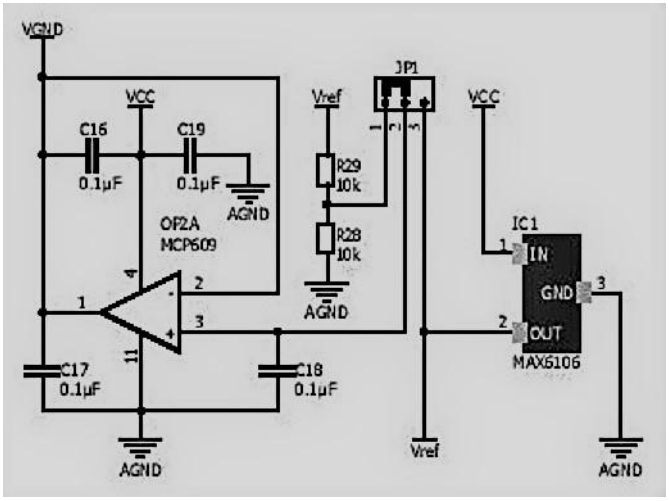


Fig. 3. Part of the click board that sets the reference mass (AGND).

EMG click gets its power supply from Arduino, who in turn is powered from the computer via USB cable.

III. Microcontroller Board

The ATmega328P microcontroller is used as a central element of the Arduino UNO board (shown on Fig. 4), which has everything needed for its support. The board contains fourteen digital input/output pins, six analog input pins, a 16 MHz crystal oscillator clock, USB 2.0 connector, power jack (if we do not want to use a computer for display) and a reset button. UNO uses another chip – ATmega16U2, for USB-to-serial conversion.

Each of the analog inputs has a 10 bit resolution (i.e. can give values from 0 to 1023), and is set to measure from ground level to 5 V; though the upper end of the range can be changed by using the AREF pin and the analogReference() function in software. The ATmega328P implements UART protocol to communicate with the computer (the speed is defined with the Serial.begin() function); it is available through digital transmit/receive pins 1 (TX) and 0 (RX). The ATmega16U2 is preprogrammed to channels the serial UART through the USB, and appears as a virtual COM port on the computer. Aside from UART, ATmega328P supports I^2C and SPI protocols.

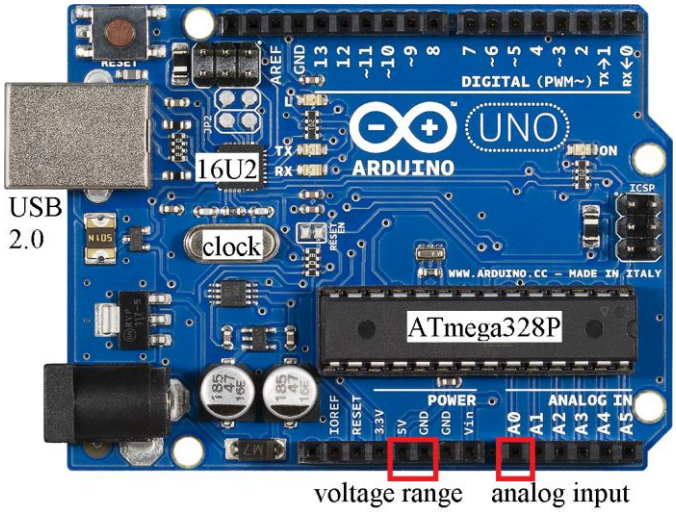


Fig. 4. Arduino UNO microcontroller board.

The Arduino software has a serial monitor which allows textual data (signal values) to be sent from the board. The two onboard LEDs connected to TX and RX pins will flash when data is being transmitted via ATmega16U2 and USB connection to the computer.

Actual pressing of the reset button is not required before an upload, because the Arduino UNO is designed to be reset by software running on a connected computer. One of the lines of the ATmega16U2 is connected to the ATmega328P's reset line with a 100 nF capacitor. When the line is asserted, the

reset line drops long enough to reset the chip. That allows the user to upload code by pressing the upload button in the Arduino software. UNO has a resettable fuse that provides an extra layer of protection for the USB port from overcurrent or shorts. If the applied current amplitude is bigger than 0.5 A, the fuse will automatically break the connection, until the overload or short is terminated.

The main purpose of the microcontroller unit is to sample the analog EMG signal coming from the click board. That can only be achieved by using its interrupt functionality. The interrupt routine is stored within the 32 KB flash program memory of the ATmega328P's CPU. The lowest addresses in the flash memory are defined as the reset and interrupt vectors. Their order determines the priority level – the lower the address the higher is the priority. RESET comes first, and we are using TIMER0_COMPA which is more than a dozen addresses higher. The interrupt vectors can be moved to the start of the boot section by setting the IVSEL bit in the CPU control register.

When an interrupt occurs the program counter will be vectored to the actual interrupt vector in order to execute its handling routine, and the corresponding interrupt flag is cleared. Interrupt flags can also be cleared by writing a logic one to the flag bit. If an interrupt conditions are met while the corresponding interrupt enable bit is cleared, the interrupt flag will be set and remembered until the interrupt is enabled, or the flag is cleared by software. Also, if interrupt condition occurs whilst the global enable bit is cleared, the corresponding interrupt flag will be set and remembered until the global enable bit is set, and will then be executed.

When the microcontroller exits an interrupt, it returns to the main program and executes one more instruction before any other pending interrupt. The cli() instruction disables interrupt immediately, and the sei() instruction enables it, although the first following instruction will be executed before any pending interrupts. The above mentioned functions are, amongst others, shown on Fig. 5.

```
int k=0, g;
volatile bool flag=false;
volatile float u;
float s;
int state=LOW;

void setup()
{
 Serial.begin(115200);
 pinMode(A0,INPUT);
 cli();
 TCCR0A = 0;
 TCCR0B = 0;
 TCNT0  = 0;
 OCR0A = 124;
 TCCR0A |= (1 << WGM01);
 TCCR0B |= (1 << CS01) | (1 << CS00);
 TIMSK0 |= (1 << OCIE0A);
 sei();
}

ISR(TIMER0_COMPA_vect){
u=(float)analogRead(A0)/1024;
flag = true;
}

void loop()
{
 while(1){
   while(k<2000){
    for(g=0, s=0; g<5; g++){
       while(!flag);
       flag = false;
       s+=u;
      }
      Serial.println(s);
      k++;
  }
  k=0;
  }
}
```

Fig. 5. Sampling instructions given to microcontroller.

The sampling frequency (*fs*) is set to 2 KHz – five times the chosen maximum frequency in our EMG signal. All five registers used inside of the setup function are 8-bit. In "clear timer on compare" (CTC) mode, which is activated with WGM01 bit, the OCR0A register is used to manipulate the counter resolution (top value) – the counter is zeroed when the counter value (TCNT0) matches the OCR0A. Output compare value is calculated as

$$OCR0A = \frac{clock}{fs \times prescale\ factor}. \qquad (1)$$

CS00 and CS01 bits set the prescaler at 64, so that the desired counter top value can fit inside the OCR0A register. All interrupts are individually masked with the timer interrupt mask register (TIMSK0), i.e. he enables the timer compare interrupt. CTC mode allows control of the compare match output frequency and simplifies the operation of counting external events.

Volatile serves to make the variable exempt from optimization by the compiler. The problem would occur when something is changed in an interrupt, while the main function tries to access the same element. The first while loop is present just to emphasize infinite repetition, it is not necessary because void loop() behaves the same on its own. Before we used registers to precisely define sampling rate, the second while loop counted 2000 samples within one second – high UART speed (115200) was used to ensure that the signal plotting happens in real-time. This loop too is unnecessary, but it does not have negative effect on code execution. ISR function enables analogRead() to take samples on 0.5 ms, but because of the for loop values will be sent to the computer every 2.5 ms, i.e. with a 400 Hz frequency. Standard UART frame length is used (10 bits; one per start and stop, and eight for the value), so the bare minimum number of needed bits per second is 4000. Since microcontroller UART module runs independently of its CPU, unnecessarily high speed is not an issue.

The noise has Gaussian distribution; mentioned in [7], so it can be filtered by arithmetic mean. When calling an interrupt, a loop is present – every five consecutive values will be averaged. Five are taken for one more reason: values coming from the A/D converter are presented in voltage quants, and it is common to display them in volts. The conversion is done by multiplying with 5 V (power supply) and dividing by 1024 (2^{10}); two fives cancel each other out. Fig. 6 shows real-time EMG signal acquisition on a serial plotter window in Arduino software. Later, if we want to show the signal on a time axis, it must be extended five times, in order to be scaled properly.

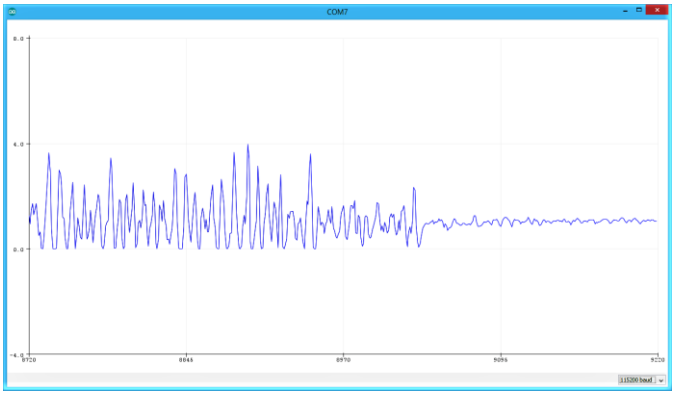


Fig. 6. Real-time EMG signal display.

IV. Conclusion

The files obtained by the Arduino serial monitor can be saved in a spreadsheet document format. From there the data can be loaded into various signal analysis programs, with tools for removing DC component and other visual disorders.

This two-part system is for academic research purposes; it is simple, relatively cheap and easy to assemble, and is a good basis for someone's further exploration of the bioelectrical signal acquisition field. Everything shown here can be translated into multi-channel measurement, and with an addition of several more auxiliary components it would give valid results for anyone's standards.

References

[1] J.D. Bronzino, "Principles of Electroencephalograpy", in *Biomedical Engineering Handbook*, vol. 1, J.D. Bronzino, 2nd ed., New York: CRC Press LLC, 2000.

[2] J. G. Webster, *Medical Instrumentation Application and Design*, New York: Wiley, 1998.

[3] M. Abeles, M. Goldstein, "Multispike Train Analysis", *Proc. IEEE*, no. 65, pp. 762-773, 1977.

[4] P. Komi, P. Tesch, "EMG frequency spectrum, muscle structure, and fatigue during dynamic contraction in man," *Eur. Jour. of Appl. Physiol. and Occ. Physiol.*, vol. 42, pp. 41-50, Sep. 1979.

[5] P. Sovilj, S. Milovančev, V. Vujičić: Digital Stochastic Measurement of a Nonstationary Signal With an Example of EEG Signal Measurement, Instrumentation and Measurement IEEE Transactions on, 2011, vol. 60 - issue 9, pp. 3230-3232, ISSN 0018-9456, DOI: 10.1109/TIM.2011.2128670

[6] M. Urekar, P. Sovilj, „EEG dynamic noise floor measurement with stochastic flash A/D converter", Biomedical Signal Processing and Control, vol. 38, pp. 337-345, Elsevier B. V, 2017, ISSN 1746-8094

[7] A. Phinyomark, C. Limsakul, P. Phukpattaranont, "EMG Feature Extraction for Tolerance of White Gaussian Noise," in *Proc. Int. Work. Symp. Science and Technology (I-SEEC 08)*, Nong Hai, Thailand, pp. 178-183, Dec., 2008.

Acquisition of BCG signal by piezoelectric sensor

Jovana Jevremov, Đorđe Novaković, *Member, IEEE*, Platon Sovilj *Member, IEEE*

***Abstract*—The main idea of this paper was to construct very simple device whose function would be an acquisition of a signal obtained by piezoelectric sensor placed on a pulsating blood vessel. The measurement results were clear and unambiguous. This would enable the simple solution for tracking the number of heartbeats in a period of time. Tracking the number of heartbeats in a period of time is a wide spread functionality used by a great number of sporting aids such as smart bracelets and smart watches, as well as many click boards and even mobile applications.**

***Index Terms*— measurement and acquisition, piezoelectric sensor, ballistocardiography, heartbeat.**

I. Introduction

The main idea of this paper was to attempt to establish a link between the mechanical movement of the pulsating blood vessel and the BCG signal by using a simple piezoelectric sensor and an adaptive circuit. Based on a similar papers [1][2], a circuit was designed in order to obtain signals from a pulsating blood vessel that would be suitable for further analysis.

Piezoelectric materials have the ability to expand or shrink when exposed to voltage and generate voltage when exposed to mechanical force or pressure.

Piezoelectric ceramics and special crystals are used mainly for the production of piezoelectric sensors (in the development project described in this paper a ceramic sensor is used). [3] Ceramics consist of fine crystals. Each crystal consists of atoms with a positive or negative electric charge that are well balanced. However, the type of dielectric ceramics, ferroelectrics, have unbalanced positive and negative electrical charges in crystals, even in natural conditions, which leads to spontaneous polarization. Immediately after exposure, the ferroelectric ceramics will develop spontaneous polarization with random polar axes. As a whole, ceramics will have well balanced positive and negative electrical charges, but with the application of high DC voltage, the polar axis generating spontaneous polarization aligns in the same direction, which cannot be canceled even if the voltage is removed. The process of aligning the polar axes of spontaneous polarization is called the polarization process.

If the polarization process is applied to ferroelectric ceramics, piezoelectric ceramics are produced. When an external voltage is applied to piezoelectric ceramics, the centers of positive and negative electric charge in the ceramic are individually attracted or repelled from external electrical charges, which leads to the spread or contraction of ceramics. [4]

On the other hand, the application of pressure on piezoelectric ceramics generates positive and negative electrical charges on the surface of piezoelectric ceramics. Conversely, if a tensile force is applied to the same material, the polarity of the electrical charge will be reversed. As described above, piezoelectric ceramics allow mutual conversion of electrical energy and mechanical energy by using polarization of crystals. The property of a piezoelectric material to translate mechanical energy into electrical is called a piezoelectric effect.

The relation between the force of action on the material and the generated electric charge does not depend on the shape of the piezoelement, nor on the size, but solely on the material from which it is made. The formula for calculating the generated charge is:

$$Q = q_{11} \cdot n \cdot F \tag{1}$$

where n is the number of connected quartz elements, F force applied to the sensor, q_{11} material constant, and Q generated electric charge. [5].

II. Electrical scheme

Figure 1 shows the electrical scheme drawn in the LTspice program. [6] From left to right, the input signal can be seen as Vin, which is led to the uninverted input of the operational amplifier with a gain of 40.

The next block is a low-frequency filter that transmits frequencies below 220 Hz and then has a signal drop of -45 dB/dec. At the end there is another voltage divider with a buffer. This components aim is to add a DC component to the signal in order to raise its mean value from the zero. Voltage supply of operating amplifiers is ± 12 V.

Jovana Jevremov is with the Faculty of Technical Sciences, University of Novi Sad, Dositeja Obradovica 6, 21000 Novi Sad, Serbia (e-mail: jevremov.jovana@uns.ac.rs).

Đorđe Novaković is with the Faculty of Technical Sciences, University of Novi Sad, Dositeja Obradovica 6, 21000 Novi Sad, Serbia (e-mail: djordjenovakovic@uns.ac.rs).

Platon Sovilj is with the Faculty of Technical Sciences, University of Novi Sad, Dositeja Obradovica 6, 21000 Novi Sad, Serbia (e-mail: platon@uns.ac.rs).

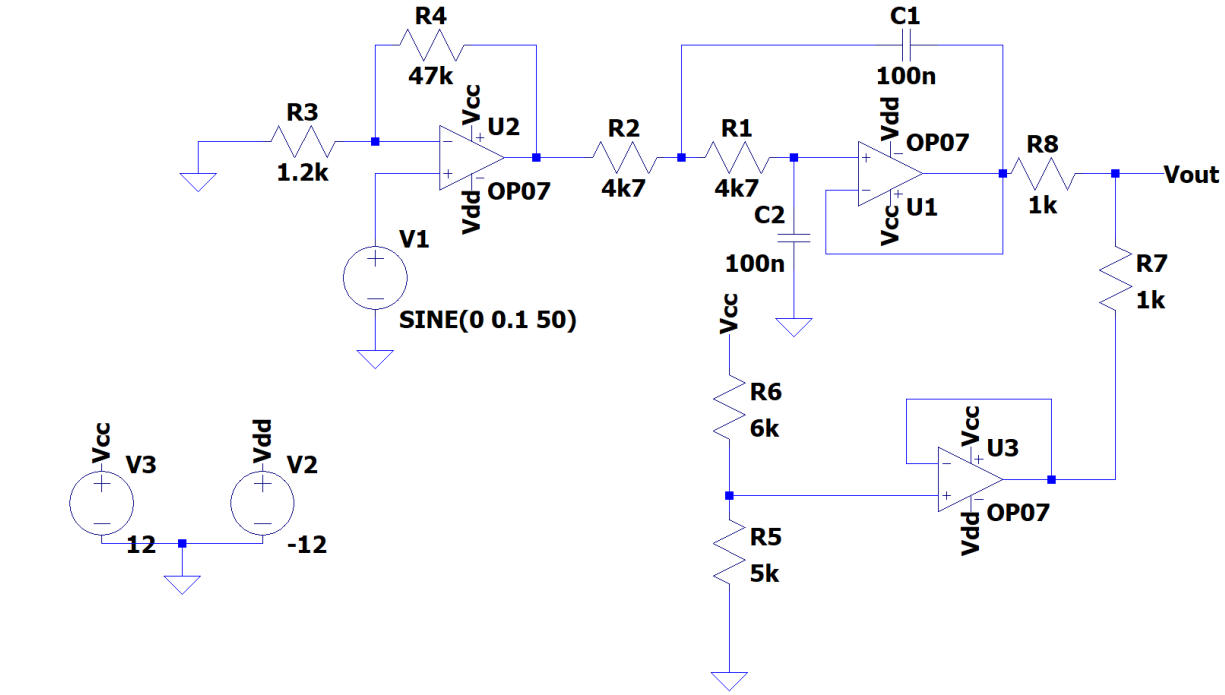


Fig. 1. Electrical scheme

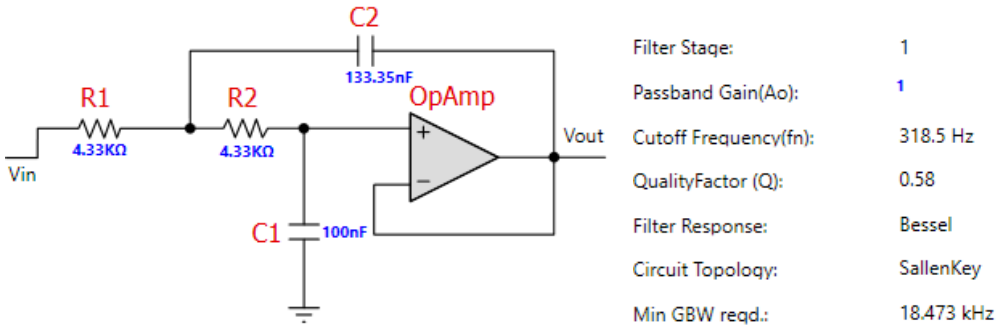


Fig. 2. Low-pass filter in FilterPro software

A. *Amplifier*

For an amplifier was chosen a non-inverting amplifier consisting of a single operational amplifier 0P07 [7] and two resistors of 47 kΩ and 1.2 kΩ. By preliminary recording of the observed signal, an oscilloscope was found its maximum amplitude of 100 mV, and therefore it was decided that the gain would be about 40 times.

B. *Low-pass filter*

The low frequency filter consists of one operational amplifier 0P07, two capacitors of 100 nF and two resistors of 4.7 kΩ. LP filtering aims to suppress high-frequency noise. The limit frequency of this filter is about 220 Hz, and then the signal is attenuated by 45 dB/dec.

The filter is designed in FilterPro, which enables the optimal solution for the parameters to be determined in a very efficient way. For the selected Sallen-Key topology, the proposed filter is given in Figure 2. Minor changes in the values of resistors and capacitors were made due to technical limitations, but topology and characteristics were retained.

C. *Offset*

To adjust the signal to an A / D converter, it was necessary to add offset to position the signal to its range. In order not to introduce additional power supplies, the existing power supply of the operational amplifiers was used, but it needed to be adjusted. A buffer block and two resistors of 1 kΩ and 1.2 kΩ represent a voltage divider that adjusts the required offset.

In Figure 3, a red signal is representing input in simulation, its amplitude is 0.1 V and frequency 50 Hz. A signal at the exit is displayed as green, amplified and raised.

Figure 4 shows how different frequency signals behave when they pass through the system. It can be noticed that the signal weakens after its frequency becomes higher than the limit.

Figure 5 shows the exact limit frequency as the spot where the signal is experiencing 3 dB attenuation.

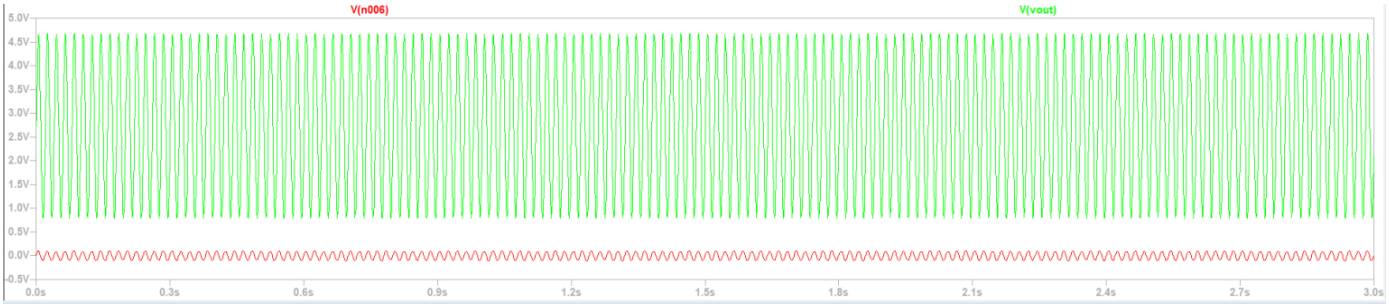


Fig. 3. Results of the transient analysis

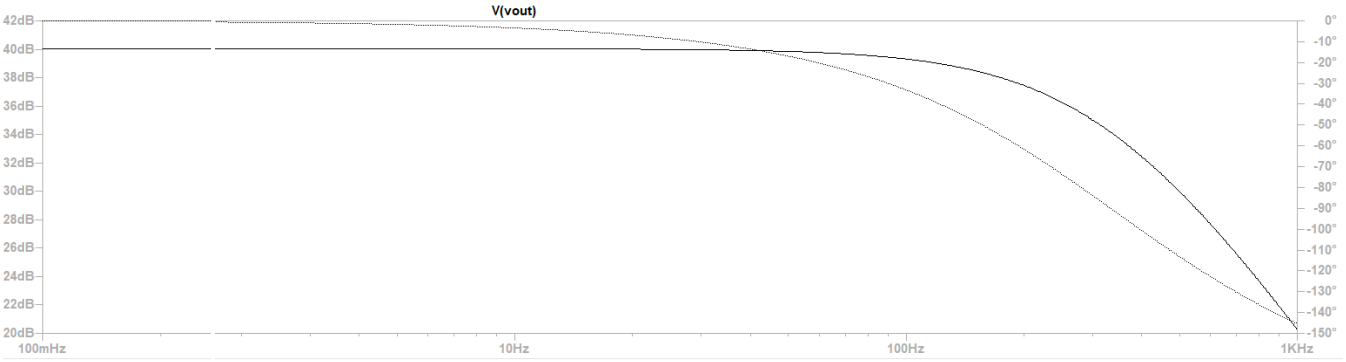


Fig. 4. Results of the AC analysis

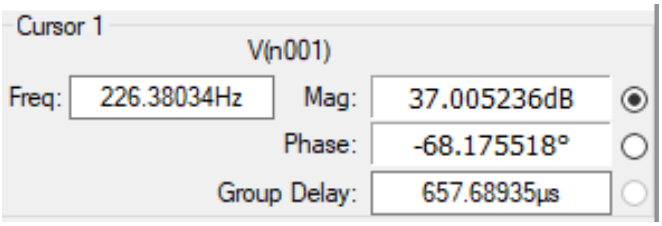


Fig. 5. Position of the cursor

III. Practical Implementation

Figure 6 shows the practical realization of the electrical scheme of Figure 1. A part of the scheme framed by a rectangle is a voltage divider with a buffer and it is adding an offset to the output signal. In the ellipse there is an amplification block, a non-inverting amplifier, and in the triangle the low-pass filter.

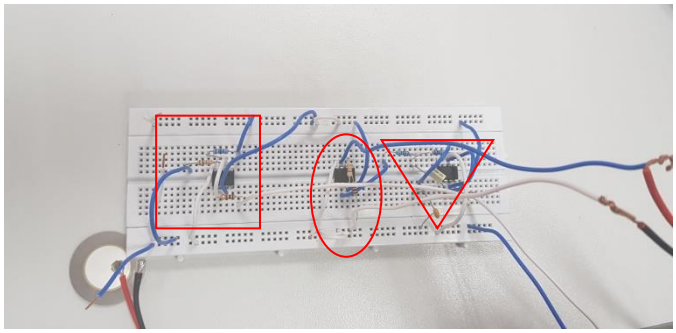


Fig. 6. Electric circuit scheme implemented on a protoboard

In addition to the highlighted parts, the sensor can also be seen in the lower left corner of the image, as well as contacts for connection with voltage +12 V, ground and voltage -12 V, along the left edge of the protoboard, respectively. On the right side of the image, there are contacts for the oscilloscope.

IV. Measurement Results

The signals recorded from the oscilloscope are preserved as a text file, and later plotted in the MATLAB. Figure 7 shows the displayed signals before and after processing. The red-colored signal represents the basic initiative obtained directly from the sensor, without any previous pre-processing.

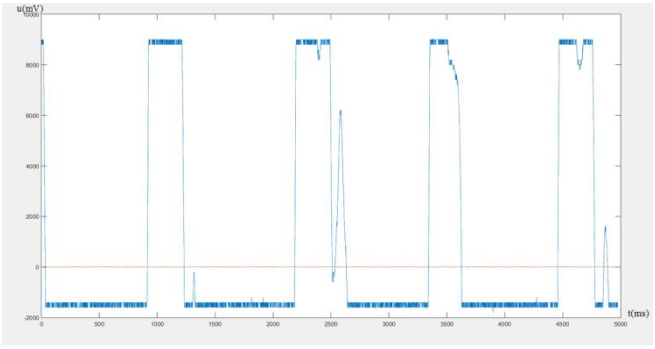


Fig. 7. Comparison of the signal from the sensor and the signal after processing

The blue signal is the same signal transmitted through the amplifier and filter system. It can be noted that the values of the original signal are very close to zero (values are of the

order mV), while the output signal is much clearer, higher values (order V) and is adjusted to further process of counting peak values of the signal.

V. Conclusion

Although it is possible to establish a connection between the contraction of individual chambers of the heart and the signals obtained by the piezoelectric sensor [2], this functionality has not been successfully demonstrated in this paper. The main problem was the hypersensitivity of the sensors and the selection of the spot of its placement on the body. On the other hand, it was shown that even though theoretical schemes cannot be performed with perfect accuracy in practice, it is possible, without significant impact on the result, to approximate the calculations. The paper presents the principle of obtaining signals from the pulsating blood vessel and its processing with the aim of determining the heart rate, but similarly, many systems used to gather signals describing other physical phenomena can be performed.

VI. References

[1] Al Ahmad M., Ahmed S. (2017). Heart-rate and Pressure-rate Determination using Piezoelectric Sensor from the Neck

[2] Taradeh N. Al, Bastaki N., Saadat I., Al Ahmad M. (2015, March 15). Non-invasive piezoelectric detection of heartbeat rate and blood pressure

[3] Description of Piezoelectric Ceramics [Web log post]. Retrieved February 2, 2019, from https://www.murata.com

[4] Keim, R. (2018, October 15). Understanding and Modeling Piezoelectric Sensors [Web log post]. Retrieved February 2, 2019, from https://www.allaboutcircuits.com/technical-articles/understanding-and-modeling-piezoelectric-sensors/

[5] Licen, H. Piezo effect and its application [Web log post]. Retrieved February 2, 2019, from http://www.trcpro.rs/dokumentacija/PDF/clanci/PiezoTehnologija.pdf

[6] Analog Devices. LT Spice [Web log post]. Retrieved February 2, 2019, from https://www.analog.com/en/design-center/design-tools-and-calculators/ltspice-simulator.html

[7] Analog Devices. Ultralow Offset Voltage Operational Amplifier, Data sheet [Web log post]. Retrieved February, 2, 2019, from https://www.analog.com/media/en/technical-documentation/data-sheets/op07.pdf

Amplifier for measurement of EMG voltage

Natalija Vukosavljević, Đorđe Novaković

***Abstract*—Measurement of low-level voltages is one of the challenges in metrology. The representatives of these voltages in biomedical measurement problems are EEG, ECG, EOG and EMG voltages. Typical approach of EMG signal measurement is measurement in a point approach. This paper proposes the design of the amplifier convenient for research system intended for measurement over an interval approach. The design is evaluated by simulation and by prototype based measurement.**

***Index Terms*—measurement; metrology; electromyography; signal conditioning.**

I. Introduction

Measurement of low-level voltages is one of the challenges in metrology. Some of the main representatives of these voltages in biomedical measurement problems are EEG (electroencephalography), ECG (electrocardiography), EOG (electrooculography) and EMG (electromiography) voltages. Electromiography is the measurement method intended for the recording of electrical activity of muscles. The advantage of electromyography is its ability to detect in vivo forces during physical activity and to quantify various pathological muscle activities. EMG signal is a nonstationary signal which can be measured by using needle electrodes or surface electrodes, the appropriate conditioning and digitizing circuits.

Standard approach of EMG signal measurement is measurement in a point approach[1]. This paper contributes with the design of the amplifier convenient for research system intended for measurement over an interval approach[2-5].

II. Model and Simulation

The designed model of theEMG amplifier consists of a protection circuit, an instrumentation amplifier, a non-inverting amplifier, a bandpass and a notch filter (Figure 1).

Fig. 1. EMG amplifier structure.

The purpose of the protection circuit (Figure 4) is to protect electrical components by limiting unwanted voltage which is a consequence of electrostatic discharge or appearance of pre-voltage.

The protection circuit consists of two resistors and of two Zener diodes connected in an anti-series way. When overvoltage appears one of Zener diodes will limit the voltage due to Zener breakdown, because of its inverse polarity, while the second diode will be directly polarized and its voltage will be 0.6 V. The chosen Zener diode has Zener breakdown for the value of voltage 4.7 V.

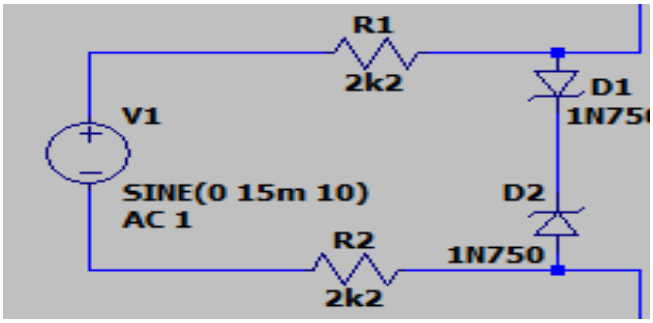


Fig. 2. Protection circuit model.

The purpose of the instrumentation amplifier is to amplify EMG signal and to eliminate common mode signal. In this design INA 122P is used, and more details about the performance of this amplifier can be found in [4].

The non-inverting amplifier (Figure 3), similar to a buffer, provides high input impedance, but also the signal amplification. This amplifier presents the second stage of amplification, and its gain is 10.

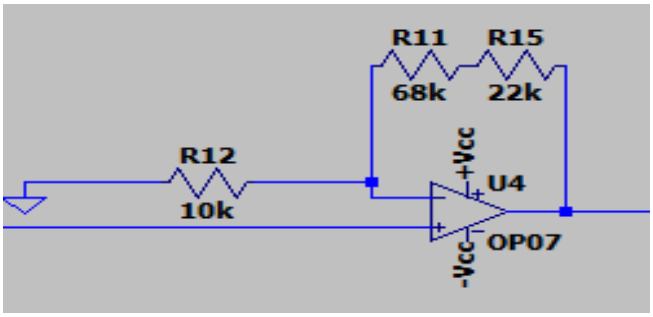


Fig. 3. The non-inverting amplifier.

A bandpass filter is obtained by connecting one active highpass filter and one active lowpass filter. The main advantage of an active filters is to isolate input from the electronic circuit, which is obtained by using operational amplifiers(OP 07).

In this implementation the bandpass filter is obtained by series connection of HP filter (cutoff frequency is 10 Hz) and LP filter (cutoff frequency is 10000 Hz). HP and LP filters are second-order filters with Butterworth design and Sallen-Key topology. The filter used in this design and its amplitude-frequency characteristic are shown in figures 4 and 6.

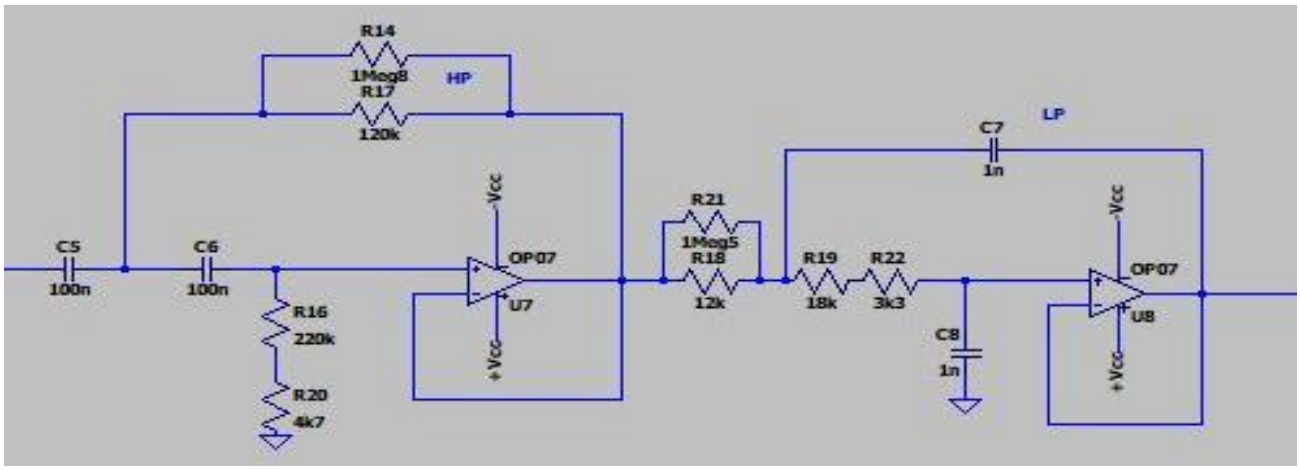


Fig. 4. Designed bandpass filter.

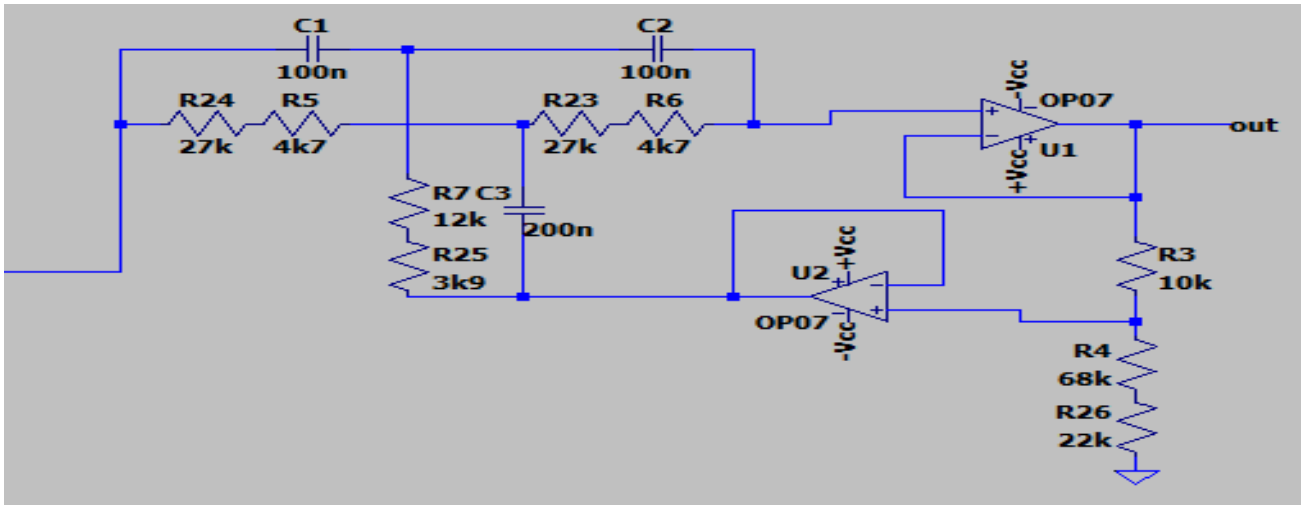


Fig.5. Active twin-t-notch filter.

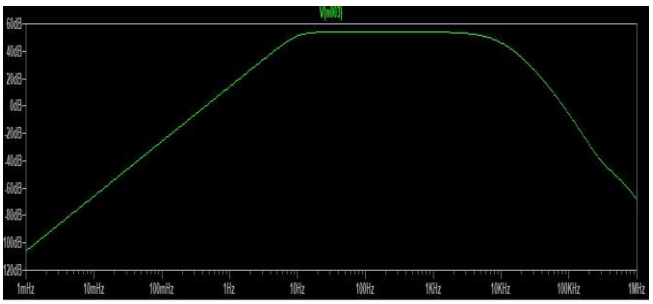


Fig. 6. The amplitude-frequency characteristic of the designed bandpass filter.

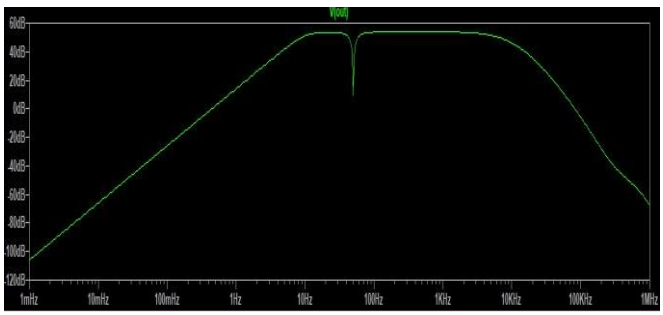


Fig.7. The amplitude frequency characteristic of an amplified signal after passing by bandpass and notch filter

The notch filter, as a special case of the bandstop filter, has the role to eliminate the line hum. It is designed as an active twin –t-notch filter (Figure 5). Amplitude -frequency characteristic after the notch and bandpass filter is shown in Figure 7.

III. PROTOTYPE AND MEASUREMENT RESULTS

Prototype is implemented on breadboard (Figure 8).

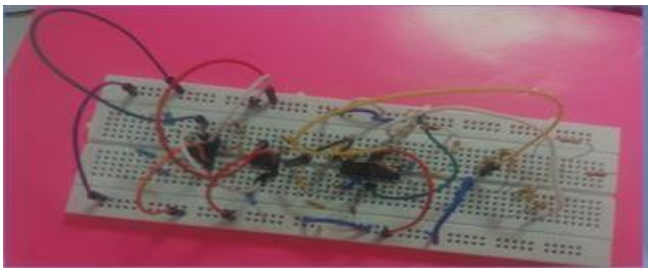


Fig.8. Prototype implementation.

The activity of muscles was measured by using surface electrodes. EMG signal measured during muscles contraction is shown at Fig. 9.

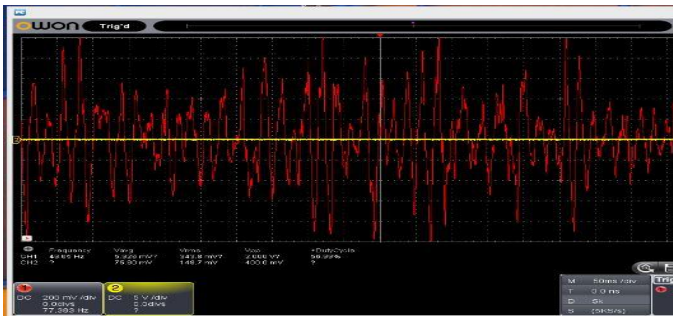


Fig. 9. EMG signal measured during muscles contraction.

IV. CONCLUSION

Typical approach of EMG signal measurement is measurement in a point approach. This paper proposes the design of the amplifier convenient for research system intended for measurement over an interval approach. The design is evaluated by simulation and by prototype based measurement. The next step is to integrate the amplifier with other circuits for measurement over an interval approach.

REFERENCES

[1] Mirjana Popović, Milica Janković, Dejan Popović, "Biomedicinska merenja i instrumentacija"

[2] Sovilj P. M., Milovančev S. S., Vujičić V.: Digital Stochastic Measurement of a Nonstationary Signal With an Example of EEG Signal Measurement, Instrumentation and Measurement IEEE Transactions on, 2011, Vol. 60 - issue 9, pp. 3230-3232, ISSN 0018-9456, DOI: 10.1109/TIM.2011.2128670

[3] P. Sovilj, M. Milovanović, D. Pejić, M. Urekar, Z. Mitrović, Influence of Wilbraham-Gibbs Phenomenon on Digital Stochastic Measurement of EEG Signal over an Interval, pp. 270-278, Measurement Science Review, Vol. 14, No. 5, 2014, ISSN 1335 – 8871

[4] Platon Sovilj, Nebojša Pjevalica, FPGA based model of processing EEG signal, 17th Telecommunications Forum TELFOR 2009, pp. 677-680, Serbia 24.-26. November 2009, ISBN: 978-86-7466-375-2

[5] Jelena Djordjevic-Kozarov, Platon Sovilj, Dejan Mitic, Vladimir Vujicic, Dragan Radenkovic, Model Development for Digital Stochastic Measurement of Noised EOG Signals, pp. 417 - 420, 49th International Scientific Conference on Information, Communication and Energy Systems and Tehnologies - ICEST 2014, Niš, 25.06.–27.06.2014, ISBN 978-86-6125-109-2

Analysis, circuit and firmware design for GSR signal acquisition

Rosa Ružičić, Đorđe Novaković

***Abstract*—In this paper we propose analysis and analog circuit design for GSR (galvanic skin response), microcontroller firmware used for data acquisition which communicates data to the computer via UART line for the further processing as well as the potential GSR use case.**

***Index Terms*—GSR; microcontroller; interrupt; UART; measurement.**

I. Introduction

Galvanic skin response, i.e. electrodermal activity (EDA) represents the change in skin humidity which comes as an effect to emotional stimulation. Skin humidity increases independently of whether it is happiness, sadness or any other emotion [1]. Increased skin humidity means better conductivity, that is, lower resistance we detect by connecting electrodes to the skin. Galvanic skin response finds its application in polygraph, popularly referred to as a lie detector.

II. About Galvanic Skin Response

Galvanic skin response (GSR) is one of the most sensitive measurements for emotional excitement. It comes from autonomic activation of sweat glands in the skin. Sweating happens just like an effect on our emotional state. Whenever we are emotionally excited our glands start excretion and we use GSR to detect it. Our body has about three million sweat glands. The density of sweat glands varies markedly across the body, being highest on the forehead and cheeks, the palms and fingers as well as on the sole of the feet.

Sweat secretion, blood pressure, heart rate, body temperature and a lot of other processes in our body we cannot control by consciousness. Our nervous system has two sections: sympathetic and parasympathetic nervous system. Sympathetic nervous system represents a rapid response, it activates when we are in stress situation, informally said "fight or flight". Parasympathetic nervous system represents slow response, it activates when we rest without excitement, and denotes „feeding and breeding" state.

Rosa Ružičić is with the Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 7, 21000 Novi Sad, Serbia (e-mail: rosaruzicic@gmail.com).
Đorđe Novaković is with the Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 7, 21000 Novi Sad, Serbia (e-mail: djordjenovakovic@uns.ac.rs).

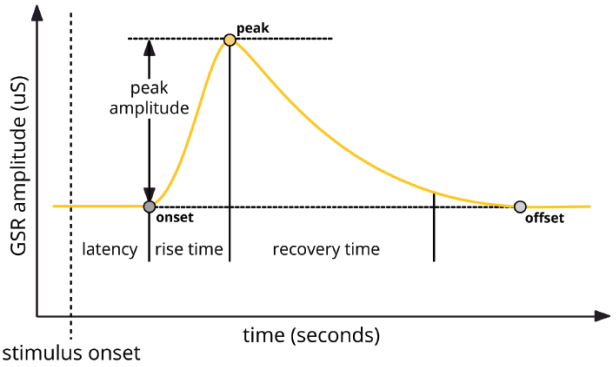


Fig. 1. GSR signal segments [2]

On the Figure 1 is shown idealized GSR signal. First part of signal (latency) includes the time duration from stimulus onset to the "GSR burst", which represents the time interval between onset and offset points in Figure 1. Peak amplitude is the GSR amplitude difference between peak and onset points. Rise time is the duration from onset to peak. Recovery time represents the duration from peak to end of "GSR burst" [3].

III. Analog circuit design

In this section, we will describe how we designed an analog circuit for GSR signal acquisition. This analog circuit is used to remove all frequencies below 0.48 Hz and above 4.8 Hz. After this step signal will be amplified by the factor of 100, then signal will be sent to the comparator, which will prepare signal for acquisition. This circuit is interesting because it detects the change in resistance. The amplitude of the signal is not of importance, we are interested in finding out whether change has happened and when its happened.

In order to build the circuit, it was necessary to draw its model. Schematic for the analog circuit was designed in LTSpice software, using model found online [2].

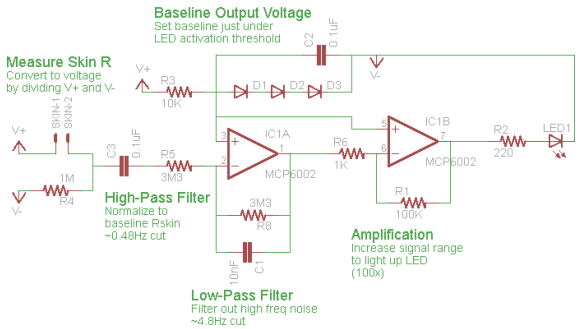


Fig. 2. Analog circuit found online

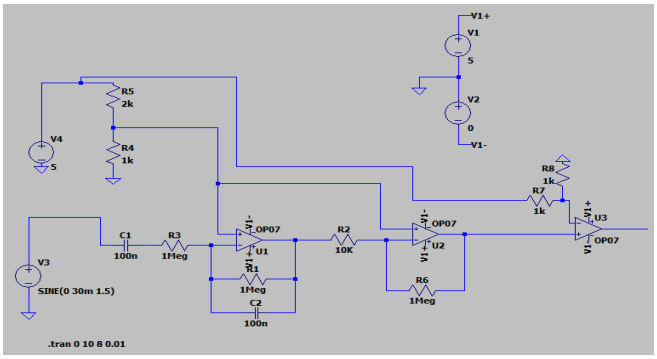


Fig. 3. LTSpice schematic

Further explained, on the input of the circuit (Figure 3) we feed AC signal, 30 mV in amplitude and of 1.5 Hz frequency. The AC signal represents measured variable skin resistance. Signal is then fed to a high-pass filter of cut-off frequency equal to 0.48 Hz. This filtered signal is sent to inverting input of a low-pass filter with cut-off frequency of 4.8 Hz, while the non-inverting input is fed with 1.6 V signal, coming from voltage divider. Amplifier‘s role is to amplify the difference of the filtered signal and the signal coming from the voltage divider, by the factor of 100. Amplified signal is then sent to a comparator, as mentioned earlier.

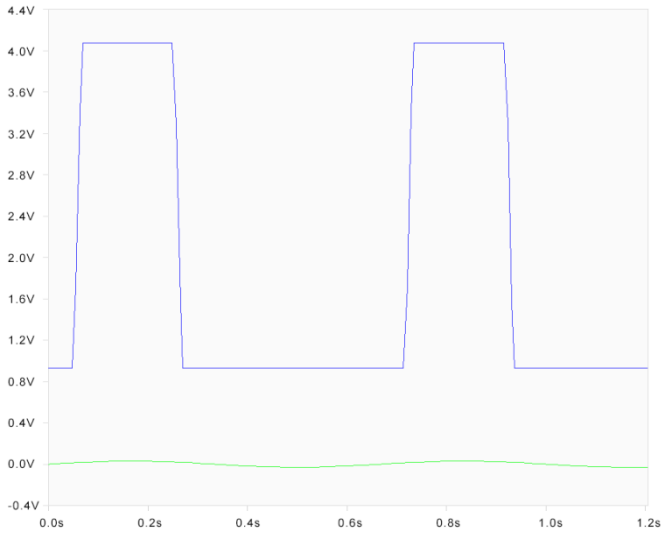


Fig.4. Circuit simulation

On the Figure 4, input signal is displayed in green while the comparator output is in blue color.

A. *Building the analog circuit*

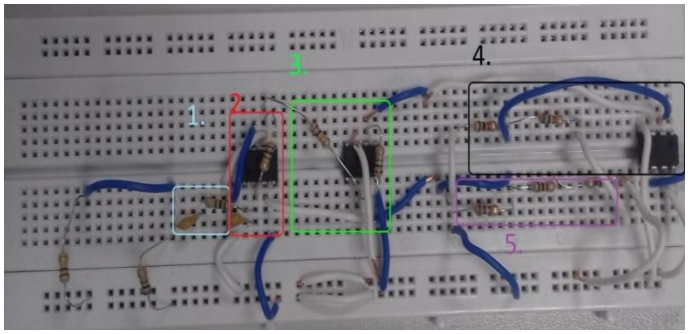


Fig. 5. Electrical circuit on protoboard.

First marked block (Figure 5) represents the high-pass filter. It consists of a C = 100 nF capacitor and a R = 1 MΩ resistor. Second marked block represents the low-pass filter, consisting of OP07 operational amplifier, R = 1 MΩ resistor and C = 100 nF capacitor which are connected in parallel and placed in a negative feedback loop of the operational amplifier. Third marked block represents amplifier consisting of OP07 and $R1$ = 10 kΩ, $R2$ = 1 MΩ resistors. Second block‘s output is fed to the inverting input of the amplifier, while the non-inverting input is fed with the fifth block‘s output.

$$A = \frac{-R2}{R1} \quad (1)$$

Fourth block represents the comparator consisting of OP07 and $R1$ = $R2$ = 1 kΩ resistors. Fifth block represents voltage divider consisting of three R = 1 kΩ resistors. The voltage is measured between the second and the third resistor of the voltage divider. The resistor in a series circuit made of the resistor and the first block is used as a reference, i.e. we use it to measure change in resistance.

B. *Simulation using oscilloscope*

The electrical circuit is supplied by +5 V DC voltage. Due to a shortage of electrodes, we used R = 10 kΩ resistor. Negative peak represents a sudden increase of resistance, while positive peak means resistance decrease (Figure 6). Interesting thing about this circuit is that it detects instant change only, after which the circuit returns to equillibrium state, i.e. it maintains constant voltage peaking 1.6 V.

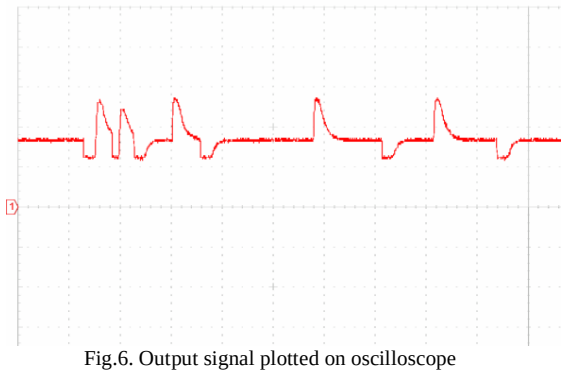


Fig.6. Output signal plotted on oscilloscope

IV. DATA ACQUISITION FIRMWARE

A. *Interrupt*

Microcontroller that we used has multiple interrupt sources and a feature which allows them to be defined as high-level or low-level priority interrupts. Events related to high-priority interrupts will intercept any lower priority interrupts that may happen in a meantime. Interrupt sources have three control bits:

1. Flag bit – Gets set when the interrupt occurs,
2. Enable bit – Enables program section execution while the flag bit is set, i.e. flag = 1,
3. Priority bit – Determines whether the bit is high or low priority. [4]

B. *UART communication*

UART (Universal asynchronous receiver/transmitter) is a computer hardware device for asynchronous serial communication in which the data format and transmission speeds are configurable. UART converts data from parallel to serial format, thus decreasing number of communication lines. Transceiver converts the parallel data into serial array of bits, while the receiver converts them back into parallel format. Two pins are used for communication, Tx and Rx, i.e. transceiver and receiver pin. UART may provide two-way communication, full duplex, which allows it to simultaneously transmit and receive data. Communication pin operates using power supply voltage levels, "1" = 5 (3, 3) V and "0" = 0 V. Easiest way to connect two microcontrollers via serial communication is to connect Tx1 to the Rx2 and Tx2 to the Rx1. It is important to define the speed of communication in order to ensure the synchronization between the transceiver and the receiver sides. Communication speed is measured in bps (bits per second) unit, and it's possible to choose among predefined values: 9600 bps, 57600 bps, 115200 bps, etc. Data package (Figure 8) starts with zero and signals the receiver side about incoming data, it's followed by the actual data and the parity bit. Parity bit is not used that often but it gets useful in applications with noisy buses. The transmitting device sets the parity bit. If the parity is set for even, the transmitting device will put a 0 in the parity bit if there is an even number of 1's in the data bits. This makes the number of 1's including the parity bit even [6]. Stop bit is always logic 0. Communication (Figure 7) is conducted via two independent lines, Tx and Rx. Data is always being transmitted in the same direction, to the computer.

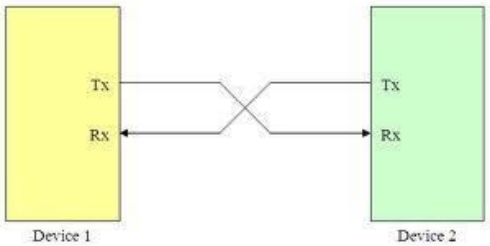


Fig. 7. Data transmission between two devices

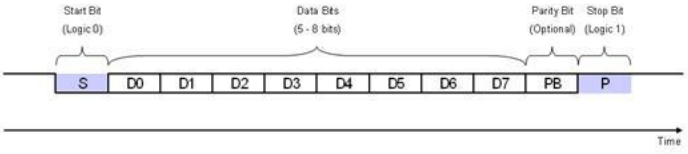


Fig. 8. Register layout

C. *Algorithm*

Firmware execution algorithm is shown in Figure 9. Program initializes variables first, then it enters super-loop, i.e. infinite loop which checks if interrupt occurred, if not, loop runs until it eventually happens. When interrupt occurs, flag is set to 0. Then it counts how many times interrupt has happened and prints change counter value in UART terminal. Figure 10 shows the execution code which runs on PIC18F45K22 microcontroller.

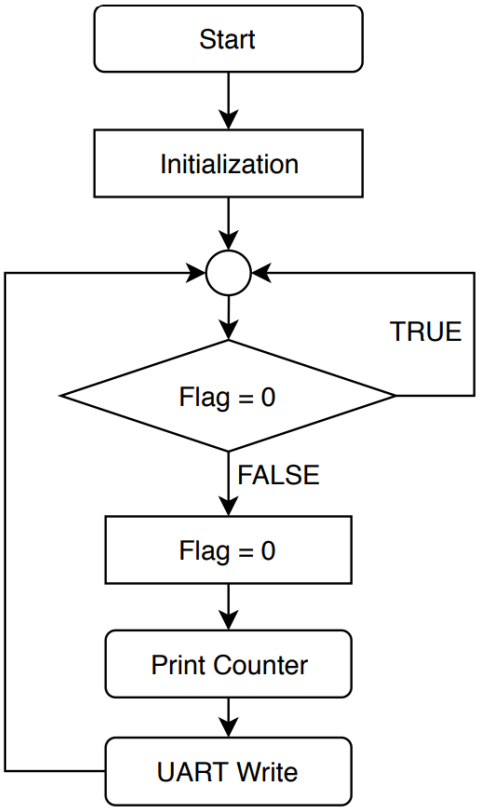


Fig. 9. Algorithm flow diagram

```
unsigned int counter = 0;
char flag;
interrupt () {
  if(INT0IF_bit) {
    INT0IF_bit = 0;
    counter++;
    flag = 1;
  }
}

static void initMain(){
  UART1_Init(9600);
  TRISB.B0 = 1;
}

void main() {
  char txt[40];

  INT0IE_bit = 1;
  GIE_bit = 1;

  // bit current_state;

  initMain();
  // current_state = 0;

  while (1) {
    while (!flag);
    flag = 0;
    sprintf(txt, "Broj promjena: %d \r\n", counter);
    UART1_Write_Text(txt);
  }
}
```

Fig. 10. Execution code, runs on PIC18F45K22 microcontroller.

V. Conclusion

We measured the changes to the skin's wetness. The effect of the increase in skin's wetness is the decrease in the resistance and we can detect change in voltage at the end of the electrical circuit. The change in resistance itself is of great importance because it shows us that there is a reaction of the human body. Then we process and amplify the signal. With a microcontroller the data is sent on the computer and the results are shown, reviewing whether change happens or not. When we connect galvanic skin response with the sphygmomanometer, heart rate monitor and breathing monitor we can make a polygraph (lie detector). Also, we can use GSR for other applications like stress level detection by using it in combination with speech [5]. In medical applications this will be very useful, because skin moisture can get a lot of information about our health. With GSR we can measure excitement levels in children with autism [1].

Acknowledgment

We would like to thank Prof. Dr. Platon Sovilj from Faculty of Technical Sciences, University of Novi Sad for his suggestion of completing this paper.

References

[1] T. Westeyn, P. Presti, and T. Starner, "ActionGSR: A Combination Galvanic Skin Response–Accelerometer for Physiological Measurements in Active Environments", 2006.
[2] http://produceconsumerobot.com/truth/
[3] https://imotions.com/wp-content/uploads/Guides/iMotions_Guide_GSR_2015.pdf
[4] http://ww1.microchip.com/downloads/en/devicedoc/39960d.pdf
[5] H. Kurniawan, A. V. Maslov, M. Pechenizkiy, "Stress Detection from Speech and Galvanic Skin Response Signals", 2013
[6] https://www.newbiehack.com/ShowVideoClip.aspx?id=851

Measurement in Fourier domain – a Natural Method of Big Data Volume Reduction

Vladimir Vujicic[1], Matija Sokola[2] Aleksandar Radonjic[3] and Platon Sovilj[4]

[1]Vladimir Vujicic - Entrepreneur Consultant in Electrical Engineering and Energetics, Novi Sad, Serbia
[2]University of Warwick, Warwick Manufacturing Group, Warwick, United Kingdom
[3]Institute of Technical Sciences of the Serbian Academy of Sciences and Arts, Belgrade, Serbia
[4]University of Novi Sad, Faculty of Technical Sciences, Novi Sad, Serbia

emails: vujicicv@uns.ac.rs, m.sokola@warwick.ac.uk, sasa_radonjic@yahoo.com, platon@uns.ac.rs

***Abstract*—The paper presents an idea of a measurement in a Fourier domain by a means stochastic digital measurement method (SDMM) as a natural and logical way to reduce the amount of big data in for processing in real time. The measurement method is explained and its application to the power and energy measurements in the power grid is briefly described.**

Keywords—Stochastic measurements, Fourier coefficients, Big Data, Signal Power.

I. Introduction

Our recent research indicates that it is possible to realize a discrete Fourier transform (DFT) processor that is capable of a fully parallel on-line computation of thousands of Fourier coefficients from an 1-bit or 2-bit array of stochastically dithered samples of the measured signal [1].

II. The Principle of Operation

The principle of operation is illustrated in Figures 1 and 2. Figure 1 shows a multiplication and accumulation (MAC) block, in which a signal f_1 (with superimposed noise n) is digitized, multiplied with a pre-stored signal f_2 and the product integrated, all within a single processor cycle. At the end of the measurement interval, division of the Counter 1 value (the integral) by the Counter 2 value (the number of samples) gives the appropriate Fourier coefficient.

Figure 2 shows a parallel processing of f_1 by $2M$ MAC blocks, thus obtaining M Fourier harmonics at the end of the measurement interval.

Every pair of coefficients (a_j, b_j) defines a signal harmonic during the measurement interval. The harmonic amplitude is an inherent harmonic descriptor for the measurement interval. The above facts have three key consequences:

1. Weierstrass approximation theorem [2] and its trigonometry polynomial are no longer relevant and are replaced by the Fourier series (integral) [3].

2. Fourier analysis is now not limited to periodic signals only, i.e. treatment of arbitrary signals becomes practically viable, and

3. There is a problem of on-line determination of the validity of a specific (a_j, b_j) pair, i.e. of a harmonic j.

Due to the consequence 1, the input signal initial value does not need to be same as the end value – these values can be arbitrary as long as they are finite. Therefore the Bernstein polynomials [4] as well as generic polynomial approximations become much less relevant.

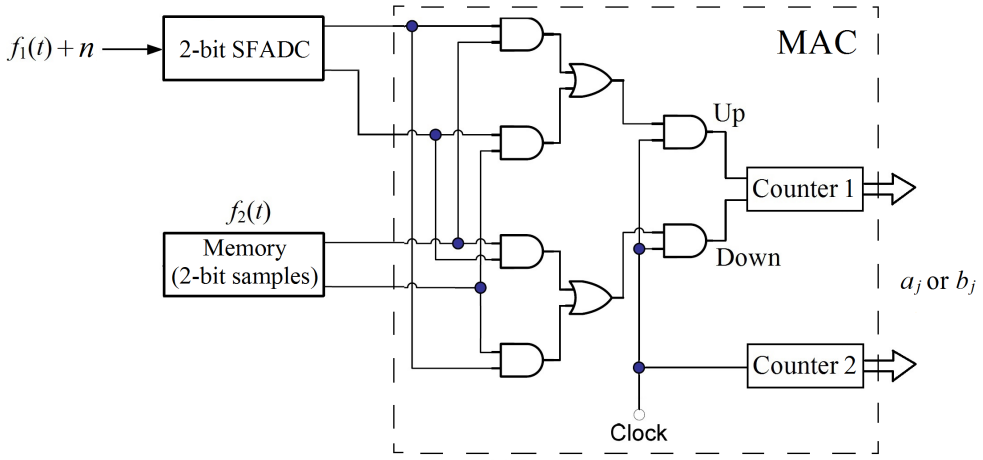


Fig. 1. An optimal scheme for measuring a single Fourier coefficient with a two-bit stochastic flash A/D converter.

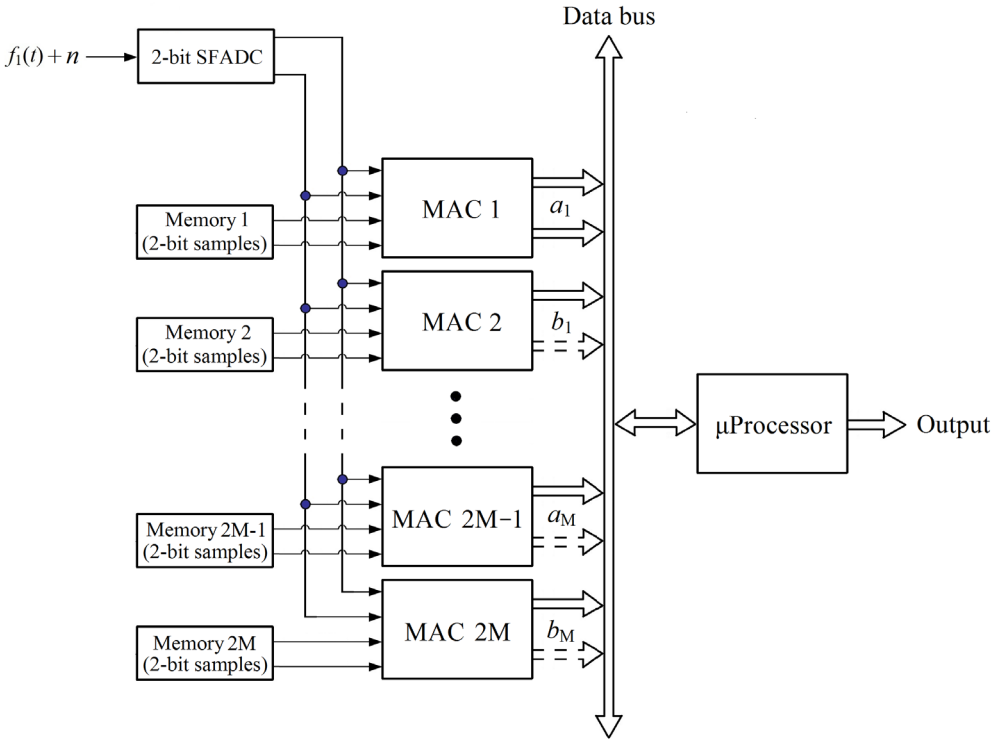


Fig. 2. A stochastic DFT processor for measuring 2*M* Fourier coefficients.

The Consequence 2 enables analysis of any continuous signal with a finite number of first-order discontinuities, via piece-wise analyses, on an arbitrary time interval [3].

The Consequence 3 indicates that an efficient and smart algorithm to quickly estimate the validity of a specific (*aj*, *bj*) pair is needed. Such an algorithm can be developed on the basis of a detailed insight into the hardware structure of the presented DFT processor. It features a simple and robust design, as well as accuracy, precision and speed [5]-[8]. The performance of first versions of such an algorithm is promising and encouraging.

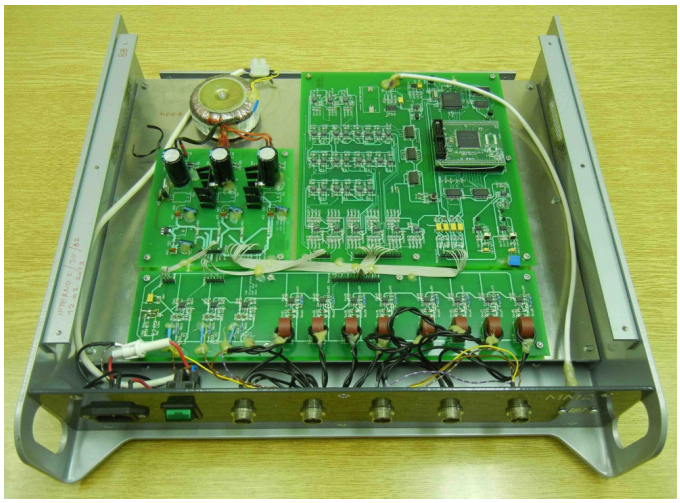


Fig. 3. A two-bit double three-phase power analyzer implemented in FPGA technology.

If such a device entailing (1-bit or 2-bit A/D converter with DFT processor and fast estimator of harmonic's validity) is to be used for processing of big data, it is natural to process and store only the important data. This means that overall volume of processing as well as memory needs to be minimised. This is not the only benefit of the proposed approach/method. The Fourier coefficients in the DFT processors, shown in Figure 2, are calculated in counters that have an inherent pipelining feature and do not suffer from carry propagation problems [9]. This can enable high data processing speeds. It is possible, therefore, to design a custom processor, tailored for the described use. The use of FPGA [10] offer many possibilities for the implementation of both prototypes, as well as a batch production. Some additional Internet of Things requirements, such as monitoring and control of household [11], may justify the use of ASIC implementations.

It should be noted that a single datum that is a measure of the observed signal (for processing and/or storage) is not a real number any more. A quantum of signal information is dominantly a complex number in a frequency domain that also contains information on its validity over the observation time interval. A finite set of such data that include the "validity" faithfully describe the signal spectrum over the measurement interval.

Signal reconstruction and/or further processing of such data is not a problem, because the equivalency of time domain and frequency (Fourier) domain is well known. The Inverse Fast Fourier Transformation (IFFT) implemented in a large number of software tools can be utilised for transition from Fourier to time domain. Our research has confirmed that use of FPGA

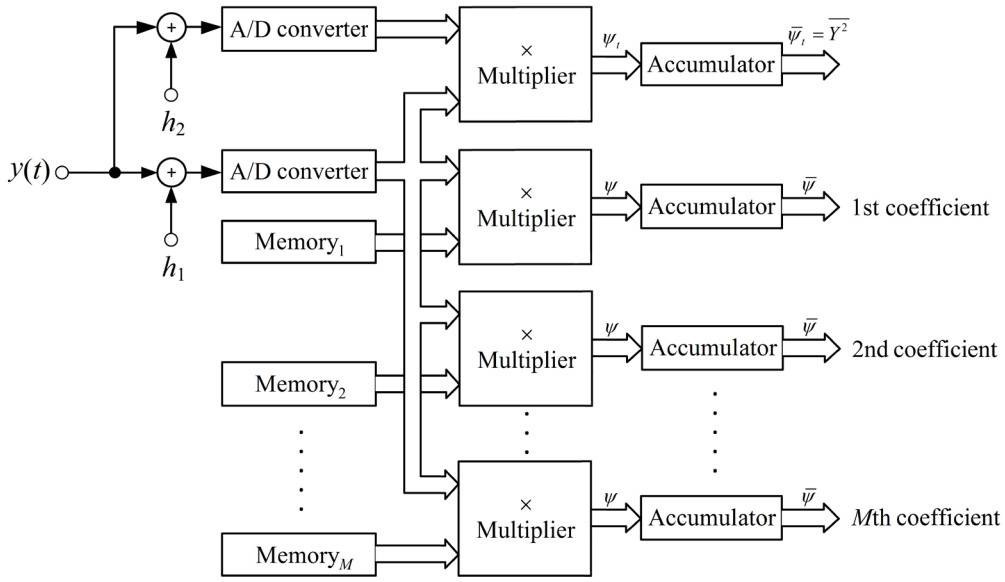


Fig. 4. A hardware scheme for on-line determination of the validity of a specific (a_j, b_j) pair, i.e. of a harmonic j.

technology enables fast and efficient experimental validation of the new approaches proposed here. Based on it, powerful measurement devices, such as the three-phase analyser and energy meter shown in Figure 3, have been developed.

Several additional fine details need to be further explored in this synergic frequency-time treatment of signal measurement and processing to make it equally efficient in both domains. We need to find answers to the following questions:

a. an optimal number of significant harmonics,
b. optimal sampling frequency of the reconstructed signal,
c. an acceptable number/size of first-order discontinuities,
d. an optimal length of the measurement interval,
e. an optimal precision of the reconstruction, etc.

The main goal of the paper is to define a role for SDDM methodology in the reduction of the volume Big Data (BD) that is independent from the BD generator, which leads to the need for a standard sampling method. The underlying motivation is based on the fact that SDDFT output can process data in real time and immediately provide naturally weighted components. In other words, non-critical results can be immediately discarded and only the critical ones are stored.

As an example of defining a specific problem and finding its optimal solution, let us consider case a): the uppermost channel in the instrument shown in Figure 4 measures the average signal power over the measurement time interval. Other channels shown in Figure 4 measure M Fourier coefficients over the same time interval. When all the coefficients are squared and then sorted in a descending order, they represent the average signal power. It is possible to approximately calculate, within a pre-defined accuracy, the average signal power by including only K ($K < M/2$) most significant harmonics. The optimisation criterion is to determine the lowest number of harmonics that satisfy the prescribed accuracy level. This criterion is common in measuring power and energy in a power grid. Similarly, problems denoted under b), c), d), e) require a clear initial brief definition in order to reach a solution. These are currently under research.

III. Conclusion

The paper presented the principle of operation of a novel approach to reducing the big data volume by means of a stochastic digital measurement method. This can be achieved by measuring harmonics that are naturally weighted, which opens a possibility to determine harmonics' validity in real time, before further processing and/or storage. This is not possible to do in time domain with data that can be obtained using the standard sampling method approach.

This novel approach opens up an array of new but solvable problems. One of these – determining an optimal number of harmonics needed in power measurements – is solved and described in the paper. The use of FPGA technology offered a great possibility for implementation and an efficient experimental verification.

The important characteristics of harmonics measurements are the adaptive precision and high accuracy, stemming from the elimination of systemic errors in hardware. Since the measurement results obtained by the proposed approach depends on the signal waveform, a further research effort will be invested in the development of applications for specific signal classes.

The use of IFFT provides simple and fast transition from frequency to the time domain – especially when a small number of significant harmonics is involved – thus enabling the use of a number of software solutions tailored for time-domain processing.

ACKNOWLEDGMENT

This work was supported in part by the Ministry of Education, Science and Technological Development of the Republic of Serbia under research grant No. TR32019 and supported in part by the Provincial Secretariat for Science and Technological Development of Autonomous Province of Vojvodina (Republic of Serbia) under research grant No. 114-451-2800-2016-02.

References

[1] D. Pejic et. *al*, "Stochastic digital DFT processor and its application to measurement of reactive power and energy," *Measurement*, vol. 124, pp. 494-504, Aug. 2018.

[2] K. Weierstrass, "Ueber die analytische Darstellbarkeit sogenannter willkuerlicher Functionen einer reellen Veraenderlichen„" Sitzungsber. Akad. Berlin, 1885, pp. 633-639, 789-805, 1885.

[3] G. H. Hardy and W. W. Rogosinski, *Fourier series*, Dover Publications, Inc., Mineola, New York, 1999.

[4] G. G. Lorentz, *Bernstein polynomials*, AMS Chelsea Publishing, 2012.

[5] V. Vujicic, "Generalized Low-Frequency Stochastic True RMS Instrument," *IEEE Trans. Instrum. Meas.*, vol. 50, no. 5, pp. 1089-1092, Oct. 2001.

[6] V. Vujicic, I. Zupunski, Z. Mitrovic, and M. A. Sokola, "Measurement in a Point versus Measurement over an Interval", Proc. IMEKO XIX World Congres, no. 480, pp. 1128-1132, Sep. 2009.

[7] V. Pjevalica and V. Vujicic, "Further Generalization of Low-Frequency True- RMS Instrument," Proc. IEEE Instrumentation and Measurement Technology Conference (IMTC) 2005, May 2005, pp. 1008-1011.

[8] J. Djordjevic-Kozarov et. al., "A Novel Method for Gibbs Phenomenon Reduction in Stochastic Measurement of EOG Signal", Proc. IcETRAN 2016, pp. MLI1.4.1-4, Zlatibor, Serbia, Jun. 2016.

[9] K. Hwang and D. DeGroot, *Parallel processing for supercomputers and artificial intelligence*, McGraw-Hill, 1989.

[10] P. A. Simpson, *FPGA Design: Best Practices for Team-based Reuse*, Springer International Publishing Switzerland 2010, 2015.

[11] N. Vandome, *Smart Homes in easy steps: Master smart technology for your home*, In Easy Steps Limited, 2018.

LabVIEW-Arduino UNO Temperature Measuring System

Josif Tomić, *Member, IEEE*, Miodrag Kušljević, Platon Sovilj, *Member, IEEE*, Vladimir Rajs, *Member, IEEE*

***Abstract*—Today's modern measuring technique is based on the implementation of microprocessor-supported measurement and information systems. The low price of computing and electronic components has led to measuring devices becoming software-oriented. The main emphasis is placed on the realization of complex mathematical algorithms, over sampled physical signals that were converted into electricity or voltage. The same case applies to temperature measurements. The temperature is undoubtedly the most widely measured physical size and there is a very large number of measuring methods and sensors that can precisely measure this size. Unfortunately, many temperature sensors have non-linear characteristics, so complex numerical formulas need to be applied to get the exact values. This paper presents a microprocessor measuring device for measuring and calibrating temperature sensors from silicon. The system is characterized by simplicity, low price and satisfactory accuracy. The device was realized with the Arduino UNO card and the program is written in the LabVIEW software package, using the LIFA library of functions.**

***Index Terms*—Temperature; Arduino UNO; LabVIEW; LIFA; Silicon temperature sensors.**

I. Introduction

Temperature is without doubt the most widely measured variable. Temperature is a measure of molecular energy, or heat energy, and the potential to transfer heat energy. Thermometers can be traced back to Galileo (1595). In the process control of chemical reactions, temperature control is of major importance, since chemical reactions are temperature-dependent. It is therefore necessary, in most cases, to measure the temperature together with the physical size, so that certain mathematical corrections can be made to achieve precise measurements.

II. Temperature Sensors

Because of the high importance of temperature information, there are a very large number of sensors and measurement methods for temperature measurement. Temperature sensors could be divided into two large groups: contact and noncontact sensors [1, 2].

Josif Tomić - Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, Novi Sad 21000, Serbia (e-mail: tomicj@uns.ac.rs).

Miodrag Kušljević – Termoelektro Enel AD,Uralska 9, Beograd 11060, Serbia (e-mail: miodrag.kusljevic@te-enel.rs).

Platon Sovilj – Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, Novi Sad 21000, Serbia (e-mail: platon@uns.ac.rs).

Vladimir Rajs – Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, Novi Sad 21000, Serbia (e-mail: vladimir@uns.ac.rs).

Noncontact temperature sensors (infrared), work mostly on the optical principle and measurement is based on the ability of all materials to emit electromagnetic radiation (infrared radiation). This radiation comes from the motion of charged particles in the atom. All matter at a temperature greater than absolute zero emits heat radiation. Infrared measuring instruments use this radiation to determine the body temperature. When a body is heated, it radiates the electromagnetic energy of different wavelengths. The advantages of contactless temperature measurements are: a) the measurement time is very short (milliseconds), b) it is possible to measure the temperature on the moving objects, and c) measurements can be performed in hazardous and physically inaccessible locations (high voltage, large distance). These devices measure without problems temperatures in the range of 1300°C to 3000°C, where contact thermometers cannot be used or have a limited lifetime. Also, energy is not taken from the measurement object so that the measurements are much more accurate.

There are several groups of contact sensors for temperature measurement, and the main groups are: a) resistance temperature detectors, b) thermistors, c) thermocouples, and d) silicon-based temperature sensors.

Platinum Resistance Temperature Detectors are characterized by a very high accuracy and precision of temperature measuring of 0.1%, in the range of -200°C to 850°C. They need to be used in conjunction with XTR1xx conditioners from Burr-Brown. They give a current excitation to the sensor, amplify and linearize the output signal. These sensors are characterized by high consistency and repeatability of measurement. Unfortunately, they also have a higher price than other temperature sensors. The Pt-100 temperature-dependent resistors have a resistance of 100Ω at a temperature of 0°C and a resistance of 138.4Ω at a temperature of 100°C. Temperature change of 1°C gives a resistance variation of 0.384Ω. Also, there are sensors with a resistance of 1000Ω at a temperature of 0°C and they have a Pt-1000 label. For these sensors, the dependence between temperature and resistance changes is mostly linear, in a low temperature range, typically between 0° C and 100° C. For accurate measurements in a wider temperature range, it is necessary to use the following equation:

$$R_T = R_0 \times (1 + A \times T + B \times T^2 + C \times T^3) \quad (1)$$

Where the coefficients R_0, A, B and C are: R_0=100Ω, A=3.9083×10^{-3} C^{-1}, B=-5.775×10^{-7} C^{-2}, C=-4.183×10^{-12} C^{-3}, bellow 0°C or C = 0, above 0°C.

Thermocouples are formed when two dissimilar metals are joined together to form a junction. A current will flow in the circuit if the two junctions are at different temperatures. The current flowing is the result of the difference in electromotive force developed at the two junctions due to their temperature difference. The voltage difference between the two junctions is measured, and this difference is proportional to the temperature. This phenomenon was discovered in 1821 by Thomas Johann Seebeck. The advantages are: low cost, simple construction, high robustness, wide temperature range (-200°C to 3000°C), short response time, good accuracy and repeatability for rough measurements (+/-2%). Their disadvantages are: low sensitivity (50μV/°C), high sensitivity to noise and insufficient accuracy for precise measurements, great nonlinearity, limited work life and definition of cold junction point. A formula that defines the dependence voltage of temperature is:

$$\nabla V = -S(T)\nabla T \tag{2}$$

Where *S(T)* is a temperature-dependent material property known as the Seebeck coefficient.

Thermistor is an element (most often a metal oxide) which, when changing temperature, changes its resistance. The temperature coefficient may be both positive (PTC) and negative (NTC), and may have a resistance variation of 10% by degrees Celsius. It is used in applications where it is necessary to detect small changes in temperature, for example, only 0.01°C. The accuracy of the thermistor is approximately ten times greater than the accuracy of the thermocouple. The formula for NTC resistors is as follows:

$$R_{NTC}=R_0 \times e^{(B/T)} \tag{3}$$

Where R_0 is the resistance at the reference temperature and B is a constant which depends on the material from which the thermistor was made. Similar formula applies to PTC resistors:

$$R_{PTC}= A + C \times e^{(B/T)} \tag{4}$$

Only the platinum RTD can have better accuracy from the thermistor. The advantages of a thermistor are small physical dimensions and high nominal resistance. Small dimensions contribute to quick response, and high resistance reduces the impact of terminal resistance. Thermistors are very non-linear and have the most common logarithmic characteristic of the third order. Thermistors can be used in a temperature range of -50°C to 100°C. Thermistors are also very sensitive to strike.

Silicon temperature sensors are devices made in much the same way as transistors and integrated circuits. A primary physical property of silicon-based temperature sensors is their extreme stability over time and over extreme environmental conditions. This is because silicon is an inherently stable element, especially in its crystalline form. The crystal structure that makes up a silicon-based temperature sensor is literally *rock-solid*. Another important consideration of these types of sensors is they are fabricated using highly repeatable manufacturing processes, using fully automatic machinery and photolithographic techniques.

Measurements made with silicon-based sensors are highly consistent and very stable, which translates to low drift in the accuracy of readings over time. In comparison with traditional temperature monitoring devices, silicon-based sensors lead to high-performance over the lifetime of the product in which they are used.

One of the most prominent silicon sensors is the KTY temperature detector, manufactured by Philips. There are a whole series of these sensors that differ in their characteristics. The formula by which the temperature resistance is calculated is:

$$R_T= R_0 \times (1+A \times T+B \times T^2) \tag{5}$$

Where R_0 is the resistance at the reference temperature, and A, B are the polynomial constants.

When high accuracy or wide measurement range is required, the temperature sensor parameters are not used from the table given by the manufacturer. This applies to all temperature sensors. The reason for this is a large error in measuring the temperature, which can be even ±20%. Therefore, individual calibration is necessary. Detailed calibration is not done by the manufacturer, as it would significantly raise the price of the temperature sensors. Calibration is done in a number of points and then numerically computes the formula by which the values of all temperatures can be obtained. This formula is most commonly of the higher order polynomial. The calibration quality depends on the precision of the reference thermometer and the mathematical algorithm used to make the curve fit. The calibration is based on several models. Most commonly, this is the *General Least Square* algorithm based on *Least Mean Squares (LMS)* algorithms or some of the exponential algorithms.

III. ARDUINO - OVERVIEW

Arduino is a prototype platform (open-source) based on an easy-to-use hardware and software. It consists of a circuit board, which can be programed a microcontroller and a ready-made software called Arduino IDE (*Integrated Development Environment*), which is used to write and upload the computer code to the physical board [3]. This software package is completely free and can be downloaded from the Internet at the address: https://www.arduino.cc/en/Main/Software. The key features of this board are: a) Arduino board is able to read analog or read/write digital signals, b) Arduino board can send a set of instructions to the microcontroller via Arduino IDE and USB cable, c) Arduino does not need an extra piece of hardware (called a boot loader) in order to load a new code onto the board, d) Arduino IDE uses a simplified version of C++, making it easier to learn to program, e) Arduino board can change the firmware to match another programming language, f) it has a very low price, about 12 €, g) it has

embedded I2C, SPI and RS232 communications, h) it has a large number of hardware add-ons for Internet and Wi-Fi communication.

The heart of Arduino UNO board is high performance, low power, RISC, Atmel AVR 8-Bit processor ATmega328, whose characteristics are: a) operating voltage of 5V, b) 131 powerful instructions, c) 16 MHz clock speed, d) 32KB flash program memory, e) 2KB SRAM, f) 1KB EEPROM, g) two 8-bit timer/counters, h) 6-PWM channels, i) 6-channel 10-bit AD converter, j) programmable serial Interface, k) 23 programmable I/O pins. The look of the Arduino UNO board is shown in Figure 1.

Fig. 1. Arduino UNO board

Lately due to the enormous popularity of the LabVIEW software package, a large number of companies create many toolkits for a variety of hardware and software interfaces. All these tools are available from the Internet site: http://www.ni.com/gate/gb/GB_EVALTLKTLVARDIO/US

As this site is continuously updated it is possible to get a list of all the available interfaces. *LabVIEW VI Package Manager* can also be accessed directly from the site: http://jki.net/vipm. These programs are in general free of charge. Most importantly, Arduino board can be programmed using only LabVIEW graphical language. In the implementation of this measuring device, the LIFA (*LabVIEW Interfaces For Arduino*) program library was used so that the entire program was realized in the LabVIEW program package. This open source toolkit is made for users to create custom programs for their sensors. In order for the Arduino card to accept LabVIEW programs it is necessary to download the *LIFA_base* firmware program, using the Arduino IDE interface [4]. This is a great relief for engineers because they do not have to write a program in two programming languages (C and LabVIEW) and it is significantly easier to test the whole program. This practically means that is possible to programming *embedded* microprocessor devices with graphical language. After writing and testing the LabVIEW program it is necessary to download program to Arduino board using a USB port. This is the process of finalizing the program, and it is only necessary to disconnect the card from the computer and plug in 9V power supply.

The LIFA software function palette automatically appears in the LabVIEW program (if the installation is done correctly) and includes a large number of functions. Some of them are at a very low level and allow access to individual pins on the Arduino card, while other functions are complete programs for measuring and acquisition signals from different types of sensors. Figure 2. shows the finished program for reading analogue values from one AD converter input.

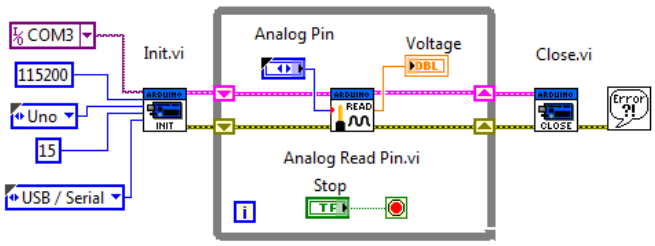


Fig. 2. Realization of LabVIEW program with LIFA interface functions

IV. Realization of Measuring System

The temperature measurement process with the silicone sensor KTY81-121 is made in the following order.

At first, calibration of the sensor is carried out by selecting the temperature range in which the unknown resistance of the sensor will be measured. The maximum temperature range is from -50°C to + 150°C. The desired temperature range needs to be divided into several parts, for example, with a step of 10°C. It is now necessary to measure the resistance of the sensor in these temperature values with a precision ohm-meter. In this way, a table is formed, which contains the values of temperature and resistance. The front panel of the finished program is shown in Figure 3. The measured values can be seen in the table on the left side of the front panel. Using the values from the table, the coefficients of the polynomial are now calculated using the method of the *Least Mean Squares* [5, 6]. As a result of this mathematical operation, *Polynomial Coefficients* are obtained, whose values are displayed on the front panel, at the right side. It can be seen that it is obtained a polynomial of a very large order. The front panel shows the graphically-fitted curve as well as the modified values of the resistance of the temperature sensor at the given temperatures. Residue returns the weighted average error of the fitted model. It seems that the fitting curve was successfully done with a very small *Mean Square Error*, *mse*=2.844. The second part of the program performs temperature measurement and is made using LIFA functions. At first, LIFA functions take the numerical value from the AD converter and calculate mathematically the resistance. The acquired resistance value is entered in the polynomial and using numerical algorithm the value of the current temperature is obtained. This value is then displayed on the front panel in numerical and graphic form.

The realized measuring device is very sensitive, so it is necessary to provide precise reference voltage for the AD converter as well as precision measuring equipment for calibration. This measuring device has a very high accuracy and precision of temperature measurement. The temperature measurement error does not exceed +/- 0.1°C throughout the measuring range, which is a very good value for this type of sensors. For high precision, sensor calibration should be performed once a year.

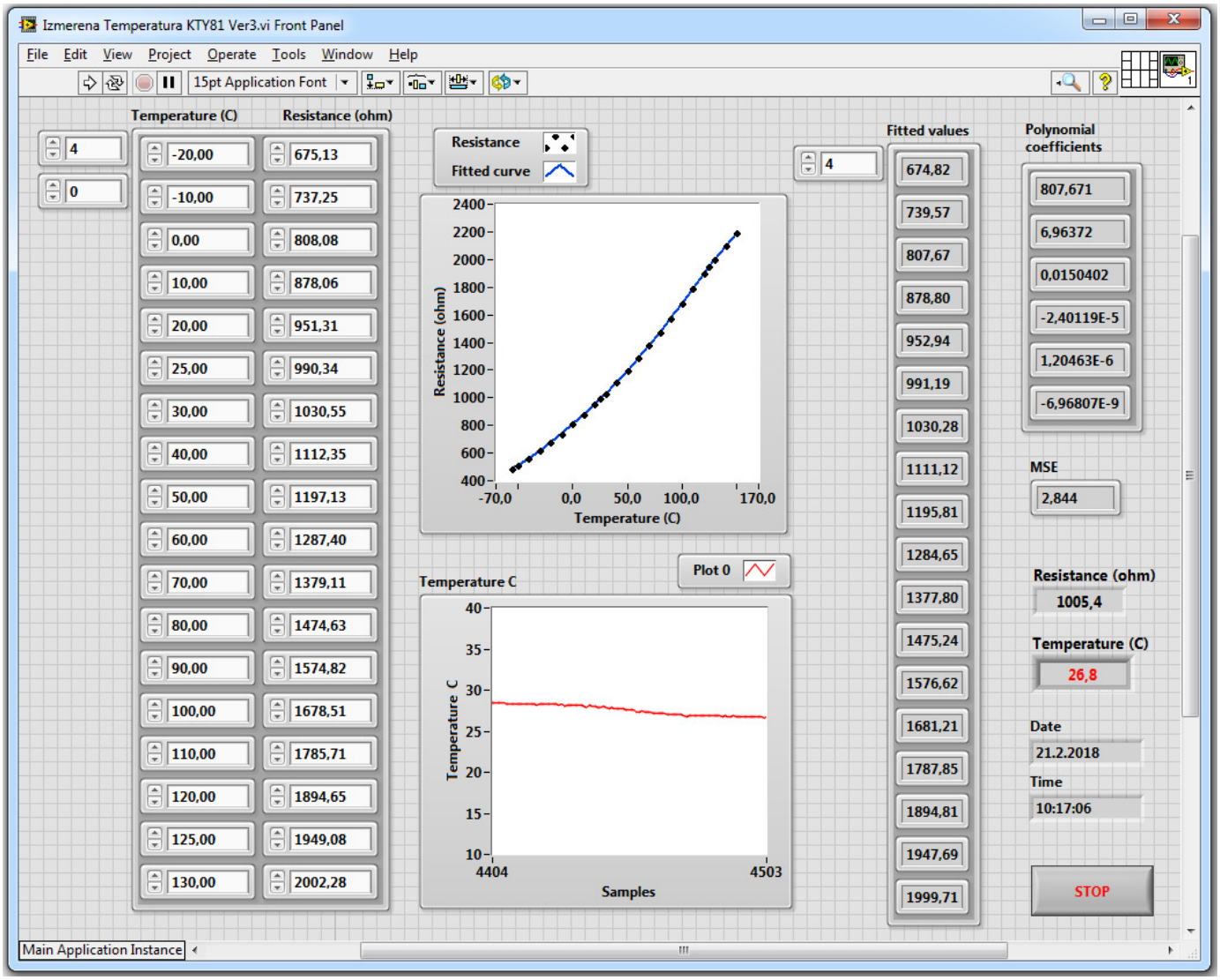


Fig. 3. Front panel of realized measurement program in LabVIEW programing package using LIFA functions

V. Conclusion

This paper presents the practical realization of a measuring device for calibration and temperature measurement using silicon sensors. With minor modifications in the program, the system could also be used to perform temperature measurements with other sensors that possess nonlinear characteristics.

Acknowledgment

This paper was financially supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia within the projects number III43008 and TR32019.

References

[1] P. R. N. Childs, J. R. Greenwood, C. A. Long, *Review of temperature measurement*, Review of Scientific Instruments, Volume 71, Issue 8, pp. 2959-2978, May 2000. doi: 10.1063/1.1305516.

[2] William C. Dunn, *Introduction to Instrumentation, Sensors, and Process Control*, Artech House, London, UK, 2006. ISBN 1-58053-011-7.

[3] Arduino, *TutorialsPoint*, [Online]. Available: https://www.tutorialspoint.com/arduino/index.htm.

[4] M. Schwartz, O. Manickum, *Programming Arduino with LabVIEW*, Packt Publishing, Birmingham, UK, 2015. ISBN 978-1-84969-822-1.

[5] J. Tomić, M. Kušljević, *Merenje i analiza signala primenom LabVIEW programa*, FTN-GRID, Novi Sad, Serbia, 2016. ISBN 978-86-7892-840-6.

[6] J. Tomić, V. Rajs, V. Milosavljević, Ž. Mihajlović, Realizacija instrumenta za merenje parametara životne sredine, 57 Konferencija ETRAN, Zlatibor, Jun 2013. ISBN: 978-86-7892-447-7.

Simulacioni model stohastičkog fleš A/D konvertora

Nikola Petrović, Dragan Pejić, Marjan Urekar, Đorđe Novaković i Nemanja Gazivoda

***Apstrakt*— Na osnovu potreba za ispitivanjem rada kreiranog sistema na Fakultetu tehničkih nauka, Katedra za električna merenja, pristupilo se projektovanju i izradi simulacionog modela postojećeg sistema stohastičkog fleš A/D konvertora. Simulacioni model je kreiran u MATLAB programskom okruženju koja je imala zadatak da simulira rad postojećeg sistema, kao i kombinacije različitih uticaja na njega.**

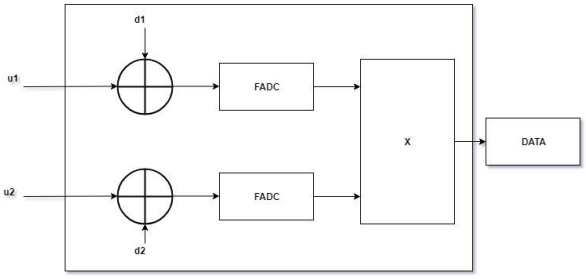
Zbog potrebe za simuliranjem velikog broja podataka, bilo je neophodno izvršiti optimizaciju modela i prilagođavanje dostupnim računarskim konfiguracijama, kao i čuvanje dobijenih podataka, parametara, rezultata i grafika.

***Ključne reči*— SFADC, Matlab, Multithreading, Big Data Analysis**

I. Uvod

Pod digitalnim merenjem danas se podrazumeva digitalno merenje standardnom sempling metodom (SSM). Unapređenje SSM se sastoji u: što tačnijem semplovanju, što je tehnološki problem, što dužoj reči A/D konvertora (ADC), što je, takođe, tehnološki problem i što većoj brzini ADC, što je isto tako tehnološki problem. Postoji metodološka kontradikcija između dužine reči ADC i njegove brzine: brzi ADC imaju kratku digitalnu reč, a ADC duge reči su spori. I ova kontradikcija se danas prevazilazi tehnološki, što je komplikovano, dugotrajno i skupo.

U pokušaju da se prevaziđe navedena metodološka kontradikcija, sredinom devedesetih godina prošlog veka, na Katedri za električna merenja FTN u Novom Sadu, uspešno je razvijena Stohastička Digitalna Merna Metoda (SDMM) [1-3]. SDMM je potpuno različita paradigma digitalnih merenja od SSM, i predstavlja metodološki iskorak u digitalnim merenjima. Naime, u SDMM se, za razliku od SSM, merena veličina ne posmatra i meri samo u tački (vremenskom trenutku), nego u skupu tačaka na intervalu na vremenskoj osi. Dalje, koriste se najbrži, fleš A/D konvertori (FADC) niske rezolucije, najčešće dvobitne. Oni su izuzetno jednostavni, robusni i pouzdani. Jednostavan FADC ima vrlo mali broj izvora sistematske greške koji se lako identifikuju i greška može da se otkloni. Prema tome, SDMM je inherentno vrlo tačna. Dvobitni FADC ima veliku grešku kvantizacije pa, na prvi pogled, izgleda da je veliki problem preciznost. Ovaj problem se efikasno rešava dodavanjem slučajnog uniformnog šuma (ditera), čime nastaje stohastički fleš A/D konvertor (SFADC), pa je SDMM i vrlo precizna. Kratka dvobitna reč SFADC ima za posledicu i vrlo jednostavan blok za osnovnu obradu – množenje i akumulaciju – (multiply and accumulate, MAC).



Sl. 1. Grafički prikaz sistema

Na slici 1. se nalazi šematski prikaz modelovanog sistema, pri čemu su u1 i u2 ulazni naponi, d1 i d2 su generisani diteri koji se dodaju na signale u1 i u2 nakon čega se vrši A/D konverzija tih signala, X predstavlja kolo za množenje dobijenih vrednosti i DATA predstavlja izlaznu brojnu vrednost.

II. Slučajne i sistematske greške

Osnovni problem kod stohastičke metode je prisustvo slučajne greške u velikom iznosu, uzrok ove greške je signal ditera koji se dodaje na ulazni signal. Usrednjavanjem signala se dolazi do zaključka da je uticaj slučajne greške obrnuto srazmeran dužini merenja signala. **Greška opada sa kvadratom dužine trajanja merenja** [4]**.**

Sistematske greške su manje od slučajne i ne zavise od trajanja merenja signala, one zavise od fizičkih osobina komponenti sistema. Sa dovoljnom količinom podataka za usrednjavanje, moguće je potisnuti slučajnu grešku i posmatrati sistematske greške. Potrebna količina podataka zavisi od frekvencije odabiranja, frekvencije signala, kao i matematičkog modela posmatrane komponente sistema.

Tokom simulacije rada celokupnog sistema i analize svih komponenata, za jednu kombinaciju parametara komponenata sistema i jedan referentni napon, vršili smo merenje signala u trajanju od 1600 sekundi.

Nikola Petrović, petrovicnikola@uns.ac.rs, Bulevar Kralja Petra I 28, Novi Sad

Dragan Pejić – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 6, 106314 Novi Sad (e-mail: pejicdra@uns.ac.rs)

Marjan Urekar – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 6, 106314 Novi Sad (e-mail: urekarm@uns.ac.rs)

Đorđe Novaković - Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 6, djordjenovakovic@uns.ac.rs

Nemanja Gazivoda - Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 6, nemanjagazivoda@uns.ac.rs

III. PROGRAMSKA SIMULACIJA

S obzirom da testiranje razvijenog sistema za jednu kombinaciju parametara sistema zahteva vreme koje nije zanemarljivo, sa kombinatorikom parametara sistema se vreme znatno produžava. Za ispitivanje svih kombinacija parametara bi sistem morao da radi više od dve godine bez prestanka.

Zbog toga je neophodno vršiti simulacije sistema koje daju znatno brže rezultate i mogućnosti simulacije različitih parametara sistema. Slika 1. prikazuje šematski prikaz sistema, pri čemu je izlaz sistema brojna vrednost DATA.

Vrednost napona U_m se izračunava na osnovu formule:

$$U_m = \sqrt{\frac{2 \times DATA}{K \times t}} \quad (1)$$

, gde je DATA suma izlaz akumulatora, K predstavlja koeficijent za određenu rezoluciju A/D konvertora, a t je vreme merenja signala.

A. Ulazni simulacioni podaci

U cilju definisanja ulaznih simulacionih podataka, neophodno je definisati sledeće vrednosti:

- Skup vrednosti rezolucija A/D konvertora
- Skup referentnih napona
- Skup vrednosti trajanja simulacionog merenja
- Broj ponavljanja simuliranog merenja
- Skup binarnih vrednosti primene algoritma svičevanja
- Skup vrednosti napona ofseta A/D konvertora
- Maksimalni napon
- Minimalni napon
- Frekvenciju odabiranja
- Frekvenciju signala
- Skup vrednosti napona ofseta D/A konvertora
- Skup rednih brojeva simuliranih problema D/A konvertora
- Skup vrednosti rezolucija D/A konvertora

Na osnovu unetih podataka se vrši kombinatorika i kreira skup mogućih kombinacija ulaznih parametara. Svakoj kombinaciji se dodeljuje unikatna šifra koja se čuva u tabeli.

B. Generisanje ditera

Jedan od ključnih delova stohastičkog fleš A/D konvertora jeste diter, nasumično generisani signal sa uniformnom raspodelom amplitude jednog praga A/D konvertora. U ovom radu, postoje četiri opcije za generisanje ditera:

1. Idealni slučaj generisanja ditera
2. Generisanje ditera sa multiplikativnim ofsetom. Svaka vrednost ditera se množi sa 1 + ofset.
3. Generisanje ditera sa kumulativnim (aditivnim) ofsetom. Svaka vrednost ditera se sabira sa ofsetom.
4. Generisanje ditera sa konačnim skupom vrednosti, određenim D/A konverzijom i njenom rezolucijom

C. Određivanje koeficijenta K

Koeficijent K je vrednost koja se menja za različit broj bita A/D konvertora i pre testiranja programa je neophodno odrediti za svaki broj bita. Postupak određivanja koeficijenta K je sledeći: Generisanje ulaznih signala sa dovoljnim brojem ponavljanja i različitim referentnim vrednostima napona; Određivanje koeficijenata K za svaki signal na osnovu (2); Čuvanje rezultata u fajlovima pod šifrovanim imenom; Prolazak kroz sačuvane rezultate za isti broj bita i usrednjavanje vrednosti koeficijenta K; Čuvanje koeficijenta K u fajlu K.mat

$$K = \frac{2 \times DATA}{U_m^2 \times t} \quad (2)$$

D. A/D konverzija

A/D konverzija predstavlja pretvaranje ulaznog analognog signala u skup brojnih vrednosti definisanih rezolucijom i naponskim opsegom A/D konvertora.

Konvencionalna metoda simulacije A/D konvertora traži veće računarske resurse i vreme u zavisnosti od rezolucije A/D konvertora. Iz tog razloga je bilo neophodno kreirati vektorizovanu simulaciju koja zahteva jednako vreme za svaku rezoluciju A/D konvertora.

1) Ofseti A/D konvertora

Ofseti A/D konvertora se definišu posebno za svaki prag i za svaki vremenski trenutak. Prethodno definisani ofseti se dodaju na ulazni signal za svaki odgovarajući prag A/D konvertora.

2) Postupak svičevanja

U cilju potiskivanja uticaja ofseta A/D konvertora, koristi se postupak svičevanja, što fizički podrazumeva zamenu ulaza A/D konvertora sa periodom koja je dva puta veća od periode ulaznog signala [5-6]. Ova ideja suzbijanja ofseta periodičnim unakrsnim preklapanjem (PUP) ulaza je poznata i široko primenjena tehnika u preciznim operacionim pojačavačima, gde se koristi za poništavanje dinamičkog ofseta.

Postupak svičevanja u simulaciji podrazumeva promenu znaka vrednosti ofseta sa periodom koja je dva puta veća od periode ulaznog signala.

IV. OPTIMIZACIJA SIMULACIJE

Mogućnosti simulacije zavise od računarske konfiguracije, kao i strukture simulacije. U cilju optimalnog iskorišćenja računarskih resursa, neophodno je izvršiti optimizaciju spram dostupnim resursima [7-8].

A. Vektorizacija

MATLAB programski jezik smešta promenljive u radnu

memoriju računara, što znači da je potrebno kreirati promenljive koje svojim sadržajem menjaju prolazak kroz petlje. Na taj način se skupovi podataka obrađuju kao celina, a ne pojedinačno. Ovaj postupak značajno ubrzava simulaciju, ali zahteva veću radnu memoriju računara kod većih količina podataka.

Prvi korak vektorizacije uzima kao promenljivu jedan jednodimenzionalni vektor signala.

Drugi korak vektorizacije uzima kao promenljivu dvodimenzionalni vektor koji u sebi sadrži jednodimenzionalne vektore signala za sva ponavljanja.

B. Planiranje

U slučaju da postoji više računara koji obrađuju podatke, neophodno je uraditi planiranje, što podrazumeva raspodelu mogućih kombinacija na sve računare spram njihovih kapaciteta i mogućnosti. Celokupan postupak se izvodi unošenjem raspoloživih računara i odnosa njihovih kapaciteta, kao i potrebnih kombinacija koje se dele.

C. Paralelno programiranje

MATLAB omogućava višejezgarnu optimizaciju petlji pomoću funkcije parfor koja paralelno pokreće onoliko paralelnih procesa koliko mikroprocesor ima jezgara. Tokom rada parfor funkcije potrebni resursi radne memorije se multipliciraju. U okviru programskog okruženja je moguće ručno precizirati koliko jezgara za koliko procesa se koristi.

D. Operacije nad velikim podacima

Ukoliko se obrađuju podaci koji zahtevaju veće resurse radne memorije nego što su dostupni, neophodno je pristupiti dodacima za obradu velikih podataka, što podrazumeva smeštanje podataka iz radne memorije na hard disk.

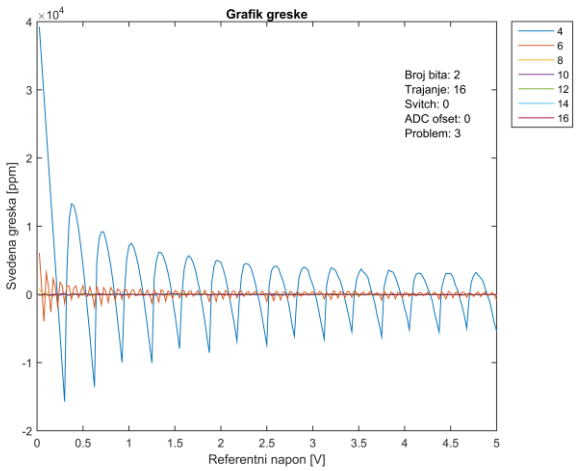
Prva mogućnost je upotreba MATLAB programskih dodataka za obradu podataka, pri čemu je tada potrebno izvršiti minorne izmene programa i odrediti lokaciju smeštenih fajlova.

Druga mogućnost je podešavanje veličine Windows Page File datoteke koja predstavlja prostor za virtuelnu radnu memoriju računara. Ova datoteka je ograničena veličinom i brzinom hard diska i preporučuje se upotreba SSD hard diska. Uz primenu ove mogućnosti nije potrebno vršiti izmenu programa.

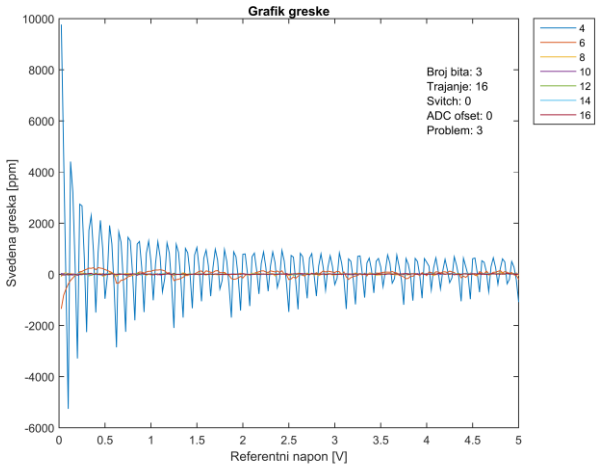
Upotreba operacija nad velikim podacima usporava simulacije za 50%, ali pruža mogućnost pokretanja simulacija na računarskim konfiguracijama koje nemaju dovoljnu količinu RAM memorije.

V. REZULTATI

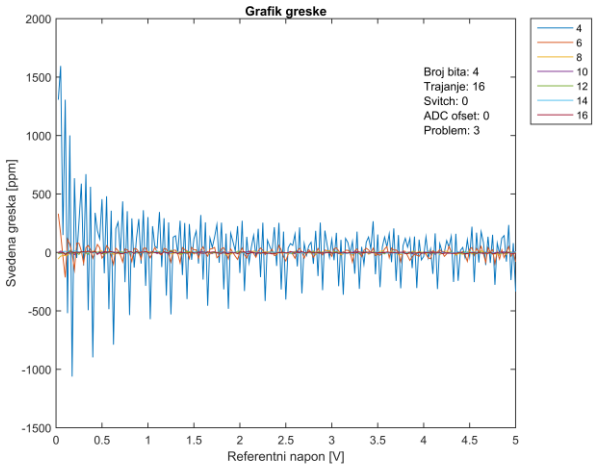
Nakon izvršenih simulacija sistema za različite uslove, svi dobijeni rezultati se čuvaju i arhiviraju u odgovarajućim datotekama. Dobijene rezultate je moguće interpretirati i porediti na različite načine u zavisnosti od potreba korisnika. Nakon postavljanja željenih poređenja, algoritmi za grafički prikaz generišu grafike dobijeni rezultata.



Sl. 2. Grafički prikaz rezultata simulacije sistema sa dvobitnim A/D konvertorom



Sl. 3. Grafički prikaz rezultata simulacije sistema sa trobitnim A/D konvertorom



Sl. 4. Grafički prikaz rezultata simulacije sistema sa četvorobitnim A/D konvertorom

Na slikama 2, 3 i 4 su prikazani rezultati za simulaciju sa rezolucijom A/D konvertora od 2, 3 i 4 bita. Na svakom

grafiku su prikazani rezultati za simulaciju sistema sa 4, 6, 8, 10, 12, 14 i 16 bita rezolucije D/A konvertora. Na ovim slikama je prikazan uticaj konačne rezolucije DACa koji generiše diter. Što je veća rezollucija ditera manji je njegov doprinos grešci i vidimo više preseka sa horizontalnom osom.

VI. ZAKLJUČAK

Trenutni sistem za simulaciju se pokazao kao uspešan uz rezultate koji su približni dobijenim rezultatima sa stvarnim sistemom. Sa trenutnim mogućnostima, sistem se prilagođava računarskim konfiguracijama na kojima se pokreće i pruža mogućnost ručno planiranog distribuiranog računanja.

Plan za dalji razvoj projekta je dodatna vektorizacija generisanih podataka, što podrazumeva kreiranje multidimenzionalne matrice podataka i njenu automatizovanu višejezgarnu obradu.

Osim toga, ideja je razvoj serverske obrade podataka uz kreiranje sopstvenih servera koji bi pružili mogućnost znatno bržeg rada celokupnog sistema.

Kao nastavak, pruža se mogućnost i kreiranja sistema mašinskog učenja, koji bi znatno olakšao generisanje podataka, određivanje koeficijenata i opisivanje sistematskih grešaka.

Kreirani sistem će se u budućnosti koristiti kao alat za ispitivanje međusobnog uticaja više problema istovremeno, što pruža mogućnost boljeg shvatanja međusobnog dejstva velikog skupa sistematskih uzročnika, što bi dalje trebalo da usmeri na načine potiskivanja njihovog uticaja, kao i višedimenzionalnu optimizaciju parametara sistema u cilju dobijanja optimalnog rešenja.

LITERATURA

[1] S. Milovančev, "Adaptivni A/D konvertor i njegova primena", doktorska disertacija, Fakultet tehničkih nauka, Novi Ssad, 1996.

[2] V. Vujičić, S. Milovančev, M. Pešaljević, I. Župunski, "Low-Frequency Stochastic True RMS Instrument", IEEE TRANS. INSTR. MEAS., VOL 48, PP 467-470, APRIL 1999.

[3] D. Pejić, V. Vujičić, Accuracy Limit Of High-Precision Stochastic Watt-Hour Meter, IEEE Trans nn Instrumentation and Measurement, Vol. 49, No 3, June 2000.

[4] V. Vujičić, "Generalized Low-Frequency Stochastic True RMS Instrument", IEEE Trans on Instrumentation And Measurement, VOL. 50., No. 5, October 2001.

[5] Pejić, Gazivoda, Ličina, Urekar, Sovilj, Vujičić, A Proposal of a Novel Method for Generating Discrete Analog Uniform Noise, Advances in Electrical And Computer Engineering, 2018, Vol 18, BR 2, 61-66

[6] Urekar, Pejić, Low Resolution Stochastic Flash ADC for High Precision Energy and RMS Voltage Measurements for Smart Grid, July 2018, DOI: 10.1109/CPEM.2018.8501106, Conference on Precision Electromagnetic Measurements CPEM 2018, Paris, France

[7] V. Pjevalica, N. Pjevalica, I. Kastelan, N. Petrovic, Acceleration of Digital Stochastic Measurement Simulation Based on Concurrent Programming, Elektronika ir elektrotechnika, vol.24, No.6, December 2018, pp 21-27, doi:10.5755/j01.eie.24.6.22284

[8] V. Pjevalica, N. Pjevalica, N. Petrović, N. Teslić, THD Factor Measurement Optimization Using Stochastic Method, Proceedings of 4th International Conference on Electrical, Electronics and Computing Engineering, IcETRAN 2017, Kladovo, Serbia, June 2017, pp. MLI1.5.1-4

ABSTRACT

Based on the needs for testing the performance of the created system at the Faculty of Technical Sciences, Department of Electrical Measurements, we designed and developed a simulation model of the existing system of stochastic flash A / D converter. A simulation was performed using the MATLAB programming language, which had the task of simulating the operation of the existing system, as well as a combination of various effects on it.

Due to the need to simulate a large number of data, it was necessary to perform model optimization and adjustment to the available computer configurations, as well as to store the obtained data, parameters, results and graphics.

Simulation model of a stochastic flash A / D converter

Nikola Petrović, Dragan Pejić, Marjan Urekar, Đorđe Novaković i Nemanja Gazivoda

SISTEM ZA DETEKCIJU POŽARA ZASNOVAN NA MIKROPROCESORSKIM MERNIM MODULIMA

Milan Šaš i Đorđe Novaković, *Member, IEEE*

***Apstrakt*—Ovaj rad prikazuje sistem za detekciju požara, zasnovan na mikroprocesorskom razvojnom sistemu EasyPIC Pro v7, sa mikrokontrolerom PIC18F87K22. Kao senzore za detekciju odabranih parametara koriste se SMOKE click (MIKROE-2560) baziran na MAX30105 senzoru za detektovanje dima, FLAME click (MIKROE-1820) baziran na PT334-6B senzoru za detektovanje plamena i CO click (MIKROE-1626) koji koristi MQ-7 senzor za merenje koncetracije ugljen-monoksida.**

***Ključne reči*— Mikroprocesorski merno-akvizicioni sistemi; Protivpožarni sistem; Dojava požara; Merenje; Senzori.**

I. UVOD

Sistem za detekciju i dojavu požara je sastavni deo svakog objetka, odnosno zgrada, kuća, stanova, hala i drugih objekata sličnih namena. Osnovni zadatak sistema je detekcija i dojava požara u ili na nekom objektu. U svetu već postoje rešenja za ovaj sistem koji se baziraju, između ostalog, na Neuronskim mrežama [1] i koriste se razni algoritmi koji služe za jasniju i precizniju detekciju požara, kako ne bi došlo do lažne uzbune [2][3].

Ovaj sistem je razvijen u sklopu predmeta mikroprocesorski merno-akvizicioni sistemi [4] na EasyPIC Pro V7 [5] razvojnom okruženju koje je razvija kompanija Mikroelektronika [6] i u čijem središtu se nalazi mikrokontrolerom PIC18F87K22 [7]. Kako svaki sistem slične namene ima određene senzore tako i ovaj sistem ima svoje senzore za detekciju požara. Odabrani senzori su senzori za dim, plamen i ugljen-monoksid koji se nalaze na zasebnim click pločicama koje je takođe razvila Mikroelektronika. Same pločice se povezuju na razvojno okruženje preko microBUS terminala. Komunikacija sa senzorima se vrši preko I2C (Inter-Intergrated Circuit) protokola ili preko interapta koji se šalje preko određenog pina na samoj pločici. U nastavku će biti razmotrene i detaljno objašnjene komponente koje su korišćenje kao senzori.

Milan Šaš – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail: milan.sas.97@gmail.com)

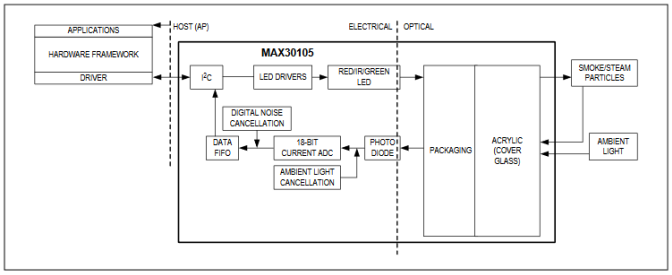
Đorđe Novaković - Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail: djordjenovakovic@uns.ac.rs).

II. SMOKE CLICK

Smoke click u sebi sadrži MAX30105 [8] visoko-osetljivi optički senzor za detekciju dima. Pločica je dizajnirana tako da može da se napaja sa 3.3V ili 5V. Komunikacija sa mikrokontrolerom se vrši preko I2C interfejsa sa dodatnom funkcijom koju omogućava INT pin. Na slici 1 se vidi sam izgled pločice:

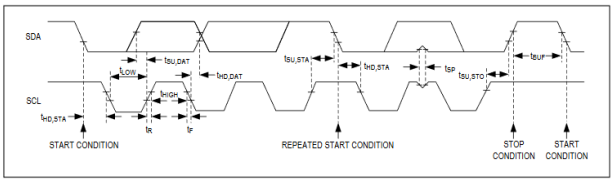
Sl. 1. Smoke click – izgled [9]

MAX30105 je čip kompanije Maxim Integrated sa integrisanim modulom za detektovanje čestica. Sadrži interni LED, foto detektore, optičke elemente i ″low noise″ elektroniku. Koristi se za detekciju dima i najveća primena je u protivpožarnim sistemima. Pošto je malih dimenzija može da se koristi i u mobilnim sistemima. Čip se napaja sa 1.8V i odvojeno sa 5V koji se koristi za napajanje LED-a. Potrošnja struje iznosi 0.7 μA. Sistemski dijagram je dat na slici 2:



Sl. 2. MAX30105 sistemski dijagram [8]

Pošto se komunikacija sa mikrokontrolerom vrši preko I2C komunikacije na sledećoj slici je dat dijagram po kome se vrši komunikacija:



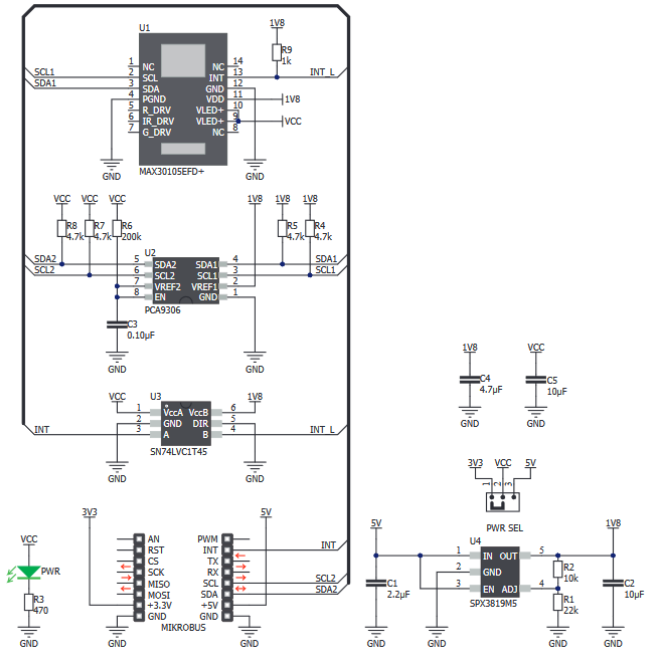
Sl. 3. I2C Komunikacija [8]

Na SL. 4 je dat Pinout dijagram pločice:

Notes	Pin	mikro BUS				Pin	Notes
	NC	1	AN	PWM	16	NC	
	NC	2	RST	INT	15	**INT**	Active-Low Interrupt (Open-Drain).
	NC	3	CS	TX	14	NC	
	NC	4	SCK	RX	13	NC	
	NC	5	MISO	SCL	12	**SCL2**	I2C Clock Input
	NC	6	MOSI	SDA	11	**SDA2**	2C Clock Data, Bidirectional (Open-Drain)
Power supply	**+3.3V**	7	3.3V	5V	10	**+5V**	Power supply
Ground	**GND**	8	GND	GND	9	**GND**	Ground

Sl. 4. Pinout dijagram smoke pločice [9]

Pored samog senzora na samoj pločici se nalaze još i PCA9306 čip [10] koji se koristi za I2C komunikaciju, SN74LVC1T45 čip [11] koji je jednobitni neinvertujući bus transrisiver i SPX3819M5 čip [12] koji je regulator pozitivnog napona. Na SL. 5. Se nalazi šema same pločice:



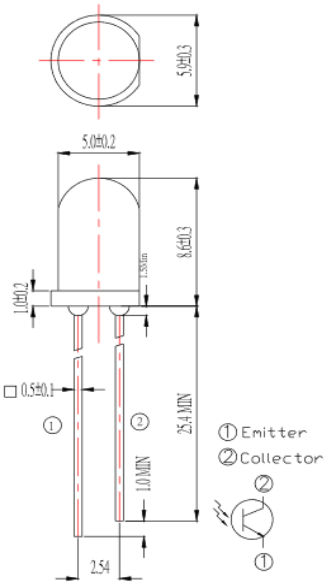
Sl. 5. Šema Smoke pločice [9]

III. FLAME CLICK

Flame click je uređaj koji na sebi ima PT334-6B [13] silikonski foto tranzistor koji je prekriven crnim epoksijem i zbog toga reaguje samo na infracrvenu svetlost. Koristi se kao alarm u slučaju vatre tako što se podešava ″threshold″ na potenciometru na samoj pločici. Kada vrednost pređe tu granicu šalje se interapt preko INT pina. Senzor takođe može da šalje i kontinualni analogni signal preko AN pina. Pločica može da se napaja sa 3.3V i 5V. Na Sl. 6. je dat prikaz pločice:

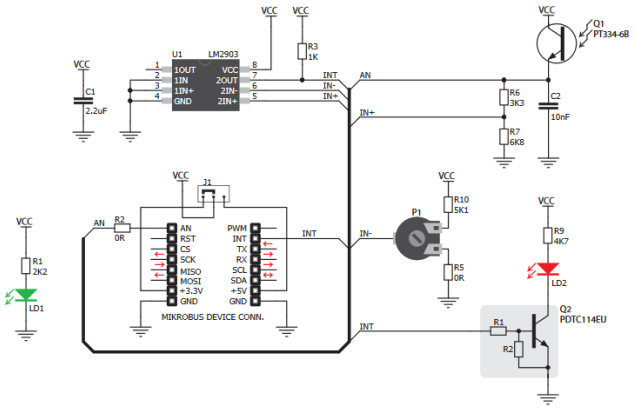
Sl. 6. Flame click – Izgled [14]

PT334-6B je veoma brz i veoma osetljiv NPN silikonski foto tranzistor koji se nalazi u standardnom 5 mm pakovanju. Pošto je prekriven crnim epoksijem, uređaj je osetljiv na infracrvenu svetlost. Brzo reaguje i veoma je osetljiv na svetlost pa se može primenjivati u infrared sistemima, kamerama, protivpožarnim alarmima kao i za hvatanje bubašvaba. Na Sl. 7. je dat izgled foto tranzistora:



Sl. 7. PT334-6B – Izgled [13]

Na pločici, osim PT334-6B, se nalazi i LM2903 [15] čip koji se koristi kao komparator za detektovanje praga provođenja. Na Sl. 8. je dat Pinout dijagram i šema Flame pločice:



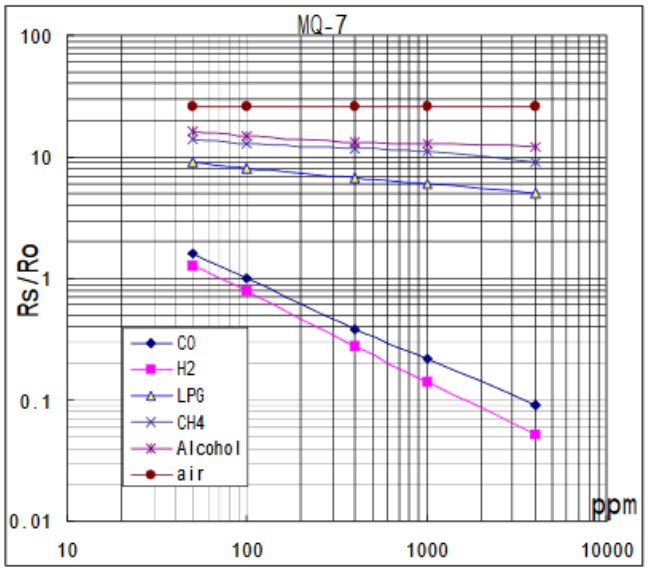
Sl. 8. Pinout dijagram i šema Flame pločice [16]

IV. CO CLICK

Ova pločica se koristi za merenje ugljen-monoksida u okolini i to postiže sa MQ-7 [17] gasnim senzorom. Na samoj pločici se nalazi potenciometar kojim se zadaje ″threshold″. Pločica se napaja sa 5V. Komunikacija sa mikrokontrolerom se vrši preko AN pina. Izgled pločice je dat na Sl. 9:

Sl. 9. CO click – izgled [18]

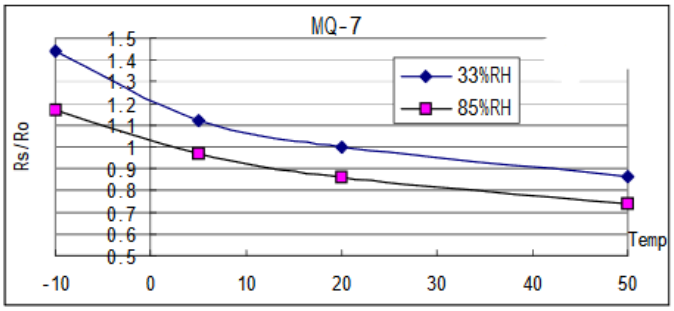
MQ-7 je gasni senzor koji služi za detekciju ugljen-monoksida (CO). Deo koji detektuje gas je napravljen od kalaj-dioksida (SnO2) koji ima manju provodnosti od vazduha. Kada se pojavi gas menja se i provodnosti. Opseg detektovanja CO je između 20ppm – 2000ppm. Kalibracija se vrši putem potenciometra koji se nalazi na pločici. Potrebno je da, kada se vrši kalibracija, senzor bude na radnoj temperaturi. Preporuka je da se upali oko 48h pre kalibracije. Na Gr. 1. je dat grafik osetljivosti senzora pri temperaturi od 20ºC, .vlažnosti vazduha od 65%, O2 koncentraciji od 21%:



Gr. 1. Karakteristika MQ-7 senzora [17]

Rs – otpornosti senzora na 100ppm CO u čistom vazduhu
Ro – otpotnost senzora na različitim koncentracijama CO

Očitavanje senzora takođe zavisi od temperature i vlažnosti vazduha. Na Gr. 2 je data zavisnost otpornosti od ovih parametara:



Gr. 2. Zavisnost otpora od temperature i vlažnosti vazduha [17]

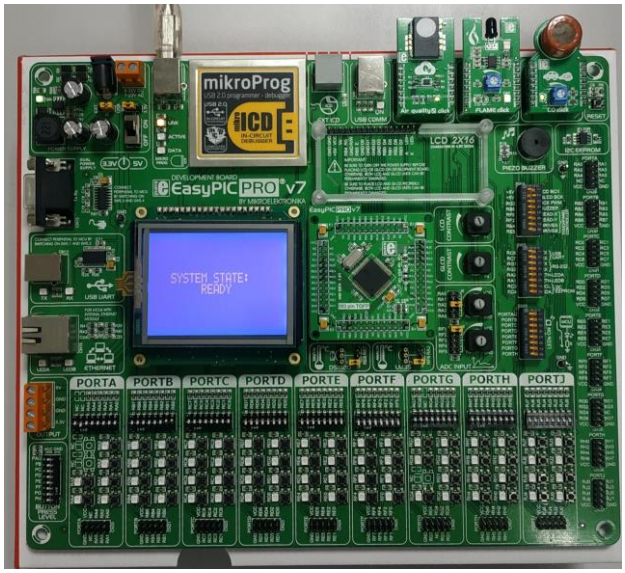
Na SL. 10 je dat Pinout dijagram pločice:

Notes	Pin	mikroBUS™				Pin	Notes
Analog output	AN	1	AN	PWM	16	NC	Not connected
Not connected	NC	2	RST	INT	15	NC	Not connected
Not connected	NC	3	CS	TX	14	NC	Not connected
Not connected	NC	4	SCK	RX	13	NC	Not connected
Not connected	NC	5	MISO	SCL	12	NC	Not connected
Not connected	NC	6	MOSI	SDA	11	NC	Not connected
Not connected	NC	7	3.3V	5V	10	+5V	Power supply
Ground	GND	8	GND	GND	9	GND	Ground

Sl. 10. Pinout dijagram CO pločice [18]

V. ZAKLJUČAK

Ovakav sistem za detekciju požara može biti veoma koristan u navedenim primenama zato što se ne oslanja na stanje odnosno prisutnost samo jedne komponente koja dovodi do aktiviranja alarma već se vrši merenje i obradu tri različite veličine koje zajedno mogu da dovedu do katastrofalnih posledica pre svega po čoveka pa i po objekte. Sistem može da se konfiguriše tako da radi kao ″Standalone″ uređaj koji bi se napajao iz baterijskog napajanja. Na slici 10. je dat izgled sistema:



Sl. 10. Izgled sistema

ZAHVALNICA

Želeo bih da zahvalim profesoru Platonu Sovilju na velikoj pomoći i podršci pri izradi ovog rada. Takođe, veliko hvala celoj Katedri za električna merena na Fakultetu tehničkih nauka u Novom Sadu. Na kraju bih se zahvalio i kompaniji Mikroelektronika na tehničkoj podršci ovom projektu.

LITERATURA

[1] F. S. a. X. Z. C. Cheng, "One fire detection method using neural networks," *Tsinghua Science and Technology*, vol. 16, no. 1, pp. 31-35, Feb. 2011.

[2] K. O. A. A. R. O. K. K. G. A. M. a. J. N. R. Sowah, "A Fire-Detection and Control System in Automobiles: Implementing a Design That Uses Fuzzy Logic to Anticipate and Respond," *IEEE Industry Applications Magazine*, vol. 25, no. 2, pp. 57-59, 2019.

[3] R. D. B. Ristić Jovan D., "Decision algorithms in fire detection systems," *Serbian Journal of Electrical Engineering*, vol. 8, no. 2, pp. 155-161, 2011.

[4] K. z. E. Merenja, "Mikroprocesorski merno - akvizicioni sistemi 1," [Online]. Available: http://kelm.ftn.uns.ac.rs/mikroprocesorski-merno-akvizicioni-sistemi-1/.

[5] MikroE, "EasyPIC PRO v7," [Online]. Available: https://www.mikroe.com/easypic-pro.

[6] MikroE, "Home Page," Mikroelektronika, [Online]. Available: https://www.mikroe.com/.

[7] Microchip, "PIC18F87K22," [Online]. Available: https://www.microchip.com/wwwproducts/en/PIC18F87K22.

[8] MikroE, "Datasheets MAX 30105," [Online]. Available: https://download.mikroe.com/documents/datasheets/max30105-datasheet.pdf.

[9] MikroE, "Smoke Click," Mikroelektronika, [Online]. Available: https://www.mikroe.com/smoke-click.

[10] T. Instruments, "Literature PCA9306," [Online]. Available: http://www.ti.com/lit/ds/symlink/pca9306.pdf.

[11] T. Instruments, "Literature SN74LVC1T45," Texas Instruments, [Online]. Available: http://www.ti.com/lit/ds/symlink/sn74lvc1t45.pdf.

[12] EXAR, "Datasheets SPX3819," Mouser, [Online]. Available: http://www.mouser.com/ds/2/146/SPX3819_DS_R200_082312-17072.pdf.

[13] Everlight, "Datasheets PT334-6B," [Online]. Available: https://media.digikey.com/pdf/Data%20Sheets/Everlight%20PDFs/PT334-6B.pdf.

[14] MikroE, "Flame Click," MIkroelektronika, [Online]. Available: https://www.mikroe.com/flame-click.

[15] T. Instruments, "Literature LM 393," Texas Instruments, [Online]. Available: http://www.ti.com/lit/ds/symlink/lm393.pdf.

[16] MikroE, "Flame Click - User Manual," Mikroelektronika, [Online]. Available: https://download.mikroe.com/documents/add-on-boards/click/flame/flame-click-manual-v100.pdf.

[17] MikroE, "Datasheets MQ 7," Mikroelektronika, [Online]. Available: https://download.mikroe.com/documents/datasheets/mq-7-datasheet.pdf.

[18] MikroE, "CO Click," Mikroelektronika, [Online]. Available: https://www.mikroe.com/co-click..

ABSTRACT

This paper presents the system that works as a fire detection system, based on development system EasyPIC Pro V7, with a PIC18F87K22 microcontroller. As sensors for detection of certain parameters system uses SMOKE click (MIKROE-2560) based on MAX30105 sensor for detection of smoke, FLAME click (MIKROE-1820) based on PT334-6B sensor for detection of flame and CO click (MIKROE-1626) based on MQ-7 carbon-monoxide concentration sensor.

System for fire detection based on microprocessor measuring modules

Milan Šaš, Đorđe Novaković

SMART Home sistem zasnovan na mikroprocesorskim mernim modulima

Duško Gajinović, Đorđe Novaković, IEEE member

***Apstrakt*—Ovaj rad prikazuje sistem, koji vrši očitavanje nekoliko senzorskih modula koristeći microBUS koji je baziran na I2C protokolu, kao i očitavanje vrednosti sa analognih senzora. Sistem je realizovan na EasyPIC Pro V7 ploči sa PIC18F87K22 mikrokontrolerom. Senzorski moduli u ovom sistemu su Weather Click koji je baziran na senzoru BME-280 za očitavanje temperature, atmosferskog pritiska i vlažnosti vazduha, kao i Proximity Click koji koristi VCNL4010 za detektovanje korisnika ispred glavne konzole. Takođe se koriste i analogni senzori LM35 za merenje temperature u pojedinačnim prostorijama. Zatim mikrokontroler vrši obradu svih očitanih signala, i prikazuje ih na 128x64 GLCD ekranu sa TouchPanel-om. Osim prikazivanja očitanih podataka, GLCD služi i za upravljanje rasvetom u pojedinačnim prostorijama kuće, kao i otključavanje i zaključavanje ulaznih vrata.**

***Ključne reči*— Mikroprocesorski merno-akvizicioni sistemi; Senzori; GLCD; Touchscreen; I2C; Temperatura; Vlažnost vazduha; Atmosferski pritisak; Kontrola električnih uređaja.**

I. Uvod

Ovaj rad je nastao na osnovu ideje da se konstruiše sistem koji bi omogućio korisniku da ima uvid u sve parametre koje očitava set odabranih senzora i mogućnost bezbednosne kontrole u vidu alarmnog sistema i elektronskih brava.

Ceo sistem je zasnovan na EasyPIC Pro V7 [1] razvojnoj ploči, na kojoj se nalazi PIC18F87K22 [2] mikrokontroler. Ova razvojna ploča i svi senzorski moduli su razvijeni i proizvedeni od strane kompanije Mikroelektronika [3]. Konkretno su u ovom projektu implementirani moduli Weather Click (MIKROE-1978) [4] za merenje temperature, atmosferskog pritiska i vlažnosti vazduha, kao i Proximity Click (MIKROE-1445) [5] za merenje udaljenosti korisnika od glavne upravljačke konzole. Osim tih modula koriste se i analogni senzori LM35 za merenje temperature. Njihove izlazne vrednosti se mere pomoću A/D konvertora koji je integrisan u sam PIC mikrokontroler. Sve prikupljene informacije se obrađuju i prikazuju na 128x64 GLCD ekranu koji se nalazi na samoj razvojnoj ploči. Interfejs je napravljen tako da prikazuje samo osnovne informacije dok je sistem u "standby" modu, a svi ostali detalji su dostupni u podmeniju, koje korisnik sam bira pritiskom na ekran.

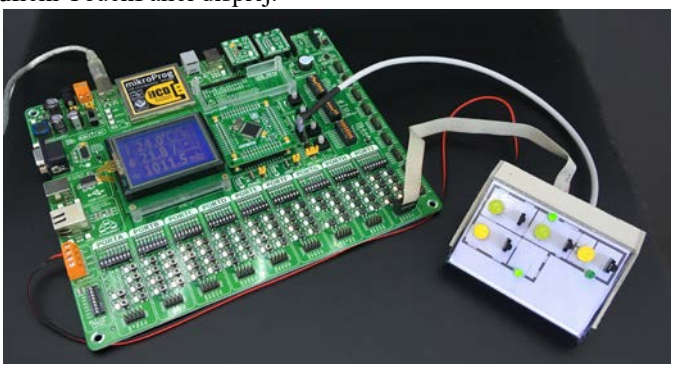
Ovaj sistem je razvijen u sklopu predmeta Mikroprocesorski merno-akvizicioni sistemi [6].

Duško Gajinović – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail: gajinovic.dusko@gmail.com)

Đorđe Novaković - Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail: djordjenovakovic@uns.ac.rs)

II. Interfejs

Početna ideja je bila da se napravi grafički intefejs pomoću kojeg se jednostavno prate očitani parametri, kao i upravljanje rasvetom i električnim bravama. U tu svrhu sam koristio grafički TouchPanel displej.



Sl. 1. SMART Home sistem

Početni ekran prikazuje osnovne informacije kao što su temperatura, vlažnost vazduha i pritisak dok se sistem nalazi u standby modu.

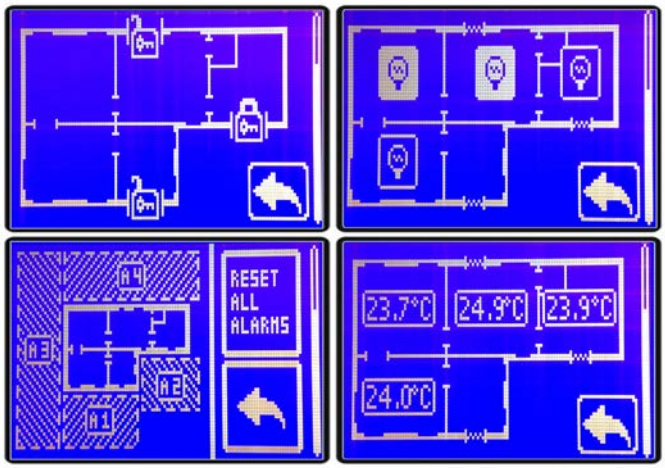
Sl. 2. "Standby" ekran

Glavni meni je dostupan pritiskom na bilo koji deo ekrana.

Sl. 3. Glavni meni

Na glavnom meniju je moguć izbor jednog od pet različitih podmenija:

1. “House Locks” meni služi za otključavanje i zaključavanje svih elektronskih barava na ulaznim vratima objekta.
2. “House Lights” meni omogućava kontrolu glavne rasvete u svakoj prostoriji.
3. “House Alarms” meni daje informaciju o aktivaciji alarmnog sistema oko objekta. Prostor oko objekta se može podeliti na više zona koje se simultano prate. Aktivacija bilo koje zone je prikazana i na “standby” meniju.
4. “House Temps” dozvoljava uvid u trenutnu temperaturu svake prostorije.
5. “System Info” daje sve neophodne informacije u vezi sistema, kao i proveru ispravnosti svih modula u sistemu.

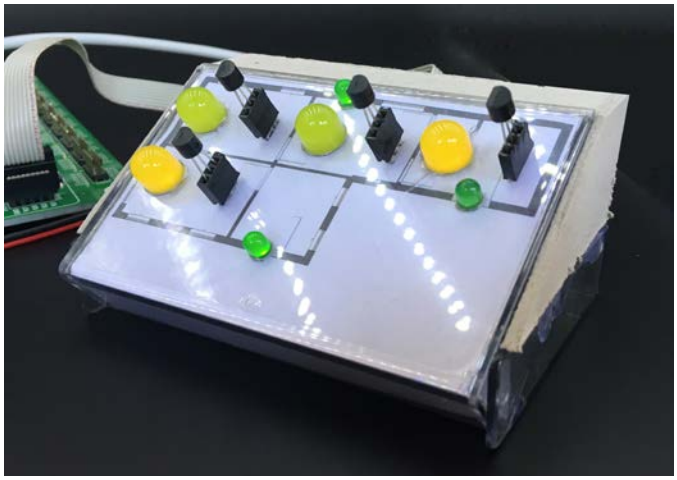


Sl. 4. Podmeniji (House Locks, House Lights, House Alarms, House Temps)

Sa desne strane svakog menija se nalazi grafički prikaz tajmera koji odborojava 30 sekundi od poslednjeg unosa ili promene. Ako korisnik duže stoji ispred glavne konzole, sensor udaljenosti će to detektovati, i tajmer neće biti aktivan.

Ovaj sistem se može jednostavno primeniti u različitim objektima, bez obzira na broj i raspored prostorija, kao i izgled i veličinu spoljnog okruženja. Primena je veoma široka, može se implementirati u sve vrste objekata (stambeni, poslovni, industrijski), u zavisnosti od potreba. Sistem se može nadograditi širokom paletom senzora i modula, koji omogućavaju preciznije i detaljnije praćenje svih željenih parametara.

Uz glavnu ploču ide i maketa objekta, koja u ovom slučaju ima četiri prostorije, u svakoj sensor za temperaturu, kao i indikator stanja rasvete i električnih brava. Maketa se može videti na slici broj 5:



Sl. 5. Maketa objekta

III. Senzorski Moduli

SMART Home sistem vrši očitavanja pomoću dva senzorska modula, Weather Click [4] i Proximity Click [5].

Weather Click (MIKROE-1978) je senzor atmosferskih veličina, koji je baziran na senzoru Bosch BME280 [7]. Ovaj senzor vrši merenja temperature, vlažnosti vazduha i atmosferskog pritiska, i dizajniran je za malu potrošnju struje i dugoročnu stabilnost. Komunikacija sa mikrokontrolerom se vrši preko I2C interfejsa.

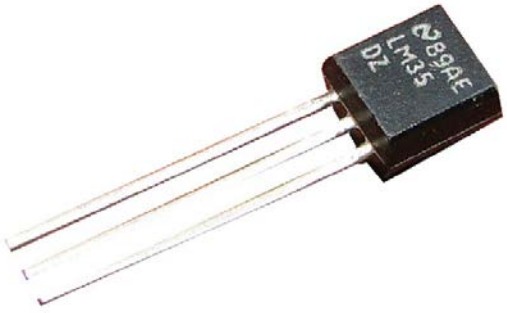
Sl. 6. Weather Click - [4]

BME280 ima visoku tačnost i pouzdanost za sva tri senzora. Senzor vlažnosti vazduha ima vreme odziva od jedne sekunde i preciznost od ± 3% RH. Senzor atmosferskog pritiska ima grešku od 0.25%. Senzori vlažnosti vazduha i pritiska mogu raditi nezavisno jedan od drugog. Najbitnija karakteristka senzora temperature je da ima visoku rezoluciju, u određenom modu, do 20 bita.

Proximity Click (MIKROE-1445) je modul koji je zasnovan na senzoru Vishay Semiconductors VCNL4010 [8], kombinuje infracrvenu LED diodu i PIN fotodiodu za merenje udaljenosti do 20 cm. Takođe može da se koristi kao senzor ambijentalnog svetla.

Sl. 7. Proximity Click – [5]

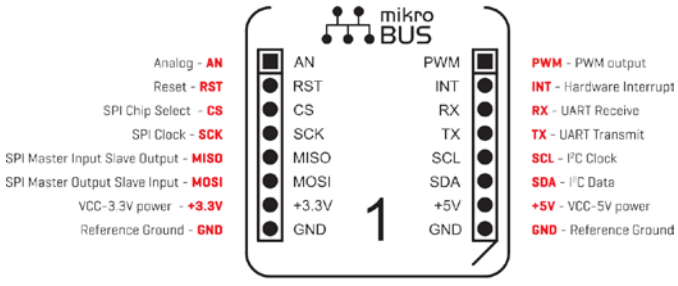
Osim senzorskih modula se koriste i analogni senzori temperature Texas Instruments LM35-DZ [9]. On se koristi za merenje pojedinačnih prostorija objekta, zbog svojih malih dimenzija i male greške merenja, od 0.5°C. Izlazni napon ovog senzora se menja za +10.0 mV/°C u odnosu na 0°C. Taj signal se pretvara u digitalni broj pomoću A/D konvertora, čija je rezolucija 12 bita, koji je integrisan u PIC mikrokontroler.



Sl. 8. Senzor LM35DZ [10]

IV. Struktura platforme

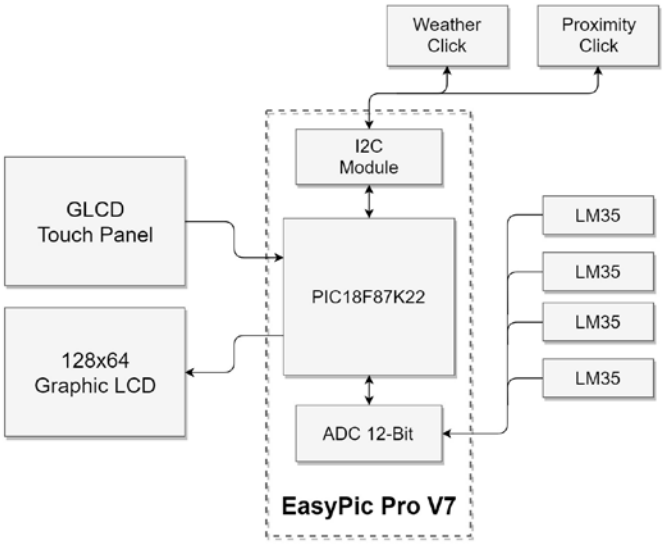
EasyPic Pro V7 je razvojni sistem koji povezuje sve periferne jedinice sa mikrokontrolerom PIC18F87K22. U samom mikrokontroleru se nalazi modul za I2C protokol, koji omogućuje komunikaciju sa senzorskim modulima preko MikroBUS-a, slika 9.



Sl. 9. MikroBUS socket – [11]

Konkretno u ovom slučaju Click pločice komuniciraju sa mikrokontrolerom isključivo preko I2C protokola. MikroBUS je pogodan i za ostale vrste senzora, posto podržava i SPI i UART protokole. Osim toga ima i pinove za PWM, analogni ulaz, kao i pristup jednom interrupt pinu po klik pločici.

Strukturu platforme je najlakše prikazati blok šemom koja je se nalazi na slici 10. Smer strelica prikazuje kako se odvija razmena podataka.



Sl. 10. Blok šema SMART Home sistema

Prilikom aktivacije sistema se inicijalizuju sve periferne jedinice. Prvo se vrši podešavanje PORT-ova mikrokontrolera, kao ulazni ili izlazni, digitalni ili analogni, podešavanje tajmerskih registara, i svih potrebnih protokola.

Nakon toga se inicijalizuje GLCD ekran kako bi sve potrebne informacije, kao i greške, mogle da se prikažu korisniku. Nakon toga se uspostavlja komunikacija sa senzorskim modulima, i na određene adrese se unose bitovi za konfigurisanje samih senzora, kao što su osetljivost, vreme odabiranja i kalibracija. Ubrzo posle toga se isti ti registri za konfigurisanje očitavaju i porede sa već zadatim vrednostima, kako bi se proverilo da li je senzor u potpunosti podešen, u slučaju da nije, ispisuje se greška na ekranu. Takođe se može pristupiti statusu svih perifernih jedinica u “System Info” meniju. LM35 senzori temperature su povezani na PORT-A mikrokontrolera, i sam port je u inicijalizaciji podešen kao analogni ulaz.

Tajmerski interapt je podešen tako da se na svakih 50ms vrši provera da li je bilo novih unosa preko touch panela. U slučaju da jeste, vrši se provera položaja unosa i da li se taj položaj poklapa sa “ikonicom” u datom meniju. Ukoliko se poklapa, poziva se odabrana funkcija. Ubrzo posle toga se radi osvežavanje ekrana, sa novim informacijama ili sa novim podmenijem. Takođe se i posle svakog unosa resetuje tajmer od 30 sekundi i kreće novo odbrojavanje. Ekran se nezavisno

od aktivnosti osvežava svake sekunde, a tome prethodi ponovno očitavanje svih vrednosti koje treba da se prikažu. Po isteku 30 sekundi od poslednjeg unosa, sistem prelazi na standby režim.

V. Zaključak

Sistem SMART Home je realizovan od komponenata koje su dostupne na domaćem tržištu, pristupačnih cena i izuzetnog kvaliteta. Veoma lako se može prilagoditi specifičnim zahtevima tržišta i sama implementacija i upotreba je jednostavna. Moguće je uz male izmene povezati ovaj uređaj sa mobilnim telefonima ili nekim web serverom, za detaljnije praćenje statusa sistema. Sistem je veoma fleksibilan zbog velikog broja senzorskih modula koji se mogu naknadno ugraditi, ukoliko se za to pokaže potreba.

Zahvalnica

Zahvaljujem se profesoru Platonu Sovilju na podršci i pomoći prilikom izrade ovog rada, kao i svim kolegama sa Katedre za električna merenja, na Fakultetu tehničkih nauka u Novom Sadu.

Literatura

[1] Mikroelektronika, "EasyPIC PRO v7", [Online]. Available: https://www.mikroe.com/easypic-pro
[2] Microchip, "PIC18F87K22", [Online]. Available: https://www.microchip.com/wwwproducts/en/PIC18F87K22
[3] Mikroelektronika, "Home Page", [Online]. Available: https://www.mikroe.com/.
[4] Mikroelektronika, "Weather Click", [Online]. Availible: https://www.mikroe.com/weather-click
[5] Mikroelektronika, "Proximity Click", [Online]. Availible: https://www.mikroe.com/proximity-click
[6] K. z. E. Merenja, "Mikroprocesorski merno - akvizicioni sistemi 1," [Online]. Available: http://kelm.ftn.uns.ac.rs/mikroprocesorski-merno-akvizicioni-sistemi-1
[7] Bosch, "Datasheet BME280", [Online]. Availible: https://download.mikroe.com/documents/datasheets/BME280.pdf
[8] Vishay, "Datasheet VCNL4010", [Online]. Availible: http://www.vishay.com/docs/83462/vcnl4010.pdf
[9] Texas Instruments, "Datasheet LM35", [Online]. Availible: http://www.ti.com/lit/ds/symlink/lm35.pdf
[10] Indiamart, "Image LM35",[Online]. Availible: https://www.indiamart.com/proddetail/lm35-lm35dz-precision-temperature-sensor-14510861388.html
[11] Mikroelektronika, "MikroBUS", [Online]. Availible: https://www.mikroe.com/mikrobus

Abstract

This project presents a system that reads multiple sensor modules using microBUS socket which is based on an I2C protocol, as well as reading values from analog sensors. This system was implemented on the EasyPIC Pro V7 board with the PIC18F87K22 microcontroller. Sensor modules for this system are Weather Click, based on the BME-280 sensor for temperature, atmospheric pressure and humidity, as well as the Proximity Click that uses the VCNL4010 sensor to detect users in front of the main console. This system also features analog LM35 sensors that are used to measure temperature in individual rooms. The microcontroller takes the readings from all of the sensors, processes them, and displays them on a 128x64 GLCD with a TouchPanel. Besides displaying the read data, GLCD also serves to manage the lighting in individual rooms of the house, as well as unlocking and locking the front door.

SMART Home system based on microprocessor measuring modules

Duško Gajinović, Đorđe Novaković, IEEE member

IMPLEMENTACIJA PID REGULATORA POMOĆU MIKROPROCESORSKIH MERNO-REGULACIONIH MODULA

Žarko Dubajić, Đorđe Novaković, IEEE member

***Apstrakt*— Ovaj rad prikazuje implementaciju PID regulatora u mikroelektronici, koristeći razvojni sistem EasyPIC Pro v7, sa mikrokontrolerom PIC18F87K22. Vrši se regulacija brzine obrtanja elektromotora. Kao senzor za detekciju okretanja osovine elektromotora koristi se inkrementalni enkoder. Za generisanje PWM signala za upravljanje koristi se DC MOTOR 4 click.**

***Ključne reči* — Mikroprocesorski merno-akvizicioni sistemi; PID; Regulacija; Merenje; Senzori; Brzina; Elektromotor.**

I. UVOD

Kontroler je deo sistema automatskog upravljanja koji obavlja upravljačku funkciju. Njegov zadatak je da, generišući upravljački signal , vodi merljivu izlanu veličinu ka referentnoj vrednosti koja je zadata signalom.

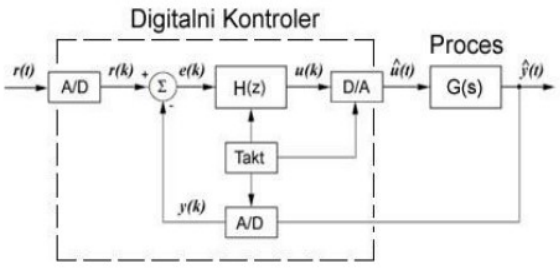
Ovaj sistem je razvijen u sklopu predmeta Mikroprocesorski merno-akvizicioni sistemi [3] na EasyPIC Pro V7 [4] razvojnom okruženju koje razvija kompanija Mikroelektronika [2] i u čijem središtu se nalazi mikrokontrolerom PIC18F87K22 [1].

II. DIGITALNI KONTROLER

Poslednjih decenija digitalna tehnika doživela je revoluciju. U sistemima automatskog upravljanja primat nad analognim preuzeli su digitalni kontroleri. Osnovni razlog je njihova dostupnost odnosno niska nabavna cena , jednostavna primena i održavanje. Digitalni kontoleri u sistemu automatskog upravljanja obavljaju istu funkciju kao i analogni a osnovna razlika između ova dva rešenja je u principu rada. Analogni konroleri izvršavaju svoj upravljački algoritam obradom vremenski neprekidnih signala. Digitalni konroleri obavljaju istu funkciju obradom digitalnih odnosno signala disrketizovanih po vremenu i po nivou. Njihova realizacija osvaruje se primenom digitalnih komponenti što ih čini neuporedivo pristupačnijim nego što su to analogni kontroleri.

Žarko Dubajić – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail: zardubajic@gmail.com)

Đorđe Novaković - Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail: djordjenovakovic@uns.ac.rs)

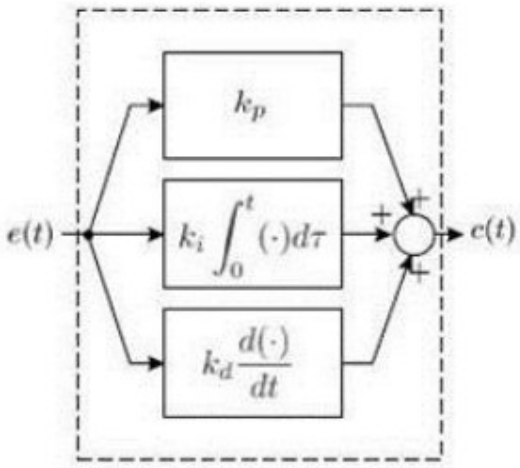


Sl. 1. Blok dijagram Sistema automatskog upravljanja sa digitalnim kontrolerom

Digitalni sistem automatskog upravljanja sadrži u opštem slučaju i analogne delove (objekat upravljanja). S obzirom da se digitalni sistemi opisuju diskretnim (diferencnim) jednačinama zadatak teorije digitalnih sitema je da nađe diskretni ekvivalent analognim delovima kako bi se vršila obrada isključivo diskretnih signala. Potrebno je naći diskretnu prenosnu funkciju sistema koja će obezbediti da sistem ima isti izlaz kao u slučaju kontinualne prenosne funkcije ali samo u trenucima odabiranja.

III. PID REGULATOR

Najčešće upotrebljavani kontroler u sistemima automatskog upravljanja je PID kontroler koji predstavlja kombinaciju Proporcionalnog Integracionog i Diferencijalng upravljanja. Razlog njegove rasprostranjenosti leži u jednostavnosti njegove primene.

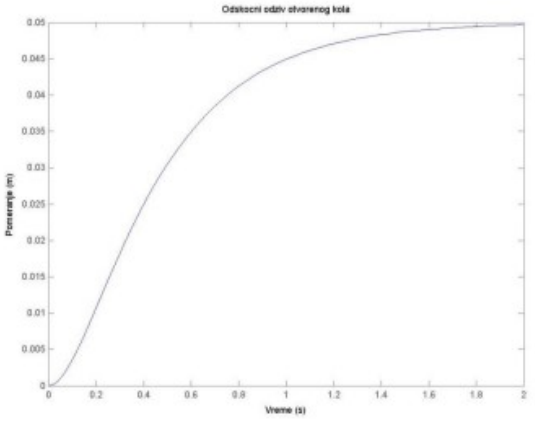


Sl. 2. Blok dijagram PID regulatora

U praksi se PID kontroleri izrađuju tako da svako od tri dejstava kontrolera bude sa podesivim parametrijma. Nakon jednostavne procedure podešavanja parametara najčešće se može postići zadovoljavajući odziv sistema čak i u slučaju kada nije poznat matematički model objekta upravljanja. U tom slučaju ovi kontroleri predstavljaju najbolje rešenje.

PID kontroleri su linearni pa daju slabe rezultate u slučaju nelinearnih sistema. Takođe su osetljivi na poremećaje visokih frekvencija što donekle može biti otklonjeno primenom nisko-propusnog filtera.

Da bi se analizirala priroda uticaja pojedinih komponti PID upravljačkog signala na sistem u celini pogodno je poznavati vremenski odziv sistema kada nema upravljanja odnono vremenski odziv otvorenog kola.



Sl. 3. Vremenski odziv otvorenog kola na jedinični odskočni signal

Jednačona PID regulatora po kojoj se računa upravljanje glasi:

$$u(t) = K_p + K_i \int_0^t e(t)dt + K_d \frac{de(t)}{dt}. \qquad (1)$$

U jednačini *u(t)* predstavlja vrednost upravljanja u vremenskom trenutku *t*. K_p, K_i i K_d su koeficijent proporcijalnog, integralnog i diferencijalnog dejstva tim redosledom, a *e(t) je* greška u vremenskom trenutku *t* dobijena razlikom referentnog i izlaznog signala.

Jednačina (1) je u kontinualnom domenu i u tom obliku nije upotrebljiva za implementaciju u mikroelektronici zato koristimo diskretnu brzinsku formu jednačine PID regulatora:

$$u(k) = u(k-1) + q_0 \cdot e(k) + q_1 \cdot e(k-1) + q_2 \cdot e(k-2). \qquad (2)$$

Za razliku od (1), u (2) se koriste diskretni odabirci u vremenu obeleženi sa *k*. Koeficijenti q_0, q_1 i q_2 se računaju kao:

$$q_0 = k_p + k_i. \qquad (3)$$

$$q_1 = k_p + k_i + 2 \cdot k_d. \qquad (4)$$

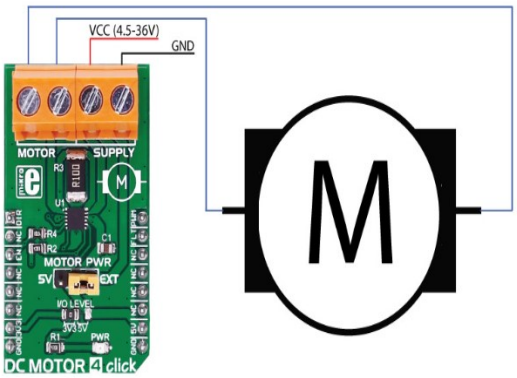
$$q_2 = k_d. \qquad (5)$$

k_p, k_i i k_d predstavljaju koeficijente proporcijalnog, integralnog i diferencijalnog upravljanja.

IV. DC MOTOR 4 CLICK

DC MOTOR 4 click [6] sadrži MAX14870 [5] motor drajver proizvođača Maxim Integrated. Click komunicira sa mikrokontrolerom preko sledećih pinova: PWM, AN, CS, i INT. Kratkospojnik na pločici se koristi za odabiranje eksternog ili internog napajanja. MAX14870 [5] motor drajver je male snage i jednostavno rešenje za upravljanje DC motorima i relejima naponima između 4.5 – 36V. Na click-u postoje dva zavrtna terminala od kojih se jedan kotisti za povezivanje motora, a drugi za eksterno napajanje ukoliko je potrebno.

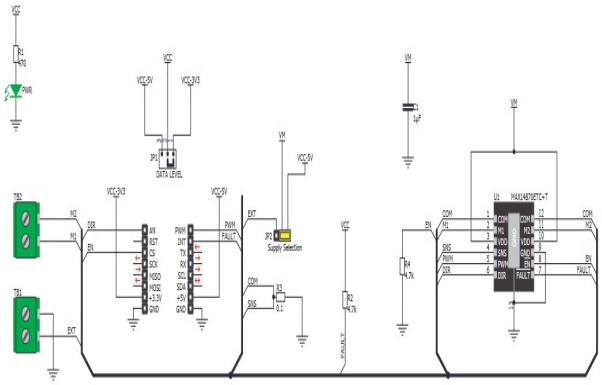
Sl. 4. DC MOTOR 4 click – izgled [6]



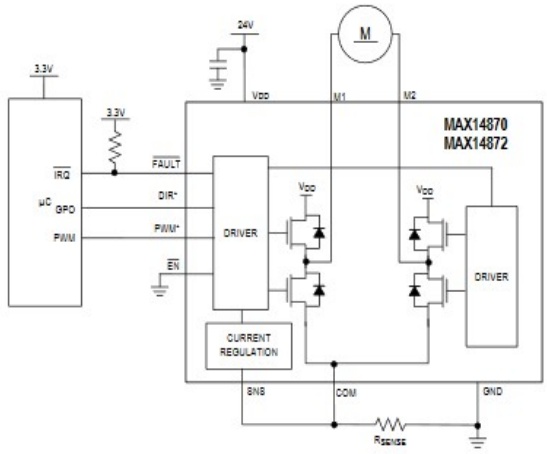
Sl. 5. DC MOTOR 4 click – način povezivanja motora i napajanja [6]

Notes	Pin	mikro BUS				Pin	Notes
Rotation direction	DIR	1	AN	PWM	16	PWM	PWM Control Logic Input
	NC	2	RST	INT	15	FLT	Open-Drain Active-Low Fault Output
Enable IC	EN	3	CS	TX	14	NC	
	NC	4	SCK	RX	13	NC	
	NC	5	MISO	SCL	12	NC	
	NC	6	MOSI	SDA	11	NC	
Power supply	+3.3V	7	3.3V	5V	10	+5V	Power supply
Ground	GND	8	GND	GND	9	GND	Ground

Sl. 6. DC MOTOR 4 click – pinout dijagram



Sl. 7. DC MOTOR 4 click – električna šema [6]



Sl. 8. MAX14870 motor drajver – šema [5]

V. Regulacija brzine obrtanja elektromotora

Za regulaciju brzine korišten je elektromotor čija je osovina bila pričvršćena za obrtni deo inkrementalnog enkodera sa 660 pozicija po krugu. Napajanje motora je vršeno koristeći eksterno napajanje od 5V preko DC MOTOR 4 click pločice koja je pretvarala DC napon u PWM signal koji je korišten za kontrolu brzine obrtanja motora. Signali sa A i B kanala enkodera su detektovani eksternim interaptom na pinovima RB0 i RB1. Pošto su signali sa enkodera samo impulsi, računanje brzine se vrši brojanjem impulsa u jedinici vremena. Referenca se zadaje putem UART-a u obrtajima po minuti.

Izračunavanje trenutne brzine obrtanja i vrednosti upravljanja se radi na svakih 100ms uz pomoć tajmerskog interapta. Za računanje upravljanja koristi se brzinska formula diskretnog PID regulatora pokazana ranije. Vrednosti upravljanja su softverski ograničene između 0 i 255 jer su to minimum i maksimum za podešavanje PWM duty cycle-a. Ispis vrednosti trenutne brzine, upravljanja i greške ce prosleđuje na UART terminal svaki sekund.

Podešavanje parametara PID regulatora vršeno je Zigler-Nicholsonovom metodom.

Brzina obrtanja elektro motora se račina preko formule:

$$obrtaji\_u\_minuti = \frac{broj\_impulsa}{660} \cdot 600 \quad (6)$$

Pod brojem impulsa misli se na impulse sa kanala A ili B inkrementalnog enkodera. Pošto se izračunavanje vrši svakih 100ms sve je pomnoženo sa 600 da bi se dobili obrtaji u minuti.

Zahvalnica

Zahvalio bih se Katedri za električna merenja, i katedri za automatiku fakulteta tehničkih nauka u Novom Sadu, za pruženo znanje potrebno za izradu ovog rada. Želim posebno da zahvalim profesoru Platonu Sovilju na pruženoj podršci u izradi ovog rada. Takođe bih da zahvalim kompaniji Mikroelektronika na tehničkoj podršci ovom projektu.

VI. Literatura

[1] Microchip, "PIC18F87K22," [Online]. Available: https://www.microchip.com/wwwproducts/en/PIC18F87K22.

[2] Mikroelektronika, "Home Page," [Online]. Available: https://www.mikroe.com/.

[3] K. z. E. Merenja, "Mikroprocesorski merno - akvizicioni sistemi 1," [Online]. Available: http://kelm.ftn.uns.ac.rs/mikroprocesorski-merno-akvizicioni-sistemi-1/.

[4] Mikroelektronika, "EasyPIC PRO v7," [Online]. Available: https://www.mikroe.com/easypic-pro.

[5] Mikroelektronika, "Datasheets," [Online]. Available: https://download.mikroe.com/documents/datasheets/MAX14870-MAX14872.pdf.

[6] Mikroelektronika, "DC MOTOR 4 click" [Online]. Available: https://www.mikroe.com/dc-motor-4-click.

Abstract

This paper presents the implementation of the PID controller in microelectronics using the EasyPIC Pro v7 development system with the PIC18F87K22 microcontroller. The rotation speed of the electric motor is controlled. An incremental encoder is used as the sensor for detecting the rotation of the shaft of the electric motor. DC MOTOR 4 click is used to generate a PWM control signal.

Implementation of PID controllers based on microprocessor measuring and control modules

Žarko Dubajić, Đorđe Novaković, IEEE member

Prilog radnom uputstvu za etaloniranje silotermometara

Ivan Gutai, *Member, IEEE*, Bojan Vujičić, *Member, IEEE*, Nemanja Gazivoda, *Member, IEEE*

***Apstrakt*—U ovom radu se daje jedan primer postupka etaloniranja silotermometara, koje se sprovodi na objektu na terenu. Prikazane su najbolje mogućnosti merenja, opisana je priprema za merenje, navedena je merna oprema koja se koristi, opisani su postupci merenja koje treba sprovesti i prikazana je obrada dobijenih rezultata.**

***Ključne reči*—Postupak etaloniranja silotermometara; ISO 17025; Metrologija**

I. Uvod

Temperatura je fizička veličina koja predstavlja stepen zagrejanosti tela. Temperatura je povezana sa termičkim kretanjem molekula ili atoma, tj. sa termodinamičkim stanjem tela i njegovom unutrašnjom energijom. Fizičke veličine kao što su masa, dužina, i druge označavaju se kao ekstenzivne ili parametarske, jer se sa povećanjem objekta povećava i njihova vrednost. Za razliku od ovih veličina, temperatura je intenzivna ili aktivna veličina. Njen intenzivni karakter ogleda se u tome što će prilikom deljenja tela na više delova svaki deo zadržati temperaturu tog tela. Drugim rečima, temperatura nema svojstvo aditivnosti. Za temperaturu se ne može izgraditi delitelj ili sabirač. Zbog toga se etalon temperature ne može praviti na način kako se to radi za ekstenzivne veličine. U skladu sa drugim zakonom termodinamike, pri uzajamnom delovanju dva tela sa različitim temperaturama prelazi toplota od tela sa većom energijom na telo sa manjom energijom. Prelaženje toplote vrši se kondukcijom, konvekcijom i radijacijom. Promena toplotnog stanja tela koja tada nastaje praćena je popratnim efektima i fenomenima, kao što su: dilatacija, ekspanzija, termoelektricitet, zračenje, itd. Zbog propratnih efekata menjaju se određena fizička svojstva tela, tj. odgovarajuće veličine: dužina, volumen, termoelektrična sila, električni otpor i dr. Ovo su termometrijske veličine, jer se njihovim direktnim merenjem dolazi do vrednosti temperature. Iz ovoga se zaključuje da je temperaturu moguće izmeriti samo posrednim putem, preko termometrijskih veličina, koje su podložne direktnom merenju i u funkcionalnoj su vezi sa temperaturom. Senzor za merenje temperature obično se naziva termometrom, a oblast merenja temperature termometrijom. Temperatura je najčešće merena veličina u tehnološkim procesima i na nju otpada oko 60% svih merenja u toj oblasti. Veliki je značaj merenja temperature i u drugim oblastima nauke i tehnike, u medicini, u svakodnevnom životu. [1].

Ovaj rad ima cilj da prikaže način na koji Laboratorija za metrologiju izvršava etaloniranja silotermometara. Svi termini i definicije su u skladu sa SRPS ISO/IEC 9000:2001, SRPS ISO/IEC 17025:2006 i Međunarodnim rečnikom osnovnih i opštih termina u metrologiji.

II. Priprema za rad

A. *Merna nesigurnost*

Merna nesigurnost iskazana u ovom radu je proširena merna nesigurnost, gde je standardna merna nesigurnost pomnožena faktorom obuhvata k = 2, što za slučaj normalne raspodele greške odgovara verovatnoći od približno 95 %. Najbolje mogućnosti etaloniranja su prikazane u Tabeli I.

TABELA I
NAJBOLJE MOGUĆNOSTI ETALONIRANJA

Veličina	Predmet etaloniranja	Opseg	Merna nesigurnost
Temperatura	Silotermometri	od -30 °C do 70 °C	≤ 0,5 °C

B. *Priprema za etaloniranje*

Priprema za etaloniranje obuhvata detaljno upoznavanje objekta etaloniranja na osnovu priložene dokumentacije, eventualnih prethodnih iskustava sa ranijih etaloniranja i ostalih informacija u kontaktima sa vlasnikom objekta, proizvođačem i/ili serviserom sistema silotermometara.

C. *Vizuelni pregled*

Vizuelnim pregledom se utvrđuje usaglašenost oznaka, opšte stanje objekta etaloniranja i konstatuju se eventualna oštećenja. Konstatuje se i da li ima ćelija bez sajli.

D. *Provera referentnih uslova etaloniranja*

Referentni uslovi za merenje su: temperatura okoline od 10 °C do 30 °C i relativna vlažnost vazduha od 40 % do 80 %.

III. Merna oprema

Oprema neophodna za etaloniranje silotermometara navedena je u Tabeli II.

Ivan Gutai – Laboratorija za metrologiju, Fakultet tehničkih nauka, Novi Sad, Srbija (e-mail: gutai@uns.ac.rs).

Bojan Vujičić – Laboratorija za metrologiju, Fakultet tehničkih nauka, Novi Sad, Srbija (e-mail: bojanvuj@uns.ac.rs).

Nemanja Gazivoda – Laboratorija za metrologiju, Fakultet tehničkih nauka, Novi Sad, Srbija (e-mail: nemanjagazivoda@uns.ac.rs).

TABELA II
OPREMA NEOPHODNA ZA ETALONIRANJE SILOTERMOMETARA

Oznaka	Naziv	Tip
MΩ	Instrument za merenje električne otpornosti izolacije	Iskra MA 2075
R_s	Dekadna kutija otpornosti	GenRad, Model 1433G
Ω	Digitalni multimetar	Fluke, Model 87V

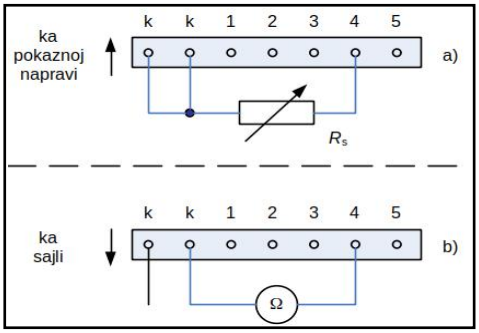
IV. POSTUPAK MERENJA

A. *Izbor sajli*

Na osnovu pokazivanja temperatura sa svih mernih mesta, konstatuju se eventualne smetnje na pojedinim kanalima (prekidi veza sa senzorima ili njihova neispravnost). Konstatuju se i eventualne velike razlike u merenim temperaturama na bliskim mernim mestima i traže se uzroci za to. Izaberu se kritične sajle i zabeleže se temperature Tm1(i) (i – nivo senzora u sajli) koje odgovaraju senzorima temperature na izabranim sajlama.

B. *Merenja na nadsilosnoj galeriji*

Sa kompetentnom osobom (serviser i/ili radnik silosa) odlazi se na nadsilosnu galeriju. Utvrđuju se merna mesta odabranih ćelija, redosled senzora, kompenzacioni otpornici, itd. Odspoji se sajla od multipleksera (pokazne naprave) i pristupa se merenjima prema šemi na slici 1.



Sl 1. Šema veza za a) etaloniranje pokazne naprave silotermometra; b) proveru ispravnosti senzora temperature

a) Na kontakte prema pokaznoj napravi priključuje se dekadna kutija otpornosti na kojoj je postavljena etalonska otpornost R_s. Imajući u vidu tip senzora temperature (Pt-100, Pt-1000, NTC, ...), vrednostima R_s dodeljuju se vrednosti temperature T_{s2}. Na pokaznoj napravi očitavaju se pokazivanje $T_{m2}(i)$. Vrednosti R_s biraju se tako da odgovaraju nazivnim temperaturama etaloniranja: 20 °C, 30 °C , 40 °C i 50 °C. Rezultati merenja se unose u Tabelu III.

b) Na kontakte prema senzoru temperature priključuje se merilo otpornosti (digitalni multimetar). Izmeri se otpornost kompenzacionog otpornika R_k. Zatim se izmere otpornosti senzora temperature $R_T(i)$ i serijski vezanog kompenzacionog otpornika R_k. Od tako izmerenih otpornosti oduzmu se vrednosti otpornosti kompenzacionog otpornika R_k. Imajući u vidu tip senzora temperature, dobijenim vrednostima pridružuju se temperature $T_{s1}(i)$. Porede se tako izračunata temperature sa temperaturama $T_{m1}(i)$, izmerenim pomoću pokazne naprave. Rezultati merenja se unose u Tabelu IV.

TABELA III
Tabela za evidenciju rezultata etaloniranja pokazne naprave silotermometra

Kompenzaciona otpornost R_k	Nivo senzora i	$R_T + R_k$	Temperatura odgovarajuća izmerenoj otpornosti T_{s1}	Očitanje pokazne naprave $T_{m1}(i)$
Ω		Ω	°C	°C
	1			
	2			
	3			
	4			
	...			

TABELA IV
Tabela za evidenciju rezultata provere ispravnosti temperaturskih sondi

Nazivna temperatura T_n	Etalonska otpornost R_s	Temperatura odgovarajuća etalonskoj otpornosti T_{s2}	Nivo senzora i	Očitanje pokazne naprave $T_{m2}(i)$	Greška pokazne naprave $T_{m2}(i)$ - T_{s2}
°C	Ω			°C	°C
20			1		
			2		
			3		
			4		
			...		
30			1		
			...		

C. *Merenje otpornosti izolacije*

Merenja otpornosti izolacije se sprovodi tako što se krajevi (dovodne žice) otpornog temperaturnog senzora kratko vežu i zajedno dovedu na jedan kraj instrumenta za merenje električne otpornosti izolacije (megaommetra). Drugi kraj megaommetra se krokodil štipaljkom spoji na uzemljenu tačku silosa. Ispitni napon megaommetra treba da bude jednosmerni napon u granicama od 10 V do 100 V.

D. *Metod etaloniranja*

Etaloniranje silotermometra vrši se metodom poređenja. Očitanja/nominalne vrednosti na instrumentu koji se etalonira porede se sa očitanjima na etalonskom instrumentu.

E. *Etaloniranje pokazne naprave*

Model izračunavanja greške izvora $G(T)$ je dat jednačinom (1):

$G(T)$ Greška pokazne naprave;

T_x Očitanje temperature na pokaznoj napravi;

δT_x Korekcija temperature očitane na pokaznoj napravi, zbog konačne rezolucije displeja pokazne naprave.

R_s Nazivna otpornost etalona otpornosti.

k_{RT} Korekcija etalonske otpornosti, iz sertifikata o etaloniranju etalon otpornika.

k_T Korekcija etalonske otpornosti, zbog promene etalonske otpornosti u zavisnosti od temperature okoline.

α Temperaturski koeficijent otpornosti.

$R(0)$ Otpornost na temperaturi od 0 °C temperaturske sonde za koju je predviđena pokazna naprava.

$$G(T) = (T_x + \delta T_x) - T_s(R_s + k_{RT} + k_T) \approx (T_x + \delta T_x) - \frac{1}{\alpha}\left(\frac{R_s + k_{RT} + k_T}{R(0)} - 1\right) \quad (1)$$

U Tabeli V je dat primer budžeta nesigurnosti etaloniranja.

TABELA V
Budžet nesigurnosti etaloniranja pokazne naprave silotermometra, sa senzorima Pt-100 - Primer

Naziv veličine	Simbol	Vrednost	Parcijalna nesigurnost	Tip nesigurnosti	Raspodela	Koeficient osetljivosti	Doprinos nesigurnosti
Očitanje na pokaznoj napravi (°C)	T_x	25,0					
Korekcija temperature, zbog konačne rezolucije displeja pokazne naprave (°C)	δT_x	0	$29\text{x}10^{-3}$	B	pravougaona	1	0,029
Nazivna otpornost etalona otpornosti (Ω)	R_s	110					

Korekcija etalonske otpornosti, iz sertifikata o etaloniranju etalon otpornika (Ω)	k_{RT}	0	6,4x10^{-3}	B	pravougaona	-2,56	0,016
Korekcija etalonske otpornosti, zbog promene etalonske otpornosti u zavisnosti od temperature okoline (Ω)	k_T	0		B	pravougaona	-2,56	0,050
Temperatura, izračunata iz otpornosti senzora (°C)		25,68	20x10^{-3}				
Greška pokazne naprave (°C)	$G(T)$	-0,7					0,060

Proširena merna nesigurnost (k = 2): 0,12 °C.

V. PRIMER REZULTATA MERENJA

Ponovljivost očitanja temperature i otpornosti svih senzora na jednoj izabranoj sajli, na sistemu silotermometara prikazana je Tabelama VI i VII.

TABELA VI
Tabela za evidenciju ponovljivosti očitanja temperature

Redni broj merenja	Nivoi					
	1	2	3	4	5	6
	T	*T*	*T*	*T*	*T*	*T*
	°C	°C	°C	°C	°C	°C
1	31,8	37,0	34,5	34,6	35,9	33,4
2	31,8	37,0	34,6	34,5	35,9	33,4
3	31,8	37,0	34,6	34,5	35,9	33,4
4	31,8	37,0	34,6	34,5	35,9	33,4

TABELA VII
Tabela za evidenciju ponovljivosti očitanja otpornosti

Redni broj merenja	Nivoi					
	1	2	3	4	5	6
	R	*R*	*R*	*R*	*R*	*R*
	Ω	Ω	Ω	Ω	Ω	Ω
1	1125,7	1144,9	1137,1	1137,3	1141,2	1131,8
2	1125,7	1144,9	1137,1	1137,3	1141,2	1131,8
3	1125,7	1144,9	1137,1	1137,3	1141,2	1131,8
4	1125,7	1144,9	1137,1	1137,3	1141,2	1131,8

LITERATURA

[1] M. Popović, Senzori i merenja, Zavod za udžbenike i nastavna sredstva, Srpsko Sarajevo, 2004

ABSTRACT

In this paper is given one of examples of silo thermometers calibration. Measurements are taking place on object on field. Best measurement abilities are shown. Measuring equipment is listed and described. Applied measuring procedures are described. Processing procedure of results is shown.

Calibration manual for silo thermometers

Ivan Gutai, Bojan Vujičić, Nemanja Gazivoda

Prilog etaloniranju pokaznih naprava termometara sa otpornim sondama

Stefan Mirković, Nemanja Gazivoda, Bojan Vujičić, Đorđe Novaković, Platon Sovilj

***Apstrakt*—U radu je izložen primer postupka etaloniranja pokaznih naprava termometara sa otpornim sondama metodom poređenja u uslovima unutar i u uslovima van metrološke laboratorije. Prikazane su najbolje mogućnosti merenja, opisana je priprema za merenje, navedena je merna oprema koja se koristi, opisani su postupci merenja koje treba sprovesti i prikazana je obrada dobijenih rezultata**

***Ključne reči*—termometar; etalonski otpornik; multimetar; etaloniranje; merna nesigurnost; budžet merne nesigurnosti; metrologija.**

I. Uvod

Ovaj rad ima cilj da prikaže jedan od načina na koji Laboratorija za metrologiju Fakulteta tehničkih nauka (FTN) u Novom Sadu izvršava etaloniranje pokaznih naprava termometara. Krajnji proizvod etaloniranja je dokument Uverenje o etaloniranju koje se izdaje klijentu saglasno dokumentovanom sistemu kvaliteta.

Postupak etaloniranja pokaznih naprava termometara sa otpornim sondama određen je radnim uputstvom izrađenim od strane tima saradnika Laboratorije za metrologiju. U radnom uputstvu, svi termini i definicije su u skladu sa SRPS ISO/IEC 9000:2015, SRPS ISO/IEC 17025:2017 i Međunarodnim rečnikom osnovnih i opštih termina u metrologiji.

U ovom radu obrađeni su termometri sa otpornim sondama (RTD). Termometri sa otpornim sondama zasnivaju se na merenju električne otpornosti ugrađene sonde koja zavisi od temperature. Etaloniranje pokazne naprave vrši se metodom poređenja. Na pokaznu napravu se, umesto temperaturske sonde, priključuje etalonski otpornik za simuliranje otpornih sondi. Očitanja na pokaznoj napravi termometra porede se sa izračunatim vrednostima temperature, koje odgovaraju zadatim etalonskim vrednostima otpornosti.

Za proračun merne nesigurnosti određivanja greške pokazne naprave ne uzimaju se u obzir otpornosti priključnih vodova.

Stefan Mirković – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg D. Obradovića 6, 21000 Novi Sad, Srbija (e-mail: mirkovicst@uns.ac.rs).

Nemanja Gazivoda – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg D. Obradovića 6, 21000 Novi Sad, Srbija (e-mail: nemanjagazivoda@uns.ac.rs).

Bojan Vujičić – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg D. Obradovića 6, 21000 Novi Sad, Srbija (e-mail: bojanvuj@uns.ac.rs).

Đorđe Novaković – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg D. Obradovića, 21000 Novi Sad, Srbija (e-mail: djordjenovakovic@uns.ac.rs).

Platon Sovilj – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg D. Obradovića, 21000 Novi Sad, Srbija (e-mail: platon@uns.ac.rs).

II. Merne mogućnosti i merna oprema

U Tabeli I prikazane su merne mogućnosti etaloniranja pokaznih naprava termometara sa otpornim sondama, raspoloživom opremom Laboratorije.

TABELA I
Merne mogućnosti

Veličina	Predmet etaloniranja	Opseg	Merna nesigurnost*
Temperatura	Pokazna naprava termometara sa otpornom sondom (unutar laboratorije)	od -200 °C do +50 °C	0,010 °C
		od +50 °C do +500 °C	0,040 °C
	Pokazna naprava termometara sa otpornom sondom (van laboratorije)	od -200 °C do +50 °C	0,15 °C
		od +50 °C do +500 °C	0,50 °C

*Merna nesigurnost je proširena merna nesigurnost, gde je standardna merna nesigurnost pomnožena faktorom obuhvata k=2, što za slučaj normalne raspodele greške odgovara verovatnoći od približno 95 %.

Merna oprema

U Tabeli II prikazani su referentni etaloni Laboratorije.

TABELA II
Referentni etaloni

Naziv	Tip
Digitalni multimetar	Hewlett Packard 3458A

Radni etaloni Laboratorije koji se koriste za etaloniranje pokaznih naprava su prikazani u Tabeli III.

TABELA III
Radni etaloni

Naziv	Tip
Digitalni multimetar	Fluke 8846A
Dekadna kutija otpornosti	GR 1433-G
	GR 1433-H

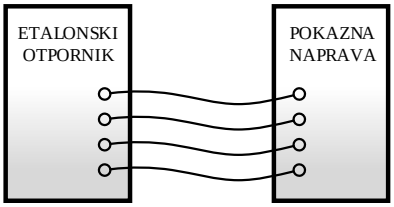
III. ETALONIRANJE UNUTAR LABORATORIJE ZA METROLOGIJU

Priprema za etaloniranje u Laboratoriji podrazumeva vizuelni pregled objekta etaloniranja, konstataciju da je objekt pripremljen za etaloniranje i proveru referentnih uslova etaloniranja. Potrebno da objekt etaloniranja bude u Laboratoriji najmanje 12 sati pre početka etaloniranja, da bi se njegova temperatura izjednačila sa temperaturom okoline. Vizuelnim pregledom se utvrđuje opšte stanje objekta etaloniranja i konstatuju se eventualna oštećenja. Utvrđuje se, takođe, i postojanje dokumentacije o objektu etaloniranja, relevantne za etaloniranje. Referentni uslovi za merenje u Laboratoriji su:

Temperatura okoline: (23 ± 2) °C;
Relativna vlažnost vazduha: (45 ± 15) %.

Referentni uslovi u Laboratoriji se održavaju stalno, i podrazumeva se ispunjenje navedenih zahteva neposredno pre merenja.

Analiziraće se slučaj kada je otporna sonda izrađena od metala (Pt-100, Pt-1000 i sl.).



Sl. 1. Blok šema etaloniranja pokazne naprave termometra sa otpornom sondom

Model izračunavanja greške E_X pokazne naprave je:

$$E_X = (T_X + \delta T_X) - T(R_S + \delta R_S) \quad (1)$$

gde je:

T_X Pokazivanje pokazne naprave;
δT_X Korekcija očitanja pokazne naprave, zbog konačne rezolucije očitavanja;
$T(R)$ Temperatura koja odgovara otpornosti R , definisana standardom ili od strane proizvođača pokazne naprave;
R_S Etalonska otpornost kojom se simulira otporna sonda;
δR_S Korekcija zbog netačnosti merenja etalonske otpornosti;

Vrednost etalonske otpornost R_S meri se digitalnim multimetrom sa granicama greške G_{Rs} određene specifikacijom proizvođača. Smatra se da gustina raspodele verovatnoće greške merenja digitalnim multimetrom kao i raspodela greške očitavanja pokazne naprave imaju ravnomernu raspodelu.

Standardna merna nesigurnost očitanja pokazne naprave zbog konačne rezolucije očitavanja određena je formulom:

$$u(\delta T_X) = \frac{LSD}{2\sqrt{3}} \quad (2)$$

gde je LSD (*Least Significiant Digit*) rezolucija displeja pokazne naprave.

Standardna merna nesigurnost koja potiče od digitalnog multimetra određena je izrazom:

$$u(\delta R_S) = \frac{G_{Rs}}{\sqrt{3}} \quad (3)$$

Koeficijenti osetljivosti su:

$$c_1 = \frac{\partial E_X}{\partial \delta T_X} = 1 \quad (4)$$

$$c_2 = \frac{\partial E_X}{\partial \delta R_S} = -\frac{\partial T(R_S + \delta R_S)}{\partial \delta R_S} \quad (5)$$

Merna nesigurnost određivanja greške etaloniranja pokazne naprave predviđene za rad sa otpornim sondama:

$$u(E_X) = \sqrt{[c_1 \cdot u(\delta T_X)]^2 + [c_2 \cdot u(\delta R_S)]^2} \quad (6)$$

Proširena merna nesigurnost U definisana je sa:

$$U = k \cdot u(E_X) \quad (7)$$

Etalonirana je pokazna naprava, za opseg temperatura od -50 °C do +600 °C, sa rezolucijom od LSD=0,01 °C , predviđena da se na nju, četvorožično, priključi otporna sonda Pt-100. Vrednost etalonskog otpornika kojim se simulira otporna sonda izmerena je digitalnim multimetrom Hewlett Packard 3458A.

TABELA IV
TABELA ZA EVIDENCIJU REZULTATA ETALONIRANJA

Očitanje pokazne naprave	Etalonska otpornost	Etalonska temperatura*	Greška pokazne naprave	Nesigurnost**	Faktor obuhvata
T_X (°C)	R_S (Ω)	$T(R_S)$ (°C)	E_X (°C)	U (°C)	k
-50,75	80	-50,771	0,021	0,0095	2
0,01	100	0,000	0,010	0,011	2
103,95	140	103,943	0,007	0,011	2
211,30	180	211,289	0,011	0,013	2
294,26	210	294,246	0,014	0,015	2
408,42	250	408,450	0,030	0,017	2
497,03	280	497,067	0,037	0,019	2
588,45	310	588,491	0,041	0,021	2

*Temperatura koja odgovara etalonskoj otpornosti R_S, prema standardu EN 60751:1996

**Merna nesigurnost je proširena merna nesigurnost, gde je standardna merna nesigurnost pomnožena faktorom obuhvata

k=2, što za slučaj normalne raspodele greške odgovara verovatnoći od približno 95 %. Tabela V ujedno prikazuje i prilog uz Uverenje o etaloniranju koje izdaje Laboratorija.

TABELA V
BUDŽET MERNE NESIGURNOSTI ETALONIRANJA UNUTAR LABORATORIJE ZA TEMPERATURU OD +25,684 °C

Naziv veličine	Simbol	Ocena	Parcijalna nesigurnost	Tip nesigurnosti	Raspodela	Koeficijent osetljivosti	Doprinos nesigurnosti
Očitanje etaloniranog termometra	T_X	25,70 °C					
Korekcije očitavanja zbog konačne rezolucije očitavanja	δT_X	0 °C	0,0029 °C	B	ravnomerna	1	0,0029 °C
Etalonska otpornost	R_S	110 Ω					
Temperatura koja odgovara etalonskoj otpornosti R_S	T_S	25,684 °C					
Korekcija etalonske otpornosti R_S zbog netačnosti merenja njene otpornosti	δR_S	0 Ω	$1{,}9\cdot10^{-3}$ Ω	B	ravnomerna	$-2{,}58$ °C·Ω^{-1}	0,0048 °C
Greška pokazne naprave	E_X	0,016 °C	Kombinovana merna nesigurnost				0,0056 °C
			Proširena merna nesigurnost (k=2)				0,012 °C

IV. ETALONIRANJE VAN LABORATORIJE ZA METROLOGIJU

Priprema za etaloniranje podrazumeva proveru metroloških karakteristika etalonskog sistema pre odlaska na mesto etaloniranja, prevoz do mesta etaloniranja, vizuelni pregled objekta etaloniranja, konstataciju da je objekt pripremljen za etaloniranje i proveru referentnih uslova etaloniranja. Da bi se smanjili rizici tokom prenošenja opreme do i od mesta etaloniranja, pre i posle prenosa vrši se provera osnovnih metroloških karakteristika etalona. Po pravilu, oprema za etaloniranje do i od mesta etaloniranja prenosi se putničkim automobilom.

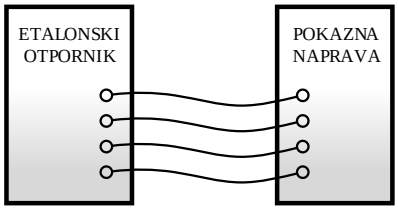
Za etaloniranja van Laboratorije je potrebno da objekt etaloniranja i etalonska oprema budu u prostoru namenjenom za etaloniranje najmanje jedan sat pre početka etaloniranja, da bi se njihova temperatura približno izjednačila sa temperaturom okoline. Za etaloniranje van Laboratorije utvrđuje se i opšte stanje prostora u kome treba da se obavi etaloniranje: prostor mora da je čist, bez prašine i drugih agensa koji mogu da deluju na mernu opremu, da bude suv i da se u njemu ne odvijaju druge aktivnosti koje mogu da nepovoljno utiču na postupak etaloniranja. Vizuelnim pregledom se utvrđuje opšte stanje objekta etaloniranja i konstatuju se eventualna oštećenja. Utvrđuje se, takođe, i postojanje dokumentacije o objektu etaloniranja, relevantne za etaloniranje.

Referentni uslovi za merenje van Laboratorije su:

Temperatura okoline: 10 °C do 30 °C;
Relativna vlažnost vazduha: 40 % do 80 %.

Neposredno pre merenja podrazumeva se ispunjenje navedenih referentnih uslova.

Analiziraće se slučaj kada je otporna sonda izrađena od metala (Pt-100, Pt-1000 i sl.).



Sl. 2. Blok šema etaloniranja pokazne naprave termometra sa otpornom sondom

Model izračunavanja greške E_X pokazne naprave je:

$$E_X = (T_X + \delta T_X) - T(R_S + \delta R_S + \delta R_{S-temp}) \qquad (8)$$

gde je:

T_X Pokazivanje pokazne naprave;
δT_X Korekcija očitanja pokazne naprave, zbog konačne rezolucije očitavanja;
$T(R)$ Temperatura koja odgovara otpornosti R , definisana standardom ili od strane proizvođača pokazne naprave;
R_S Etalonska otpornost kojom se simulira otporna sonda;
δR_S Korekcija zbog netačnosti merenja etalonske otpornosti;
$\delta R_{S\text{-temp}}$ Korekcija zbog netačnosti merenja etalonske otpornosti ako temperatura okoline izlazi van referentnog opsega temperatura etalonskog merila otpornosti;

Vrednost etalonske otpornost R_S meri se digitalnim

multimetrom sa granicama greške G_{Rs} određene specifikacijom proizvođača. Ako ambijentalna temperatura izađe van referentnog opsega multimetra, granice grešaka merenja se povećavaju zbog uticaja dodatne greške $G_{Rs\text{-}temp}$, koja je posledica ambijentalne temperature. Smatra se da gustina raspodele verovatnoće greške merenja digitalnim multimetrom kao i raspodela greške očitavanja pokazne naprave imaju ravnomernu raspodelu.

Standardna merna nesigurnost očitanja pokazne naprave zbog konačne rezolucije očitavanja određena je formulom:

$$u(\delta T_X) = \frac{LSD}{2\sqrt{3}} \tag{9}$$

Standardna merna nesigurnost koja potiče od digitalnog multimetra određena je izrazom:

$$u(\delta R_S) = \frac{G_{Rs}}{\sqrt{3}} \tag{10}$$

Standardna merna nesigurnost koja potiče od dodatne greške multimetra usled ambijentalne temperature je:

$$u(\delta R_{S-temp}) = \frac{G_{Rs-temp}}{\sqrt{3}} \tag{11}$$

Koeficijenti osetljivosti su:

$$c_1 = \frac{\partial E_X}{\partial \delta T_X} = 1 \tag{12}$$

$$c_2 = \frac{\partial E_X}{\partial \delta R_S} = -\frac{\partial T(R_S + \delta R_S + \delta R_{S-temp})}{\delta R_S} \tag{13}$$

$$c_3 = \frac{\partial E_X}{\partial \delta R_{S-temp}} = -\frac{\partial T(R_S + \delta R_S + \delta R_{S-temp})}{\delta R_{S-temp}} \tag{14}$$

Merna nesigurnost određivanja greške etaloniranja pokazne naprave predviđene za rad sa otpornm sondama:

$$u(E_X) = \sqrt{[c_1 \cdot u(\delta T_X)]^2 + [c_2 \cdot u(\delta R_S)]^2 + [c_3 \cdot u(\delta R_{S-temp})]^2} \tag{15}$$

Etalonirana je pokazna naprava, za opseg temperatura od -50 °C do +600 °C, sa rezolucijom od LSD=0,01 °C, predviđena da se na nju, četvorožično, priključi otporna sonda Pt-100. Vrednost etalonskog otpornika kojim se simulira otporna sonda izmerena je digitalnim multimetrom Fluke 8846A.

TABELA VI
BUDŽET MERNE NESIGURNOSTI ETALONIRANJA VAN LABORATORIJE ZA TEMPERATURU OD +25,684 °C

Naziv veličine	Simbol	Ocena	Parcijalna nesigurnost	Tip nesigurnosti	Raspodela	Koeficijent osetljivosti	Doprinos nesigurnosti
Očitanje etaloniranog termometra	T_X	25,70 °C					
Korekcije očitavanja zbog konačne rezolucije očitavanja	δT_X	0 °C	0,0029 °C	B	ravnomerna	1	0,0029 °C
Etalonska otpornost	R_S	110 Ω					
Temperatura koja odgovara etalonskoj otpornosti R_S	T_S	25,684 °C					
Korekcija etalonske otpornosti R_S zbog netačnosti merenja njene vrednosti	δR_S	0 Ω	$1{,}9\cdot10^{-3}$ Ω	B	ravnomerna	-2,58 °C·Ω^{-1}	0,0048 °C
Korekcija etalonske otpornosti R_S, ako temperatura okoline izlazi van referentnog opsega temperatura etalonskog merila otpornosti	$\delta R_{S\text{-}temp}$	0 Ω	$1{,}2\cdot10^{-3}$ Ω	B	ravnomerna	-2,58 °C·Ω^{-1}	0,0031 °C
Greška pokazne naprave	E_X	0,016 °C	Kombinovana merna nesigurnost				0,0064 °C
			Proširena merna nesigurnost (k=2)				0,013 °C

V. ZAKLJUČAK

Etaloniranje pokaznih naprava termometara sa otpornim sondama saglasno je dokumentovanom sistemu kvaliteta Laboratorije za metrologiju. Opisana metoda etaloniranja je jedna od metoda koje se koriste u Laboratoriji za etaloniranje pokaznih naprava termometara. Napredak u tačnosti merenja do danas dostigao je nivo da rezolucija i merna nesigurnost instrumenata prednjače ispred mogućnosti apsolutnih merenja.

Postalo je teško odgovoriti na pitanja šta nam znači merenje otpornosti sa rezolucijom 0,001 ppm kad je om (Ω) definisan sa mernom nesigurnošću reda 0,2 ppm [3]. Međutim, merna nesigurnost opisane metode ispunjava većinu zahteva klijenata Laboratorije, kada su u pitanju pokazne naprave termometara.

ZAHVALNICA

Ovaj rad je delom podržan od strane projekta ELEMEND (šifra projekta: 585681-EEP-1-2017-EL-EPPKA2-CBHE-JP).

LITERATURA

[1] "ETALONIRANJE POKAZNIH NAPRAVA TERMOMETARA SA OTPORNIČKIM SONDAMA I/ILI TERMOPAROVIMA", Radno uputstvo, Q3.JIM.19, Laboratorija za metrologiju, Fakultet tehničkih nauka, Novi Sad, 2015.

[2] "Guidelines on the Calibration of Temperature Indicators and Simulators by Electrical Simulation and Measurement", EURAMET cg-11, Version 2.0, Calibration Guide, ISBN 978-3-942992-08-4.

[3] R. Radetić, *Električna otpornost : pojava i merenje : sa originalnim rešenjima autora,* Niš, 2015, ISBN 978-86-80134-03-1.

[4] "Evaluation of measurement data – Guide to the expression of encertainty in measurement", JCGM 100:2008

[5] "International vocabulary of metrology – Basic and general concepts and associated terms (VIM)", JCGM 200:2012

[6] A. Dunjić, J. Pantelić-Babić, M. Pavićević, "Postupak etaloniranja ampermetara i kalibratora jednosmerne električne struje u dokumentovanom sistemu kvaliteta ZMDM", *Zbornik radova 50. Konferencije za ETRAN*, vol. *III*, Beograd, 2006.

[7] "HP 3458A, Operating, Programming and Configuration Manual, Hewlett Packard", Edition 1, USA, May 1988

[8] "8845A/8846A Digital Multimeter Users Manual", Fluke Corporation, USA, July 2006

ABSTRACT

The paper presents an example of the method of calibration temperature indicators with resistive probes by comparison method, in conditions inside and outside the metrology laboratory. The best measurement posibilities and measurement equipment are presented, the preparation for measurements and measurement procedures is described, and processing of the obtained results is shown.

The Contribution to The Calibration of Temperature Indicators With Resistive Probes

Stefan Mirković, Nemanja Gazivoda, Bojan Vujičić, Đorđe Novaković, Platon Sovilj

Prilog etaloniranju termometara sa direktnim očitavanjem u laboratorijskim uslovima

Marina Bulat, *Member, IEEE*, Nemanja Gazivoda, *Member, IEEE*, Ivan Gutai, *Member, IEEE*, Bojan Vujičić, *Member, IEEE*, Đorđe Novaković, *Member, IEEE* i Platon Sovilj, *Member, IEEE*

***Apstrakt*— U ovom radu je dat primer postupka etaloniranja termometara sa direktnim očitavanjem u Laboratoriji za metrologiju Fakulteta tehničkih nauka u Novom Sadu. Prikazane su najbolje mogućnosti merenja, opisana je priprema za merenje i navedena je merna oprema koja je korišćena. Opisani su postupci merenja koji treba da se sprovedu i prikazana je obrada dobijenih rezultata. Dati su primeri koji ilustruju primenu ovog uputstva. Svi termini i definicije su u skladu sa SRPS ISO/IEC 9000:2001, SRPS ISO/IEC 17025:2017 i Međunarodnim rečnikom osnovnih i opštih termina u metrologiji.**

***Ključne reči*—** merna oprema; etaloniranje; metrologija; termometar; merna nesigurnost.

I. Uvod

Uređaji za merenje temperature ili temperaturnog gradijenta nazivaju se termometri. Međusobno se razlikuju kako po principu na kojem se zasniva njihov rad, tako i prema mernom području. Proces etaloniranja koji prethodi izdavanju uverenja o etaloniranju je u upotrebi u industriji, laboratorijama, ocenjivanju usaglašenosti tela i preduzećima, kako bi zadovoljili zahteve standarda. Uverenje o etaloniranju je sredstvo kojim se obezbeđuje dokaz sledivosti merenja.

U Tabeli I su prikazane merne mogućnosti etaloniranja termometara sa direktnim očitavanjem u Laboratoriji za metrologiju Fakulteta tehničkih nauka u Novom Sadu (u daljem tekstu samo Laboratorija). Merna nesigurnost iskazana u ovom radu je proširena merna nesigurnost, kod koje je standardna merna nesigurnost pomnožena faktorom obuhvata $k = 2$, što za slučaj normalne raspodele greške odgovara verovatnoći od približno 95%.

Marina Bulat – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail: marina.bulat@uns.ac.rs). Nemanja Gazivoda – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail: nemanjagazivoda@uns.ac.rs). Ivan Gutai – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail: gutai@uns.ac.rs). Bojan Vujičić– Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail: bojanvuj@uns.ac.rs). Đorđe Novaković – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail: djordjenovakovic@uns.ac.rs). Platon Sovilj – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail: platon@uns.ac.rs).

TABELA I
NAJBOLJE MOGUĆNOSTI ETALONIRANJA U LABORATORIJI

Predmet etaloniranja	Temperaturni opseg	Merna nesigurnost
Stakleni termometri punjeni tečnošću, otpornički termometri, termometri sa termoparom	0 °C	0.05 °C
Stakleni termometri punjeni tečnošću, otpornički termometri, termometri sa termoparom	- 40 °C do 250 °C	0.10 °C
Stakleni termometri punjeni tečnošću, otpornički termometri, termometri sa termoparom	250 °C do 500 °C	2.0 °C
Termometri sa termoparom	500 °C do 1200 °C	3.5 °C

II. Priprema za rad

Priprema za etaloniranje u Laboratoriji podrazumeva vizuelni pregled objekta etaloniranja, konstataciju da je objekt pripremljen za etaloniranje i proveru referentnih uslova etaloniranja. Potrebno da objekt etaloniranja bude u Laboratoriji najmanje dvanaest sati pre početka etaloniranja, da bi se njegova temperatura izjednačila sa temperaturom okoline. Vizuelnim pregledom se utvrđuje opšte stanje objekta etaloniranja i konstatuju se eventualna oštećenja. Utvrđuje se, takođe, i postojanje dokumentacije o objektu etaloniranja, relevantne za etaloniranje.

Referentni uslovi za merenje u Laboratoriji su:

1. Temperatura okoline: (23 ± 2) °C
2. Relativna vlažnost vazduha: (45 ± 15)%.

Referentni uslovi u Laboratoriji se održavaju stalno, a provera podrazumeva ispunjenje navedenih zahteva neposredno pre merenja.

III. Merna oprema

Oprema, koja je neophodna za etaloniranje termometara sa direktnim očitavanjem, je navedena u Tabelama II, III i IV.

TABELA II
REFERENTNI ETALONI ZA ETALONIRANJE TERMOMETARA SA DIREKTNIM OČITAVANJEM

Naziv	Tip
Platinski otpornički termometar	5187 SA
Platinski otpornički termometar	5626
Etalonski termopar	S
Digitalni multimetar	HP 3458A

TABELA III
RADNI ETALONI ZA ETALONIRANJE TERMOMETARA SA DIREKTNIM OČITAVANJEM

Naziv	Tip
Platinski otporni termometar	884X-RTD
Platinski otporni termometar	884X-RTD
Etalonski termopar	S
Digitalni multimetar	8846A
Digitalni multimetar	8846A
Digitalni multimetar	8846A
Kalibrator za realizaciju 0 °C	9101
Kalibrator temperature	9103
Kalibrator temperature	9140
Kalibrator temperature	9141
Kalibrator temperature	Pegasus 1200
Drjuarova posuda za led	-
Garnitura etalonskih termometara punjenih živom	-
Garnitura otporničkih sondi Pt 100	-
Garnitura otporničkih sondi Pt 1000	-

TABELA IV
POMOĆNA OPREMA ZA ETALONIRANJE TERMOMETARA SA DIREKTNIM OČITAVANJEM

Naziv	Tip
Temperaturno kupatilo	TS-5D
Temperaturno kupatilo	MLW
Peć za etaloniranje	FTN
Lupa	-
Selektor	954
Selektor	SW142G-8-B

Tabela V daje pregled mogućnosti etaloniranja raspoloživom opremom Laboratorije u opsegu temperatura – 50 °C do 1200 °C. Plava polja ukazuju na nemogućnost pojedinih kombinacija.

TABELA V
PRIKAZ MOGUĆNOSTI ETALONIRANJA RASPOLOŽIVOM OPREMOM LABORATORIJE

Etalon / Sredina	5187 SA & HP 3458A	S(T9) & HP 3458A	884X-RTD & 8846A*	S(T1) & 8846A*	Kalibrator temperature
AVM (alkohol)	-50 °C do 25 °C				
AVM (ulje)	25 °C do 250 °C				
9101	0 °C		0 °C		0 °C
9103	-25 °C do 140 °C	-25 °C do 140 °C	-25 °C do 140 °C	-25 °C do 140 °C	-25 °C do 140 °C
9140	35 °C do 350 °C	35 °C do 350 °C	35 °C do 300 °C	35 °C do 350 °C	35 °C do 350 °C
9141	50 °C do 650 °C	50 °C do 650 °C	50 °C do 300 °C	50 °C do 650 °C	50 °C do 650 °C
Pegasus 1200	150 °C do 650 °C	150 °C do 1064 °C	150 °C do 300 °C	150 °C do 1064 °C	150 °C do 1200 °C
Peć	100 °C do 650 °C	100 °C do 1064 °C			

Najbolje mogućnosti etaloniranja u opsegu temperatura - 50 °C do 1200 °C moguće je postići kombinacijom raspoloživih etalona i sredina etaloniranja, kao što je prikazano u Tabeli VI. Plava polja ukazuju ili na nemogućnost pojedinih kombinacija ili na kombinacije koje ne daju najbolje mogućnosti etaloniranja. Neka od plavih polja mogu biti aktivirana u slučaju da se za sonde etaloniranih termometara postavljaju posebni zahtevi.

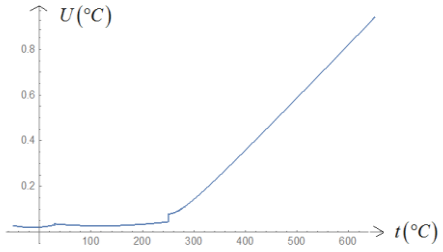
TABELA VI
PREGLED KOMBINACIJA ETALONA I SREDINA ETALONIRANJA RASPOLOŽIVOM OPREMOM LABORATORIJE KOJE DAJU NAJBOLJE MOGUĆNOSTI ETALONIRANJA

Etalon / Sredina	5187 SA & HP 3458A	S(T9) & HP 3458A	884X-RTD & 8846A*	S(T1) & 8846A*	Kalibrator temperature
AVM (alkohol)	-50 °C do 25 °C				
AVM (ulje)	25 °C do 250 °C				
9101			0 °C		
9103			-25 °C do 140 °C		
9140			140 °C do 300 °C	300 °C do 350 °C	
9141				350 °C do 650 °C	
Pegasus 1200				650 °C do 1064 °C	1064 °C do 1200 °C
Peć	250 °C do 650 °C	650 °C do 1064 °C			

Imajući u vidu metrološke karakteristike etalona i svojstva zona etaloniranja, grafički smo predstavili najbolje mogućnosti etaloniranja u funkciji temperature etaloniranja i u zavisnosti od izbora etalonskog sistema.

A. – 50 °C *do* 650 °C

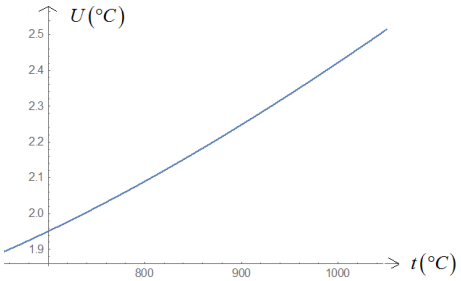
Etalonski sistem, prikazan na Sl. 1, sastoji se iz referentnog platinskog otporničkog termometra Tinsley 5187 SA i referentnog digitalnog multimetra HP 3458A. Za temperature etaloniranja u granicama – 50 °C do 25 °C, sredina etaloniranja je alkohol, u granicama 25 °C do 250 °C, sredina etaloniranja je silikonsko ulje i u granicama 250 °C do 650 °C, sredina etaloniranja je vazduh (peć).



Sl. 1. Proširena merna nesigurnost etaloniranja za faktor obuhvata $k = 2$ u zavisnosti od temperature, za etalonski sistem koji se sastoji od referentnog platinskog otporničkog termometra Tinsley 5187 SA i referentnog digitalnog multimetra HP 3458A

B. 650 °C *do* 1050 °C

Etalonski sistem, prikazan na Sl. 2, sastoji se iz referentnog termopara S tipa (Termotehna) i referentnog digitalnog multimetra HP 3458A. Sredina etaloniranja je vazduh (peć).



Sl. 2. Proširena merna nesigurnost etaloniranja za faktor obuhvata $k = 2$ u zavisnosti od temperature, za etalonski sistem koji se sastoji od referentnog termopara (Termotehna T9) i referentnog digitalnog multimetra HP 3458A

C. 1064 °C *do* 1200 °C

Etalon je suvo kupatilo Pegasus 1200. Proširena merna nesigurnost etaloniranja je ocenjena na 3.5 °C. Etaloniranje termometara vrši se metodom poređenja. Očitanja na etaloniranom termometru porede se sa očitanjima na etalonskom termometru.

IV. Etaloni

Laboratorija raspolaže opremom kojom je moguće realizovati u osnovi tri međusobno različita etalonska merna sistema.

A. Etalonski sistem sa otporničkom sondom i digitalnim multimetrom

Odstupanje (greška) pokazivanja rezultata merenja temperature etaloniranog termometra (DUT engl. *Device Under Test*) od temperature koju pokazuje etalonski merni sistem sa PRT sondom, modeluje se izrazom:

$$E_{DUT}=(T_{DUT}+\delta T_{DUT}) - \\ -(T(R_s+\delta R_s+k_{Rs})+k_{PRT}+k_{DriftPRT}+k_{Stab}+k_{HomR}+k_{HomA}),$$

gde je:

T_{DUT} Pokazivanje (aritmetička sredina pokazivanja) DUT; Analitički izraz za njenu nesigurnost je $u_{T_{DUT}}=\frac{S_{T\,DUT}}{\sqrt{n_{T\,DUT}}}$;

$S_{T\,DUT}$ Standardno odstupanje niza pojedinačnih očitanja DUT, a $n_{T\,DUT}$ broj tih očitanja;

δT_{DUT} Korekcija očitanja DUT, zbog konačne rezolucije očitavanja. Ocena korekcije je 0 °C. Standardna nesigurnost korekcije očitanja određena je formulom $u_{\delta T\,DUT}=\frac{LSD}{2\sqrt{3}}$, gde je *LSD* rezolucija displeja DUT;

R_s Otpornost etalonske sonde, PRT, izmerene etalonskim merilom otpornosti;

$T(R_s)$ Temperatura etaloniranja (aritmetička sredina niza očitanja), kao funkcija otpornosti R_s, definisana standardom, ili od strane proizvođača PRT, ili iz odgovarajućeg uverenja o etaloniranju. Njena nesigurnost je $u_{T(R_s)}=\frac{S_{Rs}}{\sqrt{n_{Rs}}}$, gde je S_{Rs} standardno odstupanje niza pojedinačnih očitanja R_s, a n_{Rs} broj tih očitanja;

δR_s Korekcija očitanja otpornosti R_s, zbog konačne rezolucije očitavanja;

k_{Rs} Korekcija merenja otpornosti R_s, zbog netačnosti merenja otpornosti. Ocena korekcije se preuzima iz uverenja o etaloniranju merila otpornosti. Nesigurnost korekcije se preuzima iz uverenja o etaloniranju ili se koriste podaci o tačnosti merila otpornosti, datih od strane proizvođača;

k_{PRT} Korekcija merenja temperature etaloniranja, zbog netačnosti PRT. Nesigurnost korekcije se preuzima iz uverenja o etaloniranju ili se koriste podaci o tačnosti merila otpornosti, datih od strane proizvođača;

$k_{DriftPRT}$ Korekcija merenja temperature etaloniranja, zbog drifta PRT;

k_{Stab} Korekcija merenja temperature etaloniranja, zbog nestabilnosti temperature etaloniranja tokom vremena etaloniranja;

k_{HomR} Korekcija merenja temperature etaloniranja, zbog radijalne nehomogenosti temperature u zoni etaloniranja;

k_{HomA} Korekcija merenja temperature etaloniranja, zbog aksijalne nehomogenosti temperature u zoni etaloniranja.

Analizu merne nesigurnosti rezultata etaloniranja DUT (merila temperature sa direktnim očitavanjem), kada se etalon sastoji od otporničke sonde (PRT) i digitalnog multimetra, prikazali smo u primerima u obliku budžeta merne nesigurnosti.

Primer 1.

Etaloniran je stakleni termometar punjen živom, mernog opsega 150 °C do 200 °C i vrednošću najmanjeg podeoka od 0.1 °C. Etalonski sistem se sastoji od platinskog otporničkog termometra, model 5187 SA, i digitalnog multimetra, model 8846A. Sredina etaloniranja je temperaturno kupatilo sa silikonskim uljem. Prikaz budžeta nesigurnosti rezultata etaloniranja prikazan je u Tabeli VII.

Budući da je objekt etaloniranja stakleni termometar punjen tečnošću, model greške termometra, koji je prikazan u (1), dopunjen je korekcijom očitanja temperature termometra, k_{DP}, zbog delimičnog potapanja termometra u ulje, kako je opisano u Referentnom dokumentu za etaloniranje ovakve vrste termometara. Kombinovana merna nesigurnost rezultata etaloniranja $u(e_{DUT})$ je 0.11 °C, a proširena merna nesigurnost rezultata etaloniranja $U(e_{DUT})$ za faktor obuhvata $k = 2$ iznosi 0.22 °C .

TABELA VII
PRIKAZ BUDŽETA MERNE NESIGURNOSTI ETALONIRANJA STAKLENOG TERMOMETRA IZ PRIMERA 1

Veličina	Simbol	Ocena	Nesigurnost	Tip	Raspodela	Osetljivost	Doprinos
Srednja vrednost očitanja temperature DUT (°C)	T_{DUT}	200.20	0	A	N	1	0.000
Korekcija očitanja temperature DUT, zbog konačne rezolucije očitanja (°C)	dT_{DUT}	0	2.9E-3	B	R	1	0.003
Korekcija očitanja temperature DUT, zbog delimičnog potapanja (°C)	k_{dp}	0.93	92.8E-3	B	R	1	0.093
Srednja vrednost otpornosti etalonskog PRT, izmerene etalonskim merilom otpornosti (Ω)	R_s	44.60841	30.0E-6	A	N	-10.4	0.000
Temperatura izmerena etalonskim termometrom (°C)	T_s	200.883					
Korekcija očitanja otpornosti R_s, zbog konačne rezolucije očitavanja (Ω)	dR_s	0	28.9E-6	B	R	-10.4	0.000
Korekcija merenja otpornosti Rs, zbog netačnosti merenja otpornosti (Ω)	k_{Rs}	0	4.9E-3	B	R	-10.4	0.051
Korekcija merenja temperature etaloniranja, zbog netačnosti PRT (°C)	k_{PRT}	0	20.8E-3	B	R	-1	0.021
Korekcija merenja temperature etaloniranja, zbog drifta PRT (°C)	$k_{DriftPRT}$	0	20.8E-3	B	R	-1	0.021
Korekcija merenja temperature etaloniranja, zbog nestabilnosti temperature etaloniranja tokom vremena etaloniranja (°C)	k_{Stab}	0	5.8E-3	B	R	-1	0.006
Korekcija merenja temperature etaloniranja, zbog radijalne nehomogenosti temperature u zoni etaloniranja (°C)	k_{HomR}	0	14E-3	B	R	-1	0.014
Korekcija merenja temperature etaloniranja, zbog aksijalne nehomogenosti temperature u zoni etaloniranja (°C)	k_{HomA}	0	14E-3	B	R	-1	0.014
Greška DUT	E_{DUT}	0.245					0.112

B. *Etalonski sistem sa termoparom i digitalnim multimetrom*

Greška e_{DUT} termometra koji se etalonira (DUT) modeluje se izrazom:

$$E_{DUT}=(T_{m\,DUT}+\delta T_{DUT}) - \\ -\left(T_s(k_{ess}+\delta e_S+ k_{es})+k_{TP}+k_{DriftTP}+k_{Stab}+k_{HomR}+k_{HomA}\right),$$

gde je:

$T_{m\,DUT}$ Rezultat (aritmetička sredina) merenja temperature DUT;

δT_{DUT} Korekcija merenja temperature DUT, zbog konačne rezolucije očitavanja;

T_s Temperatura etaloniranja, kao funkcija elektromotorne sile etalonskog termopara;

k_{ess} Korekcija očitanja elektromotorne sile e_s, zbog rasipanja rezultata merenja elektromotorne sile etalonskog termopara;

δe_S Korekcija očitanja elektromotorne sile e_s, zbog konačne rezolucije očitavanja elektromotorne sile etalonskog termopara;

k_{es} Korekcija merenja elektromotorne sile e_s, zbog netačnosti merenja elektromotorne sile etalonskog termopara. Ocena korekcije se preuzima iz uverenja o etaloniranju merila elektromotorne sile. Nesigurnost korekcije se preuzima iz uverenja o etaloniranju ili se koriste podaci o tačnosti merila elektromotorne sile, datih od strane proizvođača;

k_{TP} Korekcija merenja temperature etaloniranja, zbog netačnosti etalonskog termopara (TP). Nesigurnost korekcije se preuzima iz uverenja o etaloniranju ili se koriste podaci o tačnosti etalonskog termopara, datih od strane proizvođača;

$k_{DriftTP}$ Korekcija merenja temperature etaloniranja, zbog drifta etalonskog termopara;

k_{Stab} Korekcija merenja temperature etaloniranja, zbog nestabilnosti temperature etaloniranja tokom vremena etaloniranja;

k_{HomR} Korekcija merenja temperature etaloniranja, zbog radijalne nehomogenosti temperature u zoni etaloniranja;

k_{HomA} Korekcija merenja temperature etaloniranja, zbog aksijalne nehomogenosti temperature u zoni etaloniranja.

Analiza merne nesigurnosti rezultata etaloniranja DUT (merila temperature sa direktnim očitavanjem), kada se etalon sastoji od termopara i digitalnog multimetra, prikazana je primerima u obliku budžeta merne nesigurnosti.

Primer 2.

Etaloniran je digitalni termometar sa rezolucijom od 1 °C na temperaturi 300 °C.

T_{DUT} je srednja vrednost od pet uzastopnih očitanja DUT. Etalonski merni sistem se sastoji od etalonskog termopara (Termotehna, tip S, granice greške ± 0.5 °C) i digitalnog multimetra (8846 A, opseg 100 mV, rezolucija 10 nV, tačnosti ± (37 ppm od izmerene vrednosti + 35 ppm od mernog opsega)).

T_s je srednja vrednost od pet uzastopnih očitanja etalonskog merila temperature. Sredina je peć za etaloniranje termoparova (nestabilnost temperature etaloniranja tokom vremena iznosi 0.1 °C, radijalna nehomogenost 0.5 °C i aksijalna nehomogenost 1.0 °C). Termometar je etaloniran u Laboratoriji.

Prikaz budžeta nesigurnosti rezultata etaloniranja prikazan je u Tabeli VIII. Kombinovana merna nesigurnost rezultata etaloniranja $u(e_{DUT})$ je 1.3 °C, a proširena merna nesigurnost rezultata etaloniranja $U(e_{DUT})$ za faktor obuhvata k=2 iznosi 2.5 °C.

TABELA VIII
PRIKAZ BUDŽETA MERNE NESIGURNOSTI ETALONIRANJA TERMOMETRA IZ PRIMERA 2

Veličina	Simbol	Ocena	Nesigurnost	Tip	Raspodela	Osetljivost	Doprinos
Očitanje DUT (srednja vrednost) (°C)	$T_{m\,DUT}$	297.0	939E-3	A	N	1	939E-3
Korekcija očitanja DUT, zbog konačne rezolucije očitavanja, (°C)	dT_{DUT}	0	289E-3	B	R	1	289E-3
Ems etalonskog TP, izmerene etalonskim merilom ems (V)	e_s	2.3068E-3	72.9E-9	A	N	-110E+3	8.0E-3
Temperatura etaloniranja (°C)	T_s	298.22					
Korekcija očitanja ems e_s, zbog konačne rezolucije očitavanja (V)	de_s	0	2.9E-9	B	R	-110E+3	316E-6
Korekcija merenja ems e_s, zbog netačnosti merenja ems (V)	k_{es}	0	2.1E-6	B	R	-110E+3	227E-3
Korekcija merenja temperature etaloniranja, zbog netačnosti TP (°C)	k_{TP}	0	289E-3	B	R	-1	289E-3
Korekcija merenja temperature etaloniranja, zbog drifta PRT (°C)	$k_{DriftTP}$	0	289E-3	B	R	-1	289E-3
Korekcija merenja temperature etaloniranja, zbog nestabilnosti temperature etaloniranja tokom vremena etaloniranja (°C)	k_{Stab}	0	57.7E-3	B	R	-1	58E-3
Korekcija merenja temperature etaloniranja, zbog radijalne nehomogenosti temperature u zoni etaloniranja (°C)	k_{HomR}	0	289E-3	B	R	-1	289E-3
Korekcija merenja temperature etaloniranja, zbog aksijalne nehomogenosti temperature u zoni etaloniranja (°C)	k_{HomA}	0	577.4E-3	B	R	-1	577E-3
Greška DUT (°C)	e_{DUT}	-1.2				$u(e_{DUT})$:	1.3

C. *Kalibratori temperature*

Greška e_{DUT} termometra koji se etalonira (DUT) modeluje se izrazom:

$$E_{DUT}=(T_{m\,DUT}+\delta T_{DUT})-\left(T_s+k_{T_s}+k_{Stab}+k_{HomR}+k_{HomA}\right),$$

gde je:

$T_{m\,DUT}$ Rezultat (aritmetička sredina) merenja temperature DUT;

δT_{DUT} Korekcija merenja temperature DUT, zbog konačne rezolucije očitavanja ;

T_s Temperatura etaloniranja (pokazivanje kalibratora temperature);

k_{Ts} Korekcija zadate temperature etaloniranja, zbog netačnosti kalibratora temperature. Nesigurnost korekcije se preuzima iz uverenja o etaloniranju ili se koriste podaci o tačnosti kalibratora, datih od strane proizvođača;

k_{Stab} Korekcija merenja temperature etaloniranja, zbog nestabilnosti temperature etaloniranja tokom vremena etaloniranja;

k_{HomR} Korekcija merenja temperature etaloniranja, zbog radijalne nehomogenosti temperature u zoni etaloniranja;

k_{HomA} Korekcija merenja temperature etaloniranja, zbog aksijalne nehomogenosti temperature u zoni etaloniranja.

Analiza merne nesigurnosti rezultata etaloniranja DUT (merila temperature sa direktnim očitavanjem), kada je etalon kalibrator temperature, prikazana je primerima u obliku budžeta merne nesigurnosti.

Primer 3.

Etaloniran je digitalni termometar sa rezolucijom 0.1 °C na temperaturi – 20 °C. T_{DUT} je srednja vrednost od pet uzastopnih očitanja DUT. Etalonski merni sistem je kalibrator temperature Hart, 9103, granice greške ± 0.25 °C. T_s je temperatura zadata kalibratorom. Nestabilnost temperature etaloniranja tokom vremena iznosi 0.03 °C, radijalna nehomogenost 0.1 °C i aksijalna nehomogenost 0.1 °C. Termometar je etaloniran u Laboratoriji.

Prikaz budžeta nesigurnosti rezultata etaloniranja je prikazan u Tabeli IX. Greška DUT je – 0.20 °C, kombinovana standardna nesigurnost 0.17 °C, a proširena nesigurnost za faktor obuhvata $k = 2$ iznosi 0.35 °C.

TABELA IX
PRIKAZ BUDŽETA MERNE NESIGURNOSTI ETALONIRANJA TERMOMETRA IZ PRIMERA 3

Veličina	Simbol	Ocena	Nesigurnost	Tip	Raspodela	Osetljivost	Doprinos
Pokazivanje (aritmetička sredina pokazivanja) DUT	T_x	-20.2	0.045	A	N	1	0.045
Korekcija očitanja DUT, zbog konačne rezolucije očitavanja	δT_x	0	0.029	B	R	1	0.029
Pokazivanje kalibratora temperature	T_s	-20.0					
Korekcija zadate temperature, zbog netačnosti kalibratora temperature	k_{Ts}	0	0.144	B	R	-1	0.144
Korekcija zadate temperature etaloniranja, zbog nestabilnosti temperature etaloniranja tokom vremena etaloniranja	k_{Stab}	0	0.012	B	R	-1	0.012
Korekcija zadate temperature etaloniranja, zbog radijalne nehomogenosti temperature u zoni etaloniranja	k_{HomR}	0	0.058	B	R	-1	0.058
Korekcija zadate temperature etaloniranja, zbog aksijalne nehomogenosti temperature u zoni etaloniranja	k_{HomA}	0	0.058	B	R	-1	0.058
Greška merila temperature	E_x	-0.2					0.175

V. PROGRAM ZA OBRADU REZULTATA ETALONIRANJA

Pouzdana, jednoznačna i efikasna obrada rezultata etaloniranja svakako podrazumeva neki stepen računarske podrške. Za jednostavnije obrade, Laboratorija koristi odgovarajuće Excel-ove tabele, ali za etaloniranja koja potencijalno obuhvataju veliki broj etalona ili njihovih kombinacija, razvijanje i čuvanje za dalju upotrebu uglednih primeraka Excel-ovih tabela nije racionalno.

Primer 4.

Etaloniran je (izmišljeni) termometar sa direktnim pokazivanjem:

1) u tački – 20 °C

Srednja vrednost merenja DUT iznosi – 20.20 °C, standardna devijacija očitanja DUT 0.02 °C, broj pojedinačnih očitanja DUT je 5 i rezolucija displeja DUT je 0.01 °C. Srednja vrednost očitanja etalona iznosi 23 **Ω**, a standardna devijacija očitanja etalona 0.001 **Ω**. Broj pojedinačnih očitanja etalona je 5.

Kao etaloni su upotrebljeni PRT Tinsley 5187 SA i DMM HP 3458A, a kao sredina alkoholno kupatilo.

2) u tački 0 °C

Srednja vrednost merenja DUT je 0.1 °C, standardna devijacija očitanja DUT 0.02 °C, broj pojedinačnih očitanja DUT je 5, dok rezolucija displeja DUT iznosi 0.01 °C.

Kao etalon upotrebljen je Djuarov sud;

3) u tački 100 °C

Srednja vrednost merenja DUT je 100.1 °C, standardna devijacija očitanja DUT 0.2 °C, broj pojedinačnih očitanja DUT je 5, a rezolucija displeja DUT iznosi 0.1 °C.

Na kalibratoru Hart, 9103 zadata je temperatura od 100 °C;

4) u tački 200 °C

Srednja vrednost merenja DUT je 200.2 °C, standardna devijacija očitanja DUT 0.2 °C, broj pojedinačnih očitanja DUT je 5, a rezolucija displeja DUT iznosi 0.1 °C. Srednja vrednost očitanja etalona je 176.0 **Ω**, standardna devijacija očitanja etalona 0.005 **Ω**, broj pojedinačnih očitanja etalona je 5.

Kao etaloni upotrebljeni su PRT Fluke 884X i DMM Fluke 8846A, a u sredini Hart, 9140;

5) u tački 500 °C

Srednja vrednost merenja DUT je 500.2 °C, standardna devijacija očitanja DUT 0.2 °C, broj pojedinačnih očitanja DUT je 5, dok je rezolucija displeja DUT 0.1 °C. Srednja vrednost očitanja etalona je 4230 µV, standardna devijacija očitanja etalona 0.5 µV, broj pojedinačnih očitanja etalona je 5.

Kao etaloni upotrebljeni su TP Termotehna, S tip i DMM HP 3458A, a u sredini peć za etaloniranje;

6) u tački 1000 °C

Srednja vrednost merenja DUT je 1002 °C, standardna devijacija očitanja DUT 0.3 °C, broj pojedinačnih očitanja DUT je 5, dok je rezolucija displeja DUT 1 °C.

Na kalibratoru Isotech, Pegasus 1200 je zadata temperatura od 1000 °C.

Rezultati etaloniranja su predstavljeni na Sl. 3.

Etaloniranje termometra sa direktnim ocitavanjem

Proizvodjac: XXX
Model: ABC123
Serijski broj: 123A456
Datum etaloniranja: 30.11.2014. god.

Rezultat etaloniranja

Pokazivanje etalona °C	Ocitanje DUT °C	Greska DUT °C	Nesigurnost °C
-20.972	-20.20	0.77	0.035
0.000	0.01	0.01	0.012
100.00	100.1	0.1	0.43
200.25	200.2	-0.1	0.53
499.9	500.2	0.3	1.7
1000.0	1002.	2.	2.7

Sl. 3. Izgled priloga uz Uverenja o etaloniranju sa rezultatima etaloniranja

ZAHVALNICA

Ovaj rad je delom podržan od strane projekta ELEMEND (šifra projekta: 585681-EEP-1-2017-EL-EPPKA2-CBHE-JP).

LITERATURA

[1] DKD-R 5-1, "Kalibrierung von Widerstandsthermometern", Richtlinie, Deutscher Kalibrierdienst, Ausgabe 10/2003

[2] Q3.JIM.21, v.1, 2013, „Etaloniranje otporničkih termometara", Radno uputstvo, Laboratorija za metrologiju

[3] EURAMET cg-8, Version 2.0 (03/2011), "Calibration of Thermocouples"

[4] Q3.JIM.22, v.1, 2013, „Etaloniranje termoparova", Radno uputstvo, Laboratorija za metrologiju

[5] Q3.JIM.20, v.2, 2013, „Etaloniranje termometara sa direktnim očitavanjem", Radno uputstvo, Laboratorija za metrologiju

[6] EURAMET cg-11, Version 2.0 (03/2011), "Guidelines on the Calibration of Temperature Indicators and Simulators by Electrical Simulation and Measurement"

[7] Q3.JIM.19, v.2, 2013, „Etaloniranje pokaznih naprava termometara sa otporničkim sondama i/ili termoparovima", Radno uputstvo, Laboratorija za metrologiju

[8] J. V. Nicholas, D. R. White," Traceable Temperatures: An Introduction to Temperature Measurement and Calibration", Second Edition, JOHN WILEY & SONS, LTD.,2001

ABSTRACT

The paper presents an example of the procedure for the calibration of direct reading thermometers in the Laboratory. It demonstrates the optimal ways of measuring. In addition, it offers a description of the preparation for measuring and lists the measuring equipment that has been used. It also features a description of the measuring procedures that need to be applied and it shows the processing of the measurement results. It provides the examples illustrating the application of this instruction manual. All the terms and definitions meet the requirements of SRPS ISO/IEC 9000:2001, SRPS ISO/IEC 17025:2017 and follow International Vocabulary of Basic and General Terms in Metrology.

A Contribution to the Calibration of Direct Reading Thermometers in the Laboratory

Marina Bulat, Nemanja Gazivoda, Ivan Gutai, Bojan Vujičić, Đorđe Novaković, Platon Sovilj

Prilog etaloniranju čitača dozimetara

Marina Bulat, *Member, IEEE*, Nemanja Gazivoda, *Member, IEEE*, Ivan Gutai, *Member, IEEE*, Bojan Vujičić, *Member, IEEE*, Dragan Pejić, *Member, IEEE* i Marjan Urekar, *Member, IEEE*

***Apstrakt* — U ovom radu je dat postupak etaloniranja čitača dozimetara u Laboratoriji za metrologiju Fakulteta tehničkih nauka u Novom Sadu. Prikazane su najbolje mogućnosti merenja, opisana je priprema za merenje i navedena je merna oprema koja se koristi. Opisani su postupci merenja koje treba sprovesti, prikazana je obrada dobijenih rezultata i određena je forma procene merne nesigurnosti. Dat je primer etaloniranja čitača dozimetara. Svi termini i definicije su u skladu sa JUS ISO/IEC 9000:2001, SRPS ISO/IEC 17025:2017 i Međunarodnim rečnikom osnovnih i opštih termina u metrologiji.**

***Ključne reči* —** merna oprema; etaloniranje; metrologija; čitač dozimetara; merna nesigurnost.

I. Uvod

U Tabeli I su prikazane merne mogućnosti etaloniranja čitača dozimetara raspoloživom opremom Laboratorije za metrologiju Fakulteta tehničkih nauka u Novom Sadu (u daljem tekstu samo Laboratorija).

TABELA I
NAJBOLJE MOGUĆNOSTI MERENJA

Veličina	Predmet etaloniranja	Opseg	Merna nesigurnost
Jačina struje	Elektrometri	1 pA do 100 mA	≤ 0.05%

Posebni termini korišćeni u ovom uputstvu imaju sledeća značenja:

Čitač dozimetara – uređaj, deo dozimetra, koji meri struju odnosno naelektrisanje iz detektora i konvertuje u oblik pogodan za prikaz veličine kao što je apsorbovana doza, ekvivalentna doza, efektivna doza, kerma i odgovarajuće jačine doza ili kerme;

Marina Bulat – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail: marina.bulat@uns.ac.rs). Nemanja Gazivoda – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail: nemanjagazivoda@uns.ac.rs). Ivan Gutai – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail: gutai@uns.ac.rs). Bojan Vujičić– Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail: bojanvuj@uns.ac.rs). Dragan Pejić – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail: pejicdra@uns.ac.rs). Marjan Urekar – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail: urekarm@uns.ac.rs).

Nelinearnost - odstupanje odziva elektrometra od linearnosti. Procentualno odstupanje od linearnosti se izračunava iz relacije

$$100\cdot\left(\frac{m\cdot Q}{M\cdot q} - 1\right),$$

gde M i m predstavljaju očitavanja na elektrometru za prvu i drugu izabranu referentnu vrednost fizičke veličine, respektivno, a Q i q ulazne signale potrebne da bi proizvele referentnu vrednost za prvu i drugu izabranu vrednost fizičke veličine, respektivno. Za merenje naelektrisanja koristi se DOSE režim rada, a za merenje struje DOSE RATE;
Rezolucija prikaza - najmanja promena očitavanja na skali koja se može uočiti bez dodatne interpolacije. Za analogni prikaz odnosno displej je to najmanja frakcija intervala skale koji se može uočiti pri određenim uslovima. Kod digitalnih prikaza to je najmanji značajni inkrement očitavanja;

Odziv - odnos vrednosti prikazane na čitaču dozimetra i ulazne struje dobijene iz strujnog izvora;
Drift nule - kontinualna promena oko nulte vrednosti očitavanja na skali pri uslovima kada na elektrometru nema signala;

Šift nule - iznenadna promena u nultom očitavanju na skali elektrometra kada se režim merenja prebaci sa „nula" na „merenje" u uslovima kada nema mernog signala.

II. Opis aktivnosti

A. *Priprema za rad*

Priprema za etaloniranje podrazumeva vizuelni pregled objekta etaloniranja, konstataciju da je objekt pripremljen za etaloniranje i proveru referentnih uslova etaloniranja.

B. *Vizuelni pregled*

Vizuelnim pregledom se utvrđuje opšte stanje objekta etaloniranja i konstatuju se eventualna oštećenja. Utvrđuje se i postojanje dokumentacije o objektu etaloniranja, relevantne za etaloniranje. Čitač dozimetara (koji može biti odvojen od detektora ili njegov sastavni deo) koji se šalje na pregled mora imati stabilan displej i napunjenu bateriju, ukoliko postoji.

C. *Priprema za etaloniranje*

Potrebno je da objekt etaloniranja bude u Laboratoriji dvanaest sati pre početka etaloniranja, da bi se njegova temperatura izjednačila sa temperaturom okoline. Pre etaloniranja elektrometar treba da bude uključen najmanje dva sata.

D. *Provera referentnih uslova etaloniranja*

Referentni uslovi za merenje u Laboratoriji su:

- Temperatura okoline: (23 ± 2) °C;
- Relativna vlažnost vazduha: (45 ± 15)%.

Referentni uslovi u Laboratoriji se održavaju stalno, a provera podrazumeva ispunjenje navedenih zahteva neposredno pre merenja.

III. MERNA OPREMA

Oprema koja je neophodna za etaloniranje čitača dozimetara je navedena u Tabeli II.

TABELA II
SPISAK MERNE OPREME ZA ETALONIRANJE ČITAČA DOZIMETARA

Oznaka	Naziv	Tip
IJS	Izvor jednosmerne struje	Keithley 6220
	Etalonski otpornici	
DMM	Digitalni multimetar	HP 3458A

IV. POSTUPAK MERENJA

Etaloniranje čitača dozimetara vrši se metodom direktnog merenja. Očitavanja na čitaču su direktno srazmerna intenzitetu jačine struje koja se iz strujnog izvora primenjuje na njegove ulazne krajeve bilo direktno bilo preko etalonskog otpornika prema izrazu:

$$I=\frac{U}{R}.$$

Čitači se etaloniraju u najmanje pet tačaka koje pokrivaju najmanje 80% njihovog opsega.

Etaloniranje za merenje struje se obavlja korišćenjem strujnog izvora visoke impedanse ili merenjam pada napona na etalonskom otporniku kroz koji je propuštena poznata vrednost struje. Merni opseg struje zavisi od tipa detektora koji se koristi i može biti reda nA (1 nA do 100 nA) i reda pA (5 pA do 500 pA).

Uslovi zadavanja struje na strujnom izvoru pri etaloniranju dati su u Tabeli III.

TABELA III
USLOVI ZADAVANJA STRUJE STRUJNOG IZVORA PRI ETALONIRANJU

Opseg	Tačnost (23 °C)	Rezolucija
1 pA	0.2% + 1 pA	100 fA
10 pA	0.2% + 1 pA	100 fA
100 pA	0.2% + 1 pA	100 fA
500 pA	0.4% + 2 pA	100 fA
1 nA	0.4% + 2 pA	100 fA
2 nA	0.4% + 2 pA	100 fA
20 nA	0.3% + 10 pA	1 pA
200 nA	0.3% +100 pA	10 pA

Koeficijent etaloniranja ili kalibracioni faktor se izračunava za svaki opseg posebno kao srednja vrednost odnosa pokazivanja na čitaču i zadate struje. Koeficijent etaloniranja se izražava u jedinicama Sv/A, Gy/A, A/A ili u umnošcima pomenutih mernih jedinica. Koeficijent etaloniranja (kalibracioni faktor) se određuje za svaki merni opseg. Potrebno je proveriti da li displej pokazuje 0.000 ± 1 count. Ukoliko to nije slučaj, neophodno je podesiti nulu.

Kod merila koja imaju više opsega za svaki se određuje faktor kalibracije (koeficijent etaloniranja) koji se izražava kroz faktore promene opsega. Ti faktori su procenjeni u odnosu na opseg koji je izabran kao referentan i za koji je faktor promene opsega jednak jedinici. Kalibracioni faktori (koeficijenti etaloniranja) se međusobno ne smeju razlikovati za više od 0.5%.

U zavisnosti od detektora koji se koristi, čitači imaju različite linearne odzive. Provera linearnosti se sastoji od određivanja koeficijenta etaloniranja u odnosu na struju za različite opsege. Linearnost za svaku opseg se izražava kroz korekcioni faktor za nelinearnost, procenjen za svaki niz očitavanja u odnosu na očitavanje na referentnoj skali. Po definiciji, korekcija na nelinearnost je jednaka jedinici na referentnoj skali.

Curenje čitača se određuje posle primene spoljašnjeg izvora struje (koje simulira uslove ozračivanja detektora) i meri se pet minuta. Struja curenja ne sme biti veća od vrednosti koje je specificirao proizvođač, a ukoliko nije specificirano ne sme biti veća od 0.1 pA.

V. ETALONIRANJE ČITAČA DOZIMETARA

Vrednosti dobijene pri etaloniranju moraju biti određene sa relativnom standardnom mernom nesigurnošću i u granicama datim u Tabeli IV.

TABELA IV
GRANICE VARIJACIJE I RELATIVNA STANDARDNA MERNA NESIGURNOST RELEVANTNIH PARAMETARA

Karakteristika	Granice varijacije	Relativna standardna merna nesigurnost
Rezolucija	± 0.02%	± 0.01%
Curenje	± 0.1 pA	± 0.40%
Drift nule	± 1 count	± 0.10%
Šift nule	± 5 fA	± 0.60%
Nelinearnost	± 0.05%	± 0.01%
Promena opsega	± 0.5%	± 0.01%
Ulazna struja	± 0.2%	± 0.10%

Model izračunavanja greške etaloniranog čitača E_X:

$$E_X=(I_X + \delta I_X) - I_e,$$

gde je :

I_X – vrednost struje očitane čitačem koji se etalonira;

δI_X – greška očitavanja na čitaču koji se etalonira nastala zbog konačne rezolucije očitavanja;

I_e – vrednost struje zadate strujnim izvorom.

Koeficijenti osetljivosti su:

$$C_{\delta I_X}=\frac{\partial_{E_X}}{\partial \delta I_X}=1,\ C_{I_e}=\frac{\partial_{E_X}}{\partial I_e}=-1.$$

Merna nesigurnost određivanja greške etaloniranog čitača je:

$$u_{E_X}=\sqrt{\left(C_{\delta I_X}\cdot u_{\delta I_X}\right)^2+\left(C_{I_e}\cdot u_{I_e}\right)^2}$$

U Tabeli V je dat prikaz budžeta mernih nesigurnosti rezultata etaloniranja čitača.

TABELA V
PRIKAZ BUDŽETA MERNIH NESIGURNOSTI REZULTATA ETALONIRANJA ČITAČA

Veličina	Simbol	Merna nesigurnost	Tip merne nesigurnosti	Raspodela verovatnoće	Osetljivost c_i	Doprinos mernoj nesigurnosti
Vrednost struje očitane čitačem koji se etalonira	I_x					
Greška očitavanja, zbog konačne rezolucije očitavanja	δI_x	$\frac{I_{xr}}{2\sqrt{3}}$ Rezolucija (proizvodjač)	B	Pravougaona	1	$\frac{I_{xr}}{2\sqrt{3}}$
Vrednost struje zadate strujnim izvorom	I_e	Prema Tabeli III	B		-1	Prema Tabeli III
Greška čitača	E_x					

VI. PRIMER PROCENE MERNE NESIGURNOSTI

Merna nesigurnost tipa A, u_A, se procenjuje iz standardne devijacije srednje vrednosti ponovljenih merenja. Kombinovana merna nesigurnost je kvadratni koren zbira kvadrata merne nesigurnosti tipa A i B. Proširena merna nesigurnost je dobijena množenjem kombinovane merne nesigurnosti faktora obuhvata $k = 2$ što predstavlja interval u kome je nivo poverenja 95%.

A. Standardna merna nesigurnost tipa A

Merenje fizičke veličine X se ponovi N puta. Najbolja procenjena vrednost je aritmetička sredina svih rezultata merenja x_i:

$$\bar{x}=\frac{1}{N}\cdot\sum_{i=1}^{N}x_i$$

Za izražavanje merne nesigurnosti pojedinačnog rezultata x_i koristi se standardna devijacija:

$$\sigma_x=\sqrt{\frac{1}{N}\cdot\sum_{i=1}^{N}(x_i-\bar{x})^2}.$$

Standardna devijacija srednje vrednosti se koristi za izražavanje merne nesigurnosti najbolje procenjene vrednosti:

$$\sigma_{\bar{x}}=\frac{1}{\sqrt{N}}\cdot\sigma_x.$$

Standardna merna nesigurnost tipa A definisana je kao standardna devijacija srednje vrednosti:

$$u_A=\sigma_{x.}$$

B. Standardna merna nesigurnost tipa B

Procena merne nesigurnosti tipa B, u_B, procenjuje se iz faktora koji utiču na proces merenja, na osnovu korekcionih faktora i podataka iz literature (fizičke konstante i njihove merne nesigurnosti) kao na osnovu odabrane gustine raspodele verovatnoće.

Primer procene merne nesigurnosti tipa B

Pretpostavke:

- meri se struja I=100 pA;
- postoji donja i gornja granična vrednost za pravougaonu raspodelu $(I-\Delta,\ I+\Delta)$;
- postoji 100% verovatnoća da se tačna vrednost nađe u tom intervalu pošto je odabrana pravougaona raspodela.

Korak 1. Projektovati gustinu verovatnoće $p(x)$ za raspodelu vrednosti jačine struje:

$$p(x)=\begin{cases}C, & za\ I-\Delta\le x\le I+\Delta\\ 0\ , & u\ svim\ drugim\ slučajevima.\end{cases}$$

Vrednost integrala u granicama $(I-\Delta,\ I+\Delta)$ mora da bude jednak jedinici, stoga imamo da je:

$$\int p(x)dx=\begin{cases}C\cdot 2\Delta=1, & za\ I-\Delta\le x\le I+\Delta\\ 0, & u\ svim\ drugim\ slučajevima.\end{cases}$$

Za interval $I-\Delta\le x\le I+\Delta$ imamo da je

$$p(x)=\frac{1}{2\Delta}.$$

Korak 2. Izračunati najbolje procenjenu vrednost fizičke veličine (srednju vrednost) i varijansu v:

$$\bar{x}=\int_{-\infty}^{+\infty}x\cdot p(x)dx=\frac{1}{2\cdot\Delta}\cdot\int_{I-\Delta}^{I+\Delta}xdx=I$$

$$v=\int_{-\infty}^{+\infty}(x-\bar{x})^2\cdot p(x)dx=\frac{1}{2\cdot\Delta}\cdot\int_{I-\Delta}^{I+\Delta}(x-I)^2dx=\frac{1}{3}\cdot\Delta^2$$

$$u_B=\sqrt{v}=\frac{\Delta}{\sqrt{3}}.$$

U Tabeli VI je dat prikaz budžeta mernih nesigurnosti rezultata etaloniranja čitača (za zadatu struju 100 pA).

TABELA VI
PRIKAZ BUDŽETA MERNIH NESIGURNOSTI REZULTATA ETALONIRANJA ČITAČA(ZA ZADATU STRUJU 100 PA)

Veličina	Simbol	Merna nesigurnost	Tip merne nesigurnosti	Raspodela verovatnoće	Osetljivost c_i	Doprinos mernoj nesigurnosti
Vrednost struje očitane čitačem koji se etalonira	102 pA	0.612 pA				0.612 pA
Greška očitavanja, zbog konačne rezolucije očitavanja	100 fA	28.9 fA	B	Pravougaona	1	28.9 fA
Vrednost struje zadate strujnim izvorom	100 pA	0.05 pA	B		-1	0.05 pA
Greška čitača	E_x					0.614 pA

ZAHVALNICA

Ovaj rad je delom podržan od strane projekta ELEMEND (šifra projekta: 585681-EEP-1-2017-EL-EPPKA2-CBHE-JP).

LITERATURA

[1] Marković, M. Z., Danković, G., Živković, V.: Međunarodni rečnik osnovnih i opštih termina iz metrologije, publikacija Saveznog zavoda za mere i dragocene metale, 1996, Beograd
[2] "ISO/IEC 17025:2017 Laboratory Accreditation Program- PJLA"
[3] "SRPS ISO 9000:2001 - Sistemi menadžmenta kvalitetom - Osnove i rečnik". Retrieved 01.01.2007.

ABSTRACT

The paper presents an example of the procedure for the calibration of dosimeter readers in the Laboratory. It demonstrates the optimal ways of measuring. In addition, it offers a description of the preparation for measuring and lists the measuring equipment that has been used. It also features a description of the measuring procedures that need to be applied and it shows the processing of the measurement results. The form of the estimation of measurement uncertainty has been determined and an example of the calibration of dosimeter readers has been provided. All the terms and definitions meet the requirements of SRPS ISO/IEC 9000:2001, SRPS ISO/IEC 17025:2017 and follow International Vocabulary of Basic and General Terms in Metrology.

A Contribution to the Calibration of Dosimeter Readers

Marina Bulat, Nemanja Gazivoda, Ivan Gutai, Bojan Vujičić, Dragan Pejić, Marjan Urekar

Prilog etaloniranju pokaznih naprava termometara sa termoparovima

Stefan Mirković, Nemanja Gazivoda, Bojan Vujičić, Đorđe Novaković, Platon Sovilj

***Apstrakt*—Na osnovu metode poređenja u radu je opisan postupak etaloniranja pokaznih naprava termometara sa termoparovima unutar i van metrološke laboratorije. Obrađene su i prikazane najbolje mogućnosti merenja, postupak pripreme za etaloniranje, dokumentovana je merna oprema koja se koristi kao i postupci merenja i prikazani su specifični rezultati etaloniranja propraćeni odgovarajućim mernim nesigurnostima**

***Ključne reči*—termometar; termopar; kalibrator; multimetar; etaloniranje; merna nesigurnost; budžet merne nesigurnosti; metrologija.**

I. UVOD

Laboratorija za metrologiju Fakulteta tehničkih nauka u Novom Sadu vrši etaloniranje pokaznih naprava termometara sa termoparovima, što je opisano u ovom radu. Kao krajnji proizvod etaloniranja Laboratorija izdaje dokument Uverenje o etaloniranju koje se izdaje klijentu saglasno dokumentovanom sistemu kvaliteta. Etaloniranje pokaznih naprava termometara sa termoparovima kao i sam postupak, određen je radnim uputstvom izrađenim od strane tima saradnika Laboratorije za metrologiju, na osnovu EURAMET-ovih saveta i uputstava iz [2]. U radnom uputstvu, svi termini i definicije su u skladu sa SRPS ISO/IEC 9000:2015, SRPS ISO/IEC 17025:2017 i Međunarodnim rečnikom osnovnih i opštih termina u metrologiji.

Termometri sa termoparovima rade na principu merenja termoelektričnog napona na krajevima ugrađenog termopara, koji zavisi od temperaturnog gradijenta – Seebeck efekat, koji se može pozvati na Peltier i Thomson efekat. Etaloniranje pokazne naprave vrši se metodom poređenja. Na pokaznu napravu (DUT) se, umesto termopara, priključuje izvor etalonske elektromotorne sile (ems, EMS) za simuliranje termopara. Očitanja na pokaznoj napravi termometra porede se sa izračunatim vrednostima temperature definisanim standardom ili od strane proizvođača pokazne naprave, koje odgovaraju zadatim vrednostima elektromotornih sila.

Stefan Mirković – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg D. Obradovića 6, 21000 Novi Sad, Srbija (e-mail: mirkovicst@uns.ac.rs).

Nemanja Gazivoda – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg D. Obradovića 6, 21000 Novi Sad, Srbija (e-mail: nemanjagazivoda@uns.ac.rs).

Bojan Vujičić – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg D. Obradovića 6, 21000 Novi Sad, Srbija (e-mail: bojanvuj@uns.ac.rs).

Đorđe Novaković – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg D. Obradovića, 21000 Novi Sad, Srbija (e-mail: djordjenovakovic@uns.ac.rs).

Platon Sovilj – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg D. Obradovića, 21000 Novi Sad, Srbija (e-mail: platon@uns.ac.rs).

II. MERNE MOGUĆNOSTI I MERNA OPREMA

Merne mogućnosti u vidu mernih nesigurnosti etaloniranja pokaznih naprava termometara sa termoparovima, raspoloživom opremom Laboratorije prikazane su u tabeli ispod.

TABELA I
MERNE MOGUĆNOSTI

Veličina	Predmet etaloniranja	Opseg	Merna nesigurnost*
Temperatura	Pokazna naprava termometara sa termoparom (unutar laboratorije)	od -200 °C do +1800 °C	0,3 °C
	Pokazna naprava termometara sa termoparom (van laboratorije)	od -200 °C do +1300 °C	1,2 °C

*Merna nesigurnost je proširena merna nesigurnost, gde je standardna merna nesigurnost pomnožena faktorom obuhvata k=2, što za slučaj normalne raspodele greške odgovara verovatnoći od približno 95 %.

Merna oprema

Referentni etaloni, radni etaloni i pomoćna oprema Laboratorije koja se koristi za etaloniranje pokaznih naprava prikazani su u Tabeli II.

TABELA II
ETALONI I POMOĆNA OPREMA

Referentni etaloni	
Naziv	Tip
Digitalni multimetar	Hewlett Packard 3458A
Radni etaloni	
Naziv	Tip
Kalibrator	Time Electronics 5025
Digitalni multimetar	Fluke 8846A
Izvor jednosmerne elektromotorne sile	MB-1V
Djuarov (Dewar) sud	-
Pomoćna oprema	
Termometar/higrometar	445815

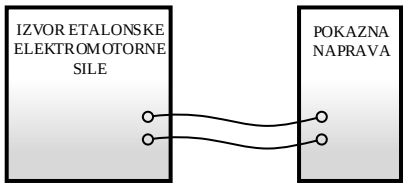
III. ETALONIRANJE UNUTAR LABORATORIJE

Vizuelni pregled objekta etaloniranja, konstatacija da je objekt pripremljen za etaloniranje i provera referentnih uslova etaloniranja unutar Laboratorije se podrazumevaju pre samog etaloniranja. Utvrđuje se opšte stanje objekta etaloniranja i konstatuju se eventualna oštećenja. Utvrđuje se, takođe, i postojanje dokumentacije o objektu etaloniranja, relevantne za etaloniranje. Objekt etaloniranja mora da bude u Laboratoriji najmanje 12 sati pre početka etaloniranja, da bi se njegova temperatura izjednačila sa temperaturom okoline. Uslovi za merenje u Laboratoriji su:

Temperatura okoline: (23 ± 2) °C;
Relativna vlažnost vazduha: (45 ± 15) %.

Ovi uslovi se održavaju stalno, i podrazumeva se ispunjenje navedenih referentnih uslova neposredno pre merenja.

Etaloniranje bez kompenzacije hladnog spoja



Sl. 1. Blok šema etaloniranja pokazne naprave termometra sa termoparom

Model izračunavanja greške E_X pokazne naprave je:

$$E_X = (T_X + \delta T_X) - T(V_S + \delta V_S) \tag{1}$$

gde je:

T_X Pokazivanje pokazne naprave;
δT_X Korekcija očitanja pokazne naprave, zbog konačne rezolucije očitavanja;
$T(V)$ Temperatura koja odgovara elektromotornoj sili V, definisana standardom ili od strane proizvođača pokazne naprave;
V_S Etalonska ems kojom se simulira termopar;
δV_S Korekcija zbog netačnosti merenja etalonske ems;

Vrednost etalonske ems V_S meri se digitalnim multimetrom sa granicama greške $\pm G_{V_S}$ određene specifikacijom proizvođača. Za gustinu raspodele verovatnoće greške merenja digitalnim multimetrom kao i gustinu raspodele greške očitavanja pokazne naprave se smatra da imaju ravnomernu raspodelu.

Standardna merna nesigurnost očitanja pokazne naprave zbog konačne rezolucije očitavanja određena je izrazom (2), gde je LSD (*Least Significiant Digit*) rezolucija displeja pokazne naprave:

$$u(\delta T_X) = \frac{LSD}{2\sqrt{3}} \tag{2}$$

Standardna merna nesigurnost koja potiče od greške multimetra određena je izrazom:

$$u(\delta V_S) = \frac{G_{V_S}}{\sqrt{3}} \tag{3}$$

Koeficijenti osetljivosti su:

$$c_1 = \frac{\partial E_X}{\partial \delta T_X} = 1 \tag{4}$$

$$c_2 = \frac{\partial E_X}{\partial \delta V_S} = -\frac{\partial T(V_S + \delta V_S)}{\partial \delta V_S} \tag{5}$$

Kombinovana standardna merna nesigurnost određivanja greške etaloniranja pokazne naprave:

$$u(E_X) = \sqrt{[c_1 \cdot u(\delta T_X)]^2 + [c_2 \cdot u(\delta V_S)]^2} \tag{6}$$

Proširena merna nesigurnost U definisana je sa:

$$U = k \cdot u(E_X) \tag{7}$$

Etalonirana je pokazna naprava, za opseg temperatura od 0 °C do +1200 °C, sa rezolucijom od LSD=0,1 °C, predviđena da se na nju priključi termopar tipa K. Vrednost etalonske elektromotorne sile kojom se simulira termopar izmerena je digitalnim multimetrom Hewlett Packard 3458A.

TABELA III
EVIDENCIJA REZULTATA ETALONIRANJA

Očitanje pokazne naprave	Etalonska ems	Etalonska temperatura*	Greška pokazne naprave	Nesigurnost**	Faktor obuhvata
T_X (°C)	V_S (mV)	$T(V_S)$ (°C)	E_X (°C)	U (°C)	k
0,0	0,000	0,00	0,0	0,058	2
199,0	8,100	199,02	0,0	0,059	2
400,0	16,400	400,07	-0,1	0,060	2
600,0	24,900	599,88	0,1	0,060	2
800,0	33,300	800,62	-0,6	0,062	2
1000,0	41,300	1000,60	-0,6	0,063	2
1200,0	48,800	1198,98	1,0	0,066	2

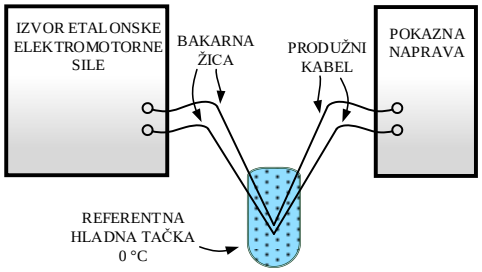
*Temperatura koja odgovara etalonskoj ems V_S, prema standardu IEC 584

**Proširena merna nesigurnost, gde je standardna kombinovana merna nesigurnost pomnožena faktorom obuhvata k=2, što za slučaj normalne raspodele greške odgovara verovatnoći od približno 95 %. Tabela III ujedno prikazuje i prilog uz Uverenje o etaloniranju koje izdaje Laboratorija.

TABELA IV
BUDŽET MERNE NESIGURNOSTI ETALONIRANJA UNUTAR LABORATORIJE ZA TEMPERATURU OD +498,96 °C

Naziv veličine	Simbol	Ocena	Parcijalna nesigurnost	Tip nesigurnosti	Raspodela	Koeficijent osetljivosti	Doprinos nesigurnosti
Očitanje etaloniranog termometra	T_X	499,5 °C					
Korekcije očitavanja zbog konačne rezolucije očitavanja	δT_X	0 °C	0,029 °C	B	ravnomerna	1	0,029 °C
Etalonska ems	V_S	20,600 mV					
Temperatura koja odgovara etalonskoj ems V_S	T_S	498,96 °C					
Korekcija etalonske ems V_S zbog netačnosti merenja njene vrednosti	δV_S	0 mV	$0{,}588 \cdot 10^{-6}$ V	B	ravnomerna	$-13{,}6$ °C·mV^{-1}	0,0080 °C
Greška pokazne naprave	E_X	0,5 °C	Kombinovana merna nesigurnost				0,030 °C
			Proširena merna nesigurnost (k=2)				0,060 °C

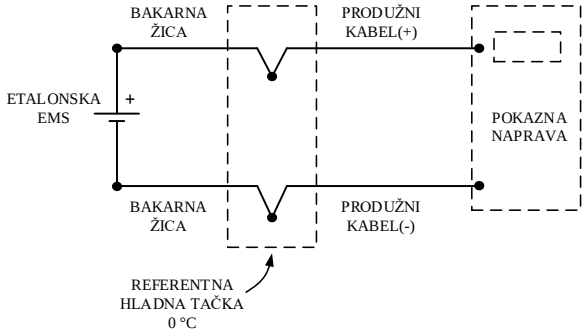
Etaloniranje sa kompenzacijom hladnog spoja



Sl. 2. Blok šema etaloniranja pokazne naprave termometra sa termoparom sa kompenzacijom hladnog spoja

Blok šema sa Sl. 2 pokazuje način povezivanja kod etaloniranja sa kompenzacijom hladnog spoja, prema [2]. Izvor etalonske ems generiše napon kojim se simulira termopar, i taj napon se preko bakarnih provodnika vodi do referentne hladne tačke (referentne temperature). Na referentnoj temperaturi bakarni provodnici su kratkospojeni sa produžnim kablovima koji je povezuju sa DUT.

Hemijski sastav produžnih kablova zavisi od tipa termopara koji se simulira, odnosno od DUT. Električna šema za etaloniranje sa kompenzacijom hladnog spoja nalazi se na Sl. 3. Ovo kolo je jedno od osnovnih kola, kako pri etaloniranju termometara, tako i pri merenju temperature termoparovima. Detaljnije objašnjenje principa funkcionisanja ovog kola može se naći u [3], [4] i [5].



Sl. 3. Električna šema etaloniranja pokazne naprave termometra sa termoparom sa kompenzacijom hladnog spoja

Model izračunavanja greške E_X pokazne naprave je:

$$E_X = (T_X + \delta T_X) - T(V_S + \delta V_S + \delta V_{S-par} + \delta V_{S-ext} + \delta V_{S-extDrift}) + \frac{S_0}{S_X}\delta T_0 \tag{8}$$

gde je:

T_X Pokazivanje pokazne naprave;
δT_X Korekcija očitanja pokazne naprave, zbog konačne rezolucije očitavanja;
$T(V)$ Temperatura koja odgovara ems V , definisana standardom ili od strane proizvođača pokazne naprave;
V_S Etalonska ems kojom se simulira termopar;
δV_S Korekcija zbog netačnosti merenja V_S;
$\delta V_{S\text{-}par}$ Korekcija ems, zbog prisustva parazitnih efekata u mernom kolu (termoelektrični, CM, strano magnetno polje);
$\delta V_{S\text{-}ext}$ Korekcija ems, zbog netačnosti karakteristika produžnih provodnika;
$\delta V_{S\text{-}extDrift}$ Korekcija ems, zbog drifta karakteristika produžnih produžnih;
δT_0 Korekcija temperature etaloniranja, zbog netačnosti temperature hladnog kraja;
S_0 , S_X Zebekovi koeficijenti, za temperature od 0 °C i temperaturu etaloniranja T_X ;

Standardne merne nesigurnosti usled:

1) Ponavljanja merenja (tip A)

$$u(T_X) = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(T_i - T_X)^2} \tag{9}$$

gde je:

- n Ukupan broj očitanja sa DUT;
- T_i Vrednost rezultata i-tog merenja očitana na DUT
- T_X Aritmetička sredina pokazivanja DUT

2) Konačne rezolucije očitanja sa DUT (tip B)

$$u(\delta T_X) = \frac{LSD}{2\sqrt{3}} \tag{10}$$

3) Netačnosti merenje etalonske ems (tip B)

$$u(\delta V_S) = \frac{G_{Vs}}{\sqrt{3}} \tag{11}$$

Sve ostale merne nesigurnosti, su merne nesigurnosti tipa B, i određuju se na osnovu specifikacija proizvođača ili rezultata zadnjeg etaloniranja. Koeficijenti osetljivoti su:

$$c_1 = \frac{\partial E_X}{\partial T_X} = 1 \tag{12}$$

$$c_2 = \frac{\partial E_X}{\partial \delta T_X} = 1 \tag{13}$$

$$c_3 = -\frac{\partial T(V_S + \delta V_S + \delta V_{S-par} + \delta V_{S-ext} + \delta V_{S-extDrift})}{\partial \delta V_S} \tag{14}$$

$$c_4 = -\frac{\partial T(V_S + \delta V_S + \delta V_{S-par} + \delta V_{S-ext} + \delta V_{S-extDrift})}{\partial \delta V_{S-par}} \tag{15}$$

$$c_5 = -\frac{\partial T(V_S + \delta V_S + \delta V_{S-par} + \delta V_{S-ext} + \delta V_{S-extDrift})}{\partial \delta V_{S-ext}} \tag{16}$$

$$c_6 = -\frac{\partial T(V_S + \delta V_S + \delta V_{S-par} + \delta V_{S-ext} + \delta V_{S-extDrift})}{\partial \delta V_{S-extDrift}} \tag{17}$$

$$c_7 = \frac{\partial E_X}{\partial T_0} = \frac{S_0}{S_X} \tag{18}$$

Kombinovana merna nesigurnost određivanja greške etaloniranja DUT:

$$u(E_X) = \sqrt{[c_1 u(T_X)]^2 + [c_2 u(\delta T_X)]^2 + [c_3 u(\delta V_S)]^2 + [c_4 u(\delta V_{S-par})]^2 + [c_5 u(\delta V_{S-ext})]^2 + [c_6 u(\delta V_{S-extDrift})]^2 + [c_7 u(\delta T_0)]^2} \tag{19}$$

TABELA V
BUDŽET MERNE NESIGURNOSTI ETALONIRANJA UNUTAR LABORATORIJE ZA TEMPERATURU OD +998,96 °C

Naziv veličine	Simbol	Ocena	Parcijalna nesigurnost	Tip nesigurnosti	Raspodela	Koeficijent osetljivosti	Doprinos nesigurnosti
Aritmetička sredina pokazivanja DUT	T_X	1001 °C	$224 \cdot 10^{-3}$ °C	A	normalna	1	0,224 °C
Korekcije očitavanja zbog konačne rezolucije očitavanja	δT_X	0 °C	$289 \cdot 10^{-3}$ °C	B	ravnomerna	1	0,289 °C
Etalonska ems	V_S	41,2 mV					
Temperatura koja odgovara etalonskoj ems V_S	T_S	998,06 °C					
Korekcija etalonske ems V_S zbog netačnosti merenja njene vrednosti	δV_S	0 mV	4,2 µV	B	ravnomerna	-25,6 °C·mV^{-1}	0,107 °C
Korekcija ems, zbog prisustva parazitnih efekata	$\delta V_{S\text{-par}}$	0 mV	2,9 µV	B	ravnomerna	-25,6 °C·mV^{-1}	0,074 °C
Korekcija ems, zbog netačnosti karakteristika produžnih provodnika	$\delta V_{S\text{-ext}}$	2,5 mV	5,0 µV	B	normalna	-25,6 °C·mV^{-1}	0,128 °C
Korekcija ems, zbog drifta karakteristika produžnih produžnih	$\delta V_{S\text{-extDrift}}$	0 mV	2,3 µV	B	ravnomerna	-25,6 °C·mV^{-1}	0,059 °C
Korekcija temeprature etaloniranja, zbog netačnosti temperature hladnog kraja	δT_0	0 mV	11,5 µV	B	ravnomerna	1,01	0,012 °C
Greška pokazne naprave	E_X	3,0 °C				Kombinovana merna nesigurnost	0,41 °C
						Proširena merna nesigurnost (k=2)	0,83 °C

Etalonirana je pokazna naprava, za opseg temperatura do +1200 °C, sa rezolucijom od LSD=1 °C , predviđena da se na nju priključi termopar tipa K. Etaloniranje je sprovedeno prema šemi na Slici 2. Tačka etaloniranja je +998,06 °C. Etalonski sistem se sastoji od kalibratora elektromotorne sile, (Time Electronics 5025), Djuartovog suda za realizaciju temperature od 0 °C i produžnih kablova za termopar tipa K.

IV. ETALONIRANJE VAN LABORATORIJE

Provera metroloških karakteristika etalonskog sistema pre odlaska na mesto etaloniranja, vizuelni pregled objekta etaloniranja, konstatacija da je objekt pripremljen za etaloniranje i provera referentnih uslova etaloniranja se podrazumevaju pre samog etaloniranja. Pre i nakon prenosa etalonskog sistema vrši se provera osnovnih metroloških karakteristika etalona, kako bi se smanjili rizici tokom prenošenja opreme do i od mesta etaloniranja. Oprema za etaloniranje do i od mesta etaloniranja bi po pravilu trebalo da se prenosi putničkim automobilom.

Etalonska oprema i objekt etaloniranja bi trebalo da budu u prostoru namenjenom za etaloniranje najmanje 1 sat pre početka etaloniranja, kako bi se njihova temperatura približno izjednačila sa temperaturom okoline. Prostor u kome treba da se obavi etaloniranje mora da bude suv, čist, bez prašine i drugih agensa koji mogu da deluju na mernu opremu, i da se u njemu ne odvijaju druge aktivnosti koje mogu loše da utiču na merenja. Utvrđuje se opšte stanje objekta etaloniranja i konstatuju se eventualna oštećenja, kao i postojanje dokumentacije o objektu etaloniranja, relevantne za etaloniranje.

Uslovi za merenje van Laboratorije su:

Temperatura okoline: 10 °C do 30 °C;
Relativna vlažnost vazduha: 40 % do 80 %.

Neposredno pre merenja podrazumeva se ispunjenje navedenih referentnih uslova.

Blok šema etaloniranja pokazne naprave van Laboratorije je ista kao šema na Slici 1.

Model izračunavanja greške E_X pokazne naprave je:

$$E_X = (T_X + \delta T_X) - T(V_S + \delta V_S + \delta V_{S-temp}) \quad (20)$$

gde je:

T_X Pokazivanje pokazne naprave;
δT_X Korekcija očitanja pokazne naprave, zbog konačne rezolucije očitavanja;
$T(V)$ Temperatura koja odgovara ems V , definisana standardom ili od strane proizvođača pokazne naprave;
V_S Etalonska ems kojom se simulira termopar;
δV_S Korekcija zbog netačnosti merenja etalonske ems;
$\delta V_{S\text{-temp}}$ Korekcija zbog netačnosti merenja etalonske ems ako temperatura okoline izlazi van referentnog opsega temperatura etalonskog merila ems;

Vrednost etalonske ems V_S meri se multimetrom sa granicama greške G_{Vs} određene specifikacijom proizvođača. U slučaju da ambijentalna temperatura izađe van referentnog temperaturnog opsega multimetra, granice greške merenja napona se povećavaju zbog uticaja dodatne greške $G_{Vs\text{-temp}}$, koja je posledica ambijentalne temperature. Za gustinu raspodele verovatnoće greške merenja digitalnim multimetrom kao i gustinu raspodele greške očitavanja pokazne naprave se smatra da imaju ravnomernu raspodelu.

Standardna merna nesigurnost očitanja pokazne naprave zbog konačne rezolucije očitavanja određena je formulom:

$$u(\delta T_X) = \frac{LSD}{2\sqrt{3}} \quad (21)$$

Standardna merna nesigurnost koja potiče od digitalnog multimetra određena je izrazom:

$$u(\delta V_S) = \frac{G_{Vs}}{\sqrt{3}} \quad (22)$$

Standardna merna nesigurnost koja potiče od dodatne greške multimetra usled ambijentalne temperature je:

$$u(\delta V_{S-temp}) = \frac{G_{Vs-temp}}{\sqrt{3}} \quad (23)$$

Koeficijenti osetljivosti su:

$$c_1 = \frac{\partial E_X}{\partial \delta T_X} = 1 \quad (24)$$

$$c_2 = \frac{\partial E_X}{\partial \delta V_S} = -\frac{\partial T(V_S + \delta V_S + \delta R V_{S-temp})}{\delta V_S} \quad (25)$$

$$c_3 = \frac{\partial E_X}{\partial \delta V_{S-temp}} = -\frac{\partial T(V_S + \delta V_S + \delta V_{S-temp})}{\delta V_{S-temp}} \quad (26)$$

Merna nesigurnost određivanja greške etaloniranja pokazne naprave predviđene za rad sa termoparovima:

$$u(E_X) = \sqrt{[c_1 \cdot u(\delta T_X)]^2 + [c_2 \cdot u(\delta V_S)]^2 + [c_3 \cdot u(\delta V_{S-temp})]^2} \quad (27)$$

Etalonirana je pokazna naprava, za opseg temperatura od 0 °C do +1200 °C, sa rezolucijom od LSD=0,1 °C, predviđena da se na nju priključi termopar tipa K. Vrednost etalonske elektromotorne sile V_S kojim se simulira termopar izmerena je digitalnim multimetrom Fluke 8846A. Temperatura etaloniranja je +499,5 °C.

TABELA VI
BUDŽET MERNE NESIGURNOSTI ETALONIRANJA VAN LABORATORIJE ZA TEMPERATURU OD +498,96 °C

Naziv veličine	Simbol	Ocena	Parcijalna nesigurnost	Tip nesigurnosti	Raspodela	Koeficijent osetljivosti	Doprinos nesigurnosti
Očitanje etaloniranog termometra	T_X	499,5 °C					
Korekcije očitavanja zbog konačne rezolucije očitavanja	δT_X	0 °C	0,029 °C	B	ravnomerna	1	0,029 °C
Etalonska ems	V_S	20,600 mV					
Temperatura koja odgovara etalonskoj ems V_S	T_S	498,96 °C					
Korekcija etalonske ems V_S zbog netačnosti merenja njene vrednosti	δV_S	0 mV	2,5 µV	B	ravnomerna	-13,6 °C·mV^{-1}	0,033 °C
Korekcija etalonske ems V_S zbog netačnosti merenja njene vrednosti, ako temperatura okoline izlazi van referentnog opsega temperatura etalonskog merila	$\delta V_{S\text{-temp}}$	0 mV	1,0 µV	B	ravnomerna	-13,6 °C·mV^{-1}	0,014 °C
Greška pokazne naprave	E_X	0,5 °C	Kombinovana merna nesigurnost				0,047 °C
			Proširena merna nesigurnost (k=2)				0,093 °C

V. ZAKLJUČAK

Etaloniranje pokaznih naprava termometara sa termoparovima saglasno je dokumentovanom sistemu kvaliteta Laboratorije za metrologiju. Opisana metoda etaloniranja je jedna od metoda koje se koriste u Laboratoriji za etaloniranje pokaznih naprava termometara i koja se poziva naa savete i uputstva iz uputstva [2]. Pored raznih softverskih i hardverskih kompenzacija hladnog kraja koje mogu da se implementiraju, merne nesigurnosti i ovako opisanih metoda ispunjavaju većinu zahteva klijenata Laboratorije, kada su u pitanju pokazne naprave termometara, kako u slučajevima etaloniranja sa tako i u slučajevima bez kompenzacije hladnog kraja.

ZAHVALNICA

Ovaj rad je delom podržan od strane projekta ELEMEND (šifra projekta: 585681-EEP-1-2017-EL-EPPKA2-CBHE-JP).

LITERATURA

[1] "ETALONIRANJE POKAZNIH NAPRAVA TERMOMETARA SA OTPORNIČKIM SONDAMA I/ILI TERMOPAROVIMA", Radno uputstvo, Q3.JIM.19, Laboratorija za metrologiju, Fakultet tehničkih nauka, Novi Sad, 2015.

[2] "Guidelines on the Calibration of Temperature Indicators and Simulators by Electrical Simulation and Measurement", EURAMET cg-11, Version 2.0, Calibration Guide, ISBN 978-3-942992-08-4.

[3] T. W. Kerlin, M. Johnson, *Practical thermocouple thermometry*, International Society of Automation (ISA), 2012, ISBN 978-1-937560-27-0.

[4] "Manual on the use of thermocouples in temperature measurement", ASTM Committee E20 on Temperature Measurement, ISBN 0-8031-1466-4

[5] D. D. Pollock, *Thermocouples : theory and properties*, State University of New York at Buffalo, ISBN 0-8493-4243-0.

[6] "Evaluation of measurement data – Guide to the expression of encertainty in measurement", JCGM 100:2008

[7] "International vocabulary of metrology – Basic and general concepts and associated terms (VIM)", JCGM 200:2012

[8] A. Dunjić, J. Pantelić-Babić, M. Pavićević, "Postupak etaloniranja ampermetara i kalibratora jednosmerne električne struje u dokumentovanom sistemu kvaliteta ZMDM", *Zbornik radova 50. Konferencije za ETRAN*, vol. *III*, Beograd, 2006.

[9] "HP 3458A, Operating, Programming and Configuration Manual, Hewlett Packard", Edition 1, USA, May 1988

[10] "8845A/8846A Digital Multimeter Users Manual", Fluke Corporation, USA, July 2006

ABSTRACT

Based on the comparison method, in this paper, the procedure of calibration of temperature indicators with thermocouples is described for conditions inside and outside the metrology laboratory. The best measurement possibilities and process of preparation for calibration is processed, as well as the used equipment. Specific calibration results were presented, accompanied by measurement uncertainties.

The Contribution to The Calibration of Temperature Indicators With Thermocouples

Stefan Mirković, Nemanja Gazivoda, Bojan Vujičić, Đorđe Novaković, Platon Sovilj

Web-bazirani merni sistemi – primer edukativnog front-enda

Ivan Gutai, *Member, IEEE,* Đorđe Novaković, *Member, IEEE,* Platon Sovilj, *Member, IEEE,* Dragan Pejić, *Member, IEEE,* Marina Bulat, *Member, IEEE,* Nemanja Gazivoda, *Member, IEEE*

***Apstrakt* — U ovom radu je prikazan primer front-end modula koji omogućava prikaz rezultata merenja sa sedam različitih senzora uz prilagodljiv dizajn koji omogućava preglednost i sa računara, ali i sa prenosnih uređaja. Svaki senzor ima zaseban panel, na visokim rezolucijama se prikazuju dva u istom redu, dok se na prenosnim uređajima slažu jedan ispod drugog. Svaki panel prikazuje listu svih rezultata merenja i vreme kada su merenja izvršena, a istovremeno se i radi lakšeg nadzora prikazuje i grafik sa svim vrednostima. Svaki od osnovnih primera koji sadrže programski kod u JavaScript-u, koji se daju kao dodatak uz ovaj rad, pored edukativnog karaktera su predviđeni i da se savladaju za manje od 15 minuta, što čini celu ovu zamisao kompatibilnom sa osnovnim principima tzv. Microlearning-a.**

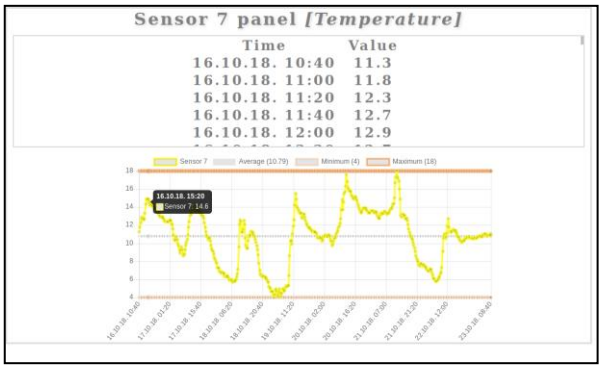
***Ključne reči*—Web-bazirani merni sistemi; Microlearning; JavaScript; HTML 5; CSS 3; Chart.js; Metrologija**

I. Uvod

Metrološki lanac, razvojem informaciono-komunikacionih tehnologija i evolucijom merne instrumentacije od analognih mernih instrumenta do složenih merno-informacionih sistema, vremenom je usložnjen. Bitni segmenti tog lanca su različiti moduli merno-informacionih sistema, uključujući i front-end module. Ovaj rad prikazuje jedan primer front-end modula web-baziranog mernog sistema, projektovanog za edukativne i razvojne svrhe.

Svedoci smo naglog uspona softverske industrije u poslednjih nekoliko decenija, čime softver postaje gradivni element većine savremenih uređaja. Neizostavni element svakog savremenog metrološkog sistema zahteva razvoj softvera, kao i korišćenje odgovarajućih alata. U metrologiji postaje sve zastupljeniji koncept udaljenih merenja, u kojem korisnik može dobiti uvid u rad određenog sistema korišćenjem mobilnog telefona, tableta, računara i sl., a takođe može i upravljati odgovarajućim procesima. Takođe, IoT (Internet of Things) koncept iziskuje pored hardverske platforme i stabilnu web aplikaciju. Kao rezultat pomenutih trendova, neophodno je napraviti edukativne platforme kako bi se studenti, kao i inženjeri, obučili za rad sa softverom, što ujedno predstavlja motivaciju za ovaj rad. Platforma koja je opisana u ovom radu primenjena je u praksi kroz realizaciju u predmetu „Web bazirani merno akvizicioni sistemi“.

Primeri koji se koriste u nastavi, namenjeni za savladavanje osnovnih tehnika front end razvoja i kompletan izvorni kod dati su na linku [1]. Primer jednog panela aplikacije je prikazan na slici 1.



Sl 1. Prikaz dinamički osvežavanih rezultata merenja u web aplikaciji

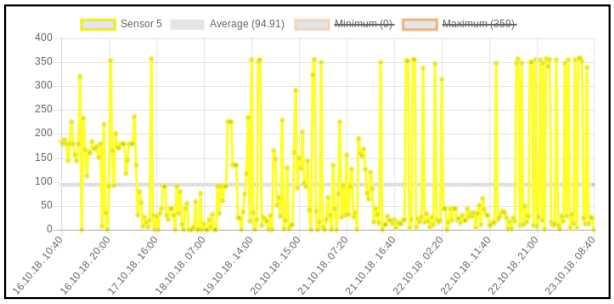
Na panelu sa slici 1. dat je primer prikazivanja rezultata u razvijenoj web aplikaciji. Panel se sastoji iz dva dela:

- grafičkog: gde su grafičkim putem prikazane vrednosti izmerene temperature u vremenu. Klikom kursora na grafičkom panelu dobijaju se podaci o odabranom odbirku.
- tekstualnog: gde su prikazani odbirci temperature u vremenu.

Razvojno okruženje koje se koristi je Microsoft Visual Studio Code, u daljem tekstu VS Code. VS Code radi na Windows-u, Mac OS-u i na Linux-u. Svi alati koji su korišćeni, besplatni su i jednostavni za konfigurisanje. HTML je korišćen za kreiranje web aplikacije, CSS za definisanje izgleda, a JavaScript za programiranje. Osim instalacije VS Code-a potrebno je instalirati Node.js®, okruženje u kom se JavaScript izvršava.

Na panelu na slici 2 je prikazana mogućnost isključivanja prikaza određene grupe podataka, npr. minimalnih i maksimalnih vrednosti, nakon čega se grafik automatski skalira po y osi. Na ovaj način dobija se na preglednosti rezultata kao i na njihovom međusobnom poređenju.

Ivan Gutai – Fakultet tehničkih nauka, Novi Sad, Srbija (e-mail: gutai@uns.ac.rs).
Đorđe Novaković – Fakultet tehničkih nauka, Novi Sad, Srbija (e-mail: djordjenovakovic@uns.ac.rs).
Platon Sovilj – Fakultet tehničkih nauka, Novi Sad, Srbija (e-mail: platon@uns.ac.rs).
Dragan Pejić – Fakultet tehničkih nauka, Novi Sad, Srbija (e-mail: pejicdra@uns.ac.rs).
Marina Bulat – Fakultet tehničkih nauka, Novi Sad, Srbija (e-mail: marina.bulat@uns.ac.rs).
Nemanja Gazivoda – Fakultet tehničkih nauka, Novi Sad, Srbija (e-mail: nemanjagazivoda@uns.ac.rs).



Sl 2. Prikaz funkcionalnosti koja omogućava selektivni grafički prikaz podataka

II. Prilagodljivi dizajn i Know-how

Savremeni internet pretraživači kao što su: Google Chrome, Mozilla Firefox, Apple Safari, Яндекс.Браузер, Microsoft Edge i razni drugi u većini slučajeva prikazuju napisanu aplikaciju baš onako kako je projektant to i zamislio. Kada se ista aplikacija otvori u staroj verziji Microsoft Internet Explorer-a, koji je bio aktuelan pre desetak godina - sledi iznenađenje. Iz navedenog razloga je pre samog kreiranja web aplikacije bitno odrediti ciljnu grupu, bilo da su to korisnici najnovijih Android ili iPhone uređaja iz 2019. ili je potrebno da web aplikacija radi besprekorno u Internet Explorer-u verzije 9 iz 2011. godine. Ciljna grupa ove web aplikacije su studenti koji aplikaciju koriste na računarima, koji imaju operativne sisteme novije generacije, a prilagodljivi dizajn je uveden da bi mogli da postave ovu aplikaciju na internet i da je slobodno koriste i na smartphone uređajima. Testiranje napisanog koda je bilo obavezno na različitim uređajima i u različitim internet pretraživačima i uvek su to radile bar dve osobe.

Web-bazirani merni sistem je razvijen kao koncept, koji zahvaljujući pomoćnim fajlovima koji se mogu preuzeti sa linka [1], omogućava studentu da relativno brzo prođe uvodnu front-end obuku, a zatim napravi web aplikaciju koja je funkcionalna i dobro izgleda na prosečnom laptop računaru sa ekranom rezolucije 1920 x 1080 px, kao i na mobilnom uređaju starije generacije koji ima rezoluciju npr. 480 x 800 px.

Telefoni iz vrhunskih serija (eng. flagship), iako imaju izuzetno velike rezolucije, umeju da stvaraju dodatne projektantske izazove. Npr. LG G3 iako ima rezoluciju od 1440 x 2560 px, nije sposoban da u portret režimu prikaže galeriju koja ima zadatu širinu od 800 px. Odgovor na to pitanje se dobije istraživanjem i informacijom da navedeni telefon ima rezoluciju od 480 x 853 dp i gustinu piksela 3. Upotreba savremenih jedinica mere kao što su dp (device independent pixels) i mnogih drugih su samo neke od stvari koje treba imati u vidu. Material.io [2] je online resurs koji između ostalog daje i smernice o dimenzijama ekrana mobilnih uređaja.

III. Korišćene tehnologije i alati za izradu materijala za vežbe

Za izradu edukativnog front-enda su korišćeni HTML 5, CSS 3 i JavaScript. Chart.js [3] je jedna u nizu besplatnih biblioteka koja je izabrana sa ciljem da omogući efektivan način za prilagodljivo iscrtavanje grafika u sklopu svakog panela. Medija upiti su upotrebljeni na takav način da se na ekranima širine do 1152 px prikazuje samo jedan panel, a preko toga, po dva panela u jednom redu. Od novijih alata, studentima je data i mogućnost da koriste i Brackets, koji omogućava brzo i lako kreiranje i testiranje web stranica i ima vrlo korisnu opciju koja se zove: "Live preview". Zbog edukativnog karaktera, veoma je bitno da u svakom trenutku HTML kod pude pravilno napisan. Da ne bi ručno tražili da li je napravljen propust u kodu, upotrebljen je efikasan alat za proveru [4], koji je kreiran od strane W3C (World Wide Web konzorcijuma). Ostavljena je i mogućnost da se prilikom dizajna izabere skladna paleta boja, uz pomoć Adobe Color Wheel online alat-a [5].

IV. Praktični razvojni primeri

Kompletan materijal pored edukativnog karaktera služi i za podsetnik za tehnički intervju. Niz primera je dat i mogu se prikazivati ili u web stranici ili u konzoli, a pošto su logički podeljeni i većina sadrži desetak linija koda, mogu se uklopiti u Microlearning. Između ostalog, za pisanje ovog rada su poslužile, dve knjige [6] i [7] i Pluralsight [8]. Funkcije i elementi JavaScript-a koji omogućavaju da se do detalja razume jedan ovakav web-bazirani merni sistem su: if else, switch case, push(), unshift(), pop(), shift(), propertiji objekta, do while petlja, while petlja, opseg funkcije, ključna reč this, prototipovi, value tipovi, reference tipovi, callback funkcije, IIFE (Immediately Invoked Function Expression), Function.apply(), Function.call(), Function.bind(), ternarni operator, setTimeout(), clearTimeout(), setInterval(), clearInterval(), Object.freeze(), Object.assign(), arrow funkcija, try catch finally, Array.reduce(), spread sintaksa (...), forEach(), map(), 4 načina za kreiranje objekta ({}, Object.create, new ključna reč, class), Closure, Promises, Iterators, Generators i Async/Await. Microlearning tehnika je korišćena iz praktičnih razloga, zato što se razumevanjem pojedinačnih funkcija, stiče i mogućnost sagledavanja šire slike i adaptacija front end rešenja.

Jedna od funkcija koja je sastavni deo web-baziranog mernog sistema je i CalculateMaxValue, koja je prikazana na slici 3.

```
function CalculateMaxValue(dataObject) {
    var maxValue;
    var dataArray = new Array;
    for (var o in dataObject) {
        dataArray.push(dataObject[o]);
    }
    maxValue = Math.max(...dataArray);
    return maxValue
}
```

Sl 3. CalculateMaxValue funkcija

Na primeru CalculateMaxValue funkcije se može videti da ona kao ulazne parametre dobija niz podataka, for petljom prolazi kroz svaku stavku, push() funkcijom ih dodaje u definisani dataArray niz, pomoću spread sintakse se niz prilagođava Math.max() funkciji, nakon čega funkcija CalculateMaxValue vraća brojnu vrednost koja predstavlja maksimalan broj od zadatog skupa brojeva.

Rezultati merenja koji se obrađuju i prikazuju, se nalaze na disku, u JSON formatu. JavaScript kod je prilagođen strukturi JSON fajla koji je prikazan na slici 4.

```
[
    {
        "Humidity": { "2018-10-18 02:20:00": 54.0,
            "2018-10-18 02:40:00": 53.0,
            "2018-10-18 03:00:00": 56.0},
        "Rainfall": {}, "WindGust": {},
        "Temperature": {}, "Pressure": {},
        "WindSpeed": {}, "WindDirection": {}
    }
]
```

Sl 4. Struktura rezultata merenja u JSON fajlu

V. PRIMER SPECIFIKACIJE ZAHTEVA

Potrebno je na jednoj stranici prikazati rezultate merenja sa tri različita senzora. Fajl sa svim rezultatima je "extractedData.json", a sadrži sledeće parametre: "Humidity, Rainfall, WindGust, Temperature, Pressure, WindSpeed i WindDirection." Obavezno je prikazati rezultate za WindGust i još sa dva senzora po izboru. Rezultate je potrebno prikazati u tabeli i na grafiku. Tabela treba da sadrži dve kolone, TIME i VALUE, a rezultati da budu prikazani u formatu, npr. "16.10.18. 11:40" i "960.40", respektivno. Prikazuju se svi rezultati, bez ikakvog filtriranja. Grafik treba da sadrži sve podatke, sortirane po vremenu. Opciono je da se prikažu minimalna, maksimalna ili srednja vrednost, na istom grafiku. Grafik treba da ima mogućnosti isključivanja prikaza seta parametara i automatsko prilagođavanje skale (Chart.js). Kao rezultat treba da se dobiju tri bloka koja u sebi sadrže i grafik i tabelu. Prilagoditi dizajn svih blokova, da budu u skladu sa prvim blokom. Izbor boja treba da se razlikuje od zadatog primera i potrebno je da bude čitko. Na rezolucijama do 1152 px je potrebno prikazivati jedan blok ispod drugog, dok je na većim rezolucijama potrebno prikazivati dva bloka, jedan pored drugog. HTML treba da bude personalizovan i da sadrži imena (polje "author") i brojeve indeksa (polje "keywords"). JavaScript fajl treba da sadrži isključivo kod koji je potreban za prikaz odabranih podataka, a nazivi promenljivih treba da budu smisleni. Zbog učitavanja fajlova sa lokalne mašine koristiti za razvoj Brackets sa opcijom "Live preview" ili neki drugi editor, uz Mozilla Firefox.

Rezultati merenja koji su obrađivani u ovom radu su preuzeti sa linka [9]. Ceo sistem ima edukativni karakter, a JavaScript je objektno orijentisan jezik, pa su u kodu date i smernice za refaktorisanje postojećeg koda i prikazane su na slici 5.

```
function Object_values(obj) {
    var vals = [];
    for (var prop in obj) {vals.push(obj[prop]);}
    return vals}

function MesurementUnit(keys, values) {
    var newUnit = {};
    this.keys = keys;
    this.values = values;
    return newUnit}

function createMyFirstObject(inputData) {
    outputData = new MesurementUnit();
    outputData.keys = Object.keys(inputData[0]);
    outputData.values = Object_values(inputData[0]);
    return outputData}
```

Sl 5.Objektno orijentisane smernice za refaktorisanje postojećeg koda

VI. PRIMENA APLIKACIJE U NASTAVI

Kako bi se spregnuo svet metrologije i softvera, na katedri za električna merenja uveden je predmet na IV godini studija pod nazivom "Web bazirani merno akvizicioni sistemi" čija je srž opisana web aplikacija. Naredni koraci predstavljaju nadograđivanje i usavršavanje metodoloških pristupa prilikom upoznavanja studenata sa novim softverskim alatima i konceptima. Pokazano je u praksi da je projektni rad takođe neizostavna celina koju je neophodno da imaju studenti inženjerskih nauka na višim godinama studija. Ovakav pristup otvara mogućnost da studenti dobiju detaljniji i sveobuhvatniji uvid u inženjerski način razmišljanja. Jedan od predloga može biti projektovanje mikrokontrolerskog sistema za merenje/nadgledanje odgovarajućeg procesa gde je neophodno da se upotrebi poznavanje hardvera, firmvera i softvera, kao i njihovu integraciju i pravilnu sinhronizaciju.

Kroz naredne ispite moguće je izvršiti detaljnije upozavanje studenata sa dodatnim razvojnima alatima. Moguće je dodati bazu podataka umesto upisivanja u postojeći fajl, merenje i prikaz rezultata u realnom vremenu, povezivanje više pomenutih mikrokontrolerskih sistema u jednu celinu itd.

Veliki broj zahteva ostavlja dovoljno prostora mentoru da na adekvatan način i izvrši ocenjivanje projektnog zadatka, gde bi svaki zahtev nosio odgovarajući broj bodova pri formiranju konačne ocene.

VII. ZAKLJUČAK

U radu je dat opis web aplikacije koja se može koristiti u različitim metrološkim konceptina poput IoT, pametne kuce, konceptu udaljenih merenja... Aplikacija predstavlja edukativnu platformu za savladavanje osnovnih softverskih web alata.

HTML 5 je upotrebljen za izradu strukture sadržaja zato što omogućava efikasno dodavanje dinamičkog sadržaja. Zbog redukovanja kompleksnosti, za stilizovanje web stranice je

umesto SASS-a korišćen CSS. Takođe, zbog redukovanja kompleksnosti aplikacija ne koristi Angular framework, već je napisana u čistom JavaScript-u (ECMAScript 6). Prilikom pisanja aplikacije akcenat je stavljen na modularnost i na upotrebu open source tehnologija. Pored mnoštva modernijih jedinica mere, za veličinu prikaza je korišćen pixel (px). Takođe, dat je i niz primera koji omogućavaju da se svako parče ove aplikacije može efikasnije analizirati. Autori se aktivno trude da koriste tzv. cutting edge tehnologije i logički dodatak na ovu aplikaciju je trenutno u fazi razvoja, gde se kao back-end koristi .NET Core tehnologija i SignalR middleware, koji pored već atraktivnog načina prikazivanja, nude i mogućnost prikazivanja rezultata u realnom vremenu. Funkcionalnost front-enda će u narednim verzijama biti proširena sa medija upitima koji su prilagođeni gledanju rezultata merenja na papiru.

LITERATURA

[1] https://github.com/IvanGutai/MicroLearning
[2] https://material.io/tools/devices
[3] https://www.chartjs.org
[4] https://validator.w3.org
[5] https://color.adobe.com/create/color-wheel
[6] https://www.oreilly.com/library/view/beginning-javascript-5th/9781118903742
[7] https://www.packtpub.com/web-development/responsive-web-design-html5-and-css3-second-edition
[8] https://www.pluralsight.com
[9] http://newcastle.urbanobservatory.ac.uk

ABSTRACT

In this paper, an example of a front-end module which shows measuring results from seven different sensors in a responsive manner is described. Responsive part enables that results can be viewed easily from PCs and from portable devices. Every sensor has a separate panel. On high resolutions, two panels are shown in one row, on portable devices panels are stacked. Each panel shows a list of measuring results with adequate time values, at the same time, for easier monitoring purposes graph with all values is shown. Each of the basic examples which contain programming code written in JavaScript, given as an addition to this paper is made in the way that can be dealt with in less than 15 minutes. Besides educational character this concept is compatible with basic Microlearning principles.

Web-based measuring system - educational front-end

Ivan Gutai, Đorđe Novaković, Platon Sovilj, Dragan Pejić, Marina Bulat, Nemanja Gazivoda

Međuprovera EMC analizatora spektra između dva etaloniranja

Aleksandar M. Kovačević, Nenad Munić, Veljko Nikolić, Ljubiša Tomić, Ivana Kostić

***Apstrakt*—Međuprovera EMC analizatora spektra između dva etaloniranja prikazana je u ovom radu. Međuprovera je bila potrebna da bi se održalo poverenje u status etaloniranja EMC analizatora spektra. Pri tome, ta provera se obavlja u skladu sa utvrđenom procedurom.**

***Ključne reči*—Međuprovera; EMC analizator spektra; poverenje u etaloniranje.**

I. Uvod

SVA oprema koja se koristi za ispitivanja, uključujući opremu za pomoćna merenja (npr. za uslove okoline), a koja bitno utiče na tačnost ili valjanost rezultata ispitivanja, mora da bude etalonirana pre puštanja u upotrebu [1]. Pri tome, laboratorija mora da uspostavi program i proceduru za etaloniranje svoje opreme. U skladu sa utvrđenim programima i procedurama, moraju se obavljati neophodne međuprovere radi održavanja poverenja u status etaloniranja [1].

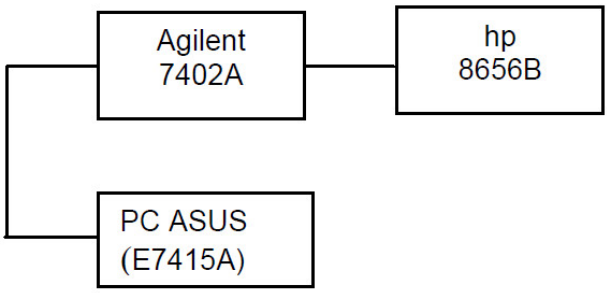
Odeljenje za elektromagnetsku kompatibilnost i uticaje okoline je akreditovano u oblasti ispitivanja elektromagnetske kompatibilnosti (u daljem tekstu Odeljenje za EMC i uticaje okoline) u okviru Centra za ispitivanje proizvoda, Tehnički opitni centar (TOC) iz Beograda [2]. Tako, Odeljenje za EMC i uticaje okoline svake godine vrši periodične preglede (međuprovere) svoje ključne opreme, a na osnovu izrađenog Plana pregleda ključne merne opreme za tu godinu. Naime, u TOC-u je izrađen dokument Uputstvo o metrološkom obezbeđenju Tehničkog opitnog centra [3] u skladu sa standardom SRPS ISO/IEC 17025:2017 [1], na osnovu koga je uspostavljen program i procedura za etaloniranje sopstvene opreme. Pri tome, Odeljenje je izradilo Plan pregleda ključne merne opreme, koja se periodično pregleda između dva etaloniranja za 2018. godinu [4]. Tako, za EMC analizator spektra E7402A, „AGILENT“, u Planu je naveden period etaloniranja od 1 godine i period internog pregleda (međuprovera) od 6 meseci, koji su određeni na osnovu dokumenata Normativ vremena baždarenja i verifikacije etalona i mernih sredstava na upotrebi u VJ [5] i Procedure za preispitivanje rokova periodičnog etaloniranja merne oprema TOC [6]. Na osnovu Plana, kao i standarda SRPS ISO/IEC 17025:2017 [1], izvršen je periodični pregled (međuprovera) EMC analizatora spektra E7402A, „AGILENT“, između dva etaloniranja (etaloniran 09.05.2018. godine, Zapisnik br. 2-138/18, akreditovana metrološka laboratorija TOC-a).

Cilj međuprovere je da, na osnovu analize dobijenih rezultata i zadatog kriterijumima, Odeljenje za elektromagnetsku kompatibilnost i uticaje okoline održi poverenje u status etaloniranja EMC analizatora spektra E7402A, „AGILENT“.

Aleksandar M. Kovačević, Nenad Munić, Veljko Nikolić, Ivana Kostić – Tehnički opitni centar, Generalštab Vojske Srbije, Vojvode Stepe 445, 11000 Beograd, Srbija (e-mail: aleksandarkovacevic1962@yahoo.com). Ljubiša Tomić – Vojnotehnički Institut, Ratka Resanovića 1, 11000 Beograd, Srbija (e-mail: ljubisa.tomic@gmail.com).

II. Uslovi međuprovere

Periodični pregled (međuprovera) EMC analizatora spektra E7402A, „AGILENT“, između dva etaloniranja (etaloniran 09.05.2018. godine, Zapisnik br. 2-138/18), izvršen je dana 15.11.2018. godine u ekranizovanoj sobi (Faradejev kavez). Pri tome, kao izvor signala korišćen je signal generator hp 8656B. Za potrebe međuprovere korišćen je stoni računar ASUS sa aplikacijom za automatizaciju merenja (EMC Measurement Application E7415A). Šema povezivanja merne opreme (merila) za međuproveru je data na Sl. 1.



Sl. 1. Šema povezivanja merne opreme za međuproveru.

Periodični pregled (međuprovera) EMC analizatora spektra E7402A je obuhvatio sledeće:

1. Provera tačnosti merenja nivoa signala, pri zadatoj vrednosti nivoa ulaznog signala i različitim vrednostima frekvencije;
2. Provera tačnosti merenja nivoa signala, pri konstantnoj vrednosti frekvencije i različitim vrednostima nivoa ulaznog signala.

Provera tačnosti merenja nivoa signala, za različite vrednosti frekvencije (videti Tabelu 1), vršena je u frekvencijskom opsegu od 100 kHz do 990 MHz, pri zadatoj

vrednosti nivoa ulaznog signala od – 10 dBm iz signal generatora.

Provera tačnosti merenja nivoa signala, za različite vrednosti nivoa ulaznog signala iz signal generatora, od – 80 dBm do + 10 dBm, vršena je na frekvencijima od 80 MHz i od 800 MHz, respektivno.

Za navedena merenja korišćena su sledeća merna sredstva i oprema:

- Signal generator hp 8656B, „HEWLETT PACKARD“, od 100 kHz do 990 MHz,
- EMC analizator spektra E7402A, „AGILENT“, od 100 Hz do 3 GHz,
- Stoni računar ASUS sa aplikacijom za automatizaciju merenja (EMC Measurement Application E7415A)
- Kabl RG-214/U.

Pri tome, karakteristike merne opreme zadovoljavaju propisani standard [7].

Uslovi okoline:

- temperatura okoline: 21 °C ± 2 °C,
- relativna vlažnost vazduha: 65 % ± 15 %.

III. Kriterijum za ocenu rezultata međuprovere

Za ocenu rezultata međuprovere korišćena je sledeća formula:

$$a_{\text{zad}} - U_m \le a_{\text{iz}} \le a_{\text{zad}} + U_m \qquad (1)$$

gde su:

a_{zad} – zadati nivo ulaznog signala (dBm),

a_{iz} – izmereni nivo ulaznog signala (dBm),

U_{m} – proširena merna nesigurnost provere tačnosti nivoa signala (za faktor proširenja ili prekrivanja $k = 2$).

Kao kriterijum za ocenu uspešnosti rezultata međuprovere uzeta je zadovoljenost formule (1).

Ukoliko je zadovoljenost formule (1) ispunjena, smatra se da je funkcionalnost EMC analizatora spektra E7402A – Agilent u periodu između dva etaloniranja očuvana, pa nema potrebe za korektivnim merama. Međutim, ukoliko zadovoljenost formule (1) nije ispunjena, potrebno je uvesti korektivne mere (vanredno etaloniranje, opravka, rashodovanje i sl.) [3].

IV. Rezultati međuprovere

Rezultati međuprovere se odnose na: 1) Proveru tačnosti merenja nivoa signala, pri zadatoj vrednosti nivoa ulaznog signala i različitim vrednostima frekvencije, 2) Proveru tačnosti merenja nivoa signala, pri konstantnoj vrednosti frekvencije i različitim vrednostima nivoa ulaznog signala.

Rezultati provere tačnosti merenja nivoa signala, za različite vrednosti frekvencije, u frekvencijskom opsegu od 100 kHz do 990 MHz, pri zadatoj vrednosti nivoa ulaznog signala od – 10 dBm iz signal generatora, dati su u Tabeli 1.

TABELA I

VREDNOSTI NIVOA SIGNALA ZA RAZLIČITE VREDNOSTI FREKVENCIJE, PRI ZADATOJ VREDNOSTI NIVOA ULAZNOG SIGNALA IZ SIGNAL GENERATORA

Zadata frekvencija [MHz]	Izmereni nivo (a_{iz}) [dBm]
0,1	-10,26
1	-10,59
50	-10,72
100	-10,22
150	-10,44
200	-10,79
300	-9,98
400	-9,92
600	-9,93
800	-9,63
900	-9,78

TABELA II

VREDNOSTI NIVOA SIGNALA NA FREKVENCIJI OD 80 MHZ, ZA RAZLIČITE VREDNOSTI NIVOA ULAZNOG SIGNALA IZ SIGNAL GENERATORA

Zadati nivo (a_{zad}) [dBm]	Izmereni nivo (a_{iz}) [dBm]
10	10,23
0	0,14
-10	-10,23
-20	-20,44
-30	-30,27
-40	-40,37
-50	-50,91
-60	-60,77
-70	-70,29
-80	-80,21

TABELA III

VREDNOSTI NIVOA SIGNALA NA FREKVENCIJI OD 800 MHZ, ZA RAZLIČITE VREDNOSTI NIVOA ULAZNOG SIGNALA IZ SIGNAL GENERATORA

Zadati nivo (a_{zad}) [dBm]	Izmereni nivo (a_{iz}) [dBm]
10	10,19
0	0,61
-10	-10,32
-20	-20,14
-30	-30,29
-40	-40,71
-50	-50,69
-60	-60,66
-70	-70,34
-80	-80,24

TABELA IV
KONAČNI PRORAČUN MERNE NESIGURNOSTI ZA SLUČAJ PROVERE TAČNOSTI MERENJA NIVOA SIGNALA

Uticajna veličina X_i	**Procena** X_i (x_i)		**Standardna nesigurnost** $u(x_i)$	**Koeficijent osetljivosti** c_i	**Doprinos standardnoj nesigurnosti** $u_i(y)=c_i\,u(x_i)$
	Vrednost [dB]	**Funkcija raspodele**			
Pokazivanje analizatora spektra	±0,1	pravougaona, k = 1,732	0,058	1	0,058
Tačnost prijema sinusnog signala na analizatoru spektra	±1,0	normalna, k = 2,000	0,500	1	0,500
Slabljenje kablova	±0,2	normalna, k = 2,000	0,1	1	0,100
Tačnost zadate vrednosti sinusnog signala iz signal generatora	±1,0	normalna, k = 2,000	0,500	1	0,500
Kombinovana standardna nesigurnost $u_c(y)$		normalna	$u_c(y) = \sqrt{\sum_i u_i^2(y)}$		0,716
Proširena merna nesigurnost U_m		normalna, k = 2	$U_m = k u_c(y)$		**1,43**

Rezultati provere tačnosti merenja nivoa signala na frekvenciji od 80 MHz, za različite vrednosti nivoa ulaznog signala iz signal generatora od – 80 dBm do + 10 dBm, respektivno, dati su u Tabeli 2. Dok su rezultati provere tačnosti merenja nivoa signala na frekvenciji od 800 MHz, za različite vrednosti nivoa ulaznog signala iz signal generatora od – 80 dBm do + 10 dBm, respektivno, dati u Tabeli 3.

Pri tome, konačni proračun merne nesigurnosti za slučaj provere tačnosti merenja nivoa signala, tip B, koji je prikazan u Tabeli 4, izračunat je u skladu sa smernicama datim u [8, 9], i odnosi se na najgori slučaj provere tačnosti merenja nivoa signala. U navedenoj tabeli, vrednosti za uticajne veličine (pokazivanje analizatora spektra, tačnost prijema sinusnog signala na analizatoru spektra, tačnost zadate vrednosti sinusnog signala iz signal generatora, slabljenje kablova) su dobijene iz proizvođačke specifikacije i kalibracionih sertifikata (uverenja o etaloniranju).

Iz Tabele 4 se vidi da proširena merna nesigurnost U_{m} iznosi 1,43 dB.

Kada se ubace vrednosti date u Tabelama 1, 2, 3, respektivno, kao i za U_{m}, u formulu (1), može se konstatovati da je zadovoljenost iste ispunjena. Na taj način, smatra se da je funkcionalnost EMC analizatora spektra E7402A – Agilent u periodu između dva etaloniranja očuvana, pa nema potrebe za korektivnim merama.

V. ZAKLJUČAK

Međuprovera merne opreme se koristi da akreditovana laboratorija održi poverenje u status etaloniranja iste.

Odeljenje za elektromagnetsku kompatibilnost i uticaje okoline iz Tehničkog opitnog centra iz Beograda, koje je akreditovano u oblasti ispitivanja elektromagnetske kompatibilnosti (EMC), svake godine vrši periodične preglede (međuprovere) svoje ključne opreme, na osnovu izrađenog Plana pregleda ključne merne opreme za tu godinu. Tako, Odeljenje je izradilo Plan pregleda ključne merne opreme, koja se periodično pregleda između dva etaloniranja za 2018. godinu, na osnovu koga je izvršen periodični pregled (međuprovera) EMC analizatora spektra E7402A, „AGILENT“, između dva etaloniranja, a u skladu sa standardom SRPS ISO/IEC 17025:2017. Pri tome, rezultati međuprovere se odnose na: 1) Proveru tačnosti merenja nivoa signala, pri zadatoj vrednosti nivoa ulaznog signala i različitim vrednostima frekvencije, 2) Proveru tačnosti merenja nivoa signala, pri konstantnoj vrednosti frekvencije i

različitim vrednostima nivoa ulaznog signala. Kao kriterijum za ocenu uspešnosti rezultata međuprovere uzeta je zadovoljenost formule (1). Cilj međuprovere je da, na osnovu analize dobijenih rezultata i zadatog kriterijumima, Odeljenje za EMC i uticaje okoline održi poverenje u status etaloniranja EMC analizatora spektra E7402A, „AGILENT“.

Kako je konstatovano da je zadovoljenost formule (1) ispunjena, smatra se da je funkcionalnost EMC analizatora spektra E7402A – Agilent u periodu između dva etaloniranja očuvana, pa nema potrebe za korektivnim merama.

ZAHVALNICA

Želeli bi da se zahvalimo Marijani Ferlan i Teodori Knežević iz TOC-a na obradi rezultata.

LITERATURA

[1] *Opšti zahtevi za kompetentnost laboratorija za ispitivanje i laboratorija za etaloniranje*, SRPS ISO/IEC 17025, ISS, 2017.

[2] http://www.toc.vs.rs.

[3] *Uputstvo o metrološkom obezbeđenju Tehničkog opitnog centra*, Interni dokument, TOC, 2009.

[4] *Plan pregleda ključne merne opreme, koja se periodično pregleda između dva etaloniranja za 2018. godinu*, Interni dokument, TOC, 2018.

[5] *Normativ vremena baždarenja i verifikacije etalona i mernih sredstava na upotrebi u VJ*, SiM-6, 1995.

[6] *Procedura za preispitivanje rokova periodičnog etaloniranja merne oprema TOC*, Interni dokument, TOC, 2005.

[7] *Specifikacija aparata i metoda za merenje radio-smetnji i imunosti – Deo 1-1: Aparati za merenje radio-smetnji i imunosti – Merni aparati*, SRPS EN 55016-1-1:2011/A1:2012/A2:2015, ISS.

[8] *Expression of the Uncertainty of Measurement in Calibration*, EA-04/02 Guide, EAL. <http://www.european-accreditation.org>.

[9] *Specifikacija aparata i metoda za merenje radio-smetnji i imunosti–Deo 4–2: Nepouzdanosti, statistike i modeliranje granica – Merna nepouzdanost instrumenata*, SRPS EN 55016-4-2:2013/A1:2014, ISS.

ABSTRACT

Intermediate check of EMC spectrum analyzer between the two calibrations is shown in this paper. Check needed to maintain confidence in the calibration status of the EMC spectrum analyzer. In addition, this check is performed in accordance to a defined procedure.

Intermediate check of EMC spectrum analyzer between the two calibrations

Aleksandar M. Kovačević, Nenad Munić, Veljko Nikolić, Ljubiša Tomić, Ivana Kostić

Automatizacija merenja nivoa imunosti na kondukcione smetnje

Nenad Munić, Aleksandar Kovačević, Vladimir Jokić, Veljko Nikolić, Ljubiša Tomić

***Apstrakt*— Merenje nivoa imunosti na kondukcione smetnje je karakteristika koja je značajna za većinu naoružanja i vojne opreme. Zahtevi i metode merenje nivoa imunosti su definisani domaćim standardima odbrane SORS 1029/89 i SORS 1762/89, respektivno, za koje postoji objektivna potreba da se osavremene u skladu sa aktuelnim međunarodnim standardima iz te oblasti. Shodno tome, koristeći standard MIL-STD-461F, u laboratoriji za EMC Tehničkog opitnog centra, pristupilo se prilagođavanju zahteva i metoda merenja nivoa imunosti na kondukcione smetnje iz navedenih domaćih vojnih standarda. Pri tome, prilagođavajući metode merenja raspoloživoj mernoj opremi laboratorije. U ovom radu će postupno biti prikazan razvoj jednog automatizovanog mernog mesta za merenje nivoa imunosti na kondukcione smetnje, kroz poseban osvrt na sledeće faze: kalibracija i ispitivanja.**

***Ključne reči*—Elektromagnetska smetnja; nivo imunosti; kondukcione smetnje.**

I. Uvod

JEDNA od karakteristika koja se meri tokom ispitivanja naoružanja i vojne opreme (NVO) u Tehničkom opitnom centru (TOC) [1] je nivo imunosti na elektromagnetske smetnje. Nivo imunosti je zapravo najviši nivo elektromagnetske smetnje koja deluje na uređaj, a pri kojem uređaj još uvek radi sa zahtevanim performansama [2]. Postoje dva načina kako se prilikom ispitivanja elektromagnetska smetnja prenosi do ispitivanog uređaja: pretežno slobodnim elektromagnetskim talasom, što predstavlja radijacione smetnje i pretežno vođenim talasom, duž provodnika (kablovima), što predstavlja kondukcione smetnje [3]. U zavisnosti od načina prenosa razlikujemo dve vrste merenja nivoa imunosti: merenja nivoa imunosti na radijacione smetnje i merenja nivoa imunosti na kondukcione smetnje.

Zahtevi i metode merenja nivoa imunosti su navedeni u domaćim standardima odbrane Republike Srbije SORS 1029/89 [4] i SORS 1762/89 [5], respektivno. Naime, to su standardi koji su usvojeni pre 30 godina i koji su izrađeni na osnovu američkog vojnog standarda MIL-STD-461B [6] iz 1980. godine. Od tada je Američki standard, usled ubrzanog razvoja električnih, elektronskih i elektromehaničkih sredstava NVO, pretrpeo brojne revizije (aktuelna „G“ revizija MIL standarda je iz 2015. godine), dok se domaći standardi nisu menjali.

Zbog toga, definisan je cilj da se, u skladu sa novijim međunarodnim vojnim standardima MIL-STD-461F [7], usklade zahtevi i metode merenja domaćih standarda SORS 1029/89 i SORS 1762/89. Ispitivanja imunosti na radijacione smetnje ili ispitivanja imunosti na polje EM smetnji [4] su u prethodnom periodu usklađena i rezultati su prikazani javnosti [8]. Takođe, potrebno je bilo to isto uraditi i po pitanju imunosti na kondukcione smetnje ili imunosti na kondukcione EM smetnje [4].

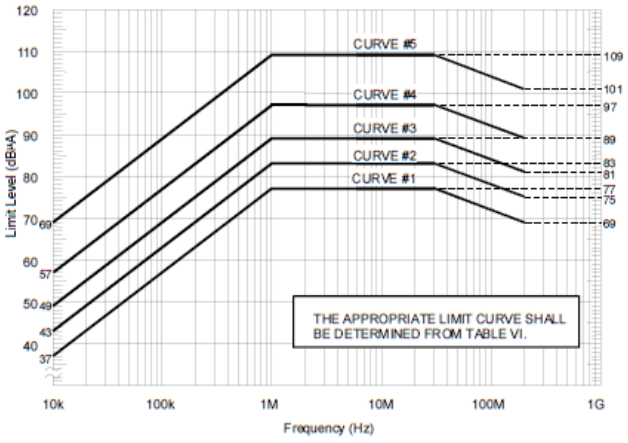
Predmet ovog rada je usavršavanje zahteva i metoda merenja nivoa imunosti na kondukcione smetnje standarda SORS 1029/89 i SORS 1762/89, sa aktuelnom verzijom standarda MIL-STD-461F, CS114 (eng. CS – conducted susceptibility). Pri tome, bilo je potrebno, u što većem obimu, automatizovati merenja, uz korišćenje postojeće merne opreme laboratorije za EMC.

Automatizacija merenja je realizovana u programskom paketu MATLAB [9] koristeći ugrađen „Instrument Control Toolbox“.

II. Merenje nivoa imunosti na kondukcione smetnje

Imunost na kondukcione smetnje se u domaćim vojnim standardima definiše kao sposobnost sredstva NVO da radi bez neželjenih odziva, pri delovanju definisanih EM smetnji, koje se prostiru duž provodnika [4].

Postojeći zahtevi i metode merenja imunosti na kondukcione smetnje su definisani u domaćim standardima SORS 1029/89 i SORS 1762/89 (zahtevi i metode za: iks1 i iks2), i njih je bilo neophodno usaglasiti sa metodom CS114 američkog vojnog standarda MIL-STD-461F.



Sl. 1. Nivoi imunosti na kondukcione smetnje (strujne smetnje) prema MIL-STD-461F, CS114 [7].

Da bi se postigao što tačniji nivo elektromagnetske smetnje, merenja nivoa imunosti smetnji (u našem slučaju kondukcionih) se rade u Faradejevom kavezu. Time se

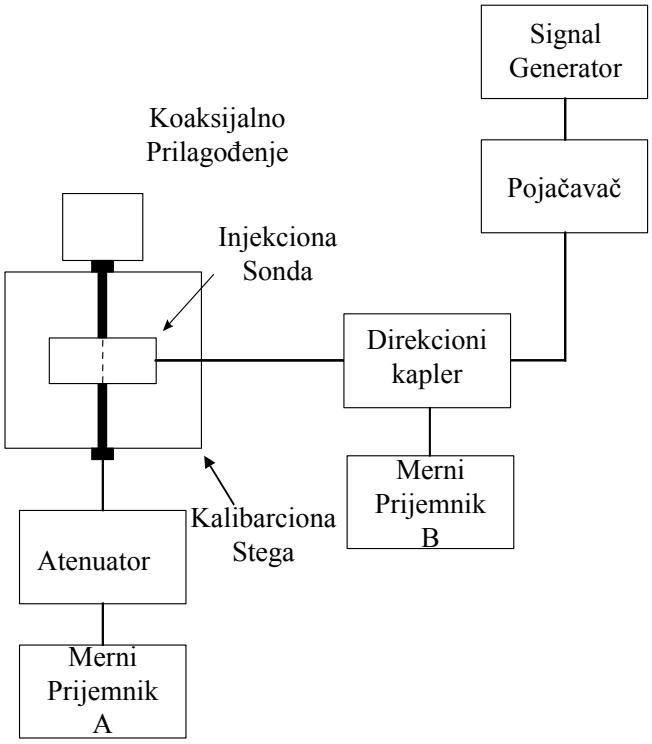
Nenad V. Munić – Tehnički opitni centar, Vojvode Stepe 445, 11000 Beograd, Srbija (e-mail: nenadmunic@yahoo.com).
Aleksandar M. Kovačević, – Tehnički opitni centar, Vojvode Stepe 445, 11000 Beograd, Srbija (e-mail: aleksandarkovacevic1962@yahoo.com).
Vladimir Jokić, –Tehnički opitni centar, Vojvode Stepe 445, 11000 Beograd, Srbija (e-mail: vlada0806@yahoo.com).
Veljko N. Nikolić, – Tehnički opitni centar, Vojvode Stepe 445, 11000 Beograd, Srbija (e-mail: veljkozmaj@yahoo.com).
Ljubiša Tomić – Vojnotehnički Institut, Ratka Resanovića 1, 11000 Beograd, Srbija (e-mail: ljubisa.tomic@gmail.com).

neutrališe eventualni uticaj elektromagnetske sprege između provodnika uređaja ili opreme koji se testira (EUT) i spoljašnjeg ambijentalnog elektromagnetskog polja. Pri ispitivanjima se kondukcione smetnje dovode u napojni kabl EUT-a. Zahtevi za nivo imunosti (struja kalibracije) su preuzeti iz standarda MIL-STD-461F, CS114, i prikazani su na sl. 1. Sa slike se može videti da postoji ukupno pet nivoa strujnih smetnji. Nivo smetnje se bira pre merenja i određuje se na osnovu namene i mesta ugradnje EUT-a. Jednom odabrani nivo se koristi i prilikom kalibracije mernog sistema i ispitivanja EUT-a.

Pri merenjima laboratorijski uslovi okoline su: temperatura okoline od 21° C ± 2° C, relativna vlažnost vazduha: 65 % ± 15 %.

III. KALIBRACIJA

Blok dijagram konfiguracije kalibracionog mernog mesta za ispitivanje imunosti na kondukcione smetnje, prema standardu MIL-STD-461F, CS114, prikazan je na sl. 2. Sa slike se vidi da se tokom kalibracije koristi direkcioni kapler, kojim se zapravo monitoriše ukupno predata RF snaga injekcionoj sondi kada je ona spregnuta u kalibracionu stegu i na izlazu te stege zatvorena sa koaksijalnim prilagođenjem, u ovom slučaju otpornikom od 50 Ω (videti sl. 3). Te vrednosti predate snage se pamte i koriste kao referentne vrednosti predajne snage, tokom ispitivanja EUT-a, kada se injekciona sonda umesto kalibracione strukture spreže sa naponskim provodnikom (kablom) EUT-a i tom prilikom se kontroliše i održava isti nivo predate snage.

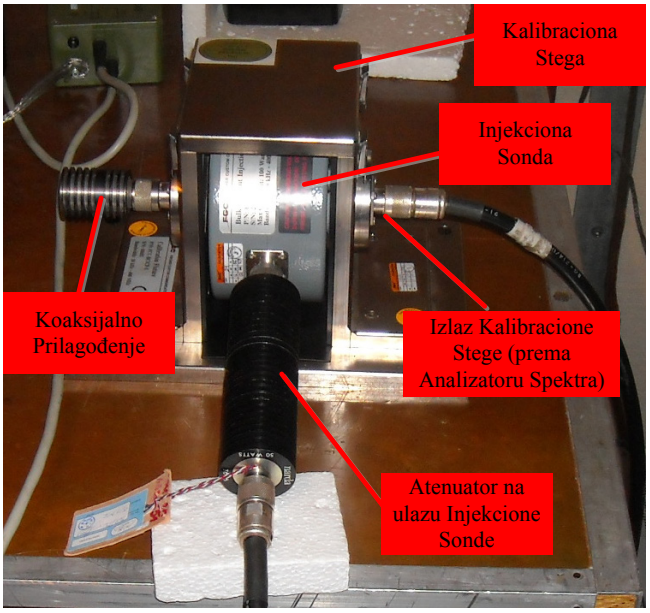


Sl. 2. Konfiguracija mernog mesta za kalibraciju, prema MIL-STD-461F CS114 [7].

Nažalost, TOC u svojoj mernoj opremi nema direkcioni kapler odgovarajuće snage i frekvencije, pa je monitorisanje predajne snage realizovano monitorisanjem nivoa izlazne snage signal generatora. Pri tome, monitorisanje se vrši za svaku frekvenciju iz mernog opsega frekvencija.

Prilikom kalibracije, za merenje nivoa injektovane struje koristi se analizator spektra (na sl. 3 označen kao Merni prijemnik A), pre kojeg je postavljen atenuator. Atenuator se koristi dvojako: radi zaštite od oštećenja analizatora spektra od previsokog ulaznog nivoa signala i radi boljeg prilagođenja analizatora spektra prema kalibracionoj strukturi.

Na analizatoru spektra je potrebno izabrati da se rezultati merenja izražavaju u jedinicama dBμA, gde se, pri tome, podrazumeva da se koristi ostala merna opremu iste karakteristične impedanse od 50 Ω, kao što je kod analizatora spektra.



Sl. 3. Injekciona sonda spregnuta u kalibracionu strukturu i zatvorena sa prilagođenjem od 50 Ω.

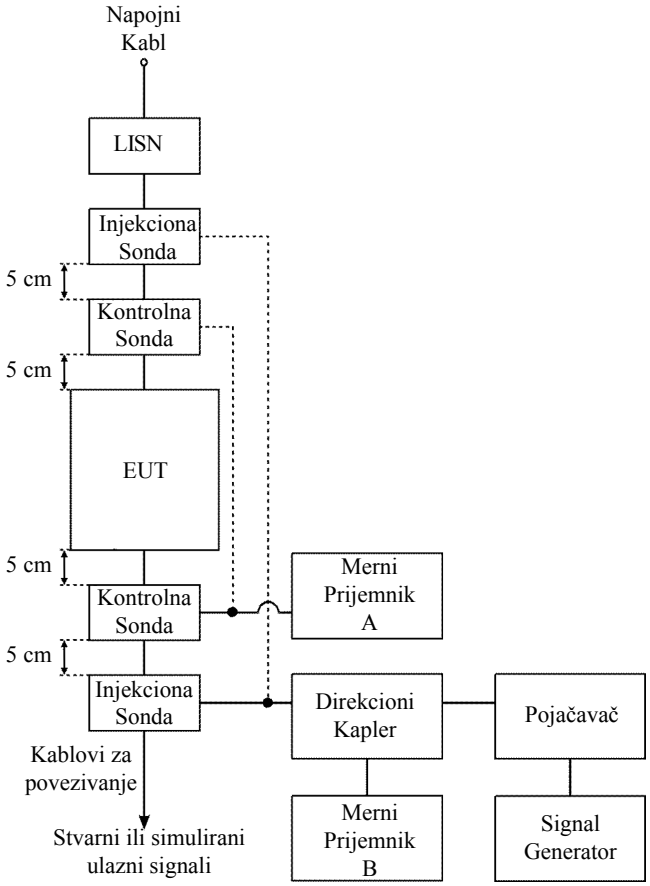
Tokom merenja, za svaku frekvenciju se očitava nivo injektovane struje i po potrebi se nivo snage signal generatora podiže ili smanjuje, tako da nivo struje bude takav da je veći od nivoa zahtevanog standardom, ali i da maksimalno premašenje tog nivoa ne bude veće od 1 dB. Nivo maksimalnog premašenja nije definisan standardom, ali je, u ovom slučaju, dobijen iskustveno, težeći da se broj iteracija prilagođavanja izlazne snage generatora prema merenom nivou injektovane struje smanji na što je moguće manji broj, čime se postiže razumno vreme trajanja kalibracije, a pri tome je maksimalna procentualna vrednost premašenja snage manja od 26%. Postignuto je da se kalibracija na svakoj frekvenciji uradi u maksimalno 3 koraka (iteracije), mada za većinu frekvencija se ona uradi u jednom koraku.

Za realizaciju ovog ispitivanja, usled velikog broja ispitnih frekvencija i složenosti merenja (paralelan rad sa više uređaja), bilo je neophodno izvršiti automatizaciju merenja. Time se trajanje ispitivanja skraćuje, a smanjuje se i mogućnost da usled ljudskog faktora dođe do greške pri merenjima.

IV. ISPITIVANJE

Za razliku od kalibracije kada se injekciona sonda povezuje na kalibracionu stegu, kod ispitivanja se ona

postavlja oko naponskog provodnika (kabla) EUT-a. Takođe, na istom provodniku, na rastojanju od 5 cm, postavlja se kontrolna sonda kojom se monitorišu nivoi injektovanih struja smetnji. Konfiguracija ovog merenja, prema standardu MIL-STD-461G, CS114, prikazana je na sl. 4. Iz istog razloga kao kod kalibracije, i kod ispitivanja EUT-a se ne koristi direkcioni kapler za monitorisanje nivoa predate snage injekcionoj sondi, već se prati izlazna snaga signal generatora. Postavljanjem kalibracionih vrednosti izlazne snage na signal generatoru se postižu ispitni uslovi, pri kojima je moguće posmatrati ili meriti funkciju EUT-a ili tražiti neki neželjeni odziv.

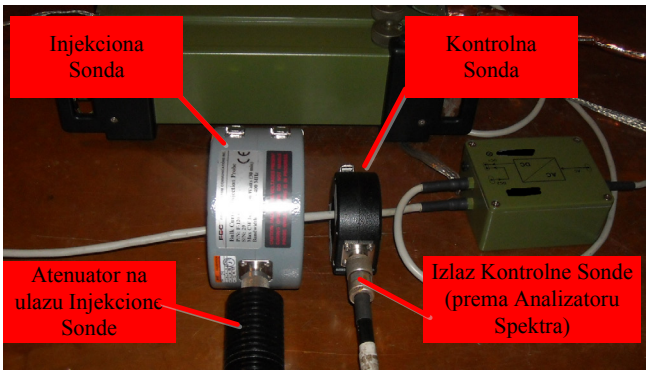


Sl. 4. Konfiguracija mernog mesta za ispitivanje uređaja koji se testira, prema MIL-STD-461F, CS114 [7].

Prilikom sprege injekcione sonde sa kalibracionom stegom, odnosno postavljanjem oko napojnog provodnika ili kabla EUT-a, javljaju se različite impedanse gledano od strane injekcione sonde prema analizatoru spektra. To može da dovede do različitih nivoa struje koja se injektuje prilikom kalibracije, odnosno ispitivanja. Naime, ako je impedansa u kolu sa EUT veća od one u kalibracionom kolu, onda nivo injektovane struje u provodniku će biti manji od one dobijene kalibracijom. Nasuprot toga, ukoliko je impedansa u kalibracionom kolu veća, onda injektovana struja može preći maksimalno dozvoljen nivo standardom i tada se u toku ispitivanja mora smanjiti snaga na signal generatoru. Tek kada se snaga dovoljno smanji, tako da vrednost injektovane struje bude u skladu sa standardom, onda se prati ispravnost funkcija EUT-a. Praćenje funkcije EUT-a podrazumeva traženje određenih indikatora degradacije ili prestanka rada EUT-a. Način kako se to utvrđuje zavisi od vrste uređaja i može se pratiti kroz: ugrađene „built-in-test“ (BIT) procedure, vizuelnim posmatranjem, slušanjem ili merenjem pojedinih izlaznih signala uređaja. Taj deo ispitivanja nije ponovljiv, tj. za svaki uređaj je različit, te se stoga on ne automatizuje. Pri tome, ispitivač ručno upisuje na kojim frekvencijama se javlja degradacija, te se na tim frekvencijama kasnije, ručnim merenjima, utvrđuje tačan nivo imunosti.

Za razliku od kalibracije kada je signal sa signal generatora sinusoidan, kod ispitivanja on je modulisan [6]. Način modulacije zavisi od tipa ispitivanog sredstva [5]. Tako na primer: za FM prijemnike se bira FM modulacija, za AM prijemnike AM modulacija, dok za većinu ostalih sredstava se bira impulsna odnosno amplitudska modulacija. Tom prilikom se u zavisnosti od parametara modulacije teži održavanju kalibracione vrednosti izlazne snage, modulisanog signala. Ukupno vreme izlaganja EUT-a na svakoj frekvenciji je ne manje od 3 s.

Tokom kalibracije i ispitivanja se vode automatizovani zapisi o nivoima injektovane struje u zavisnosti od frekvencije. Ispitivanje se ponavlja za svaki fazni napojni provodnik pojedinačno, isključujući povratni provodnik, uzemljenje i komplet napojnog kabla sa provodnicima (obuhvatom). Primer ispitivanja jednog EUT-a je prikazan na sl. 5.



Sl. 5. Deo merne postavke koji se koristi tokom ispitivanja uređaja.

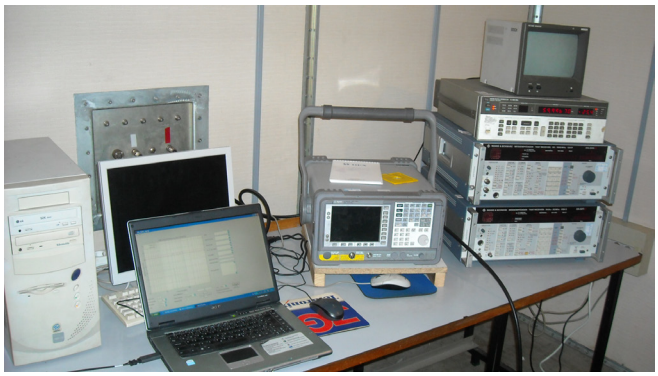
V. PRAKTIČNA REALIZACIJA AUTOMATIZOVANOG MERNOG MESTA ZA MERENJE NIVOA IMUNOSTI NA KONDUKCIONE SMETNJE

U ovom slučaju, proces merenja je automatizovan pisanjem koda u programskom paketu MATLAB [9]. Da bi se omogućila kontrola merne opreme, u MATLAB je instaliran dodatak „Instrument Control Toolbox”. Kao hardver za pokretanje MATLAB programa, moguće je koristiti PC računar ili laptop. Komunikacija između merne opreme i računara se obavlja preko HPIB (GPIB) magistrale. Za komunikaciju GPIB magistrale i USB porta računara koristi se „USB to GPIB“ adapter 82357B. Sam softver je namenski razvijen za potrebe ovog ispitivanja i omogućava realizaciju kalibracije i ispitivanja. Da bi adapter funkcionisao, mora se u operativnom sistemu instalirati odgovarajući drajveri, koji taj adapter podržavaju.

Za navedena merenja korišćena merna sredstva i oprema navedena u tabeli 1. Pri tome, karakteristike merne opreme zadovoljavaju propisani standard [10].

Pri pokretanju programa treba prvo izvršiti inicijalizaciju GPIB magistrale, zatim kreiranje objekta za rad sa GPIB magistralom, otvaranje tog objekta i zadavanje određenog

seta komandi. Treba voditi računa da stariji instrumenti ne podržavaju standardizovani set komandi (SCPI), te ako se oni ipak koriste tada nije moguća prosta promena nekog instrumenta drugim, bez prethodne izmene samog koda programa i prilagođavanja seta komandi instrumenta.



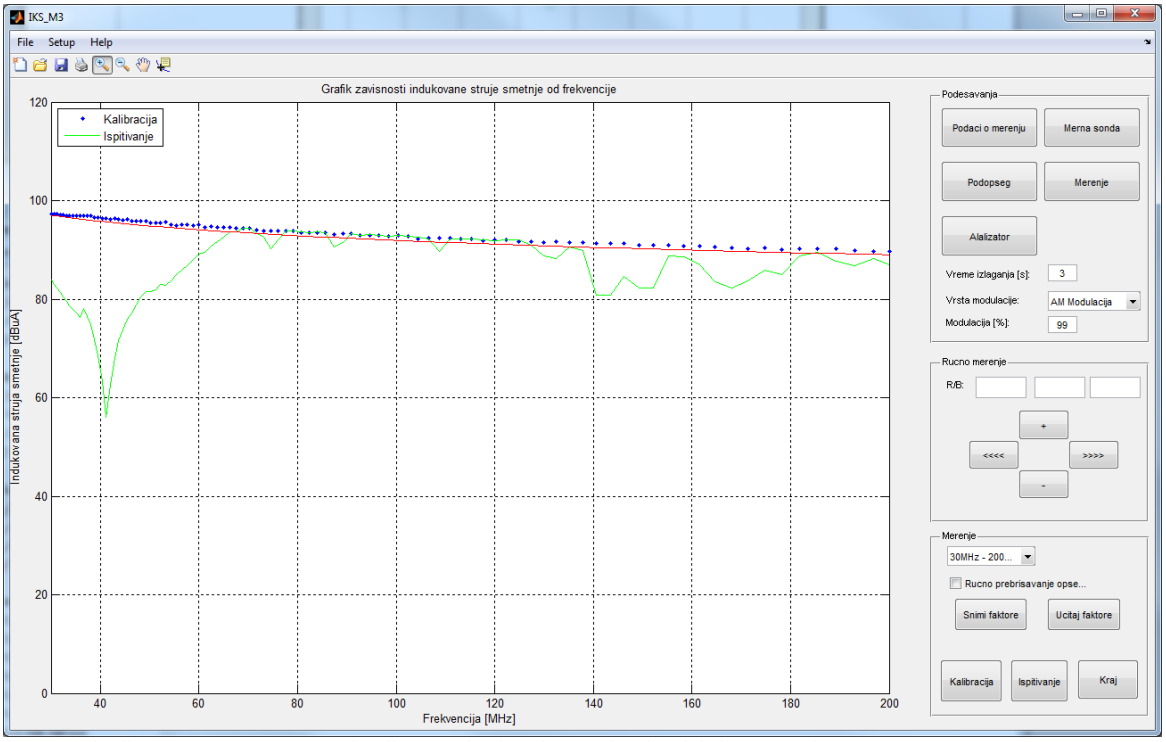
Sl. 6. Merna oprema koja se koristi za pravljenje automatizovanog mernog mesta.

U konkretnom slučaju treba da se kontrolišu dva instrumenta: signal generator i analizator spektra. Zbog toga je neophodno kreirati dva objekta kojima se ti instrumenti mogu nezavisno kontrolisati. Trebalo predvideti normalno funkcionisanje programa pri raznim slučajevima, kao na primer pri kalibraciji, kada nije moguće dobiti zahtevani nivo struje smetnji, a na signal generatoru je postavljen maksimalni nivo snage. Tada bi program očito upao u beskonačnu petlju, jer ne može da se ispuni uslov o zahtevanom nivou struje, ali se tada mogu definisati određene rutine, tako da se omogući ipak prelazak na sledeću frekvenciju, na primer tako što se za nivo imunosti na toj frekvenciji uzima maksimalno postignut nivo struje.

TABELA I
MERNA OPREMA ZA MERENJE NIVOA IMUNOSTI NA KONDUKCIONE SMETNJE

Naziv instrumenta	Proizvođač	Tip	serijski broj
Signal generator	HP	8656B	230A340387
RF pojačavač	OPHIR	5126	1020
Merni prijemnik	Agilent	E7402A	MY45119726
Razvodni transformator	Iskra	MA4801	277
Strujna sonda	Singer	94111-1	310
Strujna sonda	A.H. Systems	BCP-510	660
Strujna sonda	A.H. Systems	BCP-511	683
Injekciona strujna sonda	Fischer	F-120-6	29
Stega za kalibraciju strujne sonde	Fischer	FCC-BCICF-1	08681
RF i optički kablovi	/	RGU 214	/

Nivoi injektovane struje prilikom kalibracije i ispitivanja se pamte u promenljive i grafički se prikazuju kroz dijagram nivoa struje smetnji u zavisnosti od frekvencije (sl. 7). U slučaju neke nepravilnosti rada uređaja na nekoj frekvenciji, ta frekvencija se ručno upisuje van programa. Pri tome, po završetku prebrisavanja kompletnog podopsega, ručno vođenim postupkom (poluautomatski), kroz softverski zadate komande, dalje bi se izvršilo očitavanje minimalnog nivoa struje, gde dolazi do pogoršavanja rada EUT-a, kao i učitavanja nivoa struje kada uređaj ponovo uspostavi ispravnu funkciju. Nakon završetka ispitivanja, grafici kalibracije i ispitivanja se snimaju i postaju deo izveštaja o ispitivanju.



Sl. 7. Izgled prikaza programa za automatizaciju merenja nivoa imunosti na kondukcione smetnje.

VI. ZAKLJUČAK

Ovim radom su prezentovane osnovne bitne karakteristike merenja nivoa imunosti na kondukcione smetnje naoružanja i vojne opreme. Realizovana je implementacija standarda MIL-STD-461F, CS114, u laboratoriji za EMC u TOC-u. Pokazano je da se merenja nivoa imunosti na kondukcione smetnje mogu obavljati u Faradejevom kavezu i da je moguće umesto direkcionih kaplera, kojima se monitoriše predajna snaga injekcione sonde, da se primeni varijanta sa praćenjem nivoa snage signal generatora. Takođe, prikazana je praktična realizacija automatizovanog mernog mesta za merenje nivoa imunosti na kondukcione smetnje.

U daljem radu je potrebno na osnovu ovih rezultata pokrenuti postupak za izmenu standarda odbrane Republike Srbije SORS 1029/89 i SORS 1762/89, čime bi se oni revidirali i pratili trend svetskih standarda iz te oblasti.

LITERATURA

[1] http://www.toc.vs.rs.

[2] V. Prasad Kodali, *Engineering Electromagnetic Compatibility*, Institute of Electronics Engineers, NY, USA, 1996, ISBN 0-7803-1117-5.

[3] Antonije R. Đorđević, Dragan I. Olćan, *Ispitivanje elektromagnetske kompatibilnosti*, Elektrotehnički fakultet i Akademska misao, Beograd, ISBN 978-86-7466-446-9

[4] *Elektromagnetske smetnje, Zahtevi*, SORS 1029/89, Biro za standardizaciju i metrologiju u JNA, Beograd, 1989.

[5] *Elektromagnetske smetnje, Merenja*, SORS 1762/89, Biro za standardizaciju i metrologiju u JNA, Beograd, 1989.

[6] Electromagnetic Emission and Susceptibility Requirements for The Control of Electromagnetic Interference, MIL-STD-461B, Department of Defense, USA, 1980

[7] *Requirements for The Control of Electromagnetic Interference Characteristics of Subsystems and Equipment*, MIL-STD-461F, Department of Defense, USA, 2007.

[8] N. Munić, A. Kovačević, P. Rakonjac i V. Nikolić, *Automatizovana oprema za ispitivanje sa praktičnom realizacijom jednog mernog sistema za ispitivanje imunosti na elektromagnetsko polje smetnji*, Zbornik 60. konferencije ETRAN 2016, Zlatibor, ISBN 978-86-7466-618-0

[9] www.mathworks.com

[10] *Specifikacija aparata i metoda za merenje radio-smetnji i imunosti – Deo 1-1: Aparati za merenje radio-smetnji i imunosti – Merni aparati*, SRPS EN 55016-1-1:2011/A1:2012/A2:2015, ISS.

ABSTRACT

Conducted susceptibility is the significant feature for the most of the weapons and military equipment. Requirements and methods for measuring the immunity level are defined by domestic military standards SORS 1029/89 and SORS 1762/89, respectively and there is necessary to modernize in accordance with current international standards in this field. The requirements and methods of domestic military standards for measuring for immunity level to conducted disturbances in the laboratory for the EMC measurements of Technical Test Center have been harmonized to the MIL-STD-461F standard. The measurement methods have been adapted to the available measuring laboratory's equipment. In this paper, the development of an automatization of measurement for immunity level to conducted disturbances will be gradually shown, through a special review of the following phases: calibration and testing.

Automatization of Measurement for Immunity Level to Conducted Disturbances

Nenad Munić, Aleksandar Kovačević, Vladimir Jokić, Veljko Nikolić and Ljubiša Tomić

Linearizacija NTC termistora dvostepenim deo-po-deo linearnim A/D konvertorom kompaktnog dizajna

Jelena Jovanović, Dragan Denić

***Apstrakt*— U ovom radu predstavljen je kompaktan dizajn dvostepenog deo-po-deo linearnog A/D konvertora koji je primenjen za linearizaciju NTC termistora. Preciznije, ovim A/D konvertorom linearizuje se napon na izlazu serijsko-paralelnog razdelnika napona koji sadrži NTC termistor. Kompaktnost ovog konvertora se odnosi na upotrebu manjeg broja komparatora, što za posledicu ima smanjenje dimenzija kola i manju potrošnju energije. Kompaktnost se postiže time što oba stepena konverzije obavlja jedan isti fleš A/D konvertor sa dve različite lestvičaste mreže otpornika, svaka upotrebljena za po jedan stepen konverzije. Dakle, ovaj dvostepeni A/D konvertor ima istu rezoluciju u oba stepena konverzije. Kompenzacija nelinearnosti pomenutog napona se vrši u prvom stepenu konverzije na taj način što je prenosna funkcija prvog stepena deo-po-deo linearna aproksimacija funkcije koja je inverzna zavisnosti izlaznog napona razdelnika od temperature. Nakon primene predloženog 16-bitnog konvertora za linearizaciju termistora oznake NTSD0XV103FE1B0, proizvođača Murata, na opsegu od -40°C do 120°C nelinearnost je iznosila 0.0022%.**

***Ključne reči*— linearizacija; NTC termistor; dvostepeni deo-po-deo linearni A/D konvertor; kompaktnost dizajna; energetska efikasnost.**

I. UVOD

NTC termistor je temperaturno osetljivi otpornik sa negativnim temperaturnim koeficijentom i veoma izraženom nelinearnošću statičke prenosne funkcije (zavisnost otpornosti termistora od merene temperature je nelinearna i monotono opadajuća funkcija) [1]. U cilju dobijanja informacije o merenoj temperaturi u digitalnom formatu, potrebno je obezbediti električni signal (napon, struja) proporcionalan merenoj temperaturi. Iz pomenutog razloga, termistor se uvek postavlja u električno kolo sa konstantnim naponskim ili strujnim izvorom. Najčešće je to serijski ili serijsko-paralelni razdelnik napona [2]. Napon na izlazu ovakvog električnog kola zavisi, po nekom nelinearnom zakonu, od merene temperature jer i otpornost NTC termistora ima nelinearnu zavisnost od merene temperature. Linearizacija nelinearne zavisnosti pomenutog napona od merene temperature je problem sa kojim se autori bave u ovom radu. Kao rezultat rada na rešavanju ovog problema nastao je veliki broj metoda linearizacije, a sa njima i mogućnost da se odabere najpodesniji metod za određenu primenu.

Postoji niz tehnika linearizacije termistora koje se obavljaju pre A/D konverzije (analogne tehnike), kao i niz digitalnih tehnika linearizacije koje se obavljaju nakon A/D konverzije (primena mikrokontrolera i look-up tabela). Primera radi, u prvu grupu mogu se svrstati tehnike koje se baziraju na primeni naponskih razdelnika, poput Vitstonovog mosta i serijsko-paralelnog otpornog razdelnika napona [2], zatim kombinaciji serijsko-paralelnog razdelnika napona i operacionog pojačavača [3]. Interesantan primer predstavlja i primena aktivnog analognog kola koje se sastoji od stabilnog DC naponskog izvora, pojačavača jediničnog pojačanja, linearizacionog otpornika vezanog na red sa NTC termistorom i invertujućeg pojačavača [4]. Pojačavač jediničnog pojačanja sadrži operacioni pojačavač kojim se umanjuje pobudni napon koji može da prouzrokuje samozagrevanje termistora. Kako je pobudni napon negativan, invertujući pojačavač na svom izlazu daje pozitivan napon. Na ovaj način, obezbeđeno je da sa porastom temperature raste i izlazni napon. Adekvatnim izborom linearizacionog otpornika može se obezbediti linearna zavisnost izlaznog napona od merene temperature. U konkretnom slučaju, nelinearnost nakon izvršene linearizacije iznosi ±1% za temperaturni opseg od 30°C do 120°C.

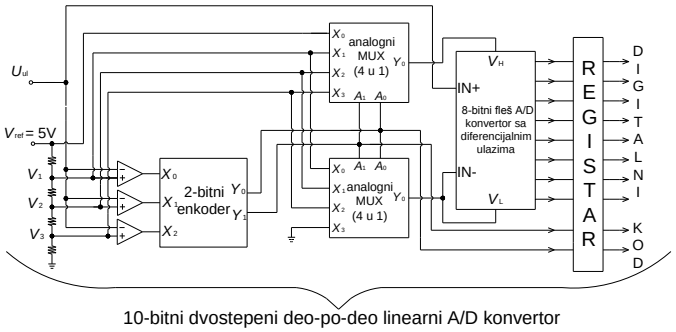
Analogne tehnike se često koriste, ali je nedostatak ovih tehnika to što dodatne komponente povećavaju i troškove realizacije i potrošnju energije. Takođe, vremenom se karakteristike analognih komponenti menjaju. Dodatno, kompenzaciju kros-osetljivosti senzora koji se linearizuje na različite parametre teško je izvesti samo primenom hardverskih komponenti.

U grupu digitalnih tehnika linearizacije spada primena look-up tabela. Ova tehnika linearizacije je često zastupljena u tzv. embeded mernim sistemima (Embedded Measurement Systems-EMS) u kojima se, za različite obrade signala, koriste mikrokontroleri. Problem sa ovim tehnikama ogleda se u tome što je potrebna veća memorija za smeštanje veće look-up tabele, pa se kod ovih tehnika uvek teži postizanju kompromisa između potrebnog memorijskog prostora i tačnosti merenja [5].

Međutim, vremenom su razvijene tzv. mešovite (analogno-digitalne) tehnike linearizacije senzora, koje se zasnivaju na primeni deo-po-deo linearnih A/D konvertora [6, 7]. Drugim rečima, linearizacija se izvodi simultano sa digitalizacijom rezultata merenja. Princip na kome se bazira linearizacija

Jelena Jovanović – Univerzitet u Nišu, Elektronski fakultet, Aleksandra Medvedeva 14, 18000 Niš, Srbija (e-mail: jelena.jovanovic@elfak.ni.ac.rs).

Dragan Denić – Univerzitet u Nišu, Elektronski fakultet, Aleksandra Medvedeva 14, 18000 Niš, Srbija (e-mail: dragan.denic@elfak.ni.ac.rs).



Sl. 1. Dvostepeni deo-po-deo linearni A/D konvertora sa 2-bitnim prvim stepenom, i 8-bitnim drugim stepenom konverzije, klasičnog dizajna

senzora primenom A/D konvertora je to što je statička prenosna funkcija A/D konvertora deo-po-deo linearna aproksimacija funkcije koja je inverzna funkciji koju je potrebno linearizovati. Najvažnija prednost ovih tehnika je to što se istovremeno, istim kolom, obavljaju dve različite obrade signala, čime se skraćuje ukupno vreme obrade signala. U odnosu na ostale tehnike linearizacije, troškovi realizacije celog mernog sistema i potrošnja energije su manji, a i postiže se veća kompaktnost dizajna što je naročito važno ako je potrebno realizovati sistem na čipu i sa velikim stepenom integrisanosti.

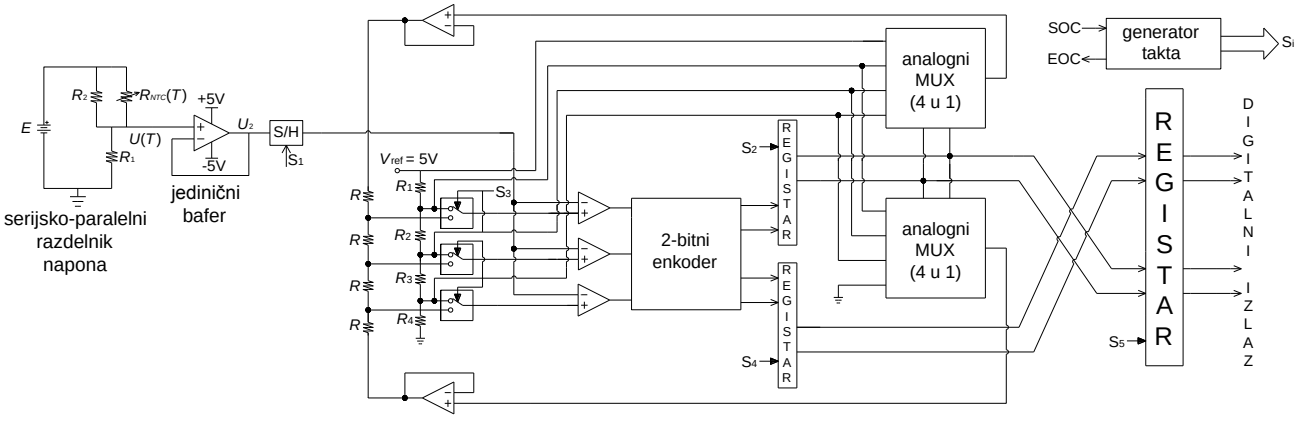
Linearizacija senzora primenom A/D konvertora obično se izvodi u dva stepena [6, 7], i to tako što se u prvom stepenu vrši kompenzacija nelinearnosti senzora, a u drugom stepenu smanjuje greška kvantizacije uneta u prvom stepenu i povećava ukupna rezolucija i preciznost merenja. Drugi stepen A/D konverzije je linearan i obično ima veću rezoluciju, za razliku od prvog koji je deo-po-deo linearan i ima manju rezoluciju (zbog kompleksnijeg dizajna). Prvi stepen konverzije obavlja fleš A/D konvertor jer je jedino u toj arhitekturi moguće izvesti deo-po-deo linearnu funkciju konverzije, dok drugi stepen može biti bilo koji tip konvertora (npr. sa sukcesivnim aproksimacijama). Još jedna velika prednost deo-po-deo linearnog A/D konvertora, u odnosu na linearni, je ta što omogućava povećanje rezolucije merenja samo u užem delu mernog opsega, čime se smanjuje prosečan broj bita za kodovanje odmeraka [6]. Primer jednog takvog A/D konvertora dat je na Sl. 1.

I pored brojnih prednosti na strani dvostepenog deo-po-deo linearnog A/D konvertora, ostaje prostora za njegovo unapređenje, pre svega u smislu jednostavnosti i kompaktnosti dizajna. U ovom radu autori primenjuju kompaktan dizajn dvostepenog deo-po-deo linearnog A/D konvertora čija oba stepena imaju istu rezoluciju, jer ih izvodi jedan fleš A/D konvertor sa dve različite lestvičaste mreže otpornika. Više o samom dizajnu i primeni ovog A/D konvertora u linearizaciji NTC termistora biće u narednom poglavlju rada.

II. KOMPAKTAN DIZAJN DVOSTEPENOG DEO-PO-DEO LINEARNOG A/D KONVERTORA

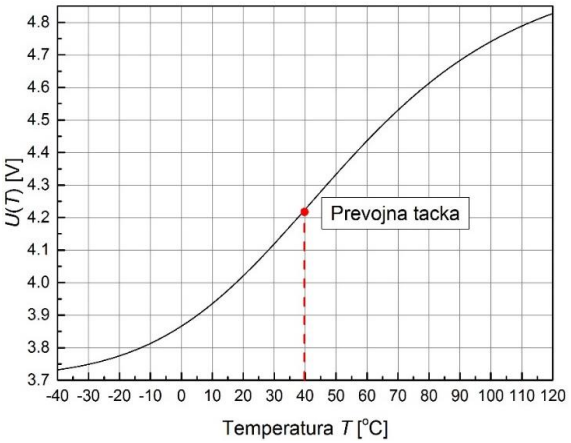
U ovom radu autori predlažu da se linearizacija NTC termistora, koji je deo serijsko-paralelnog razdelnika napona sa konstatnim naponskim izvorom, izvrši linearizacijom izlaznog napona razdelnika u dvostepenom deo-po-deo linearnom A/D konvertoru kompaktnog dizajna. Električna šema predloženog kola data je na Sl. 2. Kolo dvostepenog deo-po-deo linearnog A/D konvertora predstavlja kompaktnu verziju konvertora koju su autori koristili u radu [8]. Kompaktna verzija konvertora je razvijena sa ciljem smanjenja potrošnje energije smanjenjem broja komparatora koji učestvuju u njegovoj realizaciji, jer oba stepena konverzije obavlja jedan isti fleš A/D konvertor. Poznato je da se broj komparatora koji čine fleš A/D konvertora udvostručuje sa povećanjem rezolucije za 1 bit. Ukupna rezolucija linearizacionog kola iznosi $n=2N$, gde N predstavlja rezoluciju fleš A/D konvertora.

Zavisnost otpornosti NTC termistora od temperature može se modelovati, tj. aproksimirati tro-parametarskom Steinhart-Hart jednačinom [1, 9]. Primena ovog modela zahteva određivanje tri koeficijenta (Steinhart-Hart koeficijenti A, B i C) za šta je potrebno i dovoljno imati tri kalibracione tačke, tj. parove vrednosti (otpornost termistora, temperatura). Ove podatke smo preuzeli iz kalibracione tabele koju je obezbedio proizvođač termistora koji smo linearizovali. Reč je o NTC termistoru oznake NTSD0XV103FE1B0 proizvođača Murata [10], čiji se radni opseg prostire od -40°C do 125°C. Za proračun vrednosti otpornika R_1 i R_2 serijsko-paralelnog razdelnika napona potrebni su sledeći parametri [1]:



Sl. 2. Linearizacija NTC termistora dvostepenim deo-po-deo linearnim A/D konvertorom kompaktnog dizajna

disipaciona konstanta termistora C_d (mW/°C), otpornost termistora na 40°C (temperatura izabrane prevojne tačke, koja može uzeti bilo koju vrednost iz mernog opsega), greška merenja temperature prouzrokovana samozagrevanjem NTC termistora ΔT (°C) [1, 2, 11], i temperaturna osetljivost materijala od koga je napravljen termistor β (°K) [10]. Vrednosti ovih parametra mogu se naći u specifikacionim dokumentima koje obezbeđuje proizvođač termistora. Proračunate vrednosti otpornika R_1 i R_2 iznose: R_1=14.99 kΩ i R_2=5.22 kΩ [8]. Izlazni napon $U(T)$, na otporniku R_1, je nelinearan, a njegov oblik prikazan je na Sl. 3. Nelinearnost napona $U(T)$ je naročito izražena na granicama posmatranog temperaturnog opsega, a najbolja linearnost je u okolini prevojne tačke (40°C). Izbor prevojne tačke utiče na vrednosti otpornika R_1 i R_2, odnosno promenom vrednosti ovih otpornika menja se i oblik napona $U(T)$. U ovom radu je odabrana prevojna tačka na sredini posmatranog temperaturnog opsega.



Sl. 3. Oblik nelinearnog napona $U(T)$

Linearizacija napona $U(T)$ se izvodi dvostepenim deo-po-deo linearnim A/D konvertorom čiji je dizajn [7, 12] adaptiran specijalno za njegovu linearizaciju (refererentni naponi komparatora). U kolu prikazanom na Sl. 2 u oba stepena konverzije se koristi fleš A/D konvertor rezolucije 2 bita (N=2 bita), pa je u skladu sa konvencionalnim dizajnom fleš A/D konvertora i datom rezolucijom potrebno tri komparatora [1, 13]. Odmerak nelinearnog napona $U(T)$ i referentni naponi dovode se na ulaze komparatora radi međusobnog poređenja. Referentne napone dobijamo na otpornicima čije se otpornosti međusobno razlikuju (R_1, R_2, R_3 i R_4), jer su i referentni naponi neuniformno raspoređeni između 0 i V_{ref}. Vrednosti otpornika su odabrane tako da se na ulazima komparatora mogu podesiti vrednosti referentnih napona koje odgovaraju "break" naponima (break naponi predstavljaju granice linearnih segmenata različite širine koji čine deo-po-deo linearnu prenosnu funkciju prvog stepena konverzije). U opštem slučaju, vrednosti otpornika se moraju proračunati i podesiti unapred, kako bi linearizacija prenosne funkcije nekog senzora, ili signala sa izlaza senzora, bila moguća.

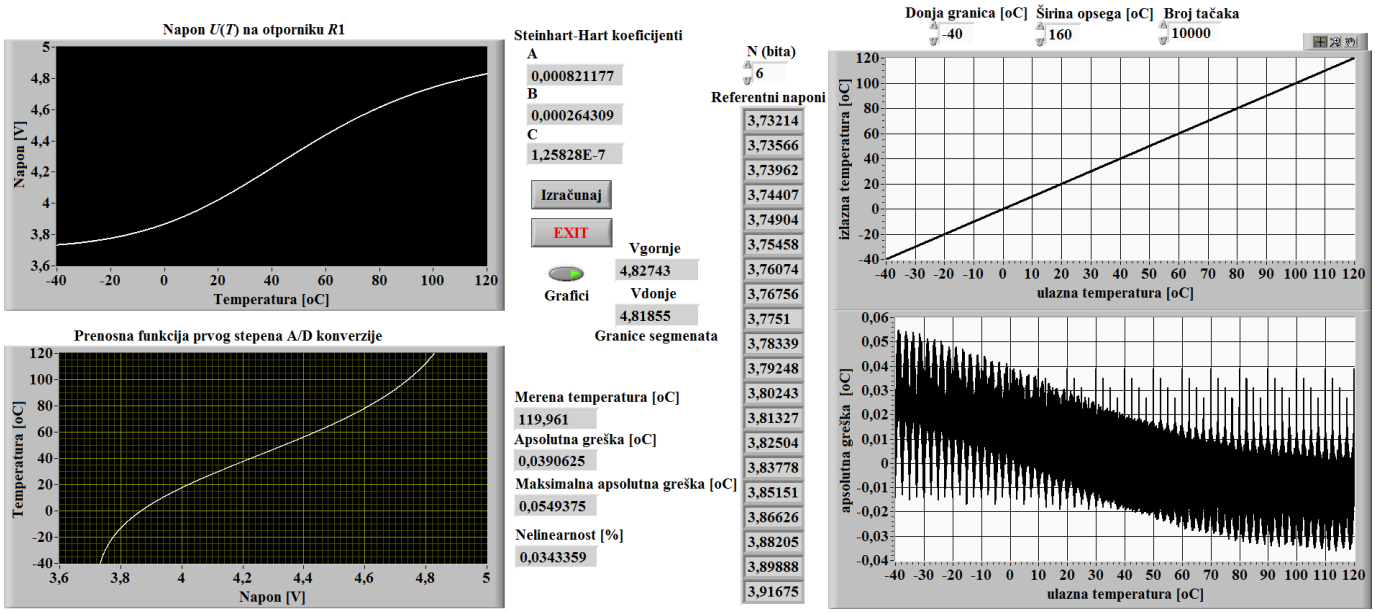
U razmatranom slučaju, signal koji se linearizuje je napon $U(T)$, pa je prenosna funkcija prvog stepena A/D konverzije deo-po-deo linearna aproksimacija njemu inverzne funkcije. Referentni naponi komparatora u prvom stepenu A/D konverzije (rezolucije 2 bita) dobijaju se tako što se temperaturni opseg podeli na 2^2=4 segmenta jednake širine (granice segmenata su T_i, i=0,1,…,2^2+1), i nađu vrednosti $U(T_k)$, k=1,…,2^2-1, pri čemu se krajnje granice temperaturnog opsega ne uzimaju u obzir (u ovom slučaju T_0 i T_5).

Rezultat prve faze konverzije su dva bita koja se smeštaju u registar primenom sinhronizacionog signala S2, a koriste se za kontrolu dva analogna multipleksera 4 u 1. Zadatak ovih multipleksera je da izdvoje granice segmenta kome pripada trenutna vrednost signala $U(T)$. U isto vreme, naponi selektovani ovim multiplekserima se vode na ulaz 2-bitnog fleš A/D konvertora sa diferencijalnim ulazima iz drugog stepena konverzije, predstavljajući granice njegovog ulaznog opsega. Sada se prekidači, kontrolisani sinhronizacionim signalom S3, zatvaraju i time zamenjuju prvu mrežu otpornika drugom mrežom, koju čine otpornici međusobno jednakih otpornosti R (jer je drugi stepen konverzije linearan). Pomenuti otpornici služe za podešavanje referentnih napona koji su uniformno raspoređeni unutar opsega definisanog referentnim, tj. "break" naponima određenim u prvom stepenu. Ovi referentni naponi se dovode na ulaze komparatora zajedno sa odmerkom napona $U(T)$. Rezultat drugog stepena konverzije predstavljaju dva bita manje težine koja se u registar smeštaju pomoću sinhronizacionog signala S4. Zajedno sa bitovima određenim u prvom stepenu konverzije, poslednja dva bita čine finalni digitalni izlaz koji ispoljava linearnu zavisnost od merene temperature.

III. NUMERIČKI REZULTATI DOBIJENI SIMULACIJOM MERNOG SISTEMA U LABVIEW SOFTVERU

Za generisanje numeričkih rezultata, koji treba da dokažu efikasnosti kola koje se predlaže za linearizaciju NTC termistora, primenjen je softverski paket LabVIEW. Na Sl. 4. je prikazan prednji panel virtuelnog instrumenta koji simulira rad celog mernog sistema uključujući kolo serijsko-paralelnog razdelnika napona i kolo 12-bitnog dvostepenog deo-po-deo linearnog A/D konvertora. Razmatran je temperaturni opseg od -40°C do 120°C.

Virtuelni instrument sadrži deo za generisanje napona $U(T)$ i njemu inverzne funkcije (deo-po-deo linearna aproksimacija ove funkcije je prenosna funkcija prvog stepena A/D konverzije). Takođe, softverski su proračunati parametri Steinhart-Hart modela termistora A, B i C, na osnovu tri poznate kalibracione tačke [1, 9]. Zatim sledi detekcija segmenta kome pripada trenutna vrednost napona $U(T)$, tj. prvi stepen A/D konverzije u kome se osim segmenta određuju i njegove granice (dva susedna referentna, tj. "break" napona). Ovi naponi predstavljaju granice ulaznog opsega drugog stepena A/D konverzije u kome se određuje uniformna ćelija kojoj pripada odmerak napona (biti manje težine). U konkretnom primeru uzeto je da oba stepena imaju rezoluciju po 6 bita, što znači da je broj neuniformnih segmenata, odnosno uniformnih ćelija 2^6-1=63.



Sl. 4. Prednji panel virtuelnog instrumenta koji simulira rad mernog sistema sa dvostepenim deo-po-deo linearnim A/D konvertorom kompaktnog dizajna

TABELA I

BROJ PRIMENJENIH KOMPARATORA U ZAVISNOSTI OD REZOLUCIJE I DIZAJNA DVOSTEPENOG DEO-PO-DEO LINEARNOG A/D KONVERTORA.

2*N* (bita)	8			12				16				
Dizajn	2x4	4+4	2+6	2x6	6+6	4+8	2+10	2x8	8+8	6+10	4+12	2+14
Broj komparatora	15	30	66	63	126	270	1026	255	510	1086	4110	16386

U realnosti, realizacija A/D konvertora ove arhitekture bila bi jako komplikovana za visoke rezolucije (tj. broj referentnih napona koje je potrebno podesiti postaje veliki), pa ni u ovom radu neće biti analizirane rezolucije veće od 8 bita po stepenu, odnosno 16 bita ukupno. Na desnoj strani prednjeg panela virtuelnog instrumenta u vidu grafika su prikazane prenosna funkcija celog mernog sistema i vrednost apsolutne greške merenja za svaku vrednost temperature (analiza je urađena sa korakom 0.016°C) unutar posmatranog opsega.

U nastavku će biti prikazani rezultati vezani za uticaj rezolucije A/D konvertora na ukupan broj komparatora, a samim tim i na energetsku efikasnost konvertora. Zaključci će biti izvedeni poređenjem predloženog, tj. kompaktnog, i klasičnog dizajna dvostepenog deo-po-deo linearnog A/D konvertora (Sl. 1), istih ukupnih rezolucija. Broj komparatora potreban za realizaciju određenog dizajna dvostepenog deo-po-deo linearnog A/D konvertora, u zavisnosti od rezolucija prvog i drugog stepena konverzije, prikazan je u Tabeli I. Razmatrani su slučajevi sa rezolucijama od 8, 12 i 16 bita. Kolone koje su u Tabeli I označene kao proizvod broja 2 i vrednosti 4, 6 ili 8 bita, odnose se na dizajn dvostepenog deo-po-deo linearnog A/D konvertora u kome jedan fleš A/D konvertor obavlja oba stepena konverzije. Ostale kolone se odnose na klasičan dizajn dvostepenog deo-po-deo linearnog A/D konvertora koji ima dva posebna fleš A/D konvertora istih ili različitih rezolucija. Posmatrajući rezultate koji su dati u Tabeli I može se zaključiti da kompaktan dizajn dvostepenog deo-po-deo linearnog A/D konvertora (2x4, 2x6 i 2x8) podrazumeva upotrebu značajno manjeg broja komparatora u poređenju sa dvostepenim deo-po-deo linearnim A/D konvertorom iste rezolucije realizovanog pomoću dva posebna fleš A/D konvertora. Nedostatak dizajna koji podrazumeva upotrebu dva posebna fleš A/D konvertora je taj što sa povećanjem rezolucije jednog fleš A/D konvertora, tako da ukupna rezolucija ostane ista (rezolucija drugog A/D konvertora je smanjena za istu vrednost), dolazi do velikog porasta u potrebnom broju komparatora (kolone 8+8, 6+10, 4+12, 2+14). Na ovaj način pokazali smo da kompaktan dizajn dvostepenog deo-po-deo linearnog A/D konvertora garantuje zauzeće manjeg prostora na integrisanoj pločici i pruža veću ekonomičnost u pogledu potrošnje energije u poređenju sa dvostepenim deo-po-deo linearnim A/D konvertorom klasičnog dizajna.

Međutim, značajnija prednost kompaktnog dizajna linearizacionog kola ogleda se u povećanju tačnosti merenja temperature. Kako bi se procenio uticaj A/D konvertora kompaktnog dizajna na tačnost merenja temperature, proračunate su vrednosti apsolutne greške merenja i nelinearnosti, primenom sledećih jednačina:

$$\Delta T(^{\circ}\mathrm{C}) = \left|T_{iz} - T_{ul}\right|, \tag{1}$$

$$\delta T(\%) = \frac{\Delta T_{\max}(^{\circ}\mathrm{C})}{\textit{širina opsega}(^{\circ}\mathrm{C})} \cdot 100\% . \tag{2}$$

TABELA II

NELINEARNOST TERMISTORA [%] NTSD0XV103FE1B0 PROIZVOĐAČA MURATA, U ZAVISNOSTI OD REZOLUCIJE DVOSTEPENOG DEO-PO-DEO LINEARNOG A/D KONVERTORA I PRIMENJENOG DIZAJNA.

	Nelinearnost δT(%)						
	Kompaktan dizajn N(bita)	Klasičan dizajn N_1(bita), N_2(bita)		Kompaktan dizajn N(bita)	Klasičan dizajn N_1(bita), N_2(bita)		
Merni opseg	$2N$=12	N_1=2, N_2=10	N_1=4, N_2=8	$2N$=16	N_1=2, N_2=14	N_1=4, N_2=12	N_1=6, N_2=10
–40°C-120°C	0.0343	4.531	0.368	0.0022	4.521	0.358	0.024
–20°C-100°C	0.0255	2.753	0.224	0.0016	2.742	0.214	0.014
0°C-80°C	0.0244	1.303	0.111	0.0015	1.291	0.101	0.007
10°C-70°C	0.0244	0.757	0.069	0.0015	0.745	0.058	0.004
20°C-60°C	0.0244	0.354	0.038	0.0015	0.343	0.027	0.002

U prethodnim izrazima figurišu sledeći parametri: T_{iz} je vrednost koja se dobija na izlazu dvostepenog deo-po-deo linearnog A/D konvertora, odnosno to je izmerena vrednost temperature, T_{ul} je vrednost na ulazu u NTC termistor, odnosno to je stvarna (tačna) vrednost temperature koja se meri, ΔT_{max} je maksimalna apsolutna greška merenja u datom opsegu merene temperature, i *širina opsega* predstavlja širinu trenutno posmatranog temperaturnog opsega. Vrednosti rezidualne nelinearnosti (nakon linearizacije) NTC termistora oznake NTSD0XV103FE1B0, proizvođača Murata, su za različite vrednosti rezolucija A/D konvertora, klasičnog i kompaktnog dizajna, prikazane u Tabeli II.

Razmatrano je više temperaturnih opsega, simetričnih u odnosu na temperaturu od 40°C. Posmatrani su slučajevi kada je ukupna rezolucija dvostepenog A/D konvertora 12, odnosno 16 bita. Iz Tabele II se jasno uočava da primena A/D konvertora kompaktnog dizajna, bez obzira na širinu mernog opsega, uvek rezultira manjom greškom nelinearnosti. Razlog za to je to što je u prvom stepenu konverzije, u kome se vrši linearizacija, velika rezoluciju (6, odnosno 8 bita), što kod dvostepenog A/D konvertora klasičnog dizajna nije uvek slučaj. Ako se zadržimo na najširi merni opseg, videćemo da se najmanja greška nelinearnosti dobija u slučaju primene 16-bitnog dvostepenog A/D konvertora kompaktnog dizajna (upotrebljeno 255 komparatora). Ova greška (0.0022%) je 10 puta manja nego u slučaju primene dvostepenog A/D konvertora klasičnog dizajna (0.024%), koji u prvom stepenu konverzije ima 6 bita i istu ukupnu rezoluciju (upotrebljeno 1086 komparatora). Ovaj rezultat je postignut zahvaljujući tome što je rezolucija prvog stepena kola kompaktnog dizajna (8 bita) veća od rezolucije prvog stepena kola klasičnog dizajna (6 bita), iako su ukupne rezolucije u oba slučaja iste (16 bita). Približno isti rezultat u okviru najšireg mernog opsega se može dobiti primenom ili 12-bitnog dvostepenog A/D konvertora kompaktnog dizajna ili 16-bitnim dvostepenim A/D konvertorom klasičnog dizajna. Dakle, u zavisnosti od željene tačnosti merenja i maksimalne rezolucije, koja je najčešće ograničena raspoloživim prostorom na integrisanoj pločici i dozvoljenom potrošnjom energije, moguće je naći kompromisno rešenje.

IV. ZAKLJUČAK

U ovom radu primenjen je kompaktan dizajn dvostepenog deo-po-deo linearnog A/D konvertora za linearizaciju NTC termistora oznake NTSD0XV103FE1B0, proizvođača Murata, na opsegu od -40°C do 120°C. NTC termistor je deo serijsko-paralelnog razdelnika napona koji na svom izlazu daje nelinearni naponski signal koji se dalje linearizuje u dvostepenom deo-po-deo linearnom A/D konvertoru. Kompaktnost ovog kola je postignuta na taj način što oba stepena konverzije obavlja jedan isti fleš A/D konvertor sa dve lestvičaste mreže otpornika: u prvom stepenu mrežu čine otpornici međusobno različitih otpornosti, dok u drugom mrežu čine otpornici međusobno jednakih otpornosti. Nakon završetka prvog stepena konverzije vrši se prebacivanje sa jedne na drugu lestvičastu mrežu otpornika. Primenom predloženog dizajna dvostepenog deo-po-deo linearnog A/D konvertora potreban broj komparatora je smanjen, pa je samim tim i potrošnja energije značajno smanjena. Takođe, i rezultati u pogledu povećanja tačnosti merenja nakon primene predloženog dvostepenog deo-po-deo linearnog A/D konvertora dokazuju njegovu efikasnost. Preciznije, greška nelinearnosti iznosi 0.0022% kada je upotrebljen dvostepeni deo-po-deo linearni A/D konvertor rezolucije 16 bita i kompaktnog dizajna (2x8), dok u slučaju primene dvostepenog deo-po-deo linearnog A/D konvertora klasičnog dizajna i iste rezolucije (6+10) nelinearnost iznosi 0.024 %, tj. 10 puta je veća. Navedeni rezultati pokazuju da se povećanje tačnosti merenja temperature, kompenzacijom nelinearnosti NTC termistora, može postići jednim energetski efikasnim, i po ceni realizacije ekonomičnim kolom.

ZAHVALNICA

Ovaj rad je podržan od strane Ministarstva prosvete, nauke i tehnološkog razvoja Republike Srbije (evidencioni broj projekta je TR 32045).

LITERATURA

[1] J. G. Webster, *The Measurement, Instrumentation and Sensors Handbook*, Boca Raton, USA: CRC Press LLC, 1999.

[2] S. B. Stankovic, P. A. Kyriacou, "Comparison of thermistor linearization techniques for accurate temperature measurement in phase change materials," *J. Phys. Conf. Ser.*, vol. 307, no. 1, pp. 1–6, 2011.
[3] A. Kumar, M. L. Singlab, A. Kumarb, J. K. Rajputc, "POMANI-Mn3O4 based thin film NTC thermistor and its linearization for overheating protection sensor," *Mater. Chem. Phys.*, vol. 156, no. 2015, pp. 150–162, 2015.
[4] A. R. Sarkar, D. Dey, S. Munshi, "Linearization of NTC thermistor characteristic using op-amp based inverting amplifier*," IEEE Sensors J.*, vol. 13, no. 12, pp. 4621-4626, 2013.
[5] L. E. Bengtsson, „Lookup table optimization for sensor linearization in small embedded systems," *Journal of Sensor Technology*, vol. 2012, no. 2, pp. 177-184, 2012.
[6] G. Bucci, M. Faccio, C. Landi, "New ADC with piecewise linear characteristic: case study-implementation of a smart humidity sensor," *IEEE Trans. Instrum. Meas.*, vol. 49, no. 6, pp. 1154-1166, 2000.
[7] A. J. Lopez-Martin, M. Zuza, A. Carlosena, "A CMOS A/D converter with piecewise linear characteristic and its application to sensor linearization," *Analog. Integr. Circ. S.*, vol. 36, no. 1, pp. 39-46, 2003.
[8] J. Jovanović, D. Denić, M. Simić, "Jedno rešenje problema nelinearnosti NTC termistora", *Proc. of 62. Conference ETRAN 2018*, Palić, Srbija, vol. 1, pp. 269-274, 11–14. jun, 2018.
[9] J. S. Steinhart, S. R. Hart, "Calibration curves for thermistors," Deep-Sea Res., vol. 15, no. 4, pp. 497–503, 1968.
[10] NTSD0XV103FE1B0 Temperature Sensor Lead Insulation Type. Datasheet for Murata. *https://www.jameco.com/Jameco/Products/ProdDS/1870999.pdf.*
[11] J. Fraden, *Handbook of Modern Sensors: Physics, Designs, and Applications*, New York, USA: Springer Science+Business Media, 2010.
[12] J. Lukić, D. Živanović, D. Denić, "A compact and cost-effective linearization circuit used for angular position sensors," *FU Aut. Cont. Rob.*, vol. 14, no. 2, pp. 123-134., 2015.
[13] R. Pallas-Areny, J. G. Webster, *Sensors and Signal Conditioning*, 2nd ed., New York, USA: John Wiley & Sons, 2001.

ABSTRACT

In this paper a compact design of a two-stage piecewise linear A/D converter is presented and applied for the NTC thermistor linearization. In precise, this A/D converter compensates nonlinearity of the output voltage of a serial-parallel voltage divider that is containing the NTC thermistor. The compactness of this converter refers to the implementation of a smaller number of comparators, resulting in reduction of the circuit dimensions and power consumption. Compactness is achieved by the fact that both conversion stages are performed by the same flash A/D converter with two different resistor ladder networks, each used for one conversion stage. Thus, this compact two-stage A/D converter has the same resolution in both conversion stages. The nonlinearity compensation of the previously mentioned output voltage is performed in the first conversion stage. The first conversion stage has transfer function that is piecewise linear approximation of the function inverse to dependence of the output voltage from the temperature. After employing the 16-bit two-stage A/D converter of a compact design for the linearization of the NTC thermistor NTSD0XV103FE1B0, manufactured by Murata, the achieved nonlinearity was 0.0022% in the range from -40°C to 120°C.

NTC thermistor linearization with a two-stage piecewise linear A/D converter of a compact design

Jelena Jovanović, Dragan Denić

Jednostavan i efikasan način za generisanje dva ansambala slučajnih brojeva uniformne raspodele sa definisanim koeficijentom korelacije

Đorđe Novaković, Dragan Pejić, Tatjana Grbić, Stefan Mirković, Marina Bulat, Nemanja Gazivoda

***Apstrakt*— U radu je dat prikaz jednostavne i efikasne metode za generisanje dva ansambla slučajnih brojeva uniformne raspodele sa zadatom vrednošću koeficijenta korelacije. Potreba za optimizacijom ovog problema se javlja u okviru primene Monte Karlo metode za određivanje merne nesigurnosti, u situacijama kada su uticajne veličine međusobno korelisane. S obzirom na vrlo veliki broj ponavljanja koja se zahtevaju radi što bolje ocene merne nesigurnosti primennom Monte Karlo metode, jasno je da svaka optimizacija doprinosi uštedi računarskih resursa i smanjenju trajanja simulacija.**

***Ključne reči*— Monte Karlo metoda, korelacija, slučajni brojevi, merna nesigurnost**

I. Uvod

Odavno je prihvaćeno da rezultat merenja nije broj, nego interval u kojem se nalazi stvarna vrednost merene veličine. Za određivanje ovog intervala je ranije korišćena teorija grešaka koja je davala mogućnost da se odrede sigurne granice greške (interval u kojem se sigurno nalazi stvarna vrednost), proceni sistematska greška, odredi nivo slučajnih grešaka, itd... U [1] (*Guide to the expression of Uncertainty in Measurement* - GUM) je definisan način iskazivanja intervala u kojem se nalazi stvarna vrednost merene veličine, uz odgovarajući nivo pouzdanosti (verovatnoću), uzimajući u obzir funkciju gustine raspodele uticajnih veličina. Opisani postupak je zasnovan na analizi propagacije mernih nesigurnosti uticajnih veličina.

Neka je veličina y određena formulom

$$y = f\left(x_1, x_2, ..., x_m\right). \tag{1}$$

Očigledno je da y zavisi od m veličina. U slučaju metrologije, y je veličina čiju mernu nesigurnost želimo da odredimo na osnovu mernih nesigurnosti izmerenih veličina ili mernih nesigurnosti poznatih veličina x_1, x_2,..., x_m.

Ako su uticajne veličine međusobno statistički nekorelisane, onda se merna nesigurnost u_y, određuje upotrebom izraza

$$u_y = \sqrt{\sum_{i=1}^{m}\left[\frac{\partial f}{\partial x_i}u_{x_i}\right]^2} \tag{2}$$

Parcijalni izvod $\partial f/\partial x_i$ i u_{x_i} predstavljaju koeficijent osetljivosti i mernu nesigurnost veličine x_i, respektivno.

U slučajevima kada su uticajne veličine međusobno korelisane, umesto (2) je neophodno koristiti

$$u_y = \sqrt{\sum_{i=1}^{m}\left[\frac{\partial f}{\partial x_i}u_{x_i}\right]^2 + 2\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}\frac{\partial f}{\partial x_i}\frac{\partial f}{\partial x_j}u_{x_i}\cdot u_{x_j}\cdot r_{ij}} \tag{3}$$

gde oznaka r_{ij} predstavlja koeficijent korelacije između veličina x_i i x_j.

Pod pretpostavkom da indirektno merena veličina y ima normalnu raspodelu, definiše se proširena merna nesigurnost, u oznaci U_y, za faktor obuhvata 2 koja predstavlja opseg u kojem se nalazi stvarna vrednost y sa verovatnoćom od 95 %, tj.

$$U_y = 2\cdot u_y. \tag{4}$$

Da bismo odredili mernu nesigurnost primenom (3) treba da odredimo sve koeficijente osetljivosti, što kod komplikovanijih zavisnosti definisanih izrazom (1) zna biti matematički zahtevno. Takođe, potrebno je da znamo ili da procenimo koeficijente korelisanosti r_{ij}. U praksi se često zanemaruje činjenica da su neke veličine međusobno korelisane pa se do vrednosti merne nesigurnosti dolazi primenom (2).

Drugi način za određivanje merne nesigurnosti je korišćenjem Monte Karlo metode (Monte Carlo Method MCM) [2]. Ulazni parametri za MKM su funkcije gustina raspodela uticajnih veličina. Vrši se veliki broj ponavljanja (10^4 do 10^6 ili više), gde se u svakom ponavljanju na slučaj, a u skladu sa definisanom gustinom raspodele, generišu veličine x_i, pa se onda na osnovu (1) određuje odgovarajuća vrednost y. Ako želimo da odredimo proširenu mernu nesigurnost za nivo pouzdanosti 95 %, potrebno je:

a) sortirati sve dobijene rezultate veličine y

b) odbaciti najmanjih i najvećih 2.5 % populacije

c) odrediti polovinu razlike najveće i najmanje preostale vrednosti populacije.

U slučaju postojanja međusobne korelacije između pojedinih uticajnih veličina, neophodno je obezbediti slučajne veličine ne samo zahtevane gustine raspodele, nego

Đorđe Novaković – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail: djordjenovakovic@ uns.ac.rs)

Dragan Pejić – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail: pejicdra@uns.ac.rs)

Tatjana Grbić – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail: tatjana@uns.ac.rs)

Stefan Mirković – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail: mirkovicst@uns.ac.rs)

Marina Bulat – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail: marina.bulat@uns.ac.rs)

Nemanja Gazivoda – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail: nemanjagazivoda@uns.ac.rs).

takve da odgovarajući koeficijenti korelacije budu jednaki procenjenim vrednostima r_{ij} u (3).

Često se vrši poređenje merne nesigurnosti određene po oba opisana postupka. Kod MKM nema nikakvih pretpostavki o obliku raspodele rezultantne veličine y i nije potrebno određivati, ponekad vrlo komplikovane, parcijalne izvode. S druge strane, za korektnu primenu MKM je neophodno obezbediti pretpostavljene raspodele i koeficijente korelacije uticajnih veličina. U nedostatku podataka oko oblika raspodele uticajne veličine, najčešće se pretpostalja uniformna raspodela. Uz ovu pretpostavku se dobijaja konzervativnija procena merne nesigurnosti, nego za većinu drugih raspodela.

II. POSTAVKA PROBLEMA

Za dve raspodele, X i Y, koeficijent korelacije, u oznaci ρ_{XY}, je definisan izrazom

$$\rho_{XY} = \frac{E(X \cdot Y) - E(X) \cdot E(Y)}{\sigma_X \cdot \sigma_Y}. \tag{5}$$

Za nekorelisane raspodele, brojilac u (5) će biti jednak nuli, te je samim tim njihov koeficijent korelacije jednak 0. Koeficijent korelacije između dve slučajne promenljive zapravo je kovarijansa između njihovih standardizovanih (normalizovanih) slučajnih promenljivih i uzima vrednosti od -1 do 1.

Neka su X i Y dve međusobno nekorelisane slučajne promenljive jedinične normalne raspodele (nulte srednje vrednosti i jedinične standardne devijacije) - $N(0,1)$. Pomoću X i Y formiramo veličinu Z

$$Z = k \cdot X + \sqrt{1-k^2} \cdot Y, \quad |k| < 1 \tag{6}$$

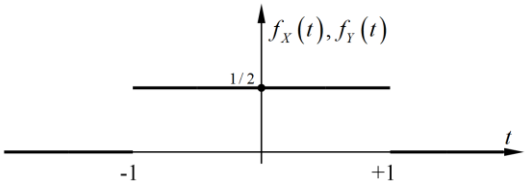
Pokazuje se da važi:

a) slučajna veličina Z ima normalnu $N(0,1)$ raspodelu,

b) koeficijent korelacije veličina X i Z će biti jednak k [3].

Autori su pokušali da primene istu metodologiju na dve veličine uniformne raspodele $U(-1,1)$.

Na slici 1. su prikazane gustine raspodele slučajnih veličina X i Y uniformne raspodele $U(-1,1)$.

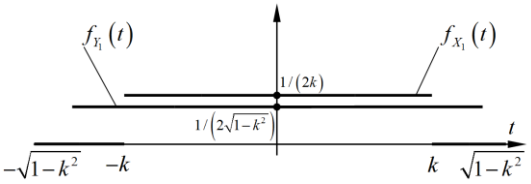


Sl. 1. Gustine raspodele veličina X i Y

Formiramo pomoćne slučajne veličine X_1 i Y_1:

$$X_1 = k \cdot X, \quad Y_1 = \sqrt{1-k^2} \cdot Y, \quad |k| < 1 \tag{7}$$

koje imaju takođe uniformne raspodele, X_1 na intervalu $(-k, k)$ i Y_1 na intervalu $\left(-\sqrt{1-k^2}, +\sqrt{1-k^2}\right)$ i njihove funkcije gustina raspodele su prikazane na slici 2.

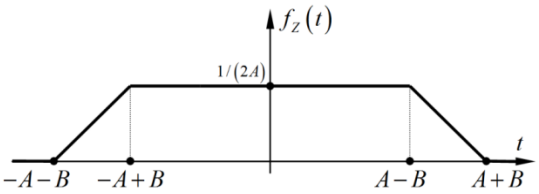


Sl. 2. Gustine raspodele veličina X_1 i Y_1

Slučajna veličina Z

$$Z = X_1 + Y_1 \tag{8}$$

ima funkciju gustine raspodele koja je prikazana na slici 3.



Sl. 3. Gustine raspodele veličine Z

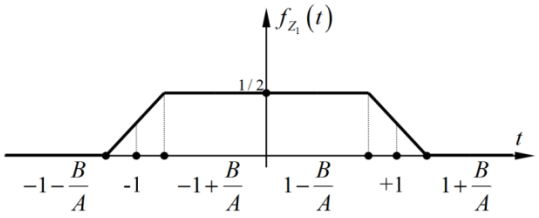
Kao u primeru sa normalnim raspodelama, dobija se da koeficijent korelacije veličine X i Z teži željenoj vrednosti k. S druge strane, kod uniformnih raspodela se dobija da rezultujuća veličina Z više nema istu gustinu raspodele kao polazne veličine X i Y. Ovde se dobija trapezna funkcija gustine raspodele. U specijalnom slučaju, kada je $k = \sqrt{1-k^2} = \sqrt{2}/2$, dobija se trougaona raspodela. Veličine A i B su određene izrazom:

$$A = \max\left(k, \sqrt{1-k^2}\right), B = \min\left(k, \sqrt{1-k^2}\right). \tag{9}$$

Za veličinu Z_1 definisanu formulom (10), prikazana je gustina raspodele na slici 4.

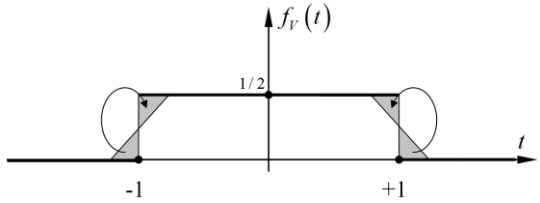
$$Z_1 = Z/A \tag{10}$$

Primetno je da se sredine bočnih strana trapezne raspodele dobijaju pri vrednostima $t = -1$, odnosno $t = +1$.



Sl. 4. Gustina raspodele veličine Z_1

Na osnovu navedenog, došli smo do ideje da polazeći od slučajne veličine Z_1, primenom geometrijskih transformacija, dobijemo slučajnu veličinu V, koja ima uniformnu $U(-1,1)$ raspodelu.



Sl. 5. Gustine raspodele veličine V

Realizacija transformacije prikazane na slici 5 se jednostavno izvodi primenom provere (11).

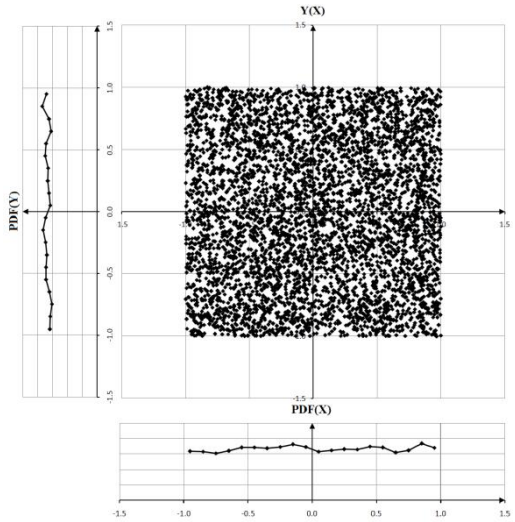
$$\begin{array}{l} \text{if } Z_1 < -1 \text{ then } V = -2 - Z_1 \\ \text{if } Z_1 > +1 \text{ then } V = +2 - Z_1 \end{array} \tag{11}$$

III. SIMULACIJE

Teorijski zaključci su potom proveravani simulacionim putem.

Dat je pseudokod Python programa koji generiše niz Z na osnovu polaznih nizova X i Y.

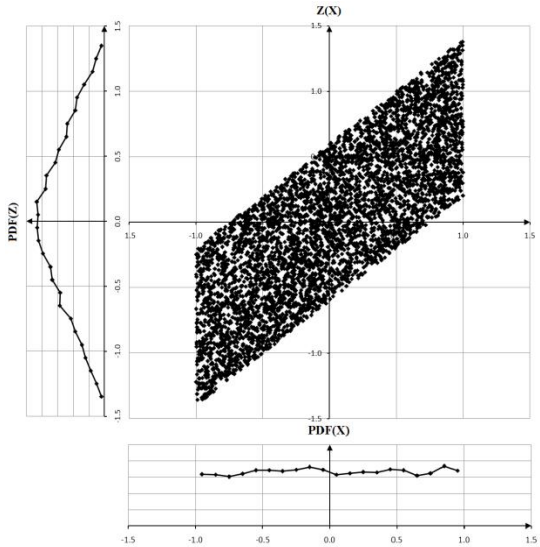
```
set k
set l to sqrt(1-k^2)
get vector_X and vector_Y from numpy.random.uniform function
set vector_V to empty array
loop through each elements in vector_X and vector_Y
   set temp_value to (k*vector_X + l*vector_Y)/max(k,l)
   if temp_value less than -1
      append (-2-temp_value) to vector_V
   else if temp_value bigger than l
      append (2-temp_value) to vector_V
   else
      append temp_value to vector_V
   end if
end loop
```



Sl. 6. *Scatter* prikaz zavisnosti *Y(X)*, prikazi histograma veličina *X* i *Y*

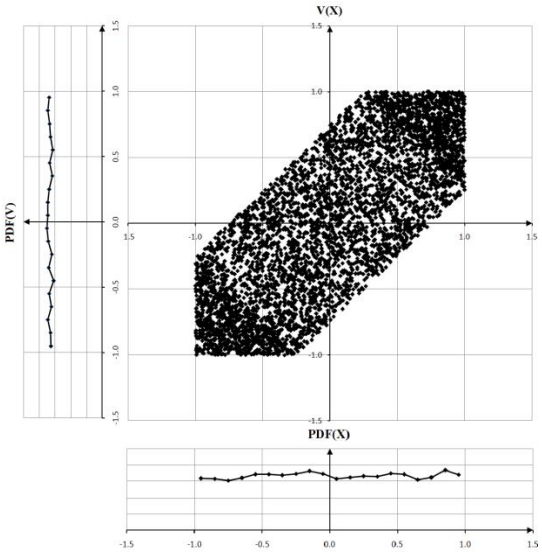
Na slikama 6, 7 i 8 su prikazane scatter zavisnosti *Y(X)*, *Z(X)* i *V(X)*, za 10000 simuliranih vrednosti, pri željenom koeficijentu korelacije 0.8, kao i dobijeni histogrami odgovarajućih veličina *X*, *Y*, *Z* i *V*.

Na slici 6 vidimo da su približno ostvarene uniformne raspodele za obe polazne veličine *X* i *Y*, a da je zbog njihove međusobne nekorelisanosti dobijeno ravnomerno pokrivanje u okviru kvadrata na grafiku *Y(X)*.



Sl. 7. Scatter prikaz zavisnosti *Z(X)*, prikazi histograma veličina *X* i *Z*

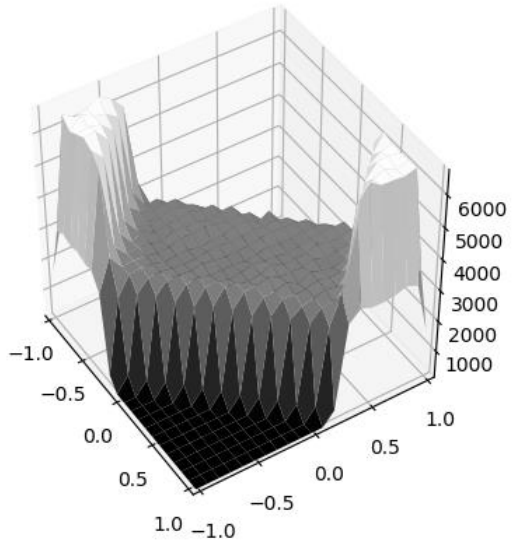
Veličina *Z*, određena na osnovu (7) ima trapeznu raspodelu, slika 7. Zavisnost *Z(X)* ima ravnomerno pokrivanje prostora, ali ne više u okviru kvadrata, nego unutar prikazanog šestougla.



Sl. 8. Scatter prikaz zavisnosti *V(X)*, prikazi histograma veličina *X* i *V*

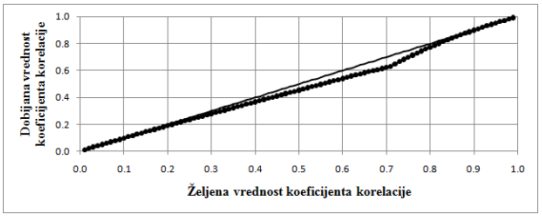
Veličina *V*, određena na osnovu (10), ima uniformu raspodelu. Na grafiku *V(X)* se uočava deo koji ima duplo veće prekrivanje, koje nastaje zbog "presavijanja" vrednosti koje su van opsega (-1,+1).

Na slici 9 je prikazan 3D histogram dobijen za milion simuliranih vrednosti *X* i *V*.



Sl. 9. 3D histogram veličina *X* i *V*

Na kraju je izvršena provera dobijenog koeficijenta korelacije ρ između *X* i *V*. Na slici 10 je prikazana zavisnost dobijenog koeficijenta korelacije ρ od željenog koeficijenta korelacije *k*, kao i prava $\rho=k$ na kojoj bi trebale da leže tačke u slučaju poklapanja ovih dveju vrednosti. Za različite vrednosti željenog koeficijenta korelacije *k* od 0 do 1, sa korakom 0.01, vršeno je ponavljanje simulacije sto puta. Na grafiku je prikazana srednja vrednost dobijenog koeficijenta korelacije ρ na osnovu vektora sa milion elemenata.



Sl. 10. Zavisnost dobijenog koeficijenta korelacije od željene vrednosti

Na slici 11 je prikazana zavisnost razlike dobijenog i željenog koeficijenta korelacije $\delta=\rho-k$. Vertikalnim stubićima je ilustrovan opseg u kojem se rasipaju rezultati.

Sl. 11. Zavisnost odstupanja dobijenog i željenog koeficijenta korelacije

Gledano očima metrologa, vidimo prisustvo slučajne i sistematske greške.

i) Slučajna greška je posledica konačnog broja elemenata i kvaliteta generatora slučajnih brojeva. Simulacije pokazuju da ova vrednost teži nuli sa povećanjem broja elemenata. Simulacije su ponavljane za tri različita generatora brojeva uniformne raspodele: Mersenne Twister, LFSR (Linear Feedback Shift Register) i kongruentnog generatora [4] [5] i nije uočena razlika u zavisnosti δ od željenog koeficijenta korelacije.

ii) Sistematska greška potiče od transformacije opisane izrazom (11). Eliminacija sistematske greške je urađena korekcijom. Neka je zavisnost dobijenog koeficijenta korelacije ρ od željenog koeficijenta korelacije k (prikazana na slici 10) data funkcijom $\rho=g(k)$. Ova funkcija nam daje vrednost koeficijenta korelacije koji ćemo dobiti ako algoritmu za generisanje dva ansambla predamo vrednost k. Inverzna funkcija nam daje vrednost koeficijenta korelacije koji treba predati algoritmu da bismo dobili baš onu vrednost koju želimo. Stoga, kada želimo da dobijemo vrednost koeficijenta korelacije ρ, algoritmu ćemo kao ulaznu veličinu proslediti $g^{-1}(\rho)$.

IV. Standardan pristup

Veličina Z, dobijena po (8) ima funkciju gustine raspodele prikazanu na slici 3. Funkcija raspodele verovatnoće veličine Z ima oblik dat u (12). Konstante navedene u (12) su pozitivne i zavise od A i B.

$$F_Z(t)=\begin{cases} 0, & t<-A-B \\ k_1\cdot t^2+k_2\cdot t+k_3, & -A-B<t\le -A+B \\ 0.5(t+1), & -A+B<t\le A-B \\ 1-k_1\cdot t^2+k_2\cdot t-k_3, & A-B<t\le A+B \\ 1, & A+B<t \end{cases} \quad (12)$$

Standardan "školski" pristup na ovom mestu bi bio primena integralne transformacije (*Probability Integral Transform* - PIT). Ovaj mehanizam se koristi za transformaciju kontinualne raspodele u uniformnu raspodelu U(0,1). Veličina K definisana izrazom (13) ima uniformnu raspodelu U(0,1).

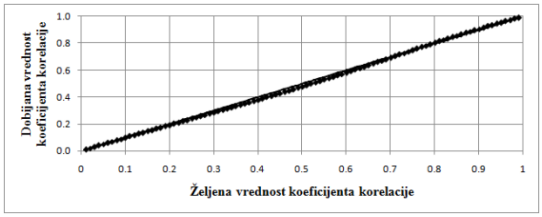
$$K=F_Z(Z), \quad K:U(0,1) \quad (13)$$

Na kraju, potrebno je veličinu K, transformisati u veličinu L primenom (14) da bi se dobila očekivana uniformna raspodela U(-1,+1).

$$L=2\cdot K-1, \quad L:U(-1,+1) \quad (14)$$

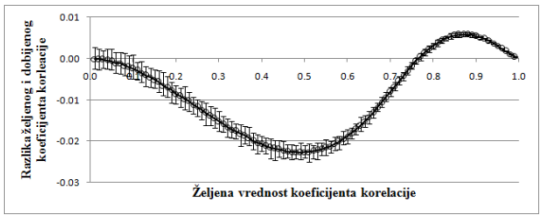
Simulacionim putem je ponovljena provera rezultata, ali ovoga puta za polaznu veličinu X i primenom PIT dobijenu veličinu L.

Na slici 12 je prikazana zavisnost dobijenog od željenog koeficijenta korelacije primenom PIT pristupa. U poređenju sa slikom 10, gde je prikazana ista zavisnost, ali za predloženu metodu, uočava se da je kod standardnog pristupa odstupanje manje.



Sl. 12. Zavisnost dobijenog koeficijenta korelacije od željene vrednosti, kada se koristi PIT

Na slici 13 prikazana zavisnost razlike dobijenog i željenog koeficijenta korelacije za standardnu PIT metodu. Sada je lakše poređenje odstupanja po predloženoj metodi (slika 11) i po standardnoj metodi (slika 13).



Sl. 13. Zavisnost odstupanja dobijenog i željenog koeficijenta korelacije, kada se koristi PIT

V. Diskusija

Prikazan je predlog nove metode za dobijanje dva ansambla uniformne raspodele i zadatog koeficijenta korelacije. Polazna osnova je postupak koji ovaj problem rešava kada je reč o jediničnim normalnim raspodelama. Prikazana je i standardna metoda za rešavanje istog problema, primenom integralne transformacije. U oba slučaja je ustanovljeno odstupanje između dobijenog i željenog koeficijenta korelacije. Kod standardne metode je ovo odstupanje oko četiri puta manje nego kod predložene metode. Vršena je provera ponašanja odstupanja dobijene i zadate vrednosti koeficijenta korelacije u zavisnosti od vrste generatora uniformnih brojeva i nije uočena nikakva zavisnost. Autori su zaključili da je odstupanje posledica transformacije trapezne raspodele uniformnom raspodelom.

Za obe metode je uočeno da je odstupanje dobijenog od željenog koeficijenta korelacije sistematska pojava. Čak i ako ne razumemo u potpunosti mehanizam delovanja sistematskog uzroka, u oba slučaja smo u stanju da otklonimo njegov uticaj. Simulacionim putem je ustanovljena zavisnost željene vrednosti ρ od zadate vrednosti k koeficijenta korelacije $\rho=g(k)$. Primenom inverzne funkcije g^{-1} je moguće izvršiti korekciju i u potpunosti otkloniti sistematsku grešku.

Po otklanjanju sistematske greške, postavlja se pitanje, koji postupak je efikasniji. Pod pojmom efikasnosti najviše se pažnja posvećuje vremenskoj efikasnosti. Primena ovog

algoritma u proračunu MKM zahteva veliki broj odbiraka čime se vreme izvršavanja MKM znatno uvećava. Ovaj algoritam bi trebao znatno ubrzati postojeće algoritme zbog svoje jednostavnosti.

VI. ZAKLJUČAK

U radu je prikazana metoda za generisanje dva ansambla uniformne raspodele sa zadatom vrednošću koeficijenta korelacije. Predloženi metod je poređen sa standardnim postupkom koji je zasnovan na primeni integralne transformacije. Standardna metoda ima četiri puta manje odstupanje dobijenog od željenog koeficijenta korelacije. Ukoliko se želi eliminisati uočeno odstupanje, u oba slučaja je neophodno vršiti korekciju, koja u potpunosti otklanja sistematsku komponentu odstupanja. Pokazano je da je predložena metoda efikasnija u smislu manjeg utroška računarskog vremena. Kod predložene metode je potrebno vršiti svega dve provere i po jedno sabiranje, dok je kod standardne metode, pored dve provere, potrebno izračunavanje transformacije koja je u dva slučaja kvadratnog tipa, a u jednom linearnog tipa. Kvadratna transformacija podrazumeva više operacija: četiri množenja i tri sabiranja, dok linearna transformacija podrazumeva po jedno sabiranje i množenje.

ZAHVALNICA

Ovaj rad je podržan od strane Ministarstva prosvete, nauke i tehnološkog razvoja Republike Srbije u sklopu projekta TR-32019, OI-174009, TR-32035.

[1] Evaluation of measurement data - Guide to the expression of uncertainty in measurement, JCGM 100:2008, GUM 1995 with minor corrections, 2008
[2] Evaluation of measurement data - Supplement 1 to the “Guide to the expression of uncertainty in measurement”-Propagation of distributions using a Monte Carlo method JCGM101:2008, 200
[3] A. Papoulis, "Probability, Random Variables, and Stochastic Processes", Mc Graw-Hill Kogakusha, Ltd., Tokyo, 1965.
[4] D. Knuth, "The Art of Computer Programming Volume 2", Seminumerical algorithms, pp. 9-38, Addison-Wesley, Massachusetts, 1998
[5] https://docs.python.org/2/library/random.html

ABSTRACT

The paper presents a simple and efficient method for generating two ensembles of random numbers of uniform distribution with the given value of the correlation coefficient. The need for optimization of this problem occurs in the application of the Monte Carlo method for the determination of measurement uncertainty, in situations where the influential quantities are correlated.

Considering the very large number of repetitions required for better estimation of uncertainty measurement by the applicable Monte Carlo method, it is clear that any optimization contributes to saving computer resources and reducing the duration of simulation.

Primena numeričkih metoda integracije na računanje efektivne vrednosti

Marina Bulat, Stefan Mirković, Dragan Pejić, Marjan Urekar, Đorđe Novaković i Nemanja Gazivoda

***Apstrakt*— Ovaj rad se bavi istraživanjem kvaliteta izračunavanja efektivne vrednosti primenom nekoliko numeričkih metoda u uslovima necelobrojnog odnosa učestanosti odmeravanja i učestanosti signala. Posmatrana su dva nezavisna problema istovremeno. Jedan je kvalitet numeričkog određivanja integrala kako bi se odredila efektivna vrednost signala, dok je drugi problem necelobrojni odnos učestanosti. Diskretizacija po amplitudi nije razmatrana.**

***Ključne reči*—odmeravanje; diskretizacija; efektivna vrednost; srednja vrednost; preciznost; trapezno pravilo; Simpsonovo pravilo.**

I. Uvod

Kod naizmeničnih napona efektivna vrednost se koristi kao osnovni parametar signala. Kvadrat efektivne vrednosti napona je srazmeran snazi koju taj napon razvija na jediničnom otporniku, pa je upravo to razlog zašto je u praksi češći slučaj da se meri efektivna vrednost nego amplituda ili trenutna vrednost napona. Stoga imamo da je

$$\frac{U_{eff}^2}{R}\cdot T=\int_0^T\frac{u^2(t)}{R}dt \qquad (1)$$

na osnovu kog se dobija analitički izraz za efektivnu vrednost napona:

$$U_{eff}=\sqrt{\frac{1}{T}\cdot\int_0^T u^2(t)\,dt}. \qquad (2)$$

Primenom analogne elektronike efektivna vrednost kontinualnog napona je određivana pomoću gore navedenog analitičkog izraza.

Pomoću analogno-digitalne konverzije analogni napon se prevodi u skup diskretnih vrednosti. Tada više nemamo na raspolaganju kontinualni signal u vremenskom domenu nego raspolažemo diskretnim skupom vrednosti. Iz ovog razloga je neophodno prilagoditi izraz za određivanje efektivne vrednosti i približno je odrediti numeričkim putem, na osnovu odbiraka koje dobijamo u procesu diskretizacije po vremenu odnosno odmeravanjem.

Ukoliko u toku jedne periode signal odmerimo u n tačaka, gde je interval integracije podeljen na n jednakih podintervala širine ΔT, uobičajeno je da se približna vrednost datog integrala proceni primenom (3):

$$U_{eff}\approx\sqrt{\frac{1}{T}\sum_{i=1}^{n}u^2(i\Delta T)\Delta T}=\sqrt{\frac{\Delta T}{T}\sum_{i=1}^{n}u^2(i\Delta T)}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}u^2(i\Delta T)} \qquad (3)$$

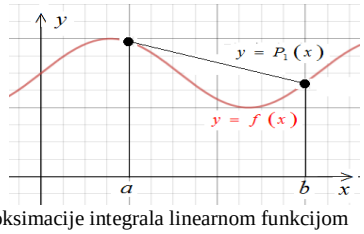
Efektivna vrednost jednaka je korenu iz srednje vrednosti kvadrata napona, pa se problem svodi na određivanje korena iz integrala kvadrata napona podeljenog sa trajanjem u kom se vrši integraljenje.

II. Numerička integracija

Numerička integracija predstavlja postupak kojim se određuje približna vrednost određenog integrala

$$I=\int_a^b f(x)\,dx,\quad a<b \qquad (4)$$

Aproksimacijom podintegralne funkcije $y=f(x)$ interpolacionim polinomom n-tog reda P_n, u zavisnosti od izbora metode.



Sl. 1. Prikaz aproksimacije integrala linearnom funkcijom

Na Sl. 1 vidimo prikaz površine određene integralom I koja je aproksimirana površinom ograničenom odsečkom funkcije $y=P_1(x)$, odsečkom ose Ox i odsečcima vertikalnih pravih u tačkama $x=a$ i $x=b$. Numeričke metode približno određuju vrednost određenog integrala tako što se data oblast podeli na što veći broj elementarnih podoblasti, koje se aproksimiraju različitim figurama, kao što su na primer pravougaonik ili trapez. Zatim se izračunaju vrednosti tih površina, a njihov zbir predstavlja aproksimaciju stvarne vrednosti određenog integrala.

A. Pravilo levih ili desnih pravougaonika

Jedna od najčešće korišćenih integracionih metoda za numeričku integraciju je pravilo levih ili desnih

Marina Bulat – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail: marina.bulat@uns.ac.rs).

Stefan Mirković – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail: mirkovicst@uns.ac.rs).

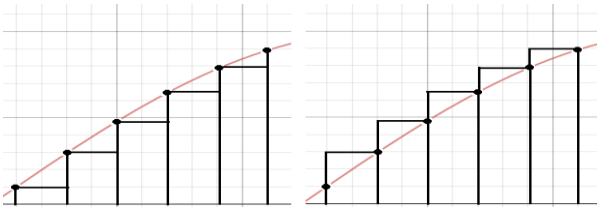
Dragan Pejić – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail: pejicdra@uns.ac.rs)

Marjan Urekar – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail: urekarm@uns.ac.rs).

Đorđe Novaković – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail: djordjenovakovic@uns.ac.rs).

Nemanja Gazivoda – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail:nemanjagazivoda@uns.ac.rs).

pravougaonika. Ove dve metode su grafički predstavljene na Sl. 2.



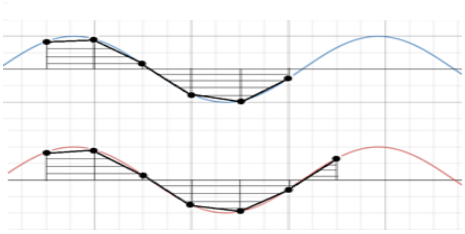
Sl. 2. Ilustracija metoda levih i desnih pravougaonika

Sa Sl. 2 se može uočiti da je uniformna dužina svih podintervala kod obe metode. Integral se aproksimira sumom površina pravougaonika gde se za jednu njegovu stranicu uzima vrednost širine podintervala ΔT, a za vrednost druge stranice pravougaonika vrednost funkcije u levoj odnosno desnoj krajnjoj tački tog podintervala. Prema ilustraciji na levoj polovini Sl. 2 možemo zaključiti da će primenom aproksimacije metodom levih pravougaonika u intervalu, gde je funkcija monotono rastuća, procena integrala biti manja od tačne vrednosti. Za opadajuću funkciju pri aproksimaciji levih pravougaonika dobili bi veću procenu integrala od stvarne vrednosti. Na desnoj polovini Sl. 2 imamo obrnutu sitaciju. Funkcija je monotono rastuća pa ćemo primenom aproksimacije desnih pravougaonika dobiti veću procenu integrala u odnosu na pravu vrednost. Ukoliko bi imali monotono opadajuću funkciju, procena integrala bi bila manja od tačne vrednosti.

1) Metoda 1

Ova metoda je standardna metoda odabiranja pravilom levih ili desnih pravougaonika koja određuje procenu efektivne vrednosti na osnovu n odbiraka dobijenih u okviru celog broja perioda, tako da su periode ograničene prolaskom napona kroz nulu, po jednačini (3).

B. Metoda 2 ili Trapezna metoda



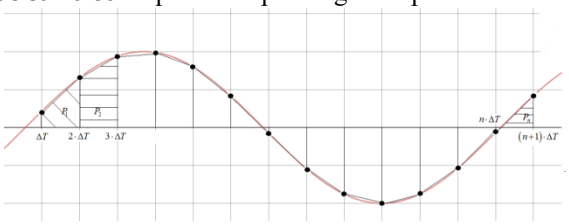
Sl. 3. Ilustracija određivanja perioda napona na osnovu prvih šest, odnosno prvih sedam tačaka

Na Sl. 3 su prikazana dva prostoperiodična napona koja su u fazi, u trajanju jedne i po periode sa istom amplitudom i učestanosti. Napon je odmeravan u šest tačaka u okviru jedne periode koja je definisana uzastopnim prolaskom rastućeg napona kroz nulti nivo. Prvih šest tačaka određuju trajanje periode koje je kraće od stvarnog trajanja periode signala. Uvođenjem sledeće, sedme tačke, dobija se opseg koji je približniji trajanju jedne periode. U zavisnosti od odnosa periode signala i periode semplovanja, perioda ograničena prvim i sedmim odbirkom će nekada biti kraća, a nekada duža od stvarne periode signala. Analize pokazuju da se dobija manja greška određivanja periode (nekada pozitivna, a nekada negativna) na osnovu sedam, nego u slučaju šest tačaka, gde se dobija veća greška uvek istog znaka (pozitivna sistematska greška).

Na Sl. 4 površina P_1 je određena donjom osnovicom ΔT, jednim krakom koji ima vrednost funkcije u trenutku odabiranja ΔT, i drugim krakom čija je vrednost funkcije u sledećem trenutku odabiranja $2\Delta T$, odnosno

$$P_1 = \frac{\Delta T}{2}\left(y(\Delta T) + y(2\Delta T)\right). \qquad (5)$$

Procena vrednosti određenog integrala I na osnovu periode koja je ograničena prvom i poslednjom, $n+1$ tačkom, računa se kao suma od n površina pravouglih trapeza.



Sl.4. Ilustracija trapezne metode

Stoga, dobijamo da je aproksimacija određenog integrala I suma od n površina definisane na datim odbircima:

$$I = \int_a^b f(x)\,dx \approx P = \sum_{i=1}^{n} P_i \qquad (6)$$

$$I \approx \frac{\Delta T}{2}\left(y(\Delta T) + y(2\Delta T)\right) + \frac{\Delta T}{2}\left(y(2\Delta T) + y(3\Delta T)\right) + \frac{\Delta T}{2}\left(y(n\Delta T) + y\left((n+1)\Delta T\right)\right)$$

$$I \approx \Delta T\left(\frac{y(\Delta T) + y\left((n+1)\cdot\Delta T\right)}{2} + \sum_{i=2}^{n} y(i\cdot\Delta T)\right). \qquad (7)$$

Ako uzmemo da je $y(i\Delta T) = y_i$, tada prethodni izraz postaje:

$$I \approx \Delta T\left(\frac{y_1 + y_{n+1}}{2} + \sum_{i=2}^{n} y_i\right). \qquad (8)$$

Na osnovu ovih proračuna pomoću izraza (3), procena efektivne vrednosti signala nad datim odbircima iznosi:

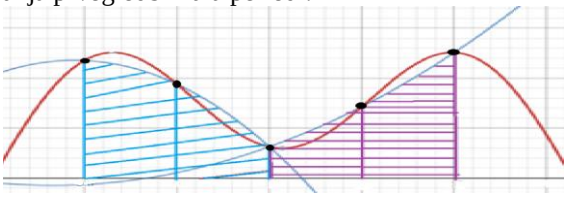
$$U_{eff} \approx \sqrt{\frac{1}{n}\left(\sum_{i=2}^{n} u_i^2 + \frac{u_1^2 + u_{n+1}^2}{2}\right)}. \qquad (9)$$

Za razliku od (3), u kojoj su svi sabirci istog značaja, u poslednjem dobijenom izrazu za efektivnu vrednost možemo uočiti da se prvi i poslednji odbirak uzimaju sa upola manjom težinom.

C. Simpsonovo pravilo

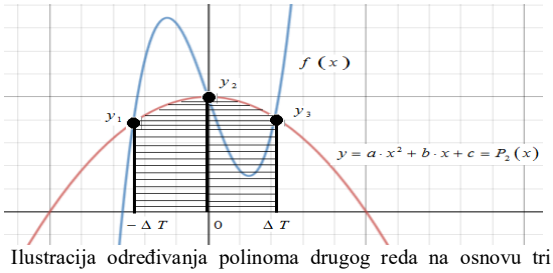
Na Sl. 5 je ilustrovana numerička integracija funkcije integracionim polinomom drugog reda. Tačna vrednost ovog integrala jednaka je površini ispod date funkcije, a približna vrednost površini ispod krive polinoma nad periodom. Simpsonovim pravilom dobija se numerička procena integrala

polinomom drugog stepena tako što se interval integracije podeli na n jednakih podintervala širine $2\Delta T$. Jedan podinterval se dobija na osnovu tri odbirka. Time je upotreba Simpsonovog pravila moguća samo u slučaja kada je broj odbiraka neparan. Ukoliko je broj dobijenih odbiraka paran, tada ova metoda ne može da se primeni. U praksi broj odbiraka dobijenih u periodi je nekad paran, a nekad neparan, što zavisi od periode signala, periode semplovanja i trenutka uzimanja prvog odbirka u periodi.



Sl. 5. Ilustracija Simpsonovog pravila

Neka je broj dobijenih odbiraka neparan u okviru jedne periode. Da bismo mogli da izvršimo procenu površine jednog podintervala, potrebno je da odredimo parabolu na osnovu tri susedne tačke funkcije. Ilustracija ovog postupka je prikazana na Sl. 6.



Sl. 6. Ilustracija određivanja polinoma drugog reda na osnovu tri tačke funkcije

$$I=\int_{-\Delta T}^{\Delta T} f(x)\,dx \approx \int_{-\Delta T}^{\Delta T} P_2(x)\,dx = \int_{-\Delta T}^{\Delta T} \left(ax^2+bx+c\right)dx = \frac{\Delta T}{3}\left(2a\Delta T^2+6c\right) \quad (10)$$

Na osnovu tačaka $(-\Delta T, y_1), (0, y_2)$ i $(\Delta T, y_3)$ dobijamo da je procena integrala nad podintervalom:

$$I \approx \frac{\Delta T}{3}\left(y_1+4y_2+y_3\right). \quad (11)$$

Na osnovu polinoma drugog reda za prve tri tačke funkcije, određuje se površina P_1, a zatim se određuje polinom drugog reda za treću, četvrtu i petu tačku, na osnovu koje se računa površina P_2. Kada se odredi površina $P_{\frac{n-1}{2}}$ na osnovu poslednje tri tačke funkcije, računa se suma svih ovih površina koja predstavlja numeričko rešenje Simpsonovog pravila. Razlika numeričkog rešenja ovog integrala i tačne vrednosti određenog integrala potiče od nepodudaranja funkcije i kvadratne aproksimacije kroz tri tačke date funkcije.

Suma površina kao aproksimacija integrala iznosi:

$$I \approx P = \sum_{i=1}^{\frac{n-1}{2}} P_i = \frac{\Delta T}{3}\left(y_1 + 4\sum_{i=1}^{\frac{n-1}{2}} y_{2\cdot i} + 2\sum_{i=1}^{\frac{n-3}{2}} y_{2\cdot i+1} + y_n\right). \quad (12)$$

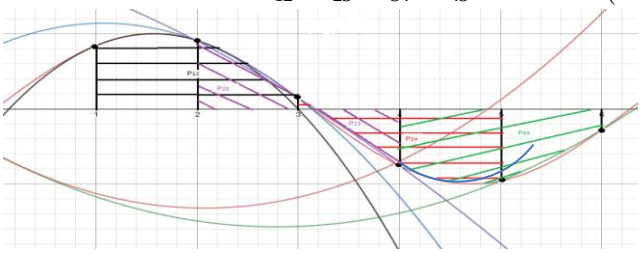
Primenjujući dalje ovaj analitički izraz u izraz za procenu efektivne vrednosti, dobijamo:

$$U_{eff} \approx \sqrt{\frac{1}{n-1}\left(\frac{u_1^2+u_n^2}{3} + \frac{4}{3}\sum_{i=1}^{\frac{n-1}{2}} u_{2\cdot i}^2 + \frac{2}{3}\sum_{i=1}^{\frac{n-3}{2}} u_{2\cdot i+1}^2\right)}. \quad (13)$$

1) Metoda 3 ili Unapređeno Simpsonovo pravilo kvadratnom funkcijom

Na Sl. 7 je prikazana ilustracija metode koja predstavlja modifikaciju Simpsonovog pravila, koju predlažu autori ovog rada. Broj semplova dobijen u okviru periode određene uzastopnim prolaskom rastućeg napona kroz nulu je šest. Jedan podinterval integracije dobija se na osnovu svake tri uzastopne tačke, na osnovu kojih se određuju polinomi drugog reda. Prvo se odredi parabola na osnovu prve, druge i treće tačke i izračuna se površina P_{12} koju ta parabola obrazuje. Zatim se na osnovu druge, treće i četvrte tačke određuje polinom drugog reda i izračuna površina P_{23}. Kada se odredi površina P_{45} na osnovu tri poslednje tačke funkcije, četvrte, pete i šeste, računa se ukupna površina kao zbir svih prethodno izračunatih površina:

$$P = P_{12} + P_{23} + P_{34} + P_{45}. \quad (14)$$



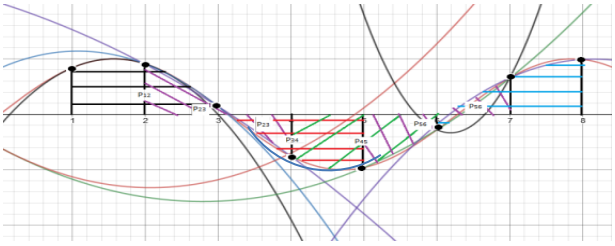
Sl.7. Ilustracija modifikacije Simpsonovog pravila

Neka su P_1 i P_2 površine, koje čine površinu P_{12}, dobijene na osnovu prve i druge odnosno druge i treće tačke i neka su P_2' i P_3 površine koje čine P_{23} dobijene na osnovu druge i treće odnosno treće i četvrte tačke. Površine P_2 i P_2' nisu jednake jer su dobijene pomoću različitih kvadratnih aproksimacija, ali su bliske po vrednostima. Ukoliko u dalji proračun uzmemo da je $P_2 \approx P_2'$, $P_3 \approx P_3'$ i $P_4 \approx P_4'$, dobijamo da aproksimacija integrala I čija je perioda određena na osnovu prve i šeste tačke postaje:

$$I \approx P = P_1 + P_2 + P_2' + P_3 + P_3' + P_4 + P_4' + P_5 \quad (15)$$

$$I \approx P_1 + 2\cdot P_2 + 2\cdot P_3 + 2\cdot P_4 + P_5 \quad (16)$$

U poslednjem izrazu možemo da uočimo da se u okviru periode dobijaju duplo veće površine P_i, osim prve i poslednje, za $i = 2, 3, 4$. Iz ovog razloga je potrebno uzeti u proračun i naredne dve tačke posle ponovnog prolaska rastućeg napona kroz nulu, tačku sedam i osam. Ilustracija ovog postupka je prikazana na Sl. 8.



Sl. 8. Ilustracija modikacije i unapređenja Simpsonovog pravila

Vrednosti funkcija u tačkama sedam i osam su bliske vrednostima funkcija u tačkama jedan i dva, pa možemo reći da su površine određene tačkom šest i sedam i tačkama sedam i osam bliske vrednostima površina P_1 i P_5.

Ukoliko prethodno uzmemo u obzir, izraz za aproksimaciju integrala I postaje:

$$I \approx P_1 + 2\cdot P_2 + 2\cdot P_3 + 2\cdot P_4 + P_5 + P_{56} + P_{67} \quad (17)$$

$$I \approx P_1 + 2\cdot P_2 + 2\cdot P_3 + 2\cdot P_4 + 2\cdot P_5 + 2\cdot P_6 + P_7 \quad (18)$$

$$P_1 \approx P_7 \quad (19)$$

$$I \approx 2\cdot\left(P_1 + P_2 + P_3 + P_4 + P_5 + P_6\right) \quad (20)$$

Na ovaj način dobijamo dva puta veću površinu od one koja nam je potrebna.

Modifikacija ove metode ogleda se u tom što se za podintervale uzimaju svake tri uzastopne tačke i u tome što je perioda ograničena prvim i drugim dodatim semplom posle ponovnog prolaska rastućeg napona kroz nulu. Parnost broja odbiraka dobijenih u okviru periode više ne igra nikakvu ulogu što predstavlja doprinos koju ova metoda daje.

Ukoliko bi za ovako definisanu periodu imali n odbiraka, tada bi površine određene na osnovu svake tri uzastopne tačke iznosile:

$$P_{12} = \frac{\Delta T}{3}\cdot\left(y_1 + 4\cdot y_2 + y_3\right) \quad (21)$$

$$P_{23} = \frac{\Delta T}{3}\cdot\left(y_2 + 4\cdot y_3 + y_4\right) \quad (22)$$

...

$$P_{n-3\,n-2} = \frac{\Delta T}{3}\cdot\left(y_{n-3} + 4\cdot y_{n-2} + y_{n-1}\right) \quad (23)$$

$$P_{n-2\,n-1} = \frac{\Delta T}{3}\cdot\left(y_{n-2} + 4\cdot y_{n-1} + y_n\right). \quad (24)$$

Primenjujući dalje ovaj analitički izraz u izraz za procenu efektivne vrednosti, dobijamo:

$$U_{eff} \approx \sqrt{\frac{1}{(n-2)}\left(\frac{u_1^2 + u_n^2}{6} + \frac{5}{6}\left(u_2^2 + u_{n-1}^2\right) + \sum_{i=3}^{n-2} u_i^2\right)}. \quad (25)$$

2) Metoda 4 ili Unapređeno Simpsonovo pravilo kubnim polinomom

Ova metoda je ekvivalent prethodnoj metodi, sa tom razlikom da su potrebna tri naredna sempla posle ponovnog prolaska rastućeg napona kroz nulu, pored semplova dobijenih u okviru jedne periode. Jedna površina se dobija na osnovu svake četiri uzastopne tačke funkcije kroz koje se aproksimira kubna funkcija. Na taj način dobija se tri puta veća površina od one površine koja je potrebna.

Da bismo mogli da izvršimo procenu površine jednog podintervala, potrebno je da odredimo kubnu funkciju na osnovu četiri tačke funkcije:

$$I = \int_{-\Delta T}^{2\cdot\Delta T} f(x)\,dx \approx \int_{-\Delta T}^{2\cdot\Delta T} P_3(x)\,dx \quad (26)$$

$$I = \int_{-\Delta T}^{\Delta T} \left(a\cdot x^3 + b\cdot x^2 + c\cdot x + d\right) dx \quad (27)$$

$$I = \frac{\Delta T}{4}\cdot\left(15\cdot a\cdot\Delta T^3 + 12\cdot b\cdot\Delta T^2 + 6\cdot c\cdot\Delta T + 12\cdot d\right). \quad (28)$$

Na osnovu tačaka $(-\Delta T, y_1), (0, y_2), (\Delta T, y_3)$ i $(2\cdot\Delta T, y_4)$ dobijamo da je procena integrala nad podintervalom:

$$I \approx \frac{3\Delta T}{8}\left(y_1 + 3y_2 + 3y_3 + y_4\right). \quad (29)$$

Na osnovu polinoma trećeg reda za prve četiri tačke funkcije, određuje se površina P_{123}, a zatim se određuje polinom trećeg reda za drugu, treću, četvrtu i petu tačku, na osnovu koje se računa površina P_{234}. Kada se odredi površina $P_{n-3\,n-2\,n-1}$ na osnovu poslednje četiri tačke funkcije, računa se suma svih ovih površina koja predstavlja numeričko rešenje ove modifikacije Simpsonovog pravila.

$$P_{123} = \frac{3\Delta T}{8}\left(y_1 + 3y_2 + 3y_3 + y_4\right) \quad (30)$$

$$P_{234} = \frac{3\Delta T}{8}\left(y_2 + 3y_3 + 3y_4 + y_5\right) \quad (31)$$

...

$$P_{n-3\,n-2\,n-1} = \frac{3\Delta T}{8}\left(y_{n-3} + 3y_{n-2} + 3y_{n-1} + y_n\right) \quad (32)$$

$$P = \frac{3\Delta T}{8}\left(y_1 + y_n + 4\left(y_2 + y_{n-1}\right) + 7\left(y_3 + y_{n-2}\right) + 8\sum_{i=4}^{n-3} y_i\right). \quad (33)$$

Primenjujući dalje ovaj analitički izraz u izraz za procenu efektivne vrednosti, dobijamo:

$$U_{eff} \approx \sqrt{\frac{1}{(n-3)}\left(\frac{u_1^2 + u_n^2}{8} + \frac{1}{2}\left(u_2^2 + u_{n-1}^2\right) + \frac{7}{8}\left(u_3^2 + u_{n-2}^2\right) + \sum_{i=4}^{n-3} u_i^2\right)}. \quad (34)$$

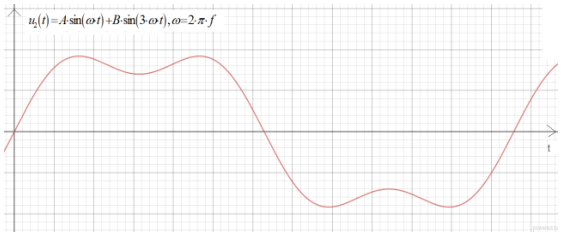
Naredne četiri metode su ekvivalent prethodno definisanim metodama, samo se ovoga puta periode definišu na osnovu prolaska rastuće ivice napona kroz prethodno određenu efektivnu vrednost napona, u čemu je doprinos ovog rada. Autori predlažu da se prvo odredi efektivna vrednost napona na osnovu prolazaka kroz nulu, a onda se nad istim odbircima izvrši računanje nove efektivne vrednosti, ali nad periodom koja je određena prolaskom napona kroz predhodno određenu efektivnu vrednost.

III. REZULTATI SIMULACIJA

Putem simulacija je proveren kvalitet numeričkog određivanja integrala kako bi se odredila efektivna vrednost

signala i uticaj necelobrojnog odnosa učestanosti odmeravanja i učestanosti signala.

Simulirana su dva signala. Jedan napon, u oznaci $u_1(t)$, je prostoperiodičan napon amplitude 1 V i učestanosti 50 Hz, a drugi napon je $u_2(t)$, predstavljen na Sl. 9.

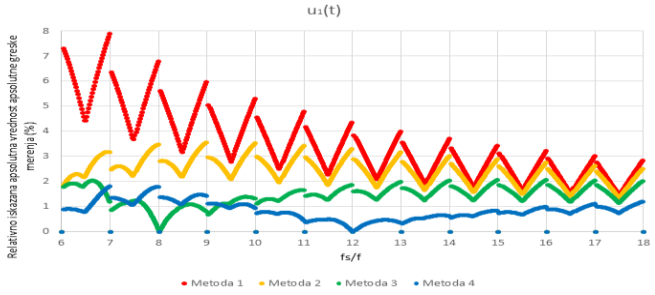


Sl. 9. Ilustracija napona $u_2(t)$

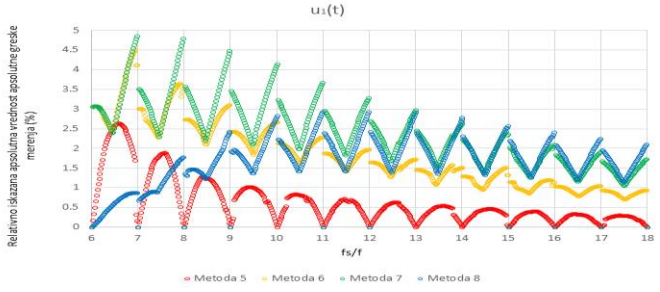
Signal $u_2(t)$ predstavlja sumu prvog i trećeg harmonika. Vrednost trećeg harmonika je odabrana tako da $u_2(t)$ prolazi kroz svoju efektivnu vrednost samo dva puta u toku jedne periode. Za veće vrednosti trećeg harmonika postojao bi problem kod određivanja periode na osnovu prolaska kroz efektivnu vrednost. U takvim slučajevima je moguće određivanje efektivne vrednosti samo primenom prve četiri metode, dok Metode 5 do 8 ne bi dale očekivani rezultat zbog pogrešnog određivanja periode signala.

U simulacijama je menjana učestanost odmeravanja sa korakom od 0.02 Hz. Ispitivano je ponašanje metoda pri odnosu učestanosti odmeravanja i učestanosti signala u granicama od 6 do 20. Da bi se verno simulirala nesinhronizovana odmeravanja, na slučajan način je uziman trenutak prvog odmeravanja gledano od početka periode signala. Greška određivanja zavisi od trenutka uzimanja prvog odbirka. Za svaku frekvenciju odmeravanja je vršeno hiljadu ponavljanja. Potom je određena najveća apsolutna vrednost greške za svaku od metoda.

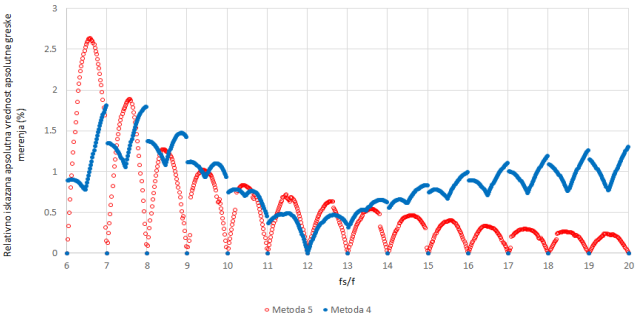
Na sledećim slikama prikazani su rezultati za oba signala za prve i druge četiri metode, respektivno, kao i rezultati poređenja metoda koje su se pokazale kao najbolje iz obe grupe.



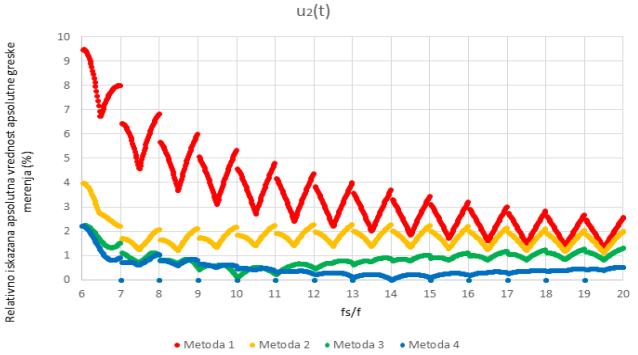
Sl. 10. Rezultati simulacije prve četiri metode za napon $u_1(t)$



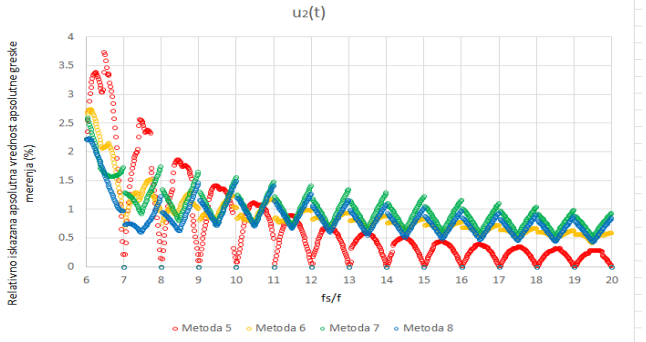
Sl. 11. Rezultati simulacije druge četiri metode za napon $u_1(t)$



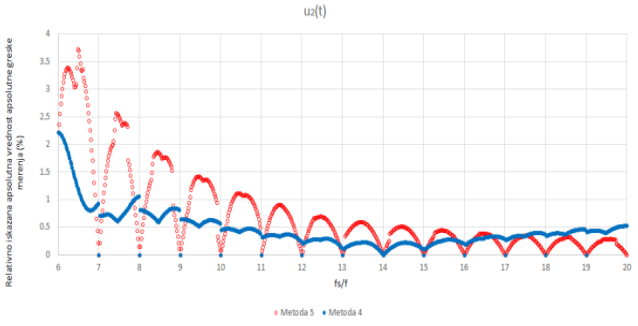
Sl. 12. Prikaz dve najbolje metode kao rezultata simulacije za napon $u_1(t)$



Sl. 13. Rezultati simulacije prve četiri metode za napon $u_2(t)$



Sl. 14. Rezultati simulacije druge četiri metode za napon $u_2(t)$



Sl. 15. Prikaz dve najbolje metode kao rezultata simulacije za napon $u_2(t)$

Sa Sl. 10 i Sl. 13 možemo da uočimo da se dobija prednost u korist svake naredne metode u odnosu na prethodnu i kod napona $u_1(t)$ i kod napona $u_2(t)$. Odnosno, kod napona kod kojih se definiše perioda uzastopnim prolaskom kroz nulu uvek se komplikovanje numeričkih metoda isplati i dobija se manja greška.

Sa Sl. 11 i Sl. 14 uočavamo obrnutu sitaciju. Komlikovanje računa za dobijanje efektivne vrednosti napona se ne isplati i vidimo da Metoda 5 ima najmanju grešku za oba signala.

Upoređivanjem dve najbolje metode iz prve i iz druge grupe serija od po četiri metode, uočavamo da je predložena Metoda 5 najefikasnija metoda kod oba napona.

IV. ZAKLJUČAK

Ovaj rad se bavi analizom metroloških osobina raznih numeričkih metoda za određivanje efektivne vrednosti napona. Cilj klasične matematike je utvrđivanje pod kojim uslovima postoji rešenje nekog problema i koje su njegove osobine, dok u numeričkoj matematici je to efektivno pronalaženje rešenja problema sa određenom tačnošću i preciznošću.

Sve formule izvedene u ovom radu imaju izraz za numeričko određivanje efektivne vrednosti koji se sastoji od korenovanja, deljenja i suma kvadrata signala pomnoženih koeficijentima, pa je vreme procesorskog rada približno jednako za sve metode.

Diskretizacija po amplitudi nije razmatrana, pa greška usled kvantizacije po vrednosti je zanemarena.

Pokazano je da se javlja problem prilikom određivanja efektivne vrednosti kada količnik učestanosti odabiranja i učestanosti ulaznog signala nije celobrojan. Zbog nesinhronog odabiranja dobija se manja efektivna vrednost kada je izvršeno više odabiranja u toku periode jer broj odbiraka u okviru jedne periode nije konstantan, već će nekad biti jedan više odbirak, a nekad jedan manje.

Urađene su simulacije za sve predložene metode. Upoređivanjem rezultata svih osam metoda, uočeno je značajno manje rasipanje rezultata i veća tačnost predložene Metode 5 u odnosu na sve ostale. U simulaciji se pokazalo da primena predložene definicije efektivne vrednosti daje tačnije i preciznije rezultate u odnosu na opštu definiciju. Kod metoda čija je perioda definisana uzastopnim prolaskom rastućeg napona kroz nulu uočena su poboljšanja ukoliko se primene druge numeričke metode u odnosu na standardnu metodu, metodu levih ili desnih pravougaonika, dok kod metoda gde je perioda ograničena uzastopnim prolaskom rastućeg napona kroz prethodno određenu efektivnu vrednost to nije slučaj. U situacijama kada se očekuje da signali imaju veliki stepen izobličenja tako da više od dva puta prolaze kroz svoju efektivnu vrednost u toku jedne periode, nije moguće primenjivati metode 5 do 8 na prikazani način. Kod ovakvih signala najbolje rezultate daje Metoda 4 – Unapređeno Simpsonovo pravilo kubnim polinomom.

Rezultati simulacija su vrlo ohrabrujući, pa bi sledeći korak provere mogućnosti predložene metode, Metode 5, mogao da bude realizovan u praksi.

LITERATURA

[1] P. Miljanic: Definitions of the average and and rms values suitable for measurement and descriptions of quasi steady state, Electronics, University of Banjaluka, Faculty of Electrical engineering, Vol. 5, No. 1-5, December 2001, pp. 18 – 20

[2] Gerson E. Mog, Eduardo P. Ribeiro: Mean and rms calculations for sampled periodic signals with non-integer number of samples per period to ac energy systems, Departamento de Engenharia Elétrica, Universidade Federal do Paraná - UFPR Centro Politécnico da UFPR - 81531-990 - Curitiba - PR – Brazil

[3] M. Albu, G. T. Heydt, : On the rms values in power quality assessment, IEEE transactions on power delivery, Vol. 18, issue 4, October 2003,pp. 1586-1587

[4] IEEE Standards Coordinating Committee 22 on Power Quality. IEEE recommended practice for monitoring electrical power quality – IEEE Std. 1159, June 1995

[5] A. Mesrobian, D.C. Rowe: The impact of variable operating frequencies and distortion on electrical power monitoring and protection systems, Petroleum and Chemical Industry Conference, Record of Conference Papers, Industry Applications Society 38th Annual, September 1991, pp. 119-122

[6] P. Miljanic, B. Stojanovic, P. Bosnjakovic: The development of a high precision power meter, IEEE Conf. Precision Electromagnetic Measur. Dig., Delft, The Netherlands, 1984, pp. 67-68

[7] N. M. Vucijak: The algorithm for phase difference deterination of low frequency sinusoidal signals, Doctoral Dissertation, University of Belgrade, School of Electrical engineering, 2015

[8] J. K. Kolanko: Accurate measurement of power, energy, and true RMS voltage using synchronous counting, IEEE Transactions on Instrumentation and Measurement, Vol. 42, No. 3, June 1993

[9] W. Guilherme, K. Ihlenfeld, G. Ramm, H. Bachmair, H Moser : Evaluation of the synchronous generation and synchronous sampling technique for the determination of low frequency AC quantities, Conference Digest, Conference on precision electromagnetic measurements, Ottawa, Ontario, Canada, June 2002

[10] M. Kampik, H. Laiz, M. Klonz, Comparison of three accurate methods to measure AC voltage at low frequencies, IEEE Trans. Instrum. Meas., Vol. 49, No. 2, 2000., pp. 429 – 433

[11] P. Petrovic, M. Stevanovic: Measuring of active power of synchronously sampled AC signals in presence of interharmonics and subharmonics, IEEE Proc. Electr. Power Appl., Vol. 153, No. 2., 2006., pp. 227 – 235

[12] P. Petrovic: New digital multimeter for accurate measurement of synchronously sampled AC signals, IEEE Trans. Instrum. Meas., Vol. 53, No. 3, 2004, pp. 716 – 725

[13] Stefan Mirković, Dragan Pejić, Marjan Urekar, Bojan Vujičić, Đorđe Novaković, "Unapređenje postojeće metode asinhronog uzorkovanja pri određivanju RMS vrednosti", ETRAN 2017, Kladovo, Srbija,2017

[14] Stefan Mirković, Dragan Pejić, Marjan Urekar, Bojan Vujičić, Đorđe Novaković, "Improvement of an Existing Method of Asynchronous Sampling for Determining RMS Value", SERBIAN JOURNAL OF ELECTRICAL ENGINEERING, Vol. 15, No. 1, February 2018

[15] "HP 3458A, Operating, Programming and Configuration Manual", Edition 1, USA, May 1988

[16] "5025 Extended Specification", V 2.7, Kent, England, 2012

ABSTRACT

This paper explores the quality of the calculation of the root mean square value using several numerical methods in the conditions of the non-integer ratio between sampling frequency and the frequency of the signal. Two independent problems were observed. One of them is the quality of the numerical assessment of integrals in order to determine root mean square. The other one is the non-integer ratio. The amplitude discretization has not been considered.

THE APPLICATION OF NUMERICAL INTEGRATION METHODS FOR DETERMINING THE ROOT MEAN SQUARE VALUE

Marina Bulat, Stefan Mirković, Dragan Pejić, Marjan Urekar
Đorđe Novaković i Nemanja Gazivoda

Prilog etaloniranju termometara sa direktnim očitavanjem u terenskim uslovima

Marina Bulat, *Member, IEEE*, Nemanja Gazivoda, *Member, IEEE*, Ivan Gutai, *Member, IEEE*, Bojan Vujičić, *Member, IEEE*, Đorđe Novaković, *Member, IEEE* i Marjan Urekar, *Member, IEEE*

***Apstrakt*— U ovom radu je dat primer postupka etaloniranja termometara sa direktnim očitavanjem van Laboratorije za metrologiju Fakulteta tehničkih nauka u Novom Sadu. Prikazane su najbolje mogućnosti merenja, opisana je priprema za merenje i navedena je merna oprema koja je korišćena. Opisani su postupci merenja koji treba da se sprovedu i prikazana je obrada dobijenih rezultata. Dati su primeri koji ilustruju primenu ovog uputstva. Svi termini i definicije su u skladu sa SRPS ISO/IEC 9000:2001, SRPS ISO/IEC 17025:2017 i Međunarodnim rečnikom osnovnih i opštih termina u metrologiji.**

***Ključne reči*—** merna oprema; etaloniranje; metrologija; termometar; merna nesigurnost.

I. Uvod

Uređaji za merenje temperature ili temperaturnog gradijenta nazivaju se termometri. Međusobno se razlikuju kako po principu na kojem se zasniva njihov rad, tako i prema mernom području. Proces etaloniranja koji prethodi izdavanju uverenja o etaloniranju je u upotrebi u industriji, laboratorijama, ocenjivanju usaglašenosti tela i preduzećima, kako bi zadovoljili zahteve standarda. Uverenje o etaloniranju je sredstvo kojim se obezbeđuje dokaz sledivosti merenja.

U Tabeli I su prikazane merne mogućnosti etaloniranja termometara sa direktnim očitavanjem, raspoloživom opremom Laboratorije za metrologiju Fakulteta tehničkih nauka u Novom Sadu (u daljem tekstu samo Laboratorija), van Laboratorije.

Marina Bulat – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail: marina.bulat@uns.ac.rs). Nemanja Gazivoda – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail: nemanjagazivoda@uns.ac.rs). Ivan Gutai – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail: gutai@uns.ac.rs). Bojan Vujičić– Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail: bojanvuj@uns.ac.rs). Đorđe Novaković – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail: djordjenovakovic@uns.ac.rs). Marjan Urekar – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail: urekarm@uns.ac.rs).

TABELA I
NAJBOLJE MOGUĆNOSTI ETALONIRANJA VAN LABORATORIJE

Predmet etaloniranja	Temperaturni opseg	Merna nesigurnost
Otpornički termometri, termometri sa termoparom	0 °C	0.06 °C
Otpornički termometri, termometri sa termoparom	25 °C do 300 °C	0.25 °C
Otpornički termometri, termometri sa termoparom	300 °C do 650 °C	3.0 °C
Termometri sa termoparom	650 °C do 1000 °C	7.5 °C

II. Priprema za rad

Priprema za etaloniranje podrazumeva proveru metroloških karakteristika etalonskog sistema pre odlaska na mesto etaloniranja, prevoz do mesta etaloniranja, vizuelni pregled objekta etaloniranja, konstataciju da je objekt pripremljen za etaloniranje i proveru referentnih uslova etaloniranja.

Da bi se smanjili rizici tokom prenošenja opreme do i od mesta etaloniranja, pre i posle prenosa vrši se provera osnovnih metroloških karakteristika etalona. Po pravilu se oprema za etaloniranje do i od mesta etaloniranja prenosi putničkim automobilom. Obezbeđuje se da temperatura prostora u kome se nalazi oprema bude u intervalu 15 °C do 30 °C.

Za etaloniranja van Laboratorije je potrebno da objekt etaloniranja i etalonska oprema budu u prostoru namenjenom za etaloniranje najmanje jedan sat pre početka etaloniranja, da bi se njihova temperatura približno izjednačila sa temperaturom okoline.

Vizuelnim pregledom se utvrđuje opšte stanje objekta etaloniranja i konstatuju se eventualna oštećenja. Utvrđuje se i postojanje dokumentacije o objektu etaloniranja, relevantne za etaloniranje.

Za etaloniranje van Laboratorije proverava se i opšte stanje prostora u kome treba da se obavi etaloniranje, a to je da prostor mora biti suv i čist, bez prašine i drugih agensa koji mogu da utiču na mernu opremu, i da se u njemu ne odvijaju

druge aktivnosti koje mogu nepovoljno da utiču na postupak etaloniranja.

Referentni uslovi za merenje van Laboratorije su:

1. Temperatura okoline: 15 °C do 30 °C,
2. Relativna vlažnost vazduha: 30% do 70%.

Provera referentnih uslova podrazumeva ispunjenje navedenih zahteva neposredno pre merenja.

III. Merna oprema

Oprema, koja je neophodna za etaloniranje termometara sa direktnim očitavanjem, navedena je u Tabeli II.

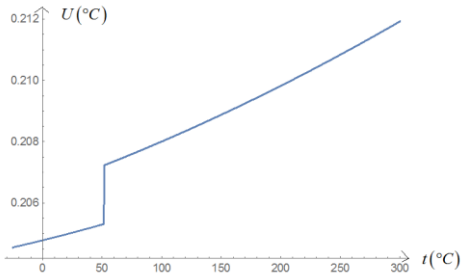
TABELA II
RADNI ETALONI ZA ETALONIRANJE TERMOMETARA SA DIREKTNIM OČITAVANJEM

Naziv	Tip	Serijski broj
Platinski otporni termometar	884X-RTD	812889
Platinski otpornički termometar	5626	812889
Etalonski termopar	S	T1-1361/1
Digitalni multimetar	8846 A 8846 A 8846 A	9310023 2090013 1918024
Kalibrator za realizaciju 0 °C	9101	B3A878
Kalibrator temperature	9103	B3A078
Kalibrator temperature	9140	B3A303
Kalibrator temperature	9141	B3A657
Kalibrator temperature	Pegasus 1200	34389/1

Imajući u vidu metrološke karakteristike etalona i svojstva zona etaloniranja, grafički su predstavljene najbolje mogućnosti etaloniranja u funkciji temperature etaloniranja i u zavisnosti od izbora etalonskog sistema.

A. – 25 °C *do* 300 °C

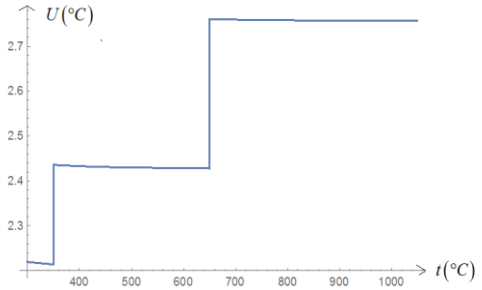
Etalonski sistem, prikazan na Sl. 1, sastoji se iz platinskog otporničkog termometra Fluke 884X-RTD i digitalnog multimetra Fluke 8846A. Za temperature etaloniranja u granicama - 25 °C do 140 °C sredina etaloniranja je suvo kupatilo Hart 9103, dok za temperature u granicama 140 °C do 300 °C sredina etaloniranja je suvo kupatilo Hart 9140.



Sl. 1. Proširena merna nesigurnost etaloniranja za faktor obuhvata $k = 2$ u zavisnosti od temperature, za etalonski sistem koji se sastoji od platinskog otporničkog termometra Fluke 884X-RTD i digitalnog multimetra Fluke 8846A

B. 300 °C *do* 1064 °C

Etalonski sistem, prikazan na Sl. 2, sastoji se iz termopara S tipa i digitalnog multimetra Fluke 8846A. Za temperature etaloniranja u granicama 300 °C do 350 °C sredina etaloniranja je suvo kupatilo Hart 9140, za temperature etaloniranja u granicama 350 °C do 650 °C sredina etaloniranja je suvo kupatilo Hart 9141, dok za temperature etaloniranja u granicama 650 °C do 1064 °C sredina etaloniranja je suvo kupatilo Pegasus 1200.



Sl. 2. Proširena merna nesigurnost etaloniranja za faktor obuhvata $k = 2$ u zavisnosti od temperature, za etalonski sistem koji se sastoji od termopara (Termotehna T1) i digitalnog multimetra Fluke 8846A

C. 1064 °C *do* 1200 °C

Za temperature etaloniranja u ovim granicama etalon je suvo kupatilo Pegasus 1200. Proširena merna nesigurnost etaloniranja je ocenjena na 3.5 °C.

Etaloniranje termometara vrši se metodom poređenja. Očitavanja na etaloniranom termometru porede se sa očitavanjima na etalonskom termometru.

IV. Etaloni

Laboratorija raspolaže opremom kojom je moguće realizovati u osnovi tri međusobno različita etalonska merna sistema. Više o njima nalazi se u nastavku rada.

A. Etalonski sistem sa otporničkom sondom i digitalnim multimetrom

Odstupanje (greška) pokazivanja rezultata merenja temperature etaloniranog termometra (DUT *engl. Device Under Test*) od temperature koju pokazuje etalonski merni sistem sa PRT sondom, modeluje se izrazom:

$$E_{DUT}=(T_{DUT}+\delta T_{DUT})-$$
$$-\left(T\left(R_s+\delta R_s+k_{Rs}+k_{Tamb}\right)+k_{PRT}+k_{DriftPRT}+k_{Stab}+k_{HomR}+k_{HomA}\right)$$

gde je:

T_{DUT} Aritmetička sredina pokazivanja DUT; Analitički izraz za njenu nesigurnost je $u_{T_{DUT}}=\frac{S_{T\,DUT}}{\sqrt{n_{T\,DUT}}}$;

$S_{T\,DUT}$ Standardno odstupanje niza pojedinačnih očitanja DUT, a $n_{T\,DUT}$ broj tih očitanja;

δT_{DUT} Korekcija očitanja DUT, zbog konačne rezolucije očitavanja. Ocena korekcije je 0 °C. Standardna nesigurnost korekcije očitanja određena je izrazom $u_{\delta T\,DUT}=\frac{LSD}{2\sqrt{3}}$, gde je *LSD* rezolucija displeja DUT;

R_s Otpornost etalonske sonde, PRT, izmerene etalonskim merilom otpornosti;

$T(R_s)$ Temperatura etaloniranja (aritmetička sredina niza očitanja), kao funkcija otpornosti R_s, definisana standardom, ili od strane proizvođača PRT, ili iz odgovarajućeg uverenja o etaloniranju. Njena nesigurnost je $u_{T(R_s)}=\frac{S_{Rs}}{\sqrt{n_{Rs}}}$, gde je S_{Rs} standardno odstupanje niza pojedinačnih očitanja R_s, a n_{Rs} broj tih očitanja;

δR_s Korekcija očitanja otpornosti R_s, zbog konačne rezolucije očitavanja;

k_{Rs} Korekcija merenja otpornosti R_s, zbog netačnosti merenja otpornosti. Ocena korekcije se preuzima iz uverenja o etaloniranju merila otpornosti. Nesigurnost korekcije preuzima se iz uverenja o etaloniranju ili se koriste podaci o tačnosti merila otpornosti, datih od strane proizvođača;

k_{Tamb} Korekcija merenja otpornosti R_s, zbog ambijentalne temperature van temperaturnog opsega digitalnog multimetra;

k_{PRT} Korekcija merenja temperature etaloniranja, zbog netačnosti PRT. Nesigurnost korekcije se preuzima iz uverenja o etaloniranju ili se koriste podaci o tačnosti merila otpornosti, datih od strane proizvođača;

$k_{DriftPRT}$ Korekcija merenja temperature etaloniranja, zbog drifta PRT;

k_{Stab} Korekcija merenja temperature etaloniranja, zbog nestabilnosti temperature etaloniranja tokom vremena etaloniranja;

k_{HomR} Korekcija merenja temperature etaloniranja, zbog radijalne nehomogenosti temperature u zoni etaloniranja;

k_{HomA} Korekcija merenja temperature etaloniranja, zbog aksijalne nehomogenosti temperature u zoni etaloniranja.

Analiza merne nesigurnosti rezultata etaloniranja DUT (merila temperature sa direktnim očitavanjem), kada se etalon sastoji od otporničke sonde (PRT) i digitalnog multimetra, prikazana je u primerima u obliku budžeta merne nesigurnosti.

Primer 1.

Etaloniran je digitalni termometar sa rezolucijom 0.1 °C na temperaturi – 20 °C. T_{DUT} je srednja vrednost od pet uzastopnih očitanja DUT. Etalonski merni sistem se sastoji od PRT (Fluke, 884X-RTD, 100 Ω, granice greške ± 0.05 °C) i digitalnog multimetra (8846 A, opseg 100 Ω, rezolucija 100 μΩ, tačnost ± (100 ppm od izmerene vrednosti + 40 ppm od mernog opsega)). R_s je srednja vrednost od pet uzastopnih očitanja etalonskog merila otpornosti. Sredina je suvi kalibrator temperature 9103 (nestabilnost temperature etaloniranja tokom vremena je 0.02 °C, radijalna nehomogenost 0.1 °C i aksijalna nehomogenost 0.1 °C). Termometar je etaloniran pri ambijentalnoj temperaturi od 15 °C. Prikaz budžeta merne nesigurnosti rezultata etaloniranja prikazan je u Tabeli III. Greška DUT je 0.1 °C, kombinovana merna nesigurnost rezultata etaloniranja 0.11 °C, a proširena merna nesigurnost za faktor obuhvata $k = 2$ rezultata etaloniranja 0.22 °C.

TABELA III
PRIKAZ BUDŽETA MERNE NESIGURNOSTI ETALONIRANJA TERMOMETRA IZ PRIMERA 1

Veličina	Simbol	Ocena	Nesigurnost	Tip	Raspodela	Osetljivost	Doprinos
Pokazivanje (aritmetička sredina pokazivanja) DUT	T_x	-19.9	44.7E-3	A	N	1	0.045
Korekcija očitanja DUT, zbog konačne rezolucije očitavanja DUT	δT_x	0	28.9E-3	B	R	1	0.029
Otpornost etalonske sonde, PRT, izmerene etalonskim merilom otpornosti	R_s	92.1520	44.7E-6	A	N	-2.57	0.000
Korekcija očitanja otpornosti, zbog konačne rezolucije očitavanja	δR_s	0	28.9E-6	B	R	-2.57	0.000
Korekcija merenja otpornosti, zbog netačnosti merenja otpornosti	k_{Rs}	0	7.6E-3	B	R	-2.57	0.020
Korekcija merenja otpornosti, zbog ambijentalne temperature van referentnog opsega	k_{Tamb}	0	1.2E-3	B	R	-2.57	0.003
Pokazivanje kalibratora temperature	T_s	-20.0					
Korekcija zadate temperature, zbog netačnosti PRT	k_{PRT}	0	28.9E-3	B	R	-1	0.029
Korekcija merenja temperature etaloniranja, zbog drifta PRT	$k_{DriftPRT}$	0	28.9E-3	B	R	-1	0.029
Korekcija zadate temperature etaloniranja, zbog nestabilnosti temperature etaloniranja tokom vremena etaloniranja	k_{Stab}	0	11.5E-3	B	R	-1	0.012
Korekcija zadate temperature etaloniranja, zbog radijalne nehomogenosti temperature u zoni etaloniranja	k_{HomR}	0	57.7E-3	B	R	-1	0.058
Korekcija zadate temperature etaloniranja, zbog aksijalne nehomogenosti temperature u zoni etaloniranja	k_{HomA}	0	57.7E-3	B	R	-1	0.058
Greška merila temperature	E_x	0.1					0.108

B. *Etalonski sistem sa termoparom i digitalnim multimetrom*

Greška e_{DUT} termometra koji se etalonira (DUT) modeluje se izrazom:

$$e_{DUT}=(T_{m\,DUT}+\delta T_{DUT})-$$
$$\left(T_s(k_{ess}+\delta e_S+k_{es}+k_{Tamb})+k_{TP}+k_{DriftTP}+k_{Stab}+k_{HomR}+k_{HomA}\right)$$

gde je:

$T_{m\,DUT}$ Rezultat (aritmetička sredina) merenja temperature DUT;

δT_{DUT} Korekcija merenja temperature DUT, zbog konačne rezolucije očitavanja;

T_s Temperatura etaloniranja, kao funkcija elektromotorne sile etalonskog termopara;

k_{ess} Korekcija očitanja elektromotorne sile e_S, zbog rasipanja rezultata merenja elektromotorne sile etalonskog termopara;

δe_S Korekcija očitanja elektromotorne sile e_S, zbog konačne rezolucije očitavanja elektromotorne sile etalonskog termopara;

k_{es} Korekcija merenja elektromotorne sile e_S, zbog netačnosti merenja elektromotorne sile etalonskog termopara. Ocena korekcije se preuzima iz uverenja o etaloniranju merila elektromotorne sile. Nesigurnost korekcije se preuzima iz uverenja o etaloniranju ili se koriste podaci o tačnosti merila elektromotorne sile, datih od strane proizvođača;

k_{Tamb} Korekcija merenja otpornosti e_S, zbog ambijentalne temperature van temperaturnog opsega digitalnog multimetra;

k_{TP} Korekcija merenja temperature etaloniranja, zbog netačnosti etalonskog termopara (TP). Nesigurnost korekcije se preuzima iz uverenja o etaloniranju ili se koriste podaci o tačnosti etalonskog termopara, datih od strane proizvođača;

$k_{DriftTP}$ Korekcija merenja temperature etaloniranja, zbog drifta etalonskog termopara;

k_{Stab} Korekcija merenja temperature etaloniranja, zbog nestabilnosti temperature etaloniranja tokom vremena etaloniranja;

k_{HomR} Korekcija merenja temperature etaloniranja, zbog radijalne nehomogenosti temperature u zoni etaloniranja;

k_{HomA} Korekcija merenja temperature etaloniranja, zbog aksijalne nehomogenosti temperature u zoni etaloniranja.

C. *Kalibratori temperature*

Greška e_{DUT} termometra koji se etalonira (DUT) modeluje se izrazom:

$$e_{DUT}=(T_{m\,DUT}+\delta T_{DUT})-(T_s+k_{T_s}+k_{Stab}+k_{HomR}+k_{HomA})$$

gde je:

$T_{m\,DUT}$ Rezultat (aritmetička sredina) merenja temperature DUT;

δT_{DUT} Korekcija merenja temperature DUT, zbog konačne rezolucije očitavanja;

T_s Temperatura etaloniranja (pokazivanje kalibratora temperature);

k_{Ts} Korekcija zadate temperature etaloniranja, zbog netačnosti kalibratora temperature. Nesigurnost korekcije se preuzima iz uverenja o etaloniranju ili se koriste podaci o tačnosti kalibratora, datih od strane proizvođača;

k_{Stab} Korekcija merenja temperature etaloniranja, zbog nestabilnosti temperature etaloniranja tokom vremena etaloniranja;

k_{HomR} Korekcija merenja temperature etaloniranja, zbog radijalne nehomogenosti temperature u zoni etaloniranja;

k_{HomA} Korekcija merenja temperature etaloniranja, zbog aksijalne nehomogenosti temperature u zoni etaloniranja.

Analiza merne nesigurnosti rezultata etaloniranja DUT (merila temperature sa direktnim očitavanjem), kada je etalon kalibrator temperature, prikazana je u primerima u obliku budžeta merne nesigurnosti.

Primer 2.

Etaloniran je digitalni termometar sa rezolucijom 0.1 °C na temperaturi –20 °C. T_{DUT} je srednja vrednost od pet uzastopnih očitanja DUT. Etalonski merni sistem je kalibrator temperature Hart, 9103, granice greške ±0.25 °C. T_s je temperatura zadata kalibratorom. Nestabilnost temperature etaloniranja tokom vremena je 0.03 °C, radijalna nehomogenost 0.1 °C a aksijalna nehomogenost 0.1 °C. Termometar je etaloniran u Laboratoriji. Prikaz budžeta merne nesigurnosti rezultata etaloniranja prikazan je u Tabeli IV. Greška DUT je 0.20 °C , kombinovana standardna nesigurnost 0.17 °C, a proširena nesigurnost za faktor obuhvata $k = 2$ je 0.35 °C.

TABELA IV
PRIKAZ BUDŽETA MERNE NESIGURNOSTI ETALONIRANJA TERMOMETRA IZ PRIMERA 2

Veličina	Simbol	Ocena	Nesigurnost	Tip	Raspodela	Osetljivost	Doprinos
Pokazivanje (aritmetička sredina pokazivanja) DUT	T_x	-20.2	0.045	A	N	1	0.045
Korekcija očitanja DUT, zbog konačne rezolucije očitavanja	δT_x	0	0.029	B	R	1	0.029
Pokazivanje kalibratora temperature	T_s	-20.0					
Korekcija zadate temperature, zbog netačnosti kalibratora temperature	k_{Ts}	0	0.144	B	R	-1	0.144
Korekcija zadate temperature etaloniranja, zbog nestabilnosti temperature etaloniranja tokom vremena etaloniranja	k_{Stab}	0	0.012	B	R	-1	0.012
Korekcija zadate temperature etaloniranja, zbog radijalne nehomogenosti temperature u zoni etaloniranja	k_{HomR}	0	0.058	B	R	-1	0.058
Korekcija zadate temperature etaloniranja, zbog aksijalne nehomogenosti temperature u zoni etaloniranja	k_{HomA}	0	0.058	B	R	-1	0.058
Greška merila temperature	E_x	-0.2					0.175

ZAHVALNICA

Ovaj rad je delom podržan od strane projekta ELEMEND (šifra projekta: 585681-EEP-1-2017-EL-EPPKA2-CBHE-JP).

LITERATURA

[1] DKD-R 5-1, "Kalibrierung von Widerstandsthermometern", Richtlinie, Deutscher Kalibrierdienst, Ausgabe 10/2003

[2] Q3.JIM.21, v.1, 2013, „Etaloniranje otporničkih termometara", Radno uputstvo, Laboratorija za metrologiju

[3] EURAMET cg-8, Version 2.0 (03/2011), "Calibration of Thermocouples"

[4] Q3.JIM.22, v.1, 2013, „Etaloniranje termoparova", Radno uputstvo, Laboratorija za metrologiju

[5] Q3.JIM.20, v.2, 2013, „Etaloniranje termometara sa direktnim očitavanjem", Radno uputstvo, Laboratorija za metrologiju

[6] EURAMET cg-11, Version 2.0 (03/2011), "Guidelines on the Calibration of Temperature Indicators and Simulators by Electrical Simulation and Measurement"

[7] Q3.JIM.19, v.2, 2013, „Etaloniranje pokaznih naprava termometara sa otporničkim sondama i/ili termoparovima", Radno uputstvo, Laboratorija za metrologiju

[8] J. V. Nicholas, D. R. White,"Traceable Temperatures: An Introduction to Temperature Measurement and Calibration", Second Edition, JOHN WILEY & SONS, LTD.,2001

[9] SP 250-23 NIST MEASUREMENT SERVICES, „Liquid-in-Glass Thermometer Calibration Service", September 1988.

ABSTRACT

The paper presents an example of the procedure for the calibration of direct reading thermometers outside the Laboratory. It demonstrates the optimal ways of measuring. In addition, it offers a description of the preparation for measuring and lists the measuring equipment that has been used. It also features a description of the measuring procedures that need to be applied and it shows the processing of the measurement results. It provides the examples illustrating the application of this instruction manual. All the terms and definitions meet the requirements of SRPS ISO/IEC 9000:2001, SRPS ISO/IEC 17025:2017 and follow International Vocabulary of Basic and General Terms in Metrology.

A Contribution to the Calibration of Direct Reading Thermometers outside the Laboratory

Marina Bulat, Nemanja Gazivoda, Ivan Gutai, Bojan Vujičić, Đorđe Novaković, Marjan Urekar

Sun and Displays: Old stories and new challenges

Branko Livada, *Member, SPIE*

***Abstract*— Display sun readability is very important in the case of the "mission critical" applications as avionic cockpit in harsh illumination environment, causing a lot of research and development leading to suitable, but expensive solutions. Wide display applications in mobile systems and vehicles require new cheap display solutions that should be readable in less demanding illumination conditions. Basic physical processes related to display readability are reviewed. The basic differences in display avionic and new mobile and vehicular applications are discussed. New application require still serious, but allow different approach and different technological solutions and measurement techniques.**

***Index Terms*— Displays, display readability, sun illumination, sun readability, automotive displays**

I. Introduction

An unbelievable "futuristic" idea about video phone from the beginning of the 20th century initiated a lot of research and developments in the various areas to become real nowadays. The advances in the digital image sensors, digital displays and mobile technology make it possible.

Digital information display technology development is diversified through different technical solutions [1]-[3], but no one of them provides satisfactory readability in the high illumination environment. The special and expensive ruggedization techniques should be applied to make them readable. The military and avionic application where display data are considered as mission critical initiated an extensive study and developments of the new technical solutions [4]-[9]. The mass application of the mobile devices [10] and nowadays application of the vehicle displays [11], [12], renew interest to sun readable display technology that use relatively simple and cheap technical solutions.

In this paper a short review of the up to date research leading to display sun readability and sun influence to the display functionality is presented, emphasizing physical processes involved, and related measurement techniques. The list of the selected references of the relevant work leading to deeper understanding is set [13]–[32] together with a short list of the relevant measurement standards [34]-[42]. The basic objective is to point out the importance of the sun readability in new practical applications, especially in the vehicle displays that could be considered as mission critical application. Also, new applications require new and cheaper technological solutions and new approach to measurements. Fortunately, there is nice scientific heritage regarding avionic display developments that could be used.

The basic properties of the sun irradiation and process involving sun radiation influence on display readability and display heating due to sun thermal load are presented as starting point. The basics of the display technology related to sun readability is described, followed by short review of the human visual system properties. Sun readability definition and related measurements are discussed to analyze display sun readability behavior in the harsh illumination environment.

Branko Livada is with the Vlatacom Institute, Milutina Milankovića 5, 11070 Belgrade, Serbia (e-mail: branko.livada@ vlatacom.com).

II. Sun Irradiation

The Sun is G class star with mean radius of about 695,000 km. The surface temperature is approximately 5900K by best fit blackbody curve, or about 5770K for temperature of a blackbody source that is the size and distance of the sun and would produce an exo-atmospheric total irradiance of 1390W/m2. The mean earth to sun distance is 149.680.106 km.

Typical Solar spectrum is shown in Fig. 1 [43], [44].

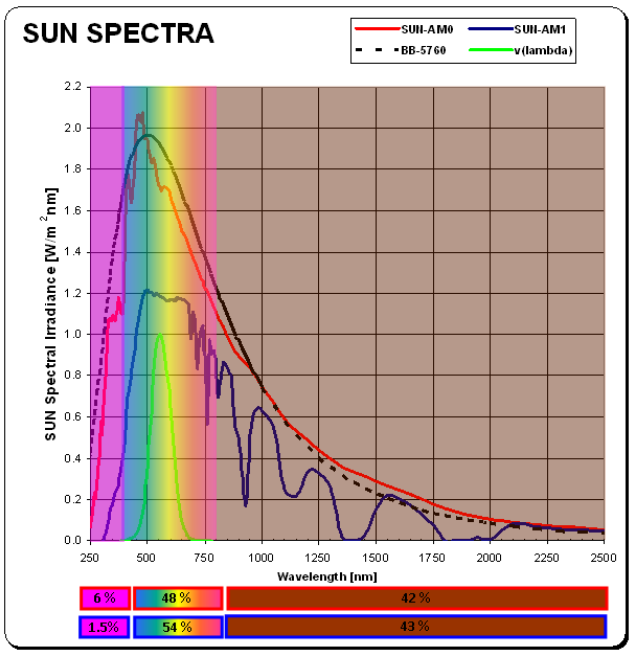


Fig. 1. Sun spectral Irradiance and spectral band flux distribution

Total solar radiation energy is distributed in different spectral regions as shown in Table I.

Mean sun irradiance outside of atmosphere is:

1390 W/m2 (mean earth-sun distance) – First Solar Constant – AM0 – Air mass zero.

As Earth trajectory is not perfect circle, Sun irradiance just outside the earth's atmosphere is:

1438 W/m2 (minimal earth-sun distance - January)
1345 W/m2 (maximal earth-sun distance - July)
Sun irradiance generates
747 W /m^2 at sea level – Second Solar Constant – AM1 – Air mass one.

Sun Illuminance outside atmosphere and on horizontal surface at sea level, maximal with sun at its zenith and clear sky is:
E = 127,823 lux = 11879 fc – Outside atmosphere
E = 86,364 lux = 8026 fc – Sea Level

TABLE I
SPECTRAL DISTRIBUTION OF SOLAR RADIATION

Spectral region (type of radiation)	Wavelength band [μm]	Relative contribution to total radiation [%]	
		Out of atm.	Sea Level
Ultraviolet - **UV**	0.29 to 0.38	**6**	**1.5**
Visible (**VIS**)	0.38 to 0.78	**48**	**54**
Near Infrared (N**IR**)	0.78 to 2.5	**42**	**43**
Far Infrared (**FIR**)	> 2.5	**4**	**1.5**

Sun irradiation influence display operation through thermal load and display image appearance (readability).

The factors influencing solar heat load thermal balance are illustrated in Fig. 2 and summarized in Table II.

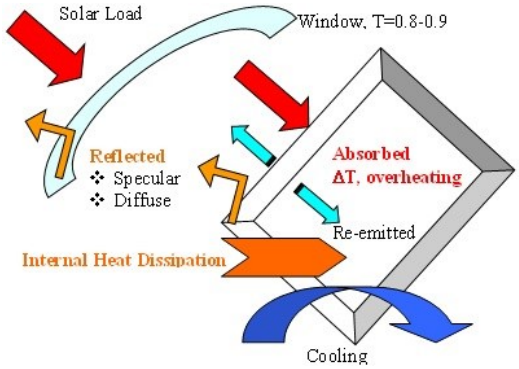


Fig. 2. Solar Load Heat balance processes

TABLE II
DISPLAY PARTS INFLUENCE ON THE THERMAL BALANCE

Display Component		Spectral Band		
		UV	***VIS***	***NIR***
External Window		A~1.0	T≅085, R=0.1	R≥0.2,
Front Glass		A~1.0	T~1.0	R≥0.2,
AMLCD		A=1.0	A≥0.7, T< 0.3	R≥0.2,
Bezel		A=1.0	A=1.0	R≥0.2,
Housing		A=1.0	A=1.0	A~1.0, E~0.9
Sun Irradiation	**OUT**	**0.06**	**0.48**	**0.42**
	SEA	**0.015**	**0.54**	**0.43**

A - Absorption; T – transmission; R - reflection

Display IR emission is comparable with solar load, and contributes to radiative cooling. The emission contribution to radiative cooling is related only to display temperature difference due to total overheating, because the part of emitted radiation is compensated from ambient. It could be considered that about 75% of solar thermal load is transferred to heat (due to solar losses in the window and NIR reflection). Radiation cooling contribution could be estimated to be about 7 W/m^2, per 1 degree K of display temperature increase.

In the case of AMLCD displays, backlight power dissipation is dominant heat source and it is the highest in high illumination conditions when high luminance level is desired.

Sun irradiation thermal load should be always considered when display is exposed to high illumination conditions. Sun related thermal load is not localized, but distributed to whole display volume. Sun load should be added to other heat sources.

There are three main types of reflection to be considered in the case of electro-optical displays [16]:
Specular reflection (surface – Fresnel's reflection) – mirror like light reflection.

Diffusive reflection (scattering and backscattering) considered as ideal Lambertian reflector – the luminance value is independent from direction

Haze reflection (organic materials – due to scattering on macromolecules) – is similar, by nature, to diffuse reflection, but spatially concentrated in selected direction, similar as specular component. Haze reflection exists between two extremes (specular and diffuse)

A more general way to describe display reflection is, so called, BRDF (Bi-Directional Reflectance Function), that involve all three (specular, diffusive and haze) components.

Display reflection or display components reflection measurements is most important part during display development phase. Also it is important part of the applied technology validation for sun readability.

Clear identification of each of three reflection types is important for display technology and structure selection for required sun readability.

Display reflection metrology should to fulfill, at least, several goals [17], [18]:

Robust: Allow some perturbations in method easing measurement process without affecting projected uncertainty.

Reproducible: Different laboratories or different operators using similar equipment should obtain same results within the requisite uncertainty.

Relevant: Must provide data comparable for different technologies and applicable in design and product improvement efforts.

Unambiguous: Must be clear for user. Method should be well documented with all relevant data on one place.

Simple: Easy and quick to perform. Avoid any un-necessary complications

Only detail BRDF measurements can to obtain full data package suitable for clear distinguishing between reflection types, but these measurements are not simple, and they are hardly applicable for fast reflection influence analysis, regarding display application and visual information extraction efficacy.

BRDF measurement results are recommended in the initial technology evaluation phase, allowing best solution selection. Because of that, several different reflection methods are developed to support sun readability evaluation.

III. Display Technology

During the second half of 20th century a lot of different display technologies were developed, as illustrated in Fig. 3.

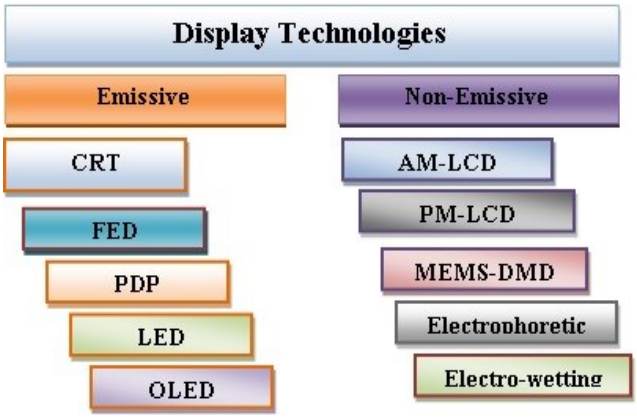


Fig.3 Display technology classification

Cathode ray tubes – CRT made a first break through in display mass application and production. Active matrix liquid crystal – AMLCD technology nowadays dominates on the market but other technologies has some advantages in selected applications (OLED - Organic Light Emitting Diode, MEMS Micro Electro Mechanical Systems). The research and development in the area of new emerging display technologies are still active.

IV. Human Visual System Basic Properties

Performance parameters of human vision are the key limiting factor for perception and extraction of information contained in the image. The visual information perception by human observer could be used directly for image quality assessment through psychophysical measurements [49]. Psychophysical measures of the image quality are too costly and time consuming for evaluation of the impact that each algorithm modification might have on image quality. On the other hand, it is convenient to have analytical model of the human vision system to be incorporated in various algorithms for image compression or processing.

The selected HVS (Human Visual System) properties describing limiting possibilities are [46]:

- Contrast sensitivity as illustrated in Fig. 4,
- Resolution power (Nyquist limit) – 56 cycles/degree,
- Visual acuity limit – 1 arc-minute, minimum perceptible limit 0.3 arc-minute,
- Dynamic Range – 10^{-6} - 10^{6} nits,
- Critical Flicker Frequency – CFF – 60 – 72 Hz.

Human visual system is adapted to be sensitive in the wide range of illumination levels - starting at less 1 mlux (night, starlight) up to more than 100 klux (direct sun illumination) for natural illumination, and up to 2klux artificial illumination (office environment).

Modeling of human vision has a long development history based on the results of psychometrics results and defined needs for aimed application. The basic principles [47], [48] are based on proper analytical modeling starting from known experimental results. One of the best known models [47] is based on the modeling of the contrast sensitivity function dependence on spatial frequency and level of illumination (see Fig. 4). Further development introduced models that involved HVS motion sensitivity (both eye motion and motion in image), temporal sensitivity and color sensitivity.

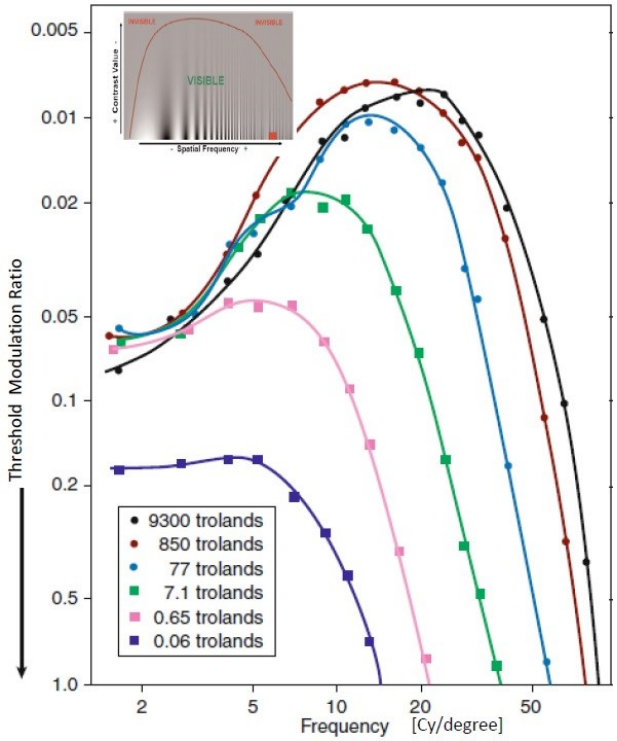


Fig. 4 HVS contrast sensitivity function (perceived threshold contrast)

V. Sun Readability

Display surface is specific light source containing information to be used by observer.

When illuminated by sun display emission is mixed with reflected sun, so information content could be severe disturbed. The root cause is reflected sun disturbance of the display visual content.

To fight sun illumination influence on the sun-readability two main approaches are used:

"Use the Sun" – Sun light illumination is used to contribute to the display luminance, as in reflective or transflective displays.

"Reject the Sun" – Techniques based on lowering reflected sun influence decreasing display reflectance, or increasing displays luminance, or both.

Sun Readability is a complex function of the whole system, in which display is used, including Display contrast and luminance, the properties of image and information content presented, the illumination environment, viewer visual properties and position.

Sun readability is mainly connected to observer's perception of visualized information. Because of that, it could be used as good display "selling point" based personal feeling as argument.

From technical point of view sun readability should be defined clearly using proper metrics and parameters that allow value understanding and repeatable measurements.

The key starting point is analysis of the display illumination environment related to system application.

The second step is definition of the display viewing envelope expected for named application.

The third, and the most important step, is defining proper sun readability metrics.

To define proper metrics required image quality and/or displayed data content, should be considered. In the case of vehicular displays the application difference between main dash-board displays and entertainment displays should be considered.

Display contrast threshold definition depending on the visual task and related illumination environment is the most common way to define sun readability criteria. This parameter is measurable, so it is fairly objective. When it is combined with display viewing and illumination conditions then we can have proper defined metrics.

Another, more accurate parameter defining sun legibility is PJND (Perceived Just Noticeable Difference) that is combined from CJND (Chrominance Just Noticeable Difference) and LJND (Luminance Just Noticeable Difference) using mathematical model to process display luminance and illumination environment spectral data following visual task requirements and observer's visual properties. Because of measurements and modeling complexity this approach is not widely accepted as legibility criteria through display specification and acceptance testing. Display complete sun readability evaluation could be complicated, and because of that, it is important to simplify measurement conditions reducing them to the worst case measurements.

Sun illumination has severe influence to display readability, including display active area sun reflections degrading display contrast (display "bleaching" or "wash out"), display chromaticity changes, and observer's eye sensitivity degradation in high direct sun illumination conditions.

System designer task is to apply all possible measures to reduce illumination (shading, windows, position selection, internal wall processing etc.) making illumination environment less severe. Also, system designer should to define display illumination environment and requirements to display design.

The large variety of illumination scenarios for sun-readable display applications is a big challenge for display system designer.

The first step is dominant source definition (direct sun, diffused – scattered illumination).

Second step is illumination source position definition related to display normal (geometry), based on illumination scenario analysis and worst case determination, applicable for test conditions definitions.

Observer eye reference points and related eye position reference boxes, defining display viewing envelope for named application is one of the possible "filters" applicable for worst case definition (see Fig. 5).

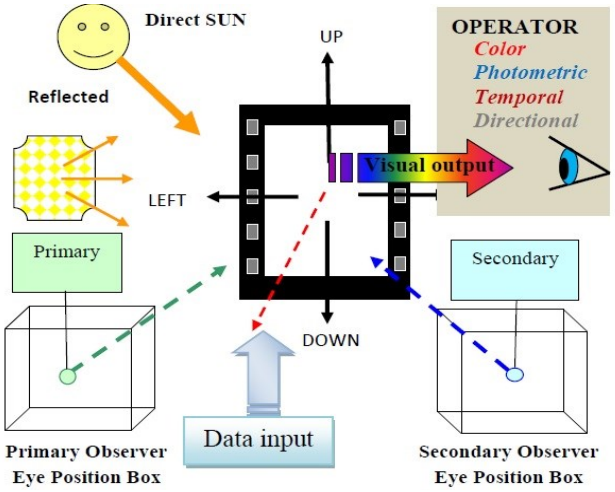


Fig. 5 Illumination environment definition

Display luminance should be in accordance with illumination scenario to allow sharp and comfortable display content perception (see illustration at Fig. 6).

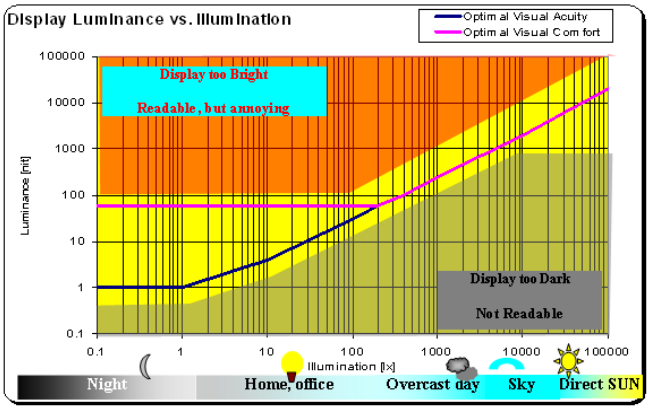


Fig. 6 Display luminance recommended level versus illumination

In the case of military or "mission critical" displays, illumination condition considered for readability evaluation, the worst illumination condition for planned application are used. Such worst illumination conditions are illustrated in Table III.

In the case of automotive displays, the key challenge is to define critical illumination conditions related to application to provide leading line for related technology development.

TABLE III
SUGGESTED VALUES FOR DISPLAY ILLUMINATION AND GLARE LUMINANCE TEST VALUES FOR AVIONIC DISPLAYS

Source Type	Sun shine ambient			
	Bubble Canopy	**Cockpit with Roof**	**Shaded**	**Enclosed cabin**
Diffusive, (Illumination)	108000 lux (10000 fc)	86000 lux (8000 fc)	3240 lux (300 fc)	540 lux (50 fc)
Specular (glare)	6800 nits (2000 fL)	6800 nits (2000 fL)	6800 nits (2000 fL)	3400 lux (1000fc)

VI. Display Measurements in Harsh Illumination Environment

Display measurements related to operation in harsh illumination environment are usually limited to display contrast measurements under simulated illumination causing specular and/or diffusive reflection.

A. *Specular Reflection* Measurements

Specular reflection (mirror like reflection) produced virtual image of the source. Specular reflection on display surface is usually connected to Fresnel reflection at boundary surface between two optically different media.

LCD Display specular reflection has higher than expected value because LCD structure is multi layered structure of glass and high index of refraction conductive ITO (Indium Tin Oxide) coatings. Haze reflection is also, higher because of light scattering on pixel patterned structure.

Practically, it is not so important to distinguish the nature of reflection because both, specular and haze reflection disturb display visual information content on the same way. There are several ways to measure specular reflection:

(A) Collimated source and narrow field of view radiation receiver: The better collimated beam and narrow as possible receiver field of view will contribute to better isolation of the haze reflection influence. This method is not widely used in display metrology.

(B) Large source specular measurements: Large uniform Lambertian source is placed under angle θs relative to display normal and narrow field of view receiver is placed at angle θr on the opposite side (see Fig. 7) and θs ≈ θr (usually 15° or 30°). The change in position size and uniformity of the source can introduce significant error. This method is excellent simulation of the real situation and it is widely used in display metrology, although do not allow clear distinguishing between specular and haze reflection. The similar case is with human eye, which count only reflected radiation in designated direction and do not recognize reflection type.

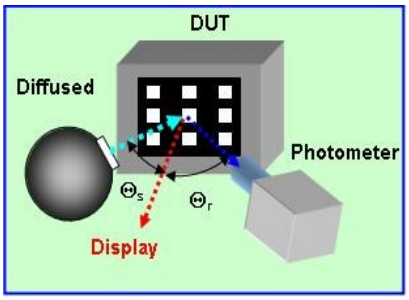


Fig. 7 Display specular reflection set up

B. *Diffusive Reflection Measurement*

In the case that concentrated - collimated incident beam (high intensity distant point source) is completely and uniformly scattered over hemisphere, one can call these type of diffuse reflection Lambertian or ideal diffuse reflection. In that case the luminance is the same in all directions.

The most common method to determine Lambertian (ideal diffuse) reflection is illustrated in Fig. 8.

Collimated beam generating predefined uniform illumination is angularly separated from receiver (low FOV photometer). The angle between incident beam direction and measurement direction is usually 30° (15°). There are two basic experimental configurations commonly used: (A) Source collimated beam optical axis is normal to display surface; (B) Photometer optical axis is normal to display surface. The incident beam illumination is determined using same experimental arrangement, but flat calibrated ideal Lambertian reflector (White reflector standard) is placed in the same plane as Display under test surface.

The display contrast measurements, under combined illumination (specular and diffusive), is illustrated Fig. 9.

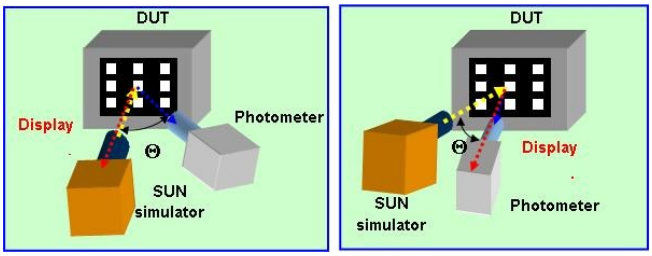


Fig. 8 Display Diffusive reflectance measurement set-up

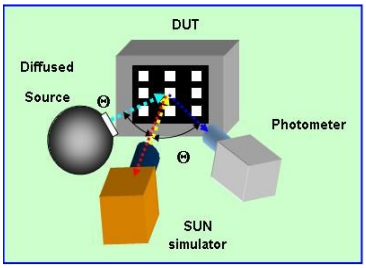


Fig. 9 Display contrast ratio measurements set- up, combined sources

VII. Conclusion

Information display mass application in mobile devices and automotive industry, require new approach to display sun readability. Sun readability was concern in "mission critical" application in avionic cockpit leading to new technical solutions according to deep knowledge of the critical processes.

Automotive display [10] applications are in the certain level mission critical (dash board displays) where sun readability is an important property. In the same time, entertainment automotive displays require good picture appearance [48], and sun readability is good "selling point".

The importance of the sun readability in the case of the automotive displays is emphasized with attention on similarities and differences with avionic displays. The scientific heritage regarding sun readability research of the avionic displays is excellent basis for new approach to sun readability measurement methods suitable for automotive displays. The basic physical processes defining sun readability are presented to support new approach on automotive displays

sun readability measurements [49] that should be based on critical illumination ambient conditions analysis. Starting from good definition of illumination environment, and display technology requirement definition, using ambient light sensors to control display operation lead to reliable and relatively cheap technical solutions.

REFERENCES

[1] R. L. Wisnieff, J. J. Ritsko, "Electronic displays for information technology", *IBM J. RES. DEVELOP*. Vol. 44 ,No. 3 May 2000

[2] Hainich, Rolf R, *Displays: fundamentals and application*, A K Peters/CRC Press, Taylor and Francis Group, LLC, 2011

[3] Janglin Chen, Wayne Cranton, Mark Fihn (Eds.), *Handbook of Visual Display Technology*, Springer-Verlag Berlin Heidelberg 2012

[4] Desjardins, Daniel D., *Military displays: technology and applications*, SPIE Press, Bellingham, 2013

[5] Branko Livada, "Avionic Displays", *Scientific Technical Review*, Vol.62,No.3-4,pp.70-79, 2012

[6] Branko Livada, Radomir Jankovic, Nikolic , "AFV Vetronics: Displays Design Criteria", *Strojniški vestnik - Journal of Mechanical Engineering (JME*), vol.58, No6, 376-385, 2012

[7] D.G. Hopper, "High Resolution Displays and Roadmap," in Proceedings of the 10th International Conference on Artificial Reality and Tele-existence (ICAT'2000)," 25-27 October 2000 in Taipei, Taiwan.

[8] Snow et al., "The AMLCD Cockpit: Promise and Payoffs" Cockpit Displays VI,: Displays for Defence Applications, Proc. SPIE vol. 3690, 1999

[9] Achintya K , Bhowmik, Zili Li, Philip Bos(Editors), *Mobile displays : technology and application*s, John Wiley & Sons Ltd, Chichester, 2008

[10] Jan Bauer, Markus Kreuzer, "Understanding the Requirements for Automotive Displays in Ambient Light Conditions, SID Proceedings , 2016 Vehicle Displays and Interfaces Symposium, September 27-28, 2016, Livonia, Michigan, USA, 2016

[11] Hai-Wei Chen, Jiun-Haw Lee, Bo-Yen Lin, Stanley Chen and Shin-Tson Wu, "Liquid crystal display and organic light-emitting diode display: present status and future perspectives", *Light: Science & Applications* No 7, 2018

[12] Shiyong Zhang and Stephen Atwood: " LCD-Luminance-Enhancement Methods for High-Ambient Applications" *Information Display* 7/06, 2006

[13] John W. Stetson: "Analog Resistive Touch Panels and Sunlight Readability*", Information Display* 12/06, 2006

[14] Brian E. Herr, Jeff Blake, and Richard D. Paynton: " Optical Enhancement and EMI Shielding for Touch Screens", *Information Display* 5/10, 2010

[15] Edward F. Kelley, George R. Jones, and Thomas A. Germer: " The Three Components of Reflection", *Information Display*, 10/98, 1998

[16] Edward F. Kelley, Max Lindfors, and John Penczek: "2.1: Invited Paper: Daylight and Sunlight Display Readability Measurement Methods", ADEAC 05, 2005

[17] Edward F. Kelley: " Reflections on Sunlight or Daylight Readability", *Information Display* l/07, 2007

[18] E. F. Kelley, M. Lindfors, and J. Penczek: "Display Daylight Ambient Contrast Measurement Methods and Daylight Readability," *J. Soc. Info. Display* vol.14, No. 11, 2006, 2006

[19] Darrel G. Hopper and Daniel D. Desjardins: "6.1: Military Cockpit Display Performance Requirements", ADEAC 04, 2004

[20] Michael Klein: "3.3: Invited Paper: Photometry and Colorimetry of Displays", ADEAC 04, 2004

[21] Michael E. Becker: "Display reflectance: Basics, measurement, and rating", *Journal of the SID* 14/11, 2006

[22] Adrianus J.S.M. de Vaan, "Competing display technologies for the best image performance", *Journal of the SID* 15/9, 2007

[23] John W. Stetson: "Analog Resistive Touch Panels and Sunlight Readability", *Information Display* 12/06, 2006

[24] Brian E. Herr, Jeff Blake, and Richard D. Paynton: "Optical Enhancement and EMI Shielding for Touch Screens", *Information Display* 5/10, 2010

[25] Phil N. Day, Jim Colville, Charlie Rohan, "An Evaluation of Sunlight-Viewable Displays", Proceedings of the 2010 British Computer Society Conference on Human-Computer Interaction, BCS-HCI 2010, Dundee, United Kingdom, 6-10 September 2010

[26] Hsi-Hao Chung, Sun Lu: „Contrast-ratio analysis of sunlight-readable color LCDs for outdoor applications", *Journal of the SID* 11/1, 2003

[27] Michael E. Becker: "Display reflectance: Basics, measurement, and rating", *Journal of the SID* 14/11, 2006

[28] Edward F. Kelley, Max Lindfors, and John Penczek: 'Daylight and Sunlight Display Readability Measurement Methods", *Journal of the SID* 14/11, 2006

[29] James D. Sampica, Joseph L. Tchon, Alyssa Butterfield, " Optical and Environmental Requirements for Display Applications", downloaded https://www.researchgate.net/publication/242699847_Optical_and_Environmental_Requirements_for_Display_Applications, Dec. 04. 2018

[30] R. Sharpe, C. Cartwright, K. Wilson, S. Riches, H. Orr, C. Bailey, C. Yin, Y. Lee: "A Usability Metric for Displays in Challenging Environments", IDW 06 , 2006

[31] Kelley, E. F., Jones, G. R., and Germer, T. A., Display reflectance model based on the BRDF*, Displays* 19 (1), pp. 27-34, June 1998.

[32] Becker, M. E., Evaluation and Characterization of Display Reflectance. *Displays* 19(1), pp. 35-54, June 1998

[33] Society of Informatioon Displays, VESA, FPDM Standard Version 1.0, Flat Panel Display Measurements Standard. May 1998

[34] Society of Information Displays, IDMS v.1.03: Information Display Measurement standard, June 01, 2012

[35] MIL-L-85762A: "LIGHTING, AIRCRAFT, INTERIOR, NIGHT VISION IMAGING SYSTEM (NVIS) COMPATIBLE", 26 August 1988

[36] MIL-HDBK-87213A (USAF) – "ELECTRONICALLY / OPTICALLY GENERATED AIRBORNE DISPLAYS, 8 February 2005

[37] MIL-STD-3009" " LIGHTING, AIRCRAFT, NIGHT VISION IMAGING SYSTEM (NVIS) COMPATIBLE" 2 February 2001

[38] SAE-AS7788, "PANELS, INFORMATION, INTEGRALLY ILLUMINATED", 1999-07

[39] SAE ARP 4260: "Photometric and Colorimetric Measurement Procedures for Airborne Flat Panel Display", 1998

[40] SAE ARP 4256 A: "Design Objectives for Liquid Crystal Displays for part 25 (Transport) Aircraft" 2008

[41] JSSG-2010-5 "CREW SYSTEMS AIRCRAFT LIGHTING HANDBOOK", 30 October 1998, JOINT SERVICE SPECIFICATION GUIDE)

[42] DEF STAN 00-970 PART 7/2 – Section 1- (LEAFLET 107 PILOT'S COCKPIT - CONTROLS AND INSTRUMENTS) – UK,

[43] W.L. Wolfe, G.J. Zissis (Editors), The Infrared Handbook", IRIA Center, Michigan 1978

[44] G.J. Zissis (Editor), *The Infrared Electro-Optical Systems Handbook*, Vol 1. "Sources of Radiation", SPIE Optical Engineering Press, Bellingham, Washington, 1993.

[45] David G. Curry, Gary Martinsen, and Darrel G. Hopper, "Capability of the human visual system", Cockpit Displays X, Darrel Hopper, Editor, Proceedings of SPIE, Vol. 5080 , 2003

[46] Barten, P.G.J., "Physical Model For The Contrast Sensitivity Of The Human Eye", Proc. SPIE 1666, pp57-72, 1992

[47] Peter G.J Barten, *Contrast sensitivity of the human eye and its effects on image quality*, Knegsel: HV Press, 1999

[48] M. F. Cowlishaw, "Fundamental Requirements for Picture Presentation", Proceedings of the SID, Vol. 26 No. 2, pp 101-107, 1985

[49] John Penczek, Edward F. Kelley Paul A. Boynton , "General framework for measuring the optical characteristics of displays under ambient illumination", *Journal of the SID* **23**/11, pp.529-542, 2015

Influence of Hydrogen Reduction on Microchannel Plate Parameters

A. Stanković, I. Zlatković, R. Nikolov, B. Brindić and D. Pantić

***Abstract*—Hydrogen reduction is a very important step in a microchannel plate (MCP) technology. This process determines the glass wafer behavior and thus defines the conductive and secondary electron emission properties. In this paper, the effect of the temperature and the time of the hydrogen reduction process on a fixed pattern noise (FPN) threshold are examined, as well as obtaining optimal resistance of MCP. In order to see the effect of reduction experiments are made with MCPs from the same batch (the same thermal history).**

***Index Terms*— Reduction, lead-silicate glasses, resistance, microchannel plate, FPN.**

I. Introduction

Microchannel plate (MCP) is a specially fabricated plate, that has several million channels and each one works as an independent electron multiplier. Fig.1 shows the MCP structure and working principle based on secondary electron emission: when incident electron enters the channel, it collides with channel inner wall surface and produces one or more electrons, that further moves down the channel by the electric field and repeatedly re-collides with channel wall, forming high-density electron cloud on output of MCP.

MCP starts as a glass tube, consisting of a core (Boron silicate- soluble glass) and cladding (lead silicate glass) that forms the microchannel structure after drawing into fibers, assembling to bundle, fusing, slicing and polishing to wafers. Next, in a chemical process, the core is removed, leaving the structure of millions of channels. The next step is hydrogen reduction in which a conducting layer is produced and a secondary electron emission layer formed on the surface of the microchannel inner wall. These surface layers can have an electron yield up to 3.5 and surface conductivity of 10^{-2} $(\Omega cm)^{-1}$ [1]. The hydrogen reduction is one of the most important processes in the technology of microchannel production where MCP gets proper electrical conductivity.

A. Stanković is with the Department of Microelectronics, Faculty of Electronic Engineering, University of Niš, Aleksandra Medvedeva 14, 18000 Niš, Serbia (e-mail: alesandra.stankovic@elfak.rs).

I. Zlatković is with the Department of Microelectronics, Faculty of Electronic Engineering, University of Niš, Aleksandra Medvedeva 14, 18000 Niš, Serbia (e-mail: ivanzlatkovic90@hotmail.com).

D. Pantić is with the Department of Microelectronics, Faculty of Electronic Engineering, University of Niš, Aleksandra Medvedeva 14, 18000 Niš, Serbia (e-mail: dragan.pantic@elfak.ni.ac.rs).

R. Nikolov is with Photon Optronics, Bulevar Svetog Cara Konstantina 80-82, 18000 Niš, Serbia (rade.nikolov@photonoptronics.rs).

Branislav Brindić is with Sova HD, Bulevar Svetog Cara Konstantina 80-82, 18000 Niš, Serbia.

Lastly, a contact layer is evaporated to provide electrical contact. Microchannel plates are widely used as detectors of charged particles, in all types of electron microscopy and image intensifier tubs for low light level imaging as electron multipliers. [3-6]

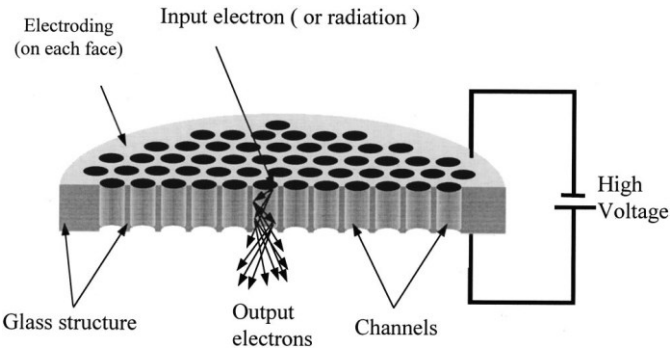


Fig.1. Structure and basic operation of MCP [2]

II. Hydrogen Reduction

Hydrogen reduction (HR) reduces the oxide of lead metal (from the surface of the glass in contact with hydrogen) and creates a thin semiconductor layer on the surface of the MCP channel. In this way, MCP activation is carried out. The electrical surface conductivity increases in several orders of magnitude in the reduction process. Generally, MCP resistance depends on the composition of the glass, the thermal history, and the reduction conditions.

The HR process is done at a given temperature in an atmosphere of hydrogen. Parameters that define the reduction process (temperature and time) depend greatly on the previous thermal processes. The thermal history of the microchannel plate begins with the first draw, and then the second. The next step in production is fusion, where interdiffusion between the core and cladding is achieved, as well as the degassing and fusion of the block at high pressure and temperature. With temperature and time of HR corrections of the desired resistance (80-200 MΩ), amplification and FPN threshold can be achieved.

Layers that are formed after reduction are [7, 8, and 9]:

1. Emission layer 20nm thick, rich in silica and alkali metals. This layer contains electrons that participate in the secondary emission of electrons.
2. A conductive layer that is rich in the lead at a depth of 100-200nm. The conductive layer function is to donate the electrons emitted by the emission layer using a strip current.

In figure 2 the cross-section of the channel wall is shown, we can see a secondary emission layer rich with silicon

dioxide and alkali metals. Alkaline elements are electropositive, which means they easily release electrons. This gives a constant source of electrons for the secondary emission of electrons. This layer determines the MCP gain.

The next layer is a conductive (resistant) layer of 50-100nm thickness. Since here lead glass is used, this layer is rich in lead. The MCP input current is of the order of 10-20 pA and the output is about 300nA (excited electrons are formed). The composition of the glass and the conditions of the reduction process (time and temperature) can control the resistance of the MCP. This resistant layer consists of a cluster (group) of lead that is used to transfer electrons. In the HR process, in the first phase, small lead grains are formed. At temperatures of 450-500°C, grains begin grouping into clusters, and the principle of conducting the current is the principle of skipping the electron from the island to adjacent islands of lead. When the electron gets out from the lattice of potassium hole remains, this place is filled with a lead electron, and then we have that second movement of current called a strip, i.e., MCP current. The most common is the output current of up to 10% of strip current. These are the electrical characteristics of the microchannel plate.

$$R = Umcp / Istrip \quad (1)$$

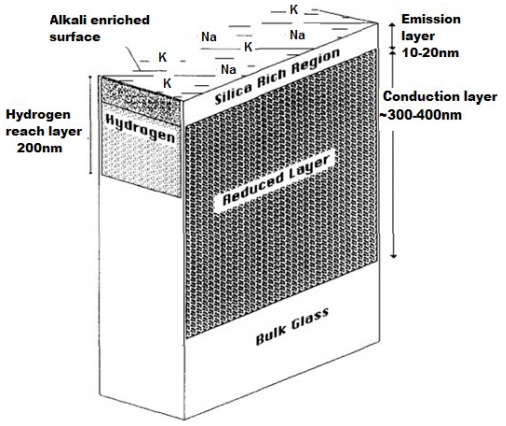


Fig. 2. MCP wall structure. [1]

On the surface of the channels, during the reduction process alkaline metal impurities (carbonates, silicates, hydrates, and chlorides) can occur due to the created conditions for the diffusion of alkaline cations to the surface through the walls. The appearance of cations on the surface is accompanied by the formation of various compounds when cations come into contact with water vapor, hydrogen, and silicates on the channel walls. They can cause excess gas, Ion feedback, and other negative effects in the work of the MCP. A thin film formed after reduction and its elimination by additional etching is shown below.

In order to achieve better results, HR is done in several stages. In the first phase, the water from the channel walls is removed, left after etching in the N_2 atmosphere.

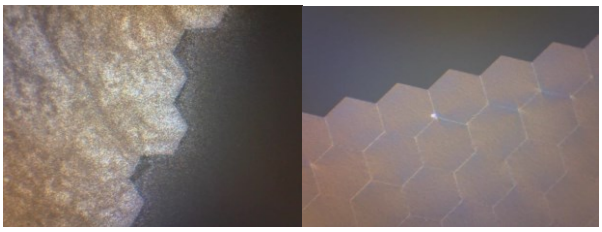


Fig. 3. Thin film layer formed after reduction (left). Removed thin film with additional etching (right).

The following stage implies the reduction of lead from the channel walls, and then the alkaline layer from the surfaces is removed by chemical treatment. The next phase is the removal of residual water (i.e. residual O2) absorbed on the surface. In the end, the formation of the conductive layer is completed, and the removal of the channel water is achieved. This improves the resistance of the MCP and the efficiency of the electron secondary emission. Cooling is performed slowly and with controlled speed to a certain temperature, followed by natural cooling. This method will improve channel purity, reduce gas content and gaseous emission, and improve the quality and reliability of the MCP.

III. Experiment

In this experiment, microchannel plates from the same block were used. Different parameters of reduction were applied. In the first case time was fixed, t=95min, while the temperature ranges between 430°C and 460°C. In the second case, the temperature is fixed, T=449°C, while the reduction time changes in the range of 35min to 155min.

In both experiments, 2-3 plates per day were used with specific reduction parameters (temperature and time) in the furnace. It is very important that the critical temperature Tg of the glass tube is not exceeded in the reduction, otherwise bending or even cracking of the microchannel plate itself can occur.

The values of the temperature and reduction time are chosen so that they are out of the optimal range, in order to create a clearer image of the consequences and characteristics of such treatment on the microchannel plate. We got either too high or too low resistance of the MCP then needed for its normal work. The second parameter that was measured is FPN threshold: it appears in the form of elements and the network. Fixed-pattern noise is usually a cosmetic blemish characterized by a faint hexagonal (honeycomb) pattern that appears throughout the viewing area. Measurements are given in tables 1 and 2, where electrical and electro-optical characteristics of the MCP are shown.

Based on good results of the first experiment related to FPN (showed in table 1) at a temperature of 449°C, the second experiment was performed with HR process at the same temperature, while reduction time was changed in the range from t = 35min to t = 155min (table 2). Figure 4 shows the dependence of the resistance and FPN threshold on the temperature and the time of reduction.

TABLE I

CHARACTERISTIC OF MICROCHANNEL PLATE AFTER REDUCTION: TEMPERATURE CHANGES FROM 430°C TO 460°C, WHILE TIME IS FIXED (95MIN)

	HYDROGEN REDUCTION PROCESS		TEST			
Blank	T(°C)	t (min)	Resistance [MΩ]	Gain [800V]	Gain [100V]	FPN threshold [nA]
MCP-1	430	95	1265.56	299	682	6
MCP-2	430	95	1226.77	444	1283	10
MCP-3	430	95	1172.5	270	632	4
MCP-4	440	95	517.92	789	3390	40
MCP-5	440	95	504.62	664	2911	40
MCP-6	440	95	486.28	198	1716	16
MCP-7	449	95	92.91	974	5111	700
MCP-8	449	95	84.28	983	5340	320
MCP-9	449	95	81.48	1083	5687	700
MCP-10	455	95	28.2	842	4702	409
MCP-11	455	95	25.93	924	5327	/9999
MCP-12	455	95	27.86	852	4690	/9999
MCP-13	460	95	16.62	587	0	178
MCP-14	460	95	16.44	541	0	150

TABLE II

CHARACTERISTICS OF MICROCHANNEL PLATE AFTER REDUCTION: TIME CHANGES FROM 35MIN TO 155MIN, WHILE THE TEMPERATURE IS FIXED (449°C)

	HYDROGEN REDUCTION PROCESS		TEST			
Blank	T(°C)	t (min)	Resistance [MΩ]	Gain [800V]	Gain [100V]	FPN threshold [nA]
MCP-15	449	35	1306.885	328	724	9
MCP-16	449	35	1504.717	342	813	7
MCP-17	449	65	406.79	709	3272	250
MCP-18	449	65	1285.64	0	0	/
MCP-19	449	95	92.91	974	5111	700
MCP-20	449	95	84.28	983	5340	320
MCP-21	449	95	81.48	1083	5687	700
MCP-22	449	125	28.96	940	5129	1000
MCP-23	449	125	26.08	1053	5861	1200
MCP-24	449	155	14.98	0	0	9999
MCP-25	449	155	15.13	0	0	9999

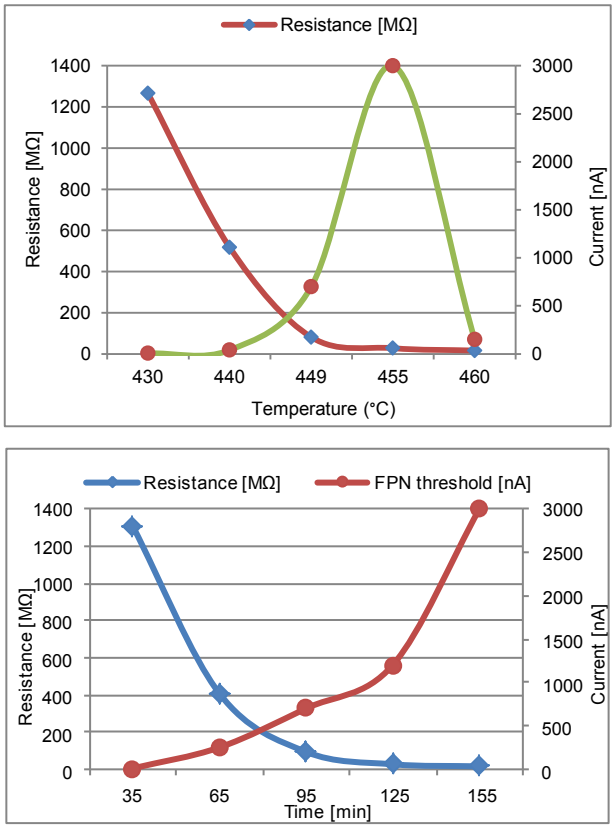


Fig. 4. Graph of resistance and FPN threshold change with the respect of reduction temperature (first graph) and time (second graph).

IV. Results and Conclusion of the Experiment

Observing experiment results of the HR process at higher temperatures, we can say that FPN disappears, i.e. it is not visible in the operating mode (U= 800V, I= 300nA), in contrary to the resistance that is too small. From graphic, we can see that for the different temperature values of 449°C and 459°C the same FPN threshold is reached, while the resistance is quite different (for 449°C, R= 92.91MΩ and 459°C is R= 18.86MΩ). When the HR temperature increases above 440ºC, the FPN threshold curve (green line) rises to a certain value and then falls. The maximum FPN threshold is at a temperature of 455°C while the resistance, R= 27.86MΩ, is outside the desired resistance value. In this experiment, we got optimal resistance and FPN threshold at the temperature of 449°C, based on that the second experiment was done where the time of reduction is changing.

The time-change experiment of the HR at constant temperature shows a clear trend of increasing the FPN threshold, with the growing time of reduction. As HR time increases, there is a significant increase in FPN threshold, it riches its maximum at 155min, but resistance gets too low.

From the above mentioned, we can conclude the following:

- When increasing the temperature and time of the reduction process, the resistance of the MCP decreases, initially significant, and then slower.
- In the initial phase of the HR process, the resistance and threshold of the FPN change simultaneously (and opposite). Intensifying the HR process by increasing the time and temperature, resistance changes gradually, while variations in the FPN threshold can be quite significant.

V. Conclusion

The reduction is a key process that is well controlled to ensure uniformity of gain and an acceptable variation between batches. The plate is reduced several times in order to produce the conductive layer beneath the surface layer. The channel surface is heated in dry hydrogen to reduce the glass-metal oxides. Production of the semi-conductive layer provides a conductive path for the electron secondary emission process. This conductive layer is one of the principal operating features of the microchannel plate. The goal of the reduction process is to create a repeatable layer throughout the channel matrix. In this paper, we examined the influence of reduction on MCP resistance and the FPN threshold. Optimal resistance and FPN threshold were obtained at a reduction temperature of 449°C and a time of 95 minutes.

Acknowledgment

We are grateful to Photon Optronics that have created conditions for performing experiments and measuring characteristics of the microchannel plate.

References

[1] A.M. Then, C.G. Pantano, "Formation and behavior of surface layers on electron emission glasses", *Journal of Non-Crystalline Solids, vol. 120, no. 1-3, pp. 178-187, Apr. 1990.*

[2] Whikun Yi, Taewon Jeong, Sunghwan Jin, SeGi Yu, Jeonghee Lee, and Jungna Heo, "Characteristic features of new electron-multiplying channels in a field emission display", *Journal of Vacuum Science & Technology B Microelectronics and Nanometer Structures, vol. 19, no. 16, pp. 2247-2251, Nov. 2001.*

[3] A. Lehmann, M. Böhm, A. Britting, W. Eyrich, M. Pfaffinger, F. Uhlig, A. Belias, R. Dzhygadlo, A. Gerhardt, K. Götzen, "Recent Developments with Microchannel-Plate PMTs", *Nucl. Instrum. Methods Phys. Res., vol. 876, pp. 42–47, Dec. 2017.*

[4] A. Hans, P. Schmidt, C. Ozga, G. Hartmann, X. Holzapfel, A. Ehresmann, A. Knie, "Extreme Ultraviolet to Visible Dispersed Single Photon Detection for Highly Sensitive Sensing of Fundamental Processes in Diverse Samples", *Materials* 2018, *11*(6), 869.

[5] G. J. Price, "Microchannel plates in astronomy," Ph.D. dissertation, Dept. of Physics and Astronomy, Univ., Leicester, U.K., 2001.

[6] J. L. Wiza, "Microchannel plate detectors", N*ucl. Instrum. Methods, vol. 162, no. 1-3, pp. 587–601, Jun 1979.*

[7] O.M. Kanunnikovaa, F.Z. Gilmutdinov, A.A. Shakov, "Interaction of lead silicate glasses with hydrogen under heating", *International Journal of Hydrogen Energy, vol. 27, no. 7-8, pp. 783-791, July-Avg. 2002.*

[8] Y. Zhang, Y. Sun, J. Wang, K. Huang, Y. Wang, J. Liu, J. Jia, B. Zhang, W. Hou, X. Lv, "Effect of Hydrogen Reduction on Properties of Lead Silicate Glass for Microchannel Plates", *IOP Conf. Ser.: Mater. Sci. Eng., vol. 423, no.1, art. no. 012167, Nov. 2018.*

[9] I. B. Kacem, L. Gautron, D. Coillot, D. R. Neuville, "Structure and properties of lead silicate glasses and melts", *Chemical Geology, Elsevier, vol. 461, pp. 104-114, Jun 2017.*

Application of a Low-Voltage Step-Up Circuit for Thermal Energy Harvesting Under Natural Convection

Jana Vračar, *Student Member, IEEE,* Miloš Marjanović, *Member, IEEE*, Aleksandra Stojković, *Student Member, IEEE*, Zoran Prijić, *Member, IEEE*, Aneta Prijić, *Member, IEEE*, Ljubomir Vračar, *Member, IEEE*

***Abstract*—This paper describes the design and testing of a step-up circuit for thermoelectric energy harvesting application under natural convection conditions. Considered thermoelectric generator with a suitable heatsink at a temperature difference of about $30\,^{\circ}\mathrm{C}$ provides a voltage of $60\,\mathrm{mV}$ at the load. The voltage is boosted using the Meissner oscillator and voltage doubler circuit. For the oscillator circuit, an analytical small signal model for the oscillation frequency estimation has been developed. The experimental characterization of the step-up circuit for thermoelectric energy harvesting was performed. It is shown that the voltage can be boosted up to $6\,\mathrm{V}$ DC.**

***Index Terms*—Step-up circuit, Meissner oscillator, Thermal Energy Harvesting, Natural Convection.**

I. Introduction

Due to industrial progress, the demand for energy consumption is rising exponentially. This will not only lead to the rapid depletion of the fossil fuel reserves but will also increase the problem of global warming. To reduce these problems and keep up with increasing energy demand, it is required to develop clean, efficient and sustainable energy production. Energy harvesting presents an appropriate solution for replacing batteries and power supplies for low power devices. Energy harvesting requires three separate technologies working together. Those are energy conversion, which transforms ambient energy into electrical energy, power management which boosts and regulates the generated energy, and energy storage [1]. Sources for energy harvesting are mainly solar radiation, temperature difference, pressure, kinetics, vibrations, or magnetic induction. In this paper, we will consider a thermoelectric generator (TEG) which, due to temperature difference, creates a voltage.

The voltage obtained by converting thermal energy into electrical one needs to be increased and stored so that it can be used to power some complex systems, such as wireless sensor network nodes. For this purpose, low voltage boost converter circuits for energy harvesting, based on Meissner oscillators [2], have been developed. These circuits are realized on a printed circuit board [3] or in integrated techniques [4]. Garcha et al. in [5] reported the lowest voltage integrated electrical startup system ($25\,\mathrm{mV}$) with on-chip magnetic elements. In [6], Macrelli et al. present design of step-up oscillator integrated circuit fabricated in $0.32\,\mu\mathrm{m}$ technology with two bond wire microtransformers with different cores. Simplified configuration of step up integrated circuit LTC3108 [7] is described in [8]. Some studies are about the low voltage step-up circuits for thermoelectric [9], [10], solar [11], mechanical [12], [13] or RF [14] energy harvesting.

In this paper, a low voltage step-up circuit for thermal energy harvesting will be presented. The characterization of an assembly incorporating a commercial TEG and a heatsink was performed in order to investigate the influence of the heatsink design on the value of the generated voltage. This voltage is sufficient to drive the oscillator circuit whose output voltage is further increased by a voltage doubler circuit. Obtained rectified voltage is stored in a large capacitor and it can be used for powering of other elements of the wireless sensor network node. The electric circuit of oscillator and voltage doubler was realized using discrete commercial electronic devices.

II. Thermoelectric Generator with Heatsink under Natural Convection

The temperature gradient from the environment has the potential to generate electrical power using a thermoelectric generator. This is a solid-state device whose working principle is based on the Seebeck effect. TEG is made of N thermoelectric couples consisting of n-type (containing free electrons) and p-type (containing free holes) box-shaped elements wired electrically in series and thermally arranged in parallel. When a temperature gradient (ΔT) is applied to the opposite sides of the TEG, the mobile charge carriers at the hot side diffuse to the cold side of the TEG. The buildup of charge carriers

Jana Vračar is with Department of Microelectronics, University of Niš, Faculty of Electronic Engineering, Aleksandra Medvedeva 14, 18000 Niš, Serbia, e-mail: jana.vracar@elfak.ni.ac.rs

Miloš Marjanović is with Department of Microelectronics, University of Niš, Faculty of Electronic Engineering, Aleksandra Medvedeva 14, 18000 Niš, Serbia, e-mail: milos.marjanovic@elfak.ni.ac.rs

Aleksandra Stojković is with Department of Microelectronics, University of Niš, Faculty of Electronic Engineering, Aleksandra Medvedeva 14, 18000 Niš, Serbia, e-mail: aleksandra.stojkovic@elfak.ni.ac.rs

Zoran Prijić is with Department of Microelectronics, University of Niš, Faculty of Electronic Engineering, Aleksandra Medvedeva 14, 18000 Niš, Serbia, e-mail: zoran.prijic@elfak.ni.ac.rs

Aneta Prijić is with Department of Microelectronics, University of Niš, Faculty of Electronic Engineering, Aleksandra Medvedeva 14, 18000 Niš, Serbia, e-mail: aneta.prijic@elfak.ni.ac.rs

Ljubomir Vračar is with Department of Microelectronics, University of Niš, Faculty of Electronic Engineering, Aleksandra Medvedeva 14, 18000 Niš, Serbia, e-mail: ljubomir.vracar@elfak.ni.ac.rs

results in a net charge at the cold side, producing the open–circuit voltage:

$$V_{TEGoc} = N\alpha\Delta T \ , \tag{1}$$

where α is Seebeck coefficient. This voltage can power an electric load through the external circuit. The commercial TEGs are usually made of doped Bi–Te semiconductor alloys having thermoelectric properties.

In our work commercial thermoelectric module GM200-127-14-16 is utilized as a generator [15]. The selected TEG has the following characteristics - dimensions: (40×40×3.4) mm, number of thermoelectric couples $N = 127$, internal electrical resistance $R_{TEG}(25\,^{\circ}\mathrm{C}) = 2.1\,\Omega$. For proper operation of the TEG the rejected heat must be removed from its cold side through a heatsink. Two types of low profile heatsinks are considered: aluminum heatsink [16] and metal (copper) foam heatsink [17]. The thermal resistance of aluminum heatsink is $9.38\,\mathrm{K/W}$, while for metal foam heatsink this value is $17.4\,\mathrm{K/W}$. The overall dimensions of both heatsinks are the same (40×40×5) mm. The aluminum heatsink has pin fin geometry, which is preferable for natural convection, while the metal foam heatsink has a form of a rectangular prism.

An aluminum heatsink is the most widely used type of heatsinks. The properties of the aluminum heatsink are excellent thermal and electrical conductivity, low weight, excellent corrosion resistance and no magnetization effect, which avoids interference of magnetic fields. The aluminum heatsink weighs approximately half as much as a copper one having the same thermal conductivity. Anodizing of the heatsink surfaces improves the strength of the natural corrosion protection.

The characteristics of the metal foam heatsink are determined by the geometry of metal foam cell structure, purity, and ductility of the metal. Low profile metal foam heatsink uses micro-porous copper foam where the interconnected pores of the foam create a large surface area. A thin hard layer of copper oxide improves the emissivity of the foam. Studies show that metal foam heatsinks offer excellent thermal performance due to their extremely high specific surface area and thermal conductivity. This is particularly important for applications which demand low volume and weight of the heatsink [18].

The mechanism of thermal energy transfer through the TEG and heatsink is conduction, while transfer of the heat from their surfaces is performed by convection and radiation. Convection is the transfer of the heat Q from a hot surface to an ambient at a lower temperature:

$$Q = h_c A(T_s - T_{amb}) \ , \tag{2}$$

where h_c is heat transfer coefficient, A is surface area, T_s is surface temperature and T_{amb} is ambient temperature. In this paper, thermal energy harvesting under natural convection conditions is considered. Natural convection is the occurrence of the air flow induced by buoyant forces. It arises from different densities of the fluid, due to the temperature variations. According to experimental research, approximately $70\,\%$ of the heat is transferred by natural convection [19]. Radiation transfers the heat between two surfaces at different temperatures in the form of electromagnetic waves.

The experimental setup, as illustrated in Fig. 1, includes cardboard box as a domain boundary, hot chuck as a heater for TEG, thermoelectric generator with heatsink and voltage step-up (boost) circuit as a device under test (DUT), oscilloscope, digital multimeter, and two PT100 temperature sensors. One PT100 sensor measures the temperature of the hot side of TEG T_{hot} and the other measures the ambient temperature. When starting the experiment, hot chuck is set to a defined temperature and appropriate voltage values are measured continuously. The voltages generated by the TEG and oscillator, as well as the output voltage are acquired by digital multimeter Agilent 34410A and through a digital oscilloscope Tektronix DPO4034. A photograph of the experimental setup located inside the domain is given in Fig. 2.

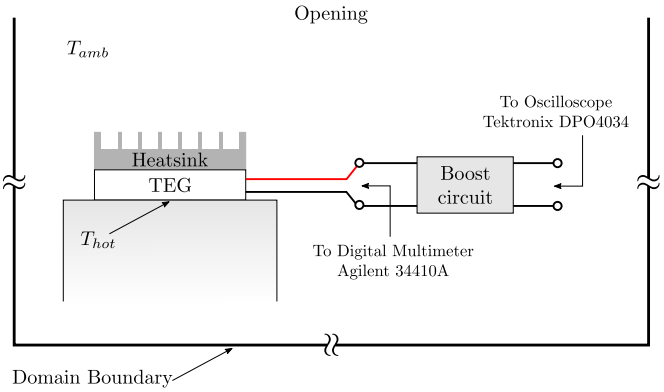


Fig. 1. Ilustration of the experimental setup. Dimensions not to scale.

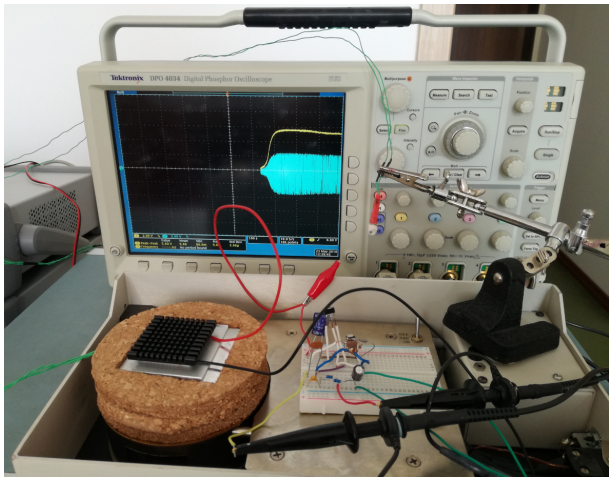


Fig. 2. Photograph of the experimental setup located inside the domain.

Prior to the experimental measurements for the step-up circuit, characterization of the assembly TEG–heatsink was performed in order to estimate values of the temperature difference needed for the generation of sufficient input voltage. The dependence of the open circuit voltage of the TEG (V_{TEGoc}) and the voltage provided by the TEG when loaded with $4\,\Omega$ resistance (V_{TEG}) as a function of temperature difference $\Delta T_{amb} = T_{hot} - T_{amb}$ for two different types of heatsinks is given in Fig. 3. In the temperature range of the interest, the open circuit voltage, as well as, the voltage under load increase

linearly with an increase of the temperature difference. The TEG with the aluminum heatsink generates higher voltages than the TEG with the metal foam heatsink. This is as expected considering thermal resistance values of the two heatsinks. For the TEG under load, the provided voltage is considerably lower than for open circuit condition due to the voltage drop at the internal resistance of the TEG and Peltier effect [20]. Also, the difference between voltages for the two heatsinks is less pronounced.

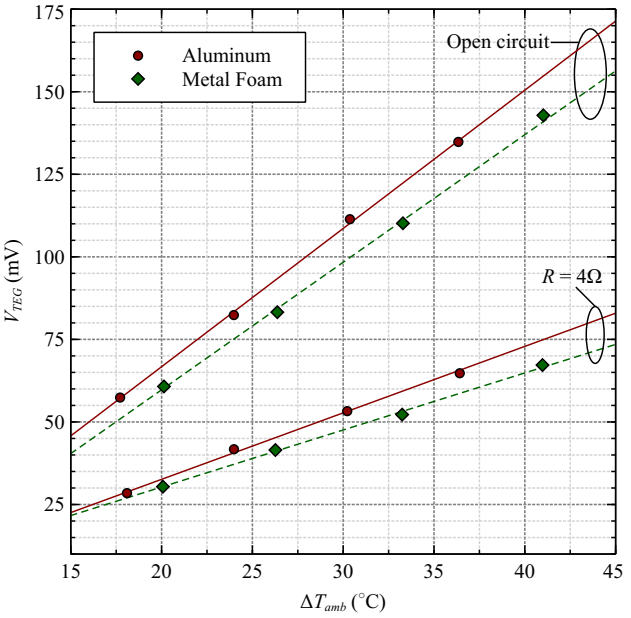


Fig. 3. Open circuit TEG voltage and loaded TEG voltage as a function of the temperature difference for an aluminum heatsink and a metal foam heatsink.

III. Design of Low-Voltage Step-Up Circuit

The most important requirements for step-up converter are as small as possible starting input voltage and as high as possible the power conversion efficiency. The voltage step-up ratio must be high enough to supply connected electronic circuitry with a required voltage value. Self-powered start-up from the applied low input voltages should ensure operation under worst case conditions, without any auxiliary power from a battery. The designed step-up converter is based on a FET-tuned oscillator with self-start up capability. The circuit relies on a Meissner-type oscillator that consists of the step-up transformer and a depletion mode n-type MOSFET [21]. A voltage doubler [22] is also included. The schematic of the circuit is shown in Fig. 4.

The depletion mode n-type MOSFET BSP149 (M_1) is chosen since it is in normally-on state at the considered low voltages due to its negative threshold voltage V_{th} [23]. The second advantage of this transistor is its small drain-source on-state resistance which ranges from $1.7\,\Omega$ to $3.5\,\Omega$. The charge pump capacitor C_1, which is an integral part of the voltage doubler circuit, has an effect on maximum output current capability. A minimum value of $1\,\text{nF}$ is recommended when operating from very low input voltages using the transformer with a ratio of $1:100$ [7]. Too large capacitor value can

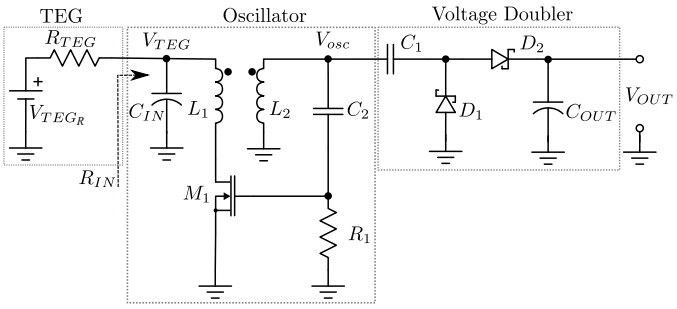


Fig. 4. Schematic of the low-voltage step-up circuit for thermal energy harvesting.

compromise circuit performance when operating at low input voltage or with high resistance sources. In this circuit, the ceramic capacitor of $C_1 = 100\,\text{nF}$ was chosen. Capacitor C_2 acts as a gate coupling capacitor. The value of this capacitor is chosen so that the oscillations start very quickly after the input voltage reaches $60\,\text{mV}$. The experimentally obtained value is $C_2 = 4.5\,\text{nF}$. Resistor R_1 provides a stable start of oscillations in the circuit. The set value of this resistor is $R_1 = 2.2\,\text{M}\Omega$. The Schottky diodes ($D_1$ and D_2) are used for the realization of the voltage doubler. These diodes are better in comparison with silicon diodes due to the lower forward voltage. For storing the harvested energy, the electrolytic capacitor C_{OUT} of $100\,\mu\text{F}$ is used.

The choice of transformer is particularly important in the design process of the circuit. The major features of the transformer for low voltage step-up oscillators are high quality factor Q, small footprint area and high turns ratio. The commercial miniature toroidal microtransformers with ferrite and magnetic low temperature co-fired ceramic cores are considered as one of the best choices for energy harvesting applications [6]. In this case, a commercial microtransformer with a ratio of $1:100$ was chosen. When the primary winding of a transformer is energized, a current flows through it. This current creates a magnetizing force H that produces the magnetic flux Φ of density B in the transformer core. When magnetizing force is increased from zero, the flux density increases up to a certain maximum value. Above this level, further increases of H results in no significant increase in B because magnetic material is saturated [24]. An ideal transformer has zero winding and core losses and unity coupling coefficient. In practice, all magnetic materials, once magnetized, retain some of their magnetization B_r even when the H is reduced to zero. The magnetizing force which must be applied to null the residual flux density is a coercive force H_c. Generally, microtransformers can experience large core losses due to eddy currents and hysteresis. The hysteresis losses result from the additional power needed to reverse the magnetic field in magnetic materials in the presence of alternating current.

The voltage source (V_{TEG}) that imposes the current through the primary winding L_1 and the normally on depletion mode MOSFET is connected to the step-up oscillator. This current induces a positive voltage at the secondary winding L_2 which increases the gate voltage of M_1 and thus primary winding current. When this current reaches the core saturation,

the secondary winding voltage starts to drop. The transistor starts turning off the current through the primary winding of the transformer which leads to reverse of the voltage at the secondary winding and the negative gate voltage. The transistor is driven near its off state, leading to a decrease of primary winding current, and thus, through the coupling of the transformer, less negative value of gate voltage. The transistor is turned on quickly and remains to conduct until the primary winding current approaches saturation so that the oscillation process starts again.

The signal from the step-up oscillator (V_{osc}) is fed to a half-wave voltage doubler. The voltage doubler circuit is composed of two sections: a clamp formed by C_1 and D_1, and a peak rectifier formed by D_2 and C_{OUT}. During the negative half cycle of an input voltage signal, diode D_1 conducts, charging capacitor C_1 up to the voltage $V_{C1} = V_{D1} - V_{OSC}$. Note that V_{OSC} marks amplitude of oscillator output voltage V_{osc}. During the positive half cycle of the input signal, diode D_1 is cut off and diode D_2 conducts charging capacitor C_{OUT}. If we adopt $V_{D1} = V_{D2} = V_D$, the output voltage will be:

$$V_{OUT} = V_{OSC} - V_{C1} - V_{D2} = 2(V_{OSC} - V_D) \; . \tag{3}$$

If diodes act as a short circuit in on state, the voltage across the capacitor C_{OUT} will discharge through the load during the negative half cycle at the input, and the capacitor is recharged up to $2V_{OSC}$ during the positive half cycle.

The small signal model of the step-up oscillator is shown in Fig. 5. The major parameters of the transistor and transformer that affect the oscillation condition are: g_m – transconductance of the MOS transistor, C_{gs} – gate-source capacitance of the MOS transistor, L_1 – primary inductance of the transformer, L_2 – secondary inductance of the transformer, R_{S1} – serial resistance of the primary winding, R_{S2} – serial resistance of the secondary winding, R_1 – external resistance, C_2 – external capacitance, M – mutual inductance of the transformer, k – coupling factor between the primary and secondary windings, where

$$M = k\sqrt{L_1 L_2} \; . \tag{4}$$

Note that $s = j\omega$. The current through the primary windings of the transformer, i.e. through the drain of the transistor is:

$$i_1 = g_m v_g \; . \tag{5}$$

Based on Kirchhoff's law, for the circuit in Fig. 5 can be written:

$$\frac{sMg_m v_g - v_g}{R_{s2} + sL_2 + \frac{1}{sC_2}} - \frac{v_g}{R_1} - v_g s C_{gs} = 0 \; . \tag{6}$$

By solving Eq. (6) we obtain:

$$-\omega C_2 + j\omega^2 C_2 M g_m = \frac{\omega R_{s2} C_2}{R_1} - \omega C_{gs}(\omega^2 L_2 C_2 - 1) + j(\omega^2 C_{gs} R_{s2} C_2) + j\frac{\omega^2 L_2 C_2 - 1}{R_1} \; . \tag{7}$$

The resonant frequency will be obtained from the condition that the real part of Eq. (7) is equal to zero:

$$\omega^2 C_{gs} C_2 L_2 - \frac{R_{s2} C_2}{R_1} - C_{gs} - C_2 = 0 \; . \tag{8}$$

By solving Eq. (8), and taking into account that $\omega = 2\pi f$, we get oscillation frequency f as:

$$f = \frac{1}{2\pi}\sqrt{\frac{R_{s2}C_2 + R_1 C_{gs} + R_1 C_2}{R_1 C_{gs} L_2 C_2}} \; . \tag{9}$$

The LCR meter Agilent 4284A was used to obtain necessary

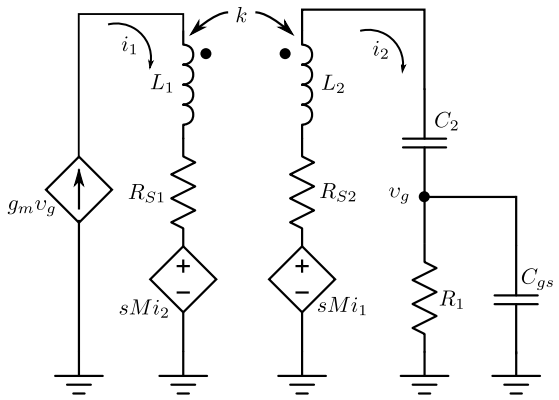


Fig. 5. Small signal equivalent model of the step-up oscillator.

data for analytical analysis of the oscillator circuit. Measured inductance and resistance values are shown in Table I. Based on MOSFET BSP149 datasheet [23] the value of the input capacitance C_{gs} is 850 pF. By replacing parameter values in Eq. (9), the frequency of the oscillator 129.9 kHz is calculated.

TABLE I
MEASURED PARAMETER VALUES OF THE TRANSFORMER AT 1 kHz

Transformer parameter	Measured value
Primary inductance - L_1	0.2 µH
Serial resistance of the primary winding - R_{S1}	7 mΩ
Secondary inductance - L_2	2.1 mH
Serial resistance of the secondary winding - R_{S2}	5.2 Ω

IV. RESULTS AND DISCUSSION

The voltage generated by the TEG with aluminum heatsink, the oscillator voltage and the voltage at the output capacitor as a function of time are presented in Fig. 6. The shown experimental results were obtained for the TEG hot side temperature $T_{hot} = 56.8\,^{\circ}\mathrm{C}$ and the ambient temperature $T_{amb} = 24.5\,^{\circ}\mathrm{C}$. In this case, for temperature difference of $\Delta T_{amb} = 32.3\,^{\circ}\mathrm{C}$, voltage of $V_{TEG} = 59\,\mathrm{mV}$ is generated. Immediately upon establishing of the voltage, the oscillations and the increase of the output voltage start. After 230 s the output voltage of 3.3 V is reached. The analysis shows that after 300 s TEG is in steady state and the output capacitor is charged to a value of 6.2 V. The experimentally obtained output voltage value is in accordance with Eq. (3), since the oscillator voltage amplitude in the positive half cycle is about 3 V.

The characteristic voltage values obtained for the TEG with metal foam heatsink resemble the ones shown in Fig. 6. However, it requires a slightly higher temperature difference ($\Delta T_{amb} = 35.9\,^{\circ}\mathrm{C}$) than the TEG with the aluminum heatsink to achieve the same voltage that drives the oscillator

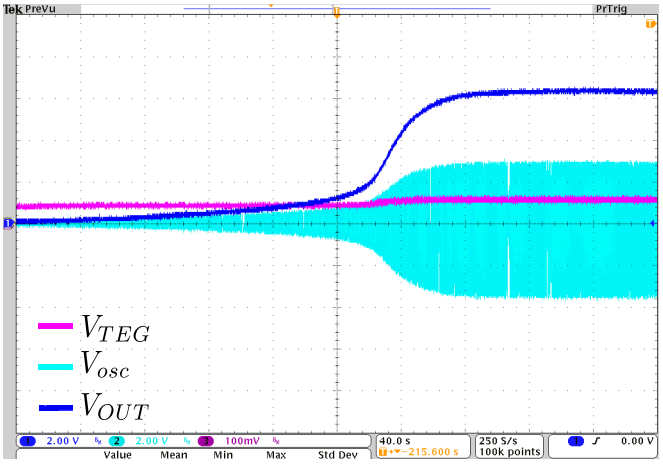


Fig. 6. Voltage generated by the TEG with aluminum heatsink, the oscillator voltage and the voltage at the output capacitor as a function of time. X-axis: 40 s/div, Y-axis: 2 V/div (for V_{osc} and V_{OUT}); 100 mV/div (for V_{TEG}).

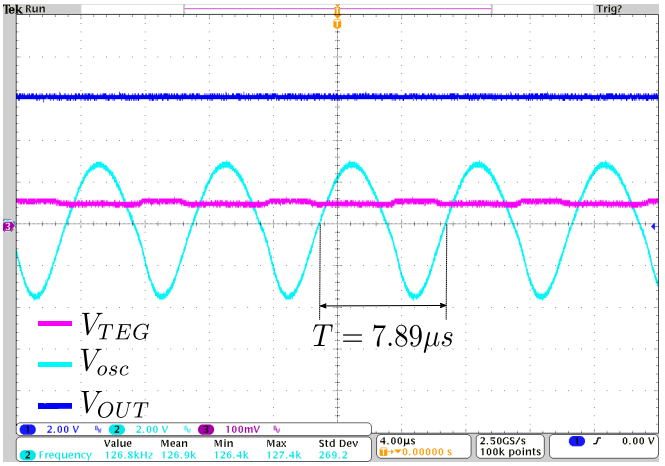


Fig. 7. Zoom In - Voltage generated by the TEG with aluminum heatsink, the oscillator voltage and the voltage at the output capacitor as a function of time. X-axis: 4 µs/div, Y-axis: 2 V/div (for V_{osc} and V_{OUT}); 100 mV/div (for V_{TEG}).

circuit. In this case, the necessary hot side temperature is $T_{hot} = 60.4\,°\mathrm{C}$ for the same ambient temperature. These results are in compliance with characterization data for the two TEG–heatsink assemblies presented in Fig. 3. It can be concluded that metal foam heatsinks can be also incorporated into the design of thermal harvesting systems under appropriate thermal conditions. Even though their cooling performances are somewhat lower than that of aluminum heatsinks, their advantages are compact form, ultra low profile, and low weight.

Based on the data from Fig. 3, the value of the input resistance of the step-up circuit R_{IN} while charging output capacitance is estimated to be approximately $4\,\Omega$. Analysis of oscillator voltage using the oscilloscope (Fig. 7) found that the frequency of oscillation is 126.9 kHz. It can be concluded that the results obtained from the analytical model and experiments are in excellent agreement.

Considering the power conversion efficiency, presented circuitry enables around 20% of the input electrical power obtained by the TEG to be transferred to the output capacitor. Having in mind that the commercial integrated step-up converter and power manager like LTC3108 [7] has efficiency in the range 5–40% depending on the input voltage value, our obtained value is quite satisfactory.

During the analysis of the circuit, it was noticed that the oscillation stops for a period of time much longer than the period of oscillations (Fig. 8). This phenomenon is called squegging [7]. It occurs when a charge builds up on the capacitor C_2, such that DC bias point shifts and oscillations are extinguished for a certain period of time. When the charge on the capacitor leaks out, oscillations resume. It is difficult to predict when this phenomenon will occur. While squegging is not harmful, it reduces the average output current capability. Squegging can easily be avoided by the addition of a leakage resistor in parallel with C_2 capacitor. Resistor values in the range of $100\,\mathrm{k}\Omega$ to $1\,\mathrm{M}\Omega$ are sufficient to eliminate squegging without having any negative impact on the performance.

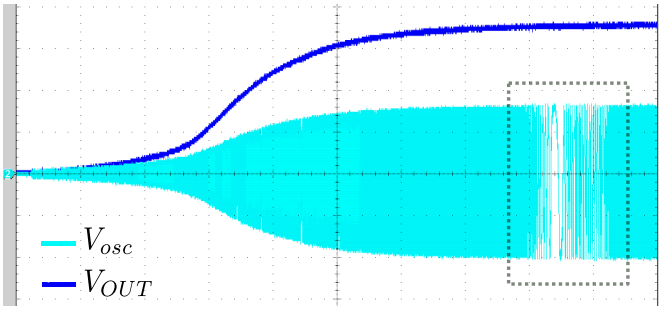


Fig. 8. Illustration of the squegging phenomenon.

V. Conclusion

This paper describes low-voltage step-up circuit design for thermal energy harvesting application under natural convection. Analytically determined frequency of the Meissner oscillator, based on the derived model for small signals, showed a good agreement with experimentally obtained results. By analyzing the voltage vaweforms at characteristic points in the circuit, it was found that at hot side–ambient temperature difference of about 30 °C, TEG with heatsink generates a voltage sufficient to start the oscillator (about 60 mV). After about 4 min, the output voltage at the constant temperature difference can be boosted to 3.3 V. The maximum obtained DC voltage is about 6 V. This circuit can be used to boost low voltage from different energy harvesting sources.

Acknowledgement

This work was supported in part by the Serbian Ministry of Education, Science and Technological Development under Grant TR32026 and in part by Ei PCB Factory, Niš, Serbia.

References

[1] J. K. Hart and K. Martinez, "Environmental sensor networks: A revolution in the earth system science," *Earth-Science Reviews*, vol. 78, no. 3-4, pp. 177–191, oct 2006.

[2] P. Garcha, "Fully integrated ultra low voltage cold start system for thermal energy harvesting," Master's thesis, Massachusetts Institute of Technology, 2016.

[3] P. Woias, M. Islam, S. Heller, and R. Roth, "A low-voltage boost converter using a forward converter with integrated Meissner oscillator," *Journal of Physics: Conference Series*, vol. 476, p. 012081, dec 2013.

[4] P.-H. Chen, K. Ishida, K. Ikeuchi, X. Zhang, K. Honda, Y. Okuma, Y. Ryu, M. Takamiya, and T. Sakurai, "Startup techniques for 95 mV step-up converter by capacitor pass-on scheme and Vth-tuned oscillator with fixed charge programming," *IEEE Journal of Solid-State Circuits*, vol. 47, no. 5, pp. 1252–1260, may 2012.

[5] P. Garcha, D. El-Damak, N. Desai, J. Troncoso, E. Mazotti, J. Mullenix, S. Tang, D. Trombley, D. Buss, J. Lang, and A. Chandrakasan, "A 25 mV-startup cold start system with on-chip magnetics for thermal energy harvesting," in *ESSCIRC 2017 - 43rd IEEE European Solid State Circuits Conference*. IEEE, sep 2017, pp. 127–130.

[6] E. Macrelli, A. Romani, R. P. Paganelli, A. Camarda, and M. Tartagni, "Design of low-voltage integrated step-up oscillators with microtransformers for energy harvesting applications," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 62, no. 7, pp. 1747–1756, jul 2015.

[7] *LTC3108 ultralow voltage step-up converter and power manager, Linear Technology Corporation, 2010, data sheet.* [Online]. Available: https://www.infineon.com

[8] Z. Prijić, L. Vračar, and A. Prijić, "Design and characterization of thermoelectric energy harvesting systems for wireless sensor network nodes," in *Proc. 5th International Conference on Electrical, Electronic and Computing Engineering – IcETRAN*, 2018, pp. 930–936.

[9] N. V. Desai, Y. Ramadass, and A. P. Chandrakasan, "A bipolar ±40 mV self-starting boost converter with transformer reuse for thermoelectric energy harvesting," in *Proceedings of the 2014 international symposium on Low power electronics and design - ISLPED '14*. ACM Press, 2014, pp. 221–226.

[10] P. Woias, "Thermoelectric energy harvesting from small and variable temperature gradients," 2015.

[11] A. Shrivastava, N. E. Roberts, O. U. Khan, D. D. Wentzloff, and B. H. Calhoun, "A 10 mV-input boost converter with inductor peak current control and zero detection for thermoelectric and solar energy harvesting with 220 mV cold-start and -14.5 dBm, 915 MHz RF kick-start," *IEEE Journal of Solid-State Circuits*, vol. 50, no. 8, pp. 1820–1832, aug 2015.

[12] S. Ahmed, K. Mamun, A. Barua, J. Sikder Joy, and R. Chakma, "Design & implementation of controller based buck-boost converter for small wind turbine," *IOSR Journal of Electrical and Electronics Engineering (IOSR-JEEE)*, vol. 10, pp. 44–50, 11 2015.

[13] M. Arifujjaman, M. Iqbal, J. Quaicoe, and M. Khan, "Modeling and control of a small wind turbine," in *Canadian Conference on Electrical and Computer Engineering*. IEEE, 2005., pp. 778–781.

[14] S.-E. Adami, V. Marian, N. Degrenne, C. Vollaire, B. Allard, and F. Costa, "Self-powered ultra-low power DC-DC converter for RF energy harvesting," in *2012 IEEE Faible Tension Faible Consommation*. IEEE, jun 2012.

[15] *Thermoelectric generator module GM200-127-14-16, Adaptive, datasheet.* [Online]. Available: https://www.europeanthermodynamics.com/products/datasheets/6-GM200-127-14-16%20(2).pdf

[16] *CCI11 - Heat Sink, BGA, FPGAs, datasheet.* [Online]. Available: http://www.farnell.com/datasheets/317989.pdf

[17] *Low Profile Metallic Foam Heat Sinks, Versarien Technologies, 2015, datasheet.* [Online]. Available: http://www.versarien-technologies.co.uk/wp-content/uploads/2015/04/VTL-LowProfileHeatsink-DATASHEET-MARCH2015fv.pdf

[18] T. ur Rehman, H. M. Ali, A. Saieed, W. Pao, and M. Ali, "Copper foam PCMs based heat sinks: An experimental study for electronic cooling systems," *International Journal of Heat and Mass Transfer*, vol. 127, pp. 381–393, dec 2018.

[19] *Heat Sink Design Facts and Guidelines for Thermal Analysis, technical brief, Wakefield-vette*. [Online]. Available: https://www.digikey.com/en/pdf/w/wakefield-thermal-solutions/heat-sink-design-for-thermal-analysis

[20] S. Dalola, M. Ferrari, V. Ferrari, M. Guizzetti, D. Marioli, and A. Taroni, "Characterization of thermoelectric modules for powering autonomous sensors," *IEEE Journal of Instrumentation and Measurement*, vol. 58, pp. 99–107, Jan. 2009.

[21] J. Damaschke, "Design of a low-input-voltage converter for thermoelectric generator," *IEEE Transactions on Industry Applications*, vol. 33, no. 5, pp. 1203–1207, 1997.

[22] A. Sedra and K. Smith, *Microelectronic circuits*. Oxford University Press, 2004.

[23] *BSP149 Small-Signal-Transistor, Infineon, 2012, data sheet.*, Infineon, 2012. [Online]. Available: http://www.linear.com

[24] M. E. Schultz, *Grob's basic electronics*. McGraw-Hill, 2011.

MICRO ELECTROMECHANICAL SYSTEMS (MEMS) BASED MICROFLUIDIC PLATFORMS (INVITED PAPER)

Dana Vasiljević-Radović, ICTM CMT, University of Belgrade, Serbia
Milena Rašljić, ICTM CMT, University of Belgrade, Serbia
Milče Smiljanić, ICTM CMT, University of Belgrade, Serbia
Žarko Lazić, ICTM CMT, University of Belgrade, Serbia
Katarina Radulović, ICTM CMT, University of Belgrade, Serbia
Katarina Cvetanović-Zobenica, ICTM CMT, University of Belgrade, Serbia

ABSTRACT

In this work an overview of Micro Electromechanical Systems (MEMS)-based microfluidic platforms for different applications is presented. Microfluidics refers to a set of technologies that control the flow of liquids or gases through miniaturized systems in typical amounts of nano- and pico- liters. Microfluidic devices are characterized by microchannels with characteristic dimensions in the micrometer range. The main techniques, technologies and materials used for fabrication of MEMS microfluidic devices and systems are presented. The used materials and their properties are very important for the final characteristics and functionalities of devices. As an example, the design and fabrication of our opto-fluidic lab-on-a-chip device based on silicon and pyrex glass is given.

Analysis of the Fundamental Detection Limit in Microfluidic Chemical and Biological Sensors

Ivana Jokić, Katarina Radulović, Miloš Frantlović, Zoran Djurić, Katarina Cvetanović Zobenica, Predrag M. Krstajić

***Abstract*—Detection limits in microfluidic chemical and biological sensors, which determine the range of analyte concentrations reliably detectable by the sensor, are important sensor parameters. The lower limit of detection, defined as the lowest concentration that can be distinguished from noise, has its minimum determined by the fundamental adsorption-desorption (AD) noise, inevitable in adsorption-based devices. In this work, we analyze this fundamental detection limit, particularly considering the influence of mass transfer processes in microfluidic devices. For that purpose, we derive the expression for the sensor's signal-to-noise ratio (SNR), which takes into account the AD noise, and then the equation for the minimal analyte concentration at which the SNR has a sufficiently high value for reliable analyte detection. Subsequently, we analyze the mass transfer influence on the sensor's maximal achievable signal-to-noise ratio and on the fundamental detection limit. The results of the analysis show a significant mass transfer influence on these important sensor performance metrics. They also provide guidelines for achieving the sensor's best possible detection performance through the optimization of the sensor design and operating conditions.**

***Index Terms*—Microfluidic sensor; biosensor; chemical sensor; detection limit; mass transfer; signal-to-noise ratio.**

I. Introduction

Detection of chemical substances or biological species (viruses, bacteria, DNA, proteins, cells), and measurement of their concentration in samples are performed in order to monitor the pollution present in the environment, food and water, or the health condition of living organisms. Therefore, their significance in environment protection, agriculture, food industry and healthcare is high. In these fields it is especially important to obtain reliable measurement results in a short period of time, and often at locations which are far from laboratories where conventional sample analysis can be performed. The concentrations to be detected are typically very low, requiring very sensitive measurement equipment and methods. Current research activities are aimed toward the development of highly sensitive and portable chemical and biological sensors based on micro- and nanotechnologies, which are recognized as promising for the mentioned applications. Among them are microfluidic adsorption-based sensors [1, 2].

In microfluidic devices a sensing element is placed in a reaction chamber, which is a part of a microfluidic system for sample delivery and analyte detection. The principle of operation of adsorption-based chemical and biological sensors is based on the adsorption-desorption (AD) process of analyte particles (gas atoms or molecules, biomolecules or micro-organisms) on the sensor's sensing surface, which causes a measurable change of some of the sensor's parameters, and yields the sensor's output signal. Therefore, the output signal is determined by the number of adsorbed analyte particles, which depends on the analyte concentration in the analyzed sample.

The instantaneous number of adsorbed particles depends on AD and mass transfer (convection and diffusion) processes of analyte particles, by which they are transported through the sensor's reaction chamber to and from the adsorption sites on the sensing surface where they bind and unbind. The random nature of these coupled processes results in inevitable stochastic fluctuations of the number of adsorbed particles, and therefore in stochastic fluctuations of the sensor's response, known as AD noise. This fundamental noise sets the ultimate noise performance of adsorption-based sensors, determining the maximal achievable signal-to-noise ratio (SNR), and the minimal detectable analyte concentration. The AD noise analysis is, therefore, significant for optimization of adsorption-based sensors in terms of improved sensing performance. It is based on the stochastic analysis of fluctuations of the number of adsorbed analyte particles.

Stochastic mathematical models for the analysis of fluctuations of the number of adsorbed particles that take into account mass transfer processes, are rare [3-5]. While those presented in [3, 4] consider coupled AD process and diffusion of analyte particles, the stochastic model presented in [5] takes into account the coupling of AD process, diffusion and

Ivana Jokić is with the Institute of Chemistry, Technology and Metallurgy – Center of Microelectronic Technologies, University of Belgrade, Njegoševa 12, 11000 Belgrade, Serbia (e-mail: ijokic@nanosys.ihtm.bg.ac.rs).

Katarina Radulović is with the Institute of Chemistry, Technology and Metallurgy – Center of Microelectronic Technologies, University of Belgrade, Njegoševa 12, 11000 Belgrade, Serbia (e-mail: kacar@nanosys.ihtm.bg.ac.rs).

Miloš Frantlović is with the Institute of Chemistry, Technology and Metallurgy – Center of Microelectronic Technologies, University of Belgrade, Njegoševa 12, 11000 Belgrade, Serbia (e-mail: frant@nanosys.ihtm.bg.ac.rs).

Zoran Djurić is with the Serbian Academy of Sciences and Arts, and the Institute of Technical Sciences of SASA, Knez Mihailova 25, 11000 Belgrade, Serbia (e-mail: zoran.djuric@itn.sanu.ac.rs).

Katarina Cvetanović Zobenica is with the Institute of Chemistry, Technology and Metallurgy – Center of Microelectronic Technologies, University of Belgrade, Njegoševa 12, 11000 Belgrade, Serbia (e-mail: katarina@nanosys.ihtm.bg.ac.rs).

Predrag Krstajic is with the Institute of Chemistry, Technology and Metallurgy – Center of Microelectronic Technologies, University of Belgrade, Njegoševa 12, 11000 Belgrade, Serbia (e-mail: pkrstajic@nanosys.ihtm.bg.ac.rs).

convection, which corresponds to realistic operating conditions in microfluidic sensors. In the mentioned literature the stochastic models were used for the analysis of the expected value and variance of the number of adsorbed particles, which reveal the stochastic response kinetics and sensor AD noise, as well as for the analysis of the maximal possible sensor's SNR. The influence of mass transfer on sensor's SNR and detection limit has not been quantitatively analyzed in particular until now.

In this work, starting from the stochastic response model of adsorption-based sensors, which considers the influence of coupled stochastic AD and mass transfer processes on the change of the number of adsorbed particles [5], we derive the expression for the sensor's SNR that takes into account the AD noise. Subsequently we obtain the equation for determination of the minimal analyte concentration at which the SNR has a sufficiently high value for reliable analyte detection. After that, we analyze the SNR and the minimal concentration level of the analyte in microfluidic chemical and biological sensors, considering the influence of mass transfer of particles on these important sensor performance metrics. In numerical simulations that we perform, we consider the range of mass transfer coefficients typical for macromolecules that are detected by state-of-the-art biosensors. The presented results enable the analysis of the dependence of the minimal detectable signal on the sensor design parameters, and are useful for achieving the improved sensing performance.

II. THEORETICAL ANALYSIS

The fluctuations of the output signal of adsorption-based chemical and biological sensors result from various kinds of noise, i.e. the noise originating from the stochastic nature of physical processes (AD and mass transfer) which are essential for analyte detection by the sensing element, and the noises from the sensor transduction mechanism and the read-out circuitry. The former, known as AD noise, poses the fundamental detection and quantification limits of analyte concentration, inherent to all adsorption-based devices. In some cases, AD noise can dominate compared to other noise sources [3]. Assuming that the transducer noise and the noise of read-out circuitry are minimized by applying appropriate techniques, reduction of the inevitable AD noise remains a means of achieving the sensor's best possible detection performance, i.e. maximization of sensor's signal-to-noise ratio and approaching its lowest possible detection limit. Therefore, the analysis of the sensor SNR excluding the transducer and read-out circuitry noise yields the guidelines for lowering the fundamental detection limit.

By assuming a linear relation between the sensor response and the number of adsorbed particles, the expected value of stochastic response is $\langle r\rangle = m\langle N\rangle$ (where $\langle N\rangle$ is the expected value of the number of adsorbed analyte particles, and m is the proportionality factor equal to the average contribution of a single particle adsorption to the sensor response), sensor's AD noise is $\sigma_{r,AD}^2 = m^2\sigma_N^2$ (where σ_N^2 is the variance of the number of adsorbed particles), so the sensor's maximal SNR (the ratio of signal power and noise power), which is determined only by the fundamental AD noise, is

$$SNR = \frac{< N >^2}{\sigma_N^2} \tag{1}$$

Here we consider the SNR in the steady state (established after all transient processes have ended), in which the measurements are preferably carried out in practice. We use the stochastic model presented in [5], describing the number of adsorbed particles that randomly fluctuates due to coupled AD and mass transfer processes. Based on the model, the steady-state expected value and variance of the number of adsorbed particles are given by expressions

$$\langle N\rangle = \frac{N_m k_a C}{k_a C + k_d + \frac{k_a k_d}{k_m A}} \tag{2}$$

$$\sigma_N^2 = k_d\langle N\rangle \frac{\left[1 + (N_m - \langle N\rangle)\frac{k_a}{k_m A}\right]^2}{k_a C + k_d + \frac{k_a k_d N_m}{k_m A}} \tag{3}$$

respectively, where C is the concentration of target particles in the analyzed sample, N_m is the number of adsorption sites on the sensing surface, k_a and k_d are the adsorption and desorption rate constants, k_m is the mass transfer coefficient which models the combined effect of diffusion and convection on particle transport to the surface adsorption sites and from them, and A is the sensing surface area. k_m depends on geometrical parameters of the sensing system, the flow rate of the sample, and the diffusivity of the analyte [6]. Based on Eqs. (1)-(3) we obtain

$$SNR = \frac{k_a C N_m}{k_d} \cdot \frac{\left[k_a C + k_d\left(1 + \frac{k_a}{k_m A}\right)\right]\left[k_a C + k_d\left(1 + \frac{k_a N_m}{k_m A}\right)\right]}{\left[k_a C + k_d\left(1 + \frac{k_a}{k_m A}\right)\left(1 + \frac{k_a N_m}{k_m A}\right)\right]^2} \tag{4}$$

This expression clearly shows the dependence of the sensor's maximal achievable SNR on the mass transfer coefficient, and thus enables the analysis of the influence of mass transfer process on the SNR.

The dynamic range, defined as the range of analyte concentrations reliably detectable by the sensor, is also one of important sensor characteristics. The upper limit of the dynamic range is determined by the sensor saturation, i.e. the limited number of adsorption sites on the sensing surface. It will not be considered in this paper. The lower limit of the

dynamic range is the lowest concentration that can be distinguished from noise. This detection limit is determined as the analyte concentration level at which the sensor's SNR is equal to a minimal acceptable value for reliable detection. Let us assume that it is a value F. The condition $SNR=F$, where the SNR is given by Eq. (4), yields the equation for the fundamental detection limit C_{min}

$$\begin{aligned}&(k_aC_{\min})^3\frac{N_m}{k_d}+(k_aC_{\min})^2\left[N_m\left(2+\frac{k_a}{k_mA}+\frac{k_aN_m}{k_mA}\right)-F\right]\\&+(k_aC_{\min})\left[(N_m-2F)k_d\left(1+\frac{k_a}{k_mA}\right)\left(1+\frac{k_aN_m}{k_mA}\right)\right]\\&-Fk_d^2\left(1+\frac{k_a}{k_mA}\right)^2\left(1+\frac{k_aN_m}{k_mA}\right)^2=0\end{aligned} \tag{5}$$

By numerically solving the Eq. (5) for C_{min} for a series of the coefficient k_m values, the analysis can be performed of the mass transfer effect on the minimal concentration that can be reliably detected at the required SNR value F.

III. Results and Discussion

By using the derived expressions (Eqs. (4) and (5)) we analyze the SNR and the minimal concentration level at which the SNR has the required value that ensures reliable analyte detection in microfluidic chemical and biological sensors. We particularly consider the influence of mass transfer of particles on these important performance characteristics, after the steady state of all relevant transient processes has been reached.

A biosensor for detection of macromolecules (proteins), with the sensing surface area $A=10^{-11}$ m^2, and the adsorption sites surface density $n_m=N_m/A=3\cdot10^{17}$ 1/m^2 is used in the analysis. The parameters of the AD process are $k_a=1.33\cdot10^{-19}$ m^3/s and $k_d=0.08$ 1/s. The range of the mass transfer coefficient k_m is from 10^{-6} m/s to 10^{-1} m/s.

Fig. 1 shows the biosensor's SNR as a function of the mass transfer coefficient at the analyte concentration $C=6\cdot10^{16}$ 1/m^3 (the solid curve). The dashed curve represents the SNR obtained by using the stochastic model that neglects the mass transfer influence [5], according to which

$$SNR_i=\frac{k_aCN_m}{k_d} \tag{6}$$

The diagram in Fig. 1 shows that the SNR value is lower for lower mass transfer rates (i.e, for lower k_m). Hence, slow mass transfer reduces the maximal possible sensor's SNR. In the considered range of k_m values SNR exhibits a change of almost three orders of magnitude. As k_m value increases, SNR monotonically increases and reaches its maximal value equal to that obtained by using the model that does not take into account mass transfer (Eq. (6)). Therefore, when mass transfer is sufficiently fast, its influence becomes negligible. In the analyzed case, mass transfer with the coefficient greater than 10^{-2} m/s does not lead to the decrease of the sensor's SNR.

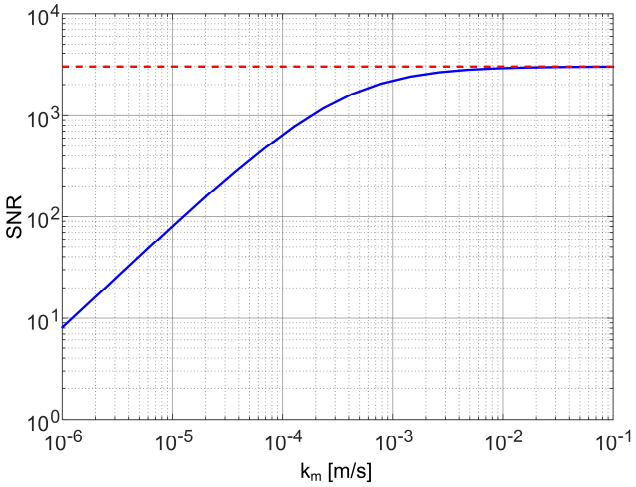


Fig. 1 The sensor's maximal achievable SNR depending on the mass transfer coefficient (solid line), and SNR obtained by neglecting the mass transfer influence (dashed line).

The minimal concentration level for $SNR=9$ (a value considered as required for reliable analyte detection [7]) is shown as a function of k_m in Fig. 2 (solid line). C_{min} obtained from Eq. (6) for $SNR_i=F$

$$C_{\min,i}=\frac{Fk_d}{k_aN_m} \tag{7}$$

i.e. by the analysis that neglects the mass transfer influence, is represented by a dashed line in the same diagram, also for $F=9$.

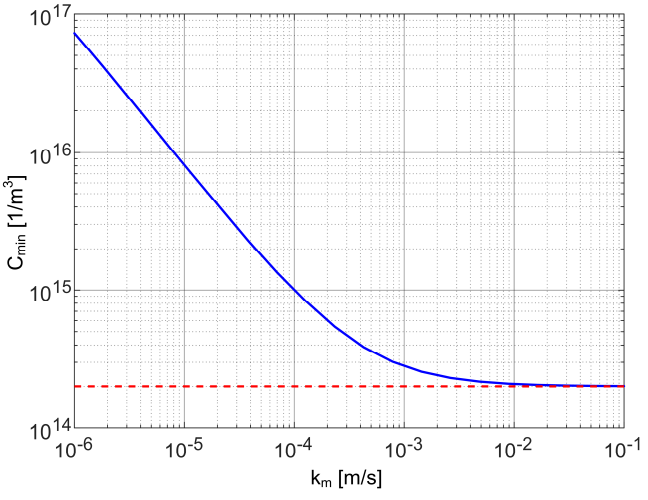


Fig. 2 The dependence of the sensor's fundamental detection limit on the mass transfer coefficient (solid line), and the same parameter obtained by neglecting the mass transfer influence (dashed line).

Fig. 2 shows that mass transfer significantly influences C_{min} in such a way that slow mass transfer increases the

concentration value and thus degrades the sensor performance. When the mass transfer is sufficiently fast, its effect becomes negligible, and C_{min} reaches the lowest value equal to $C_{min,i}$.

The presented analysis shows that it is necessary to consider the mass transfer influence when the optimization of microfluidic sensors and their operating conditions is performed aiming to achieve the required SNR value for reliable analyte detection, i.e. to ensure analyte detection in a certain concentration range.

IV. Conclusions

In this paper the theory is presented that enables the analysis of mass transfer influence on the maximal achievable SNR, and of the lowest detectable analyte concentration in microfluidic sensors. The analysis has shown that a slow mass transfer degrades these important sensor performance parameters.

The mass transfer coefficient is a known function of the sensing system geometrical parameters, the flow rate of the sample through the microfluidic reaction chamber, and the diffusivity of the analyte. Therefore, the presented theory enables the analysis of the dependences of the considered sensor performance metrics on sensor design parameters and operating conditions, and thus provides the means for sensor optimization.

Acknowledgment

This work has been funded by the Serbian Ministry of Education, Science and Technological Development (Project TR 32008) and by the Serbian Academy of Sciences and Arts (Project F-150).

References

[1] K.-K. Liu, R.-G. Wu, Y.-J. Chuang, H. S. Khoo, S.-H. Huang, F.-G. Tseng, "Microfluidic Systems for Biosensing," *Sensors*, vol. 10, pp. 6623-6661, 2010.

[2] E. K. Sackmann, A. L. Fulton, D. J. Beebe, "The present and future role of microfluidics in biomedical research," *Nature*, vol. 507, pp. 181-189, 2014.

[3] A. Hassibi, H. Vikalo, A. Hajimiri, "On noise processes and limits of performance in biosensors," *J. Appl. Phys.*, vol. 102, 014909 1-12, 2007.

[4] G. Tulzer, C. Heitzinger, "Fluctuations due to association and dissociation processes at nanowire-biosensor surfaces and their optimal design," *Nanotechnology*, vol. 26, pp. 025502 1-9, 2015.

[5] I. Jokić, Z. Djurić, K. Radulović, M. Frantlović, P. M. Krstajić, K. Cvetanović Zobenica, "Steady-State Analysis of Stochastic Time Response of Chemical and Biological Microfluidic Sensors," Proc. 5th International Conference on Electrical, Electronic and Computing Engineering, IcETRAN 2018, Palić, Serbia, pp. 943-948, June 11–14, 2018,

[6] D. G. Myszka, X. He, M. Dembo, T. A. Morton, B. Goldstein, "Extending the Range of rate constants available from BIACORE: interpreting mass transport-influenced binding data," *Biophys. J.*, vol. 75, pp. 583–594, 1998.

[7] A. Shrivastava, V. B. Gupta, "Methods for the determination of limit of detection and limit of quantification of the analytical methods," *Chronicles of Young Scientists*, vol. 2, pp. 21-25, 2011.

A consideration of the use of ICTM SP-12 pressure sensor for ultrasound sensing

Jelena Stevanović, Žarko Lazić, Milče M. Smiljanić, Katarina Radulović, Danijela Randjelović, *Member, IEEE*, Miloš Frantlović and Milija Sarajlić, *Member, IEEE*

***Abstract*— A consideration study for the application of the pressure sensor SP-12 developed and produced by ICTM CMT as an ultrasound sensor is given. The interaction of ultrasound with the sensor's membrane was analytically described, but for the initial examination of its performance, Finite Elements Method simulation was applied. The sensor SP-12 has eigenfrequencies in the range from 200 kHz to the frequencies higher than 2 MHz. The amplitude of the output signal, which is proportional to Von Mises stress, is highest for the lowest frequency, and it exponentially decreases as the eigenfrequencies increase. This makes the sensor suitable for the ultrasound measurements in the range of hundreds of kHz.**

***Index Terms*—pressure sensor; ultrasound; Von Mises stress; piezoresistor; eigenfrequencies.**

I. Introduction

Ultrasound measurement and detection have many important applications of everyday life and industry. Navigation of vehicles [1], medical examination [2], materials testing [3] and sonication (ultrasound processing of liquids) [4] are some of the applications. Sensors and detectors for ultrasound comprise different models of operation, for instance capacitive transducers [5], Fiber Bragg Grating [6] or Spherical-Omnidirectional Ultrasound Transducers [7].

One possibility for the ultrasound detection is measurement of the pressure differences it makes on the membrane of the pressure sensor. For this purpose, device originally developed as a pressure sensor can serve as an ultrasound sensor. At ICTM CMT in Belgrade, Serbia, there has been a long history of pressure sensors research and development, from the model SP-6 developed in 1980s to the model SP-12, which is currently in production [8]. This work examines possibility of application of the SP-12 pressure sensor for ultrasound sensing and gives brief overview on analytical model of membrane under mechanical stress together with numerical simulations. A proposal of experimental procedure is also given. In the presented analysis it is investigated whether the SP-12 can serve as an ultrasound sensor at the frequencies equal to its eigenfrequencies.

Jelena Stevanović is with the ICTM CMT, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia (e-mail: jelena@nanosys.ihtm.bg.ac.rs)

Žarko Lazić is with the ICTM CMT, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia (e-mail: zlazic@nanosys.ihtm.bg.ac.rs)

Milče M. Smiljanić is with the ICTM CMT, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia (e-mail: smilce@nanosys.ihtm.bg.ac.rs)

Katarina Radulović is with the ICTM CMT, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia (e-mail: kacar@nanosys.ihtm.bg.ac.rs)

Danijela Randjelović is with the ICTM CMT, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia (e-mail: danijela@nanosys.ihtm.bg.ac.rs)

Miloš Frantlović is with the ICTM CMT, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia (e-mail: frant@nanosys.ihtm.bg.ac.rs)

Milija Sarajlić is with the ICTM CMT, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia (e-mail: milijas@nanosys.ihtm.bg.ac.rs)

II. Analytical model of SP-12 membrane under mechanical stress

A. Description of sensor SP-12

The SP-12 sensor is fabricated on double-side polished single crystal n-type 3`` silicon wafer, with the resistivity of 3-5 Ωcm [8]. Four piezoresistors are formed by the photolithography process and thermal diffusion of boron at 920°C. The effective length and width of each piezoresistor are 135 µm and 5 µm, respectively. The concentration of dopants is between $1.5 \cdot 10^{20}$ cm^{-3} and $2 \cdot 10^{20}$ cm^{-3} [8].

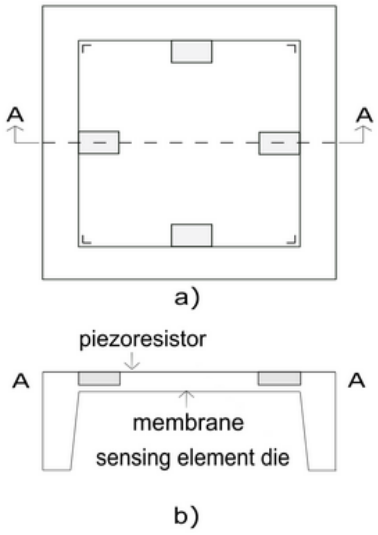


Fig 1. Schematic illustration of the SP-12 sensing element die: a) top view of the sensor; b) cross section through the middle of the diaphragm

Positions of piezoresistors on the silicon membrane are important for the sensitivity of the device. All of them are located near the edge of the membrane, two in parallel and two in the transversal direction, as it is shown in Fig. 1. They are connected in the Wheatstone bridge. The resistor positions are optimal in terms of the highest possible sensitivity and linearity of the output signal. The membrane is square of dimensions 1000 µm × 1000 µm, fabricated by wet anisotropic etching of silicon of the bottom side on the wafer [8]. The obtained thickness of the membrane is 18

μm. Metallization is done by aluminum sputtering under the base pressure of $2 \cdot 10^{-6}$ mbar. The overall size of the sensing element die is 2000 μm × 2000 μm × 380 μm [8]. After the fabrication of the die, it is anodically bonded to a 1.7 mm thick glass support [8].

B. Analytical model

The behavior of the sensor's membrane under mechanical stress was analytically modeled by using an approximation of the classical plate theory. According to this theory, differential equation for anisotropic clamped rectangular thin plate is [9]:

$$D_x \frac{\partial^4 \omega}{\partial x^4} + 2H \frac{\partial^4 \omega}{\partial x^2 \partial y^2} + D_y \frac{\partial^4 \omega}{\partial y^4} = q \qquad (1)$$

where ω is the deflection of the plate midsurface (displacement in the z direction), q is the intensity of uniform load on the plane, H is the parameter highly dependent on symmetry, and for materials with orthotropic elasticity values like single-crystal (100) silicon [10], it has a value of $\sqrt{(D_x D_y)}$ and D_x and D_y are the flexural rigidities in two orthogonal directions, and for orthotropic materials it is given by [11]:

$$D_{x/y} = \frac{E_{x/y} h^3}{12(1 - \nu_{xy} \cdot \nu_{yx})}. \qquad (2)$$

Here E is the modulus of elasticity in tension and compression, ν is the Poisson's ratio and h is the plate thickness. The boundary conditions for a clamped thin plate are [9]:

$$\omega = 0, \frac{\partial^2 \omega}{\partial x^2} = 0, \frac{\partial^2 \omega}{\partial y^2} = 0, \qquad (3)$$

for $x = 0$ and $x = a$, and $y = 0$ and $y = b$, where a and b indicate the plate edge lengths. In other words, the bending moment at the edge of the membrane is zero. For the square plate, a and b are equal. If the load q is represented in the form of a double trigonometric series, a solution of the differential equation (1) can be presented in the form of [9]:

$$\omega = \sum_{m=1,3,5\ldots}^{\infty} \sum_{n=1,3,5\ldots}^{\infty} a_{mn} \sin \frac{m\pi x}{a} \sin \frac{n\pi y}{b}. \qquad (4)$$

By substituting this solution in Eq. (1), the expression for coefficients a_{mn} is found. In the case of uniform load the deflection surface is symmetrical with respect to the axes $x = a/2$ and $y = b/2$. For that reason all terms with even numbers for m or n in series (4) do not exist. Hence, the final solution of Eq. (1) is [9]:

$$\omega = \frac{16q}{\pi^6} \sum_{m=1,3,5\ldots}^{\infty} \sum_{n=1,3,5\ldots}^{\infty} \frac{\sin \frac{m\pi x}{a} \sin \frac{n\pi y}{b}}{mn\left(\frac{m^4}{a^4} D_x + \frac{2m^2 n^2}{a^2 b^2} H + \frac{n^4}{b^4} D_y\right)}. \qquad (5)$$

For orthotropic (100) silicon membrane, values of the modulus of elasticity are $E_x = E_y = 169$ *GPa* [12], while the values of the Poisson's ratio, when the extension is applied along x, i.e. y direction, are $\nu_{xy} = \nu_{yx} = 0.064$ [12]. By including these values in Eq. (2), flexural rigidities D_x and D_y are calculated to be $8{,}25 \cdot 10^{-5}$ *Pa·m*^{3}. We assume that amplitude of the ultrasound pressure is 100 Pa. Deflection at the center of the orthotropic plate with $a = b$ and $D_x = D_y$ can be expressed by the formula [9]:

$$\omega = 0.00407 \cdot \frac{q \cdot b^4}{D_y}, \qquad (6)$$

and is estimated to *4,9 nm*. It is important to mention that this calculated deflection refers to the membrane's center when static pressure is applied. That means resonant frequency (eigenfrequency) contribution is not included in the calculation. In order to obtain values of deflection for membrane's resonant frequency when harmonic perturbation like ultrasound is applied, further analysis is needed, whose complexity overcomes the scope of this paper.

III. Finite Elements Method Model of Eigenfrequencies

The eigenfrequencies of the SP-12 membrane can be found from the FEM model by solving the eigenvalue problem that arises from the equations if velocity is considered to be unknown. A standard matrix form of the dynamic equation of motion can be written as [13]:

$$[M]\frac{d^2 y}{dt^2} + [B]\frac{dy}{dt} + [K]y = 0, \qquad (7)$$

where y is the vector of nodal displacement under the external force vector, and [M], [B] and [K] are the element matrices for mass, damping and stiffness for the whole structure. Under free vibration, the eigenfrequencies and the mode shapes of a multiple degree of freedom system are the solutions of the eigenvalues problem.

FEM codes are designed to solve systems of equations like Eq. (6), with one equation for each of the relevant planes. It is possible to determine the eigenvalues and eigenvectors after integrating the approximate solution and forming the matrices.

The finite element modeling was performed using COMSOL Multiphysics software [14] which provides data about the interaction between the ultrasound and silicon membrane. A quarter of the whole tested diaphragm is chosen for the model and appropriate boundary conditions and symmetry are defined in order to simplify the construction of the model and subsequent calculation. The intensity of ultrasound pressure in boundary load is set to value of 100 Pa. In order to obtain structural response of harmonic load on membrane, Frequency Domain Study was applied. Range of ultrasound frequencies is selected to include the values of the eigenfrequencies, obtained as results of Eigenfrequency Analysis performed before the mentioned study. The maximum of displacements and Von Mises stresses appear on frequencies values that correspond to the eigenfrequencies of the tested membrane.

Results of the simulation are shown in Figs. 2-6 for the first five eigenfrequencies. Values of Von Mises stresses are listed in the Table 1 only for the positions of interest, which coincide with position of SP-12 piezoresistors. These positions are marked in Figs. 2-6 with the white arrows.

TABLE I
NUMERICAL RESULTS OF THE SIMULATION

eigenfrequency (Hz)	Von Mises stress (Pa)	membrane central point amplitude (μm)
$2.59 \cdot 10^5$	$1.13 \cdot 10^9$	10.89
$9.35 \cdot 10^5$	$8.45 \cdot 10^8$	1.03
$9.39 \cdot 10^5$	$4.45 \cdot 10^7$	0.59
$1.55 \cdot 10^6$	$4.06 \cdot 10^6$	$36 \cdot 10^{-3}$
$1.80 \cdot 10^6$	$1.72 \cdot 10^5$	$44 \cdot 10^{-6}$

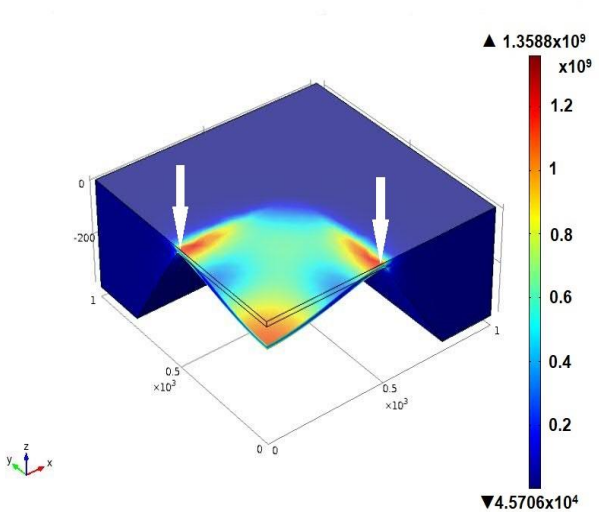


Fig 2. Mode shape of tested silicon membrane for eigenfrequency $2.59 \cdot 10^5$ Hz

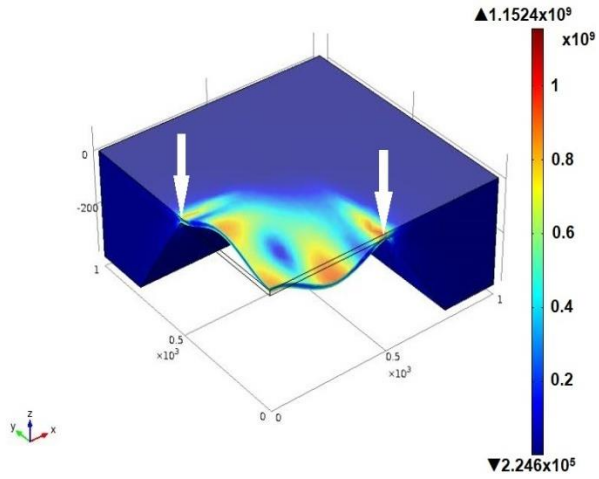


Fig 3. Mode shape of tested silicon membrane for eigenfrequency $9.35 \cdot 10^5$ Hz

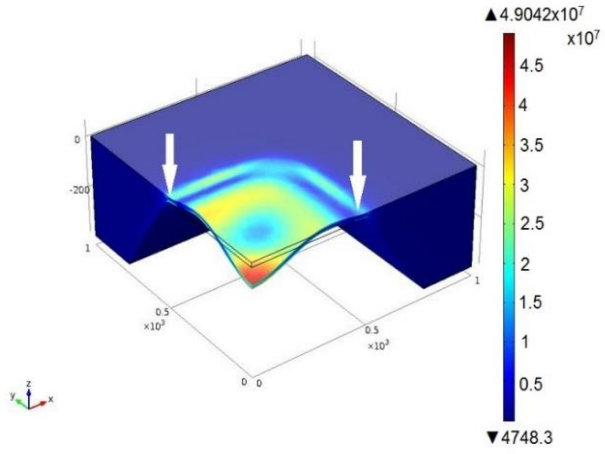


Fig 4. Mode shape of tested silicon membrane for eigenfrequency $9.39 \cdot 10^5$ Hz

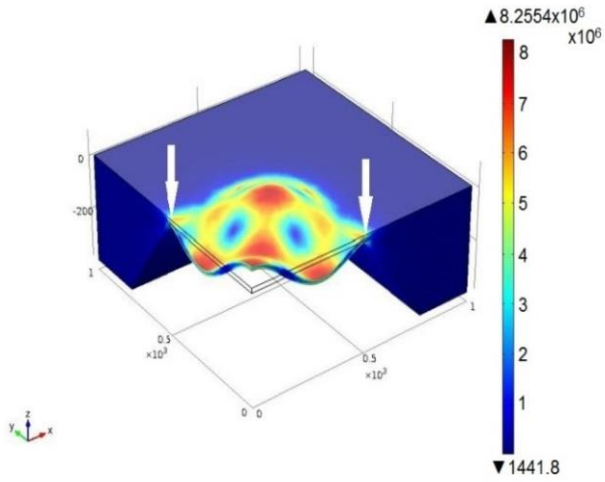


Fig 5. Mode shape of tested silicon membrane for eigenfrequency $1.55 \cdot 10^6$ Hz

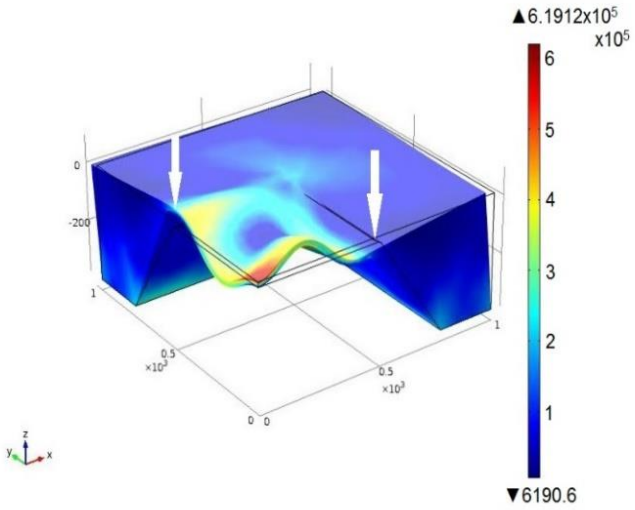


Fig 6. Mode shape of tested silicon membrane for eigenfrequency $1.80 \cdot 10^6$ Hz

IV. Proposal of the Experiment

In the proposed experimental set-up (Fig. 7), the ultrasound source will be a specifically shaped material from the group of piezoelectric ceramics. The SP-12 sensor will be connected to a constant current source which provides the excitation current of 5 mA. The output of the SP-12 sensor is connected to a spectrum analyzer. A maximum amplitude position in a digital record of the analyzer is expected to coincide with simulation results from Comsol. A potential problem of the experimental set-up would be the choice of the appropriate ultrasound source, whose range of soundwave frequencies should also include the resonant frequencies of the tested silicon membrane.

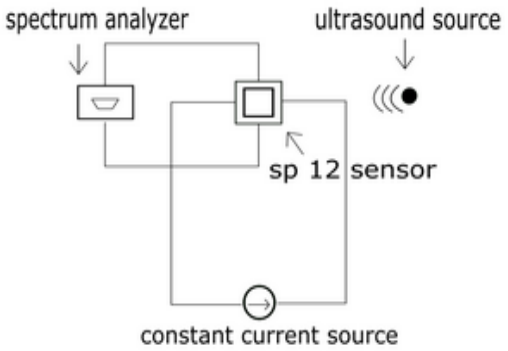


Fig 7. Schematic illustration of the proposed experiment

V. Conclusion

The purpose of this consideration study, was to examine the suitability of the ICTM CMT pressure sensor SP-12 for ultrasound detection. The interaction of mechanical pressure with the sensors's membrane was analytically described and numerically simulated. It was shown that this type of sensor has a potential to be used as ultrasound sensor. The range of ultrasound frequencies that can be probed corresponds to the silicon membrane eigenfrequencies.

It was noticed that stress at the positions of the piezoresistors has the highest value for the first two mode shapes. Therefore, eigenfrequencies that correspond to those mode shapes will be of our interest for the future ultrasound detection and measurement. Experimental testing is in preparation, where SP-12 will be connected in electrical circuitry in the similar way as pressure sensor, but the read-out will be performed by spectrum analyzer. The degree of agreement between analytical and numerical predictions with the experimental results will be investigated.

ACKNOWLEDGMENT

This work was supported by the Ministry of education, science and technological development of the Republic of Serbia, within the Project TR32008.

References

[1] J. Hyo Rhee and J. Seo, "Low-Cost Curb Detection and Localization System Using Multiple Ultrasonic Sensors," *Sensors*, vol. 19, 21 March 2019.

[2] T. L. Szabo, "*Diagnostic Ultrasound Imaging: Inside Out Biomedical Engineering*," 2th ed. revised Academic Press, ISBN 9780123965424, 2013.

[3] C. H. Chen, "*Ultrasonic And Advanced Methods For Nondestructive Testing And Material Characterization*," University of Massachusetts, USA, World Scientific, ISBN 9814476404, 2007.

[4] J. L. Capelo-Martínez, "*Ultrasound in Chemistry: Analytical Applications*," Wiley-VCH Verlag GmbH & Co. KgaA, Weinheim, Germany, ISBN 978-3-527-31934, 2009.

[5] A. Pirouz and F. Levent Degertekin, "An Analysis Method for Capacitive Micromachined Ultrasound Transducer (CMUT) Energy Conversion during Large Signal Operation Sensors," *Sensors*, vol. *19*, 20 February 2019.

[6] Ch. Rao and L. Duan, "Bidirectional, Bimodal Ultrasonic Lamb Wave Sensing in a Composite Plate Using a Polarization-Maintaining Fiber Bragg Grating," *Sensors*, vol. *19*, 19 March 2019.

[7] S. Sadeghpour, S. Meyers, J. P. Kruth, J. Vleugels, M. Kraft and R. Puers, "Resonating Shell: A Spherical-Omnidirectional Ultrasound Transducer for Underwater Sensor Networks," *Sensors*, vol. *19*, 13 February 2019.

[8] M. Frantlović, I. Jokić, Ž. Lazić, M. M. Smiljanić, M. Obradov, B. Vukelić, Z. Jakšić and S. Stanković, "A method enabling simultaneous pressure and temperature measurement using a single piezoresistive MEMS pressure sensor," *Meas. Sci. Technol.* vol. *27*, 21 October 2016.

[9] S. Timoshenko, S. Woinowsky-Krieger, "Bending of Anisotropic Plates" in *Theory of plates and shells*, 2th ed. McGraw-Hill Book Company, USA, ISBN 0-07-064779-8, 1989.

[10] Z. Qin, Y. Gao, J. Jia, X. Ding, L. Huang and H. Li, "The Effect of The Anisotropy of Single Crystal Silicon on the Frequency Split of Vibrating Ring Gyroscopes," *Micromachines*, vol. *10*, 14 February 2019.

[11] A. F. Johnson and A. Woolf, "Deflection and stress analysis of orthotropic plates in flexure," *Computers & Structures*, vol. *18*, pp *911-919*, 1984.

[12] E. V. Thomsen, K. Reck, G. E, Skands, C. V. Bertelsen, O. Hansen, "Silicon as anisotropic mechanical material: Deflection of thin crystal planes," *Sensors and Actuators A: Physical*, vol. *220*, pp. *347-364*, 2014.

[13] A. Belhadj, R. Cheesewright and C. Clark, "The simulation of coriolis meter response to pulsating flow using a general purpose F. E. code," *Journal of Fluids and Structures*, vol. *14*, pp. *613-634*, 2000

[14] Comsol software https://www.comsol.com/ Accessed 18.1.2019.

Consideration of Thin Film Ionization Vacuum Pressure Sensor

Marko Bošković, Danijela Randjelović, *Member, IEEE*, Milena Rašljić, Katarina Cvetanović-Zobenica, Žarko Lazić, Milče M. Smiljanić, and Milija Sarajlić, *Member, IEEE*

***Abstract*— A novel concept of vacuum pressure sensor based on thin film technology is presented. The sensor is designed as a 1 μm thick aluminium film patterned as a structure of wedges facing each other along a sharp tip. The distance between the wedge tips is 3 μm. This structure is obtained by laser writing in vector mode. Parts of the sensor structure are fabricated and measured. Analytical consideration of the proposed structure is given together with the concept of the experimental set up for testing of the sensor.**

***Index Terms*—Pressure sensor; DC discharge; Electrical conductivity of gases.**

I. Introduction

Vacuum technology is very important in many scientific disciplines and industrial processes [1]. Measurement of the vacuum pressure depends on the range of vacuum, namely rough vacuum, medium vacuum, high or ultra-high vacuum (UHV). For measurements of rough vacuum, from atmospheric pressure down to 0.1 Pa, gauges like Pirani or mechanical manometers are used [2, 3, 4]. For high vacuum, or pressure lower than 0.1 Pa ionization gauges are used [5].

In this work, an ionization sensor of vacuum is designed, modelled by analytical formulas, and part of it is fabricated. Fabrication in the proposed technology of direct laser writing is feasible but it takes relatively long time for the machine to finish one production cycle.

Aim of this work is to examine whether it is possible to determine gas pressure by measuring the conductivity of gases with a microdevice. In the following, proposed sensor design and discussion of the proposed design pattern are given. Characteristics of fabricated sensor part are examined by Atomic Force Microscopy measurements. Afterwards, some theoretical background of electrical conductivity in gases is considered and calculations for proposed structure are performed. Last part of paper gives proposal for measurement set-up and schematic for calibration of the sensor.

Marko Bošković is with the ICTM-CMT, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia (e-mail: boskovic@nanosys.ihtm.bg.ac.rs).

Danijela Randjelović is with the ICTM-CMT, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia (e-mail: danijela@nanosys.ihtm.bg.ac.rs).

Milena Rašljić is with the ICTM-CMT, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia (e-mail: milena@nanosys.ihtm.bg.ac.rs).

Katarina Cvetanović-Zobenica is with the ICTM-CMT, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia (e-mail: katarina@nanosys.ihtm.bg.ac.rs).

Žarko Lazić is with the ICTM-CMT, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia (e-mail: zlazic@nanosys.ihtm.bg.ac.rs).

Milče M. Smiljanić is with the ICTM-CMT, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia (e-mail: smilce@nanosys.ihtm.bg.ac.rs).

Milija Sarajlić is with the ICTM-CMT, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia (e-mail: milijas@nanosys.ihtm.bg.ac.rs).

II. Sensor Design and Fabrication

The sensor is designed in four bunches, each with two pads. Each bunch consists of four stripes on one pad and five on another, with wedges which make a fringe pattern between them, Fig. 1. Each stripe is patterned on both sides with a series of wedges which are facing the wedges of the opposite stripe along the sharp tip. Enlarged detail of fringes (red rectangle in Fig. 1) is given in Fig. 2. The shortest distance between the tips is 3 μm.

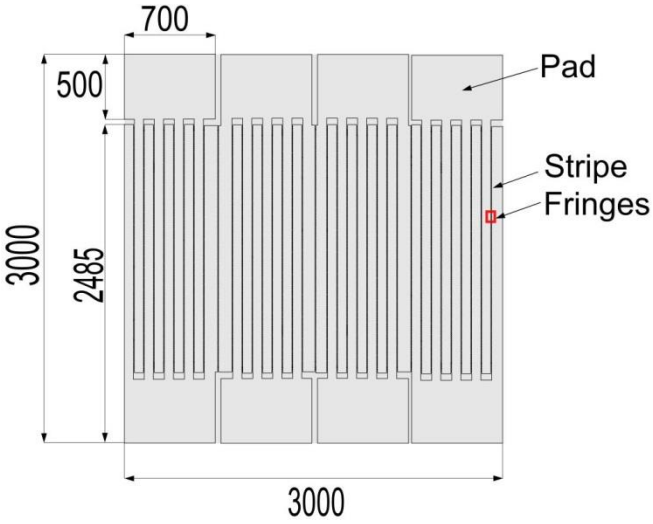


Fig. 1. Proposed design of the vacuum pressure sensor. Red rectangle is shown in Fig. 2. All units are in micrometers.

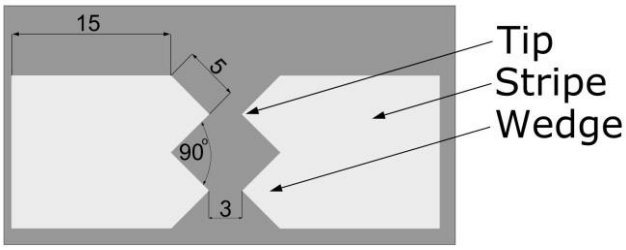


Fig. 2. Enlarged detail of the proposed design (red rectangle in Fig.1). Units are in micrometers.

The technology needed for the fabrication of this type of the designed structures was already developed at ICTM-CMT [6]. With this technology, it was possible to fabricate the lines with the period of 6 μm and clearance between 2 μm and 3 μm using Laser Writer (LW405, MicroTech, Italy) in vector mode. Vector mode enables a machine to draw continuous lines and the width of the line is defined by

the time of photoresist exposure and subsequent development.

A part of the proposed pattern for the vacuum pressure sensor is shown in Fig. 3.

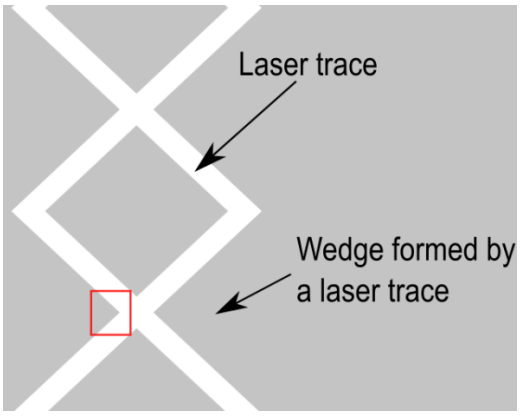


Fig. 3. A detail of proposed pattern and marked area where AFM scan was performed (red square).

The white line in Fig. 3 represents areas of wafer which will be exposed to laser radiation. After photolithography, we get structure similar to the structure shown in Fig. 2. If lines are 2 µm wide and are crossing each other on 90 degrees angle, calculated distance between wedge tips (electrodes) is 2.83 µm. Pattern was drawn using CleWin software [7].

The sensor was fabricated in a planar technology used for microelectronic devices. On a 3 inch wafer, 380 µm thick, <100> orientation, 3-5 Ωcm resistivity, n-type, one side polished, SiO_2 layer was grown by thermal oxidation at 1100 ℃ for 105 minutes. After thermal oxidation aluminium with 1% of silicon (Al 1% Si) was sputtered by DC Magnetron Sputtering. The wafer was then coated with photoresist AZ 1505 (MicroChemicals, Germany), 0.5 µm thick by spin coating. The wafer prepared in this way was then exposed to laser radiation using direct laser writer (LW405, MicroTech, Italy, 405 nm wavelength) in vector mode. After exposure, photoresist was developed using MIF 726 (MicroChemicals, Germany) developer for 25 seconds. Afterwards, photoresist was baked at 115ºC for 50 seconds. The next procedure was removal of aluminium from the exposed areas. This was done with a solution made of acetic acid, phosphoric acid and nitric acid. After this, the remained photoresist was removed with acetone.

In order to explore characteristics of lines fabricated in vector mode, Atomic Force Microscopy (AFM) measurement was performed. The structure on which AFM measurement was performed is marked as red square in Fig. 3. A 3D AFM scanning along the transition area between sputtered Al 1% Si and SiO_2 is shown in Fig. 4. Fig. 5 gives 2D image of the same structure. The profile of the transition (red line in Fig. 5) is shown in Fig. 6.

Thickness of sputtered Al 1% Si is estimated to be around 0.7 µm from Fig. 6 and surface roughness is around 73 nm, Fig 6.

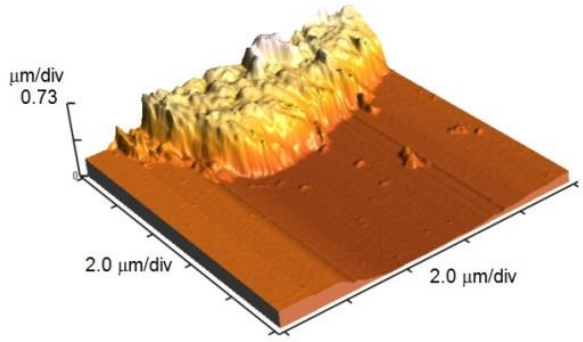


Fig. 4. 3D AFM picture of the obtained structure

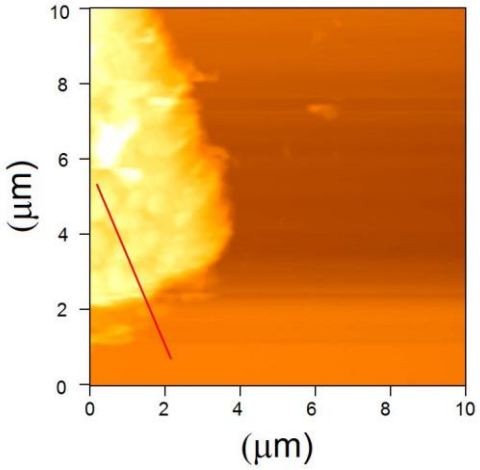


Fig. 5. 2D AFM image with the profile line marked red.

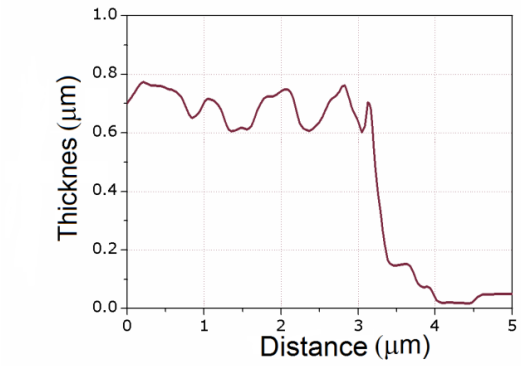


Fig. 6. profile of the characteristic line.

III. Analytical Model of Sensor Functionality

Under normal conditions gas is behaving as an electrical insulator. Under certain conditions gas can conduct electricity [8]. The flow of the electrical current through gases is known as electrical discharge. There are three different types of electrical discharge – Townsend (dark) discharge, glow discharge and arc discharge. These types can be distinguished by current-voltage characteristics.

For the functionality of this sensor, the region of interest

is Townsend discharge. In this region the applied voltage accelerates free ions and electrons so that certain number of them will reach the electrodes. The current is proportional to the applied voltage and gas acts as an Ohmic resistor. Further increasing of voltage allows all charged particles to reach the electrode and current enters the saturation. If the applied voltage is further increased, the current rapidly increases. This is a consequence of creating new ions in gas by collisions.

Ionization in this region is a consequence of three processes [9]:

-While moving toward anode electrons collide with gas particles and generate more electrons and ions;

-Positive ions moving toward the cathode collide with gas particles, ionize and also generate certain numbers of electrons and positive ions;

-Particles such as positive ions strike the cathode to emit secondary electrons.

Each of these processes is quantitatively characterized by three Townsend coefficients: α, β and γ respectively. Townsend coefficient α is the electron ionization coefficient for the volume of gas. If the number of accelerated electrons is n_{e0} and the gap between the electrodes is x then the number of produced electrons is [9]:

$$n_e = n_{e0}e^{\alpha x}, \tag{1}$$

and discharge current reaching the anode is [9]:

$$i_e = i_{e0}e^{\alpha x}. \tag{2}$$

Coefficient α depends on the mean free path of electrons, λ_e, and intensity of the electric field E. It can be calculated using equation [9]:

$$\alpha = \frac{1}{\lambda_e}\exp\left(-\frac{V_i}{E\lambda_e}\right), \tag{3}$$

where V_i is ionization energy of gas component.

Mean free path is the average distance travelled by particle between colliding. With approximation of the ideal gas state, mean free path can be calculated using equation [10]:

$$\lambda_e = \frac{kT}{\sqrt{2}\,p\pi d^2}, \tag{4}$$

where k is the Boltzmann constant, T is the gas temperature, p is the gas pressure and d is the effective diameter of the particle.

Equations (3) and (4) show that, for a specific gas α is a function of the gas pressure and the electric field intensity.

Coefficient β is a positive ion ionization coefficient. Required energy for positive ions to ionize neutral particles is thousands of electron volts. In a normal discharge process ions do not have this energy so a contribution of β is negligible [9].

Coefficient γ is the electrode surface ionization coefficient for positive ions. It shows how many electrons are emitted from the cathode when positive ion strikes it. If $n_0(e^{\alpha x}-1)$ positive ions strike a cathode, the number of electrons emitted from cathode will be $\gamma n_0(e^{\alpha x}-1)$. Coefficient γ is a function of cathode material and kinetic energy of impact ions. Ions must have enough energy to overcome the electron binding energy of cathode material. Given that the kinetic energy of ions is determined by the intensity of the electric field and gas pressure, one can tell that coefficient γ is a function of gas pressure, intensity of the electric field and cathode material. Coefficient γ can be calculated using Baragiola empirical equation [11]:

$$\gamma = 0.032(0.78V - 2\varphi), \tag{5}$$

or Hagstrum's semi empirical equation [11]:

$$\gamma = \frac{0.2(0.8V-2\varphi)}{\varepsilon_F}. \tag{6}$$

In both equations, V is the energy of incident ion, φ is the work function, and ε_F is the Fermi energy.

Total discharging current is [9]:

$$i_e = \frac{i_{e0}e^{\alpha x}}{1-\gamma(e^{\alpha x}-1)} \tag{7}$$

Total discharge current is a function of Townsend coefficients, thus, it is a function of pressure and intensity of electric field.

Total discharge current as a function of pressure was calculated for various values of the electric field intensity and for various gases. Total discharge current was calculated using (7). The work function of aluminium is 4.25 eV [12] so electrode surface ionization coefficient (γ) contribution is negligible for the considered voltages. Electron ionization coefficient was calculated using (3), and mean free path was calculated using (4).

Total discharge current versus pressure for argon and molecules of nitrogen and oxygen is shown in Fig. 7. Pressure range is between $1{\cdot}10^5$ Pa and $7{\cdot}10^5$ Pa. Ionization energies are 15.763 eV for argon [13], 12.0697 eV for oxygen [14] and 15.581 eV for nitrogen [15]. Kinetic diameters are 340 pm for argon [16], 346 pm for oxygen [17], and 364 pm for nitrogen [17].

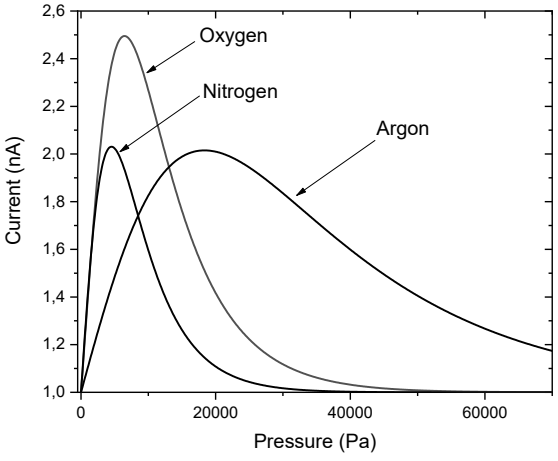


Fig. 7. Current versus pressure for O_2, N_2 and Ar for intensity of electric field of $1{\cdot}10^7$ V/m.

Total discharge current for various intensities of the electric field is shown in Fig. 8.

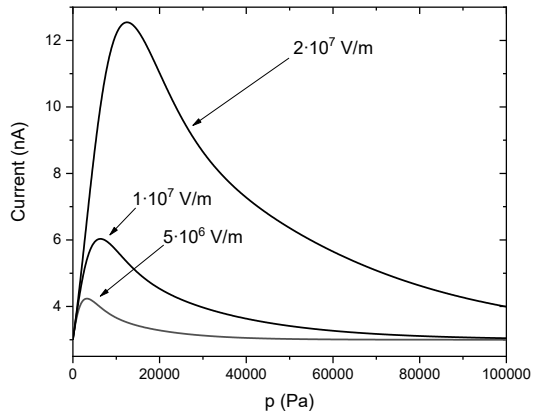


Fig. 8. Total discharge current versus pressure for different intensity of the electric field for Ar, O_2, and N_2 all together.

Influence of the intensity of the electric field (i.e. applied voltage) can be understood from the plot of total current vs. pressure for various intensities of electric field, Fig. 8.

Fig. 8. shows that current has maximum values on 2300 Pa, 4500 Pa and 9000 Pa for intensity of the electric field of $5{\cdot}10^6$ V/m, $1{\cdot}10^7$ V/m, and $2{\cdot}10^7$ V/m, respectively.

Applying different voltage on sensor pads can provide required pressure range.

Current in Fig. 7, and Fig. 8 is calculated for one pair of wedges (Fig. 2). In a single bunch there are 2110 pairs of wedges and in whole sensor, there are 9240 pairs of wedges, so current has to be multiplied by these factors.

IV. Proposal of the Experiment

Fig. 9 shows simplified scheme of measurement set-up.

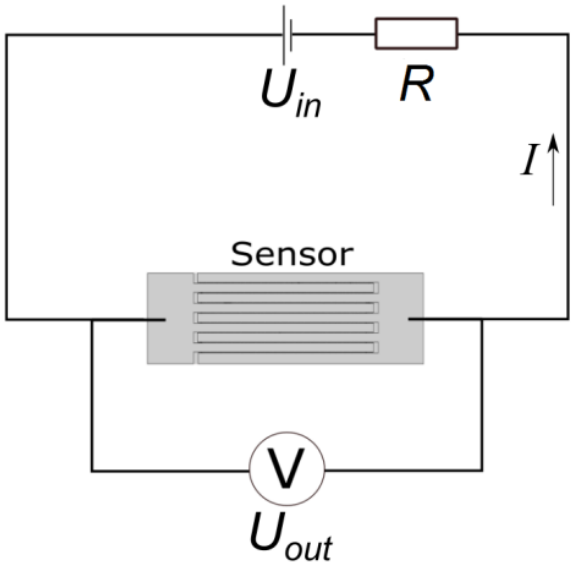


Fig. 9. Simplified scheme of the measurement principle.

Principle of operation can be understood from Fig. 9. Voltage is applied on pads of one bunch. If, for given voltage, gas pressure is above pressure upper limit for which gas between wedges is conductive there will be no current flow through a circuit (resistance of gas between opposite wedges is infinite) and output voltage will be the same as the input voltage, $U_{out} = U_{in}$. If the applied voltage is increased, or if the pressure is reduced, gas between wedges will become conductive and there will be current flow through the circuit. The output voltage will be lower and could be calculated using the equation:

$$U_{out} = U_{in} - IR, \tag{8}$$

where I is the current through the circuit and R is the resistance, Fig. 9. Current in a circuit is a total discharge current (7). Further reduction of the pressure will alter Townsend coefficients and change discharge current and, according to (8), the change in current will give different output voltage.

In the proposed design sensor consists out of four bunches identical with one on Fig. 9 (labelled as sensor). These bunches can be connected in parallel if the current in one bunch is not large enough to be detected.

Schematic for calibration of the sensor is shown in Fig. 10.

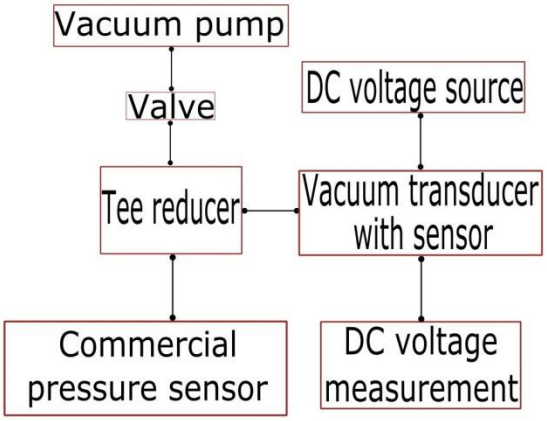


Fig. 10. Block scheme of sensor calibration experiment.

Calibration of the sensor could be done by a recording of the sensor voltage drop as a function of the pressure for a given intensity of the electric field. The pressure should be slowly reduced using a vacuum pump and control valve. The pressure value can be controlled by the commercial pressure sensor. Voltage drop should be measured for different pressure values using appropriate electronics. As a result, the voltage vs. pressure curve could be obtained which may serve as a calibration curve for the sensor.

V. Conclusion

Design and analytical model for planar pressure sensor have been presented. It was shown that it is feasible to fabricate the desired structure using direct laser writing. Calculations based on the presented analytical model have been done for different gases and for various intensities of the electric field. Calculations showed that measurable current-dependence of pressure can be accomplished for a wide range of pressure by applying different voltages. The proposed sensor design seems to be a competitive microscopic device for measuring pressure in a wide vacuum range.

ACKNOWLEDGMENT

This work was supported by the Serbian Ministry of Education, Science and Technological Development under the project TR32008.

REFERENCES

[1] J. H. Leck, *Total and Partial Pressure Measurement in Vacuum Systems*, 1st edition, London, England: Springer Science & Business Media, 2012.

[2] D. Randjelović, A. Petropoulos, G. Kaltsas, M. Stojanović, Ž. Lazić, Z. Djurić, M. Matić, "Multipurpose MEMS Thermal Sensor Based on Thermopiles", Sensors and Actuators A, *Sensors and Actuators A: Physical*, Vol. 141, Issue 2, pp. 404-413, February, 2008.

[3] D. V. Randjelović, M. P. Frantlović, B. L. Miljković, B. M. Popović, Z. S. Jakšić, "Intelligent Thermal Vacuum Sensors Based on Multipurpose Thermopile MEMS Chips", *Vacuum*, Vol. 101, pp. 118-124, March, 2014.

[4] D. Randjelović, V. Jovanov, Ž. Lazić, Z. Djurić, M. Matić, "Vacuum MEMS Sensor Based on Thermopiles – Simple Model and Experimental Results", Proc. 26th International Conference on Microelectronics MIEL 2008, Niš, Serbia, vol. 2, pp. 367-370, 11.-14.5.2008.

[5] G.J. Schulz, A.V. Phelps, ''Ionization Gauges for Measuring Pressures up to the Millimeter Range,'' *Review of Scientific Instruments*, vol. 28, no. 12, 1051-1054, December, 1957.

[6] M. Sarajlić, M. M. Smiljanić, Ž. Lazić, K. Cvetanović-Zobenica, D. Randjelović and D. Vasiljević-Radović, "Direct Laser Writing of micro-structures in vector mode for chemical sensors" Proc. 5th International conference IcETRAN, Palic, Serbia, pp. 949-952, 11.-14.06.2018.

[7] Clewin Software https://wieweb.com/site/ , Accessed 15.02.2019.

[8] J.J. Thomson, *Conduction of Electricity through Gases*, 2nd edition, London, England: Cambridge: At the university press, 1906.

[9] D. Xiao, *Gas discharge and gas insulation*, 1st edition, Shanghai, China: Shanghai Jiao Tong University Press, Springer, 2016.

[10] R. D. Levine, *Molecular Reaction Dynamics*, 1st edition, New York, United States of America: Cambridge University press, 2005.

[11] Y. Yamauchi, R. Shimizu, ''Secondary Electron Emission from Aluminium by Argon and Oxygen Ion Bombardment below 3 keV,'' *Japanese Journal of Applied Physics*, vol. 22, no. 4, pp. 227-229, April, 1983.

[12] E. William, J. Mitchell and J. W. Mitchell, ''The work function of copper, silver and aluminium'' *The Royal Society publishing*, vol. 120, pp. 70-84, December, 1951.

[13] K. M. Weitzel, J. Mahnert, M. Penno, ''ZEKE-PEPICO investigations of dissociation energies in ionic reactions,'' *Chemical physics letters*, vol. 224, pp. 371-380, July, 1994.

[14] R. G. Tonkyn, J. W. Winniczek, M. G. White, ''Rotationally resolved photoionization of O_2 near threshold'', *Chemical Physics Letters*, vol. 164, pp. 137-142, December, 1989.

[15] T. Trickl, E. F. Cromwell, Y. T. Lee, A. H. Kung, ''State-selective ionization of nitrogen in the X_2=0 and v=1 states by two-color (1+1) photon excitation near threshold'', *Journal of Chemical Physics*, vol. 91, pp. 6006-6012, November, 1989.

[16] D. W. Breck, *Zeolite Molecular Sieves: Structure, Chemistry and Use*, 1st edition, New York: Wiley, 1973.

[17] A. F. Ismail, K.C. Khulbe, T. Matsuura, *Gas separation membranes*, 1st edition, Basel, Switzerland: Springer International Publishing Switzerland, 2015.

Etched Parallelogram Patterns with Sides Along <100> and <n10> Directions in 25 wt % TMAH

Milče M. Smiljanić, Žarko Lazić, Branislav Radjenović, Marija Radmilović-Radjenović, Vesna Jović
Milena Rašljić, Katarina Cvetanović Zobenica, Ana Filipović,

***Abstract*— In this paper, we present and analyze etching of parallelogram patterns in the masking layer on a (100) silicon in 25 wt % TMAH water solution at the temperature of 80 ℃. Sides of parallelogram islands in the masking layer are designed along <n10> and <100> crystallographic directions. A 3D simulation of the profile evolution from these patterns during etching of silicon using the level set method is also presented. We determined all crystallographic planes that appear during etching in the experiment and obtained simulated etching profiles of these 3D structures. A good agreement between dominant crystallographic planes through experiments and simulations is obtained.**

***Index Terms*—silicon; wet etching; TMAH; simulation; level set method.**

I. Introduction

Because of its advantages to other etchants (high selectivity to thermal oxide, very smooth etching surface, integrated circuits process compatibility), wet etching of a (100) silicon substrate in tetramethylammonium hydroxide (TMAH) water solution was intensively studied [1-26]. Etched silicon shapes are limited by the mask pattern designs and the etching anisotropy of TMAH water solution. Because of the differences in etch rates [2-8] during anisotropic wet etching some crystallographic planes will appear, while others will disappear. The most used etching patterns in the masking layer in the fabrication of various sensors and actuators were rectangular patterns with sides along <110> and <100> crystallographic directions. In the previous studies of silicon wet etching [1-26], processes were conducted using various etching solutions of TMAH at different temperatures and silicon wafers of various crystallographic orientations. Etching of square patterns with sides along <100> crystallographic directions was analyzed in [9,25]. Etching of octagonal patterns with sides along <210>, <310> and <410> crystallographic directions was discussed in [9] for TMAH water solutions. In our previous paper [25], we studied silicon etching of square and circle patterns in the masking layer when 25 wt % TMAH water solution is used at the temperature of 80 ℃. The sides of square patterns in the masking layer were designed along predetermined <n10> crystallographic directions. Authors in [26] explored etching of a (110) silicon using parallelogram patterns with sides along <110> and <210> crystallographic directions.

This paper presents our further work on a (100) silicon etching in 25 wt % TMAH water solution at the temperature of 80 ℃. For the first time etching of parallelogram patterns in the masking layer with sides that are designed along determined crystallographic directions <n10> (1<n<10) and <100> is analyzed. A 3D simulation of the profile evolution from these patterns islands during etching of silicon based on the level set method is presented. The level set method for evolving interfaces belongs to the geometric type of methods, and it is specially designed for profiles that can develop sharp corners, change of topology and undergo orders of magnitude changes in speed. All simulations are performed using a three-dimensional (3D) anisotropic etching simulator based on the sparse field method for solving the level set equations, described in our previous publications [27-33]. Pictures of the simulated etching profiles are rendered by Paraview visualization package [34]. We presented the simulated etching profiles and SEM micrographs to demonstrate all exposed crystallographic planes. Our aim is to observe and analyze the appearance of crystallographic planes and verify agreement of simulation with experimental results. Knowing the evolution of crystallographic planes during etching enables easy fabrication of various 3D silicon structures that can be used in the design of sensors and actuators.

II. Experimental work

Phosphorus-doped {100} oriented 3” silicon wafers (Wacker, SWI) with mirror-like single or double side polished

Milče M. Smiljanić is with the Institute of Chemistry, Technology and Metallurgy-Centre of Microelectronic Technologies (IHTM-CMT), University of Belgrade, Njegoševa 12, 11000 Belgrade, Serbia (e-mail: smilce@nanosys.ihm.bg.ac.rs).

Žarko Lazić is with the Institute of Chemistry, Technology and Metallurgy-Centre of Microelectronic Technologies (IHTM-CMT), University of Belgrade, Njegoševa 12, 11000 Belgrade, Serbia (e-mail: zlazic@nanosys.ihm.bg.ac.rs).

Branislav Radjenović is with the Institute of Physics, University of Belgrade, Pregrevica 118, 11080 Belgrade, Serbia (e-mail: bradjeno@ipb.ac.rs).

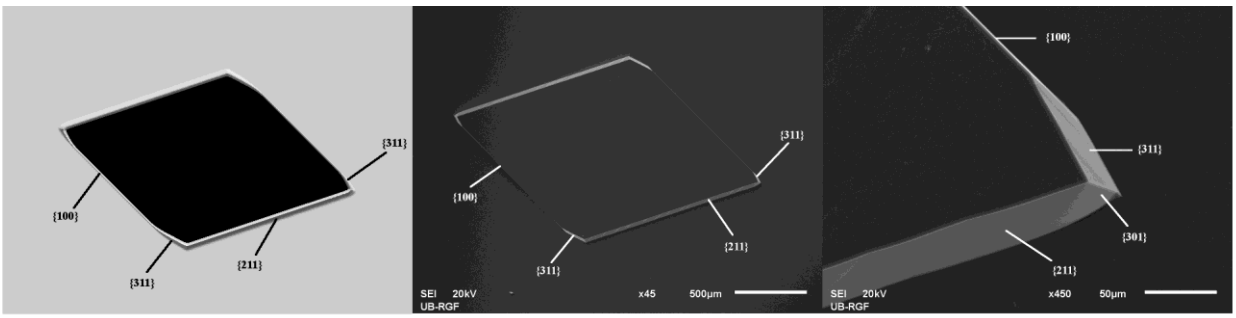
Marija Radmilović-Radjenović is with the Institute of Physics, University of Belgrade, Pregrevica 118, 11080 Belgrade, Serbia (e-mail: marija@ipb.ac.rs).

Vesna Jović is with the Institute of Chemistry, Technology and Metallurgy-Centre of Microelectronic Technologies (IHTM-CMT), University of Belgrade, Njegoševa 12, 11000 Belgrade, Serbia (e-mail: vjovic@nanosys.ihm.bg.ac.rs).

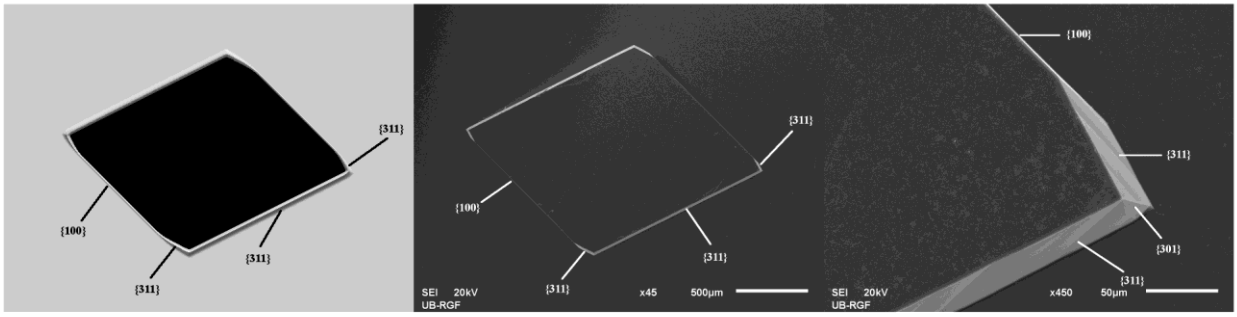
Milena Rašljić is with the Institute of Chemistry, Technology and Metallurgy-Centre of Microelectronic Technologies (IHTM-CMT), University of Belgrade, Njegoševa 12, 11000 Belgrade, Serbia (e-mail: milena@nanosys.ihm.bg.ac.rs).

Katarina Cvetnović Zobenica is with the Institute of Chemistry, Technology and Metallurgy-Centre of Microelectronic Technologies (IHTM-CMT), University of Belgrade, Njegoševa 12, 11000 Belgrade, Serbia (e-mail: katarina@nanosys.ihm.bg.ac.rs).

Ana Filipović is with the Institute of Chemistry, Technology and Metallurgy-Centre of Microelectronic Technologies (IHTM-CMT), University of Belgrade, Njegoševa 12, 11000 Belgrade, Serbia (e-mail: ana@nanosys.ihm.bg.ac.rs).

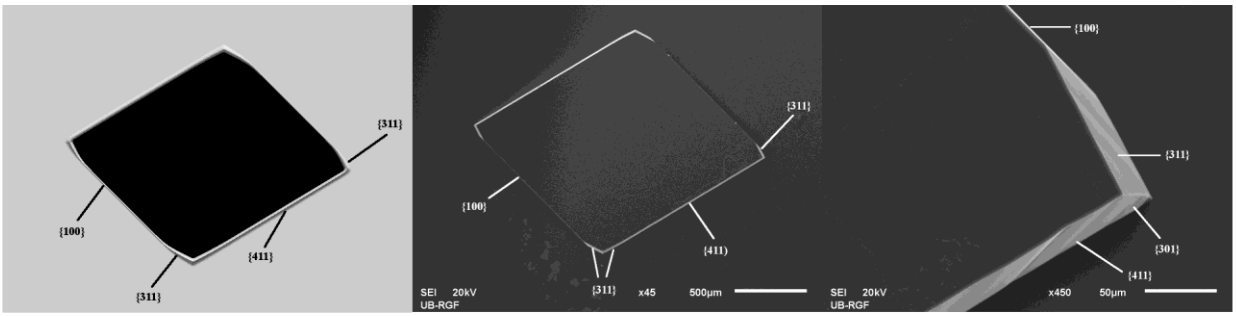


(a)

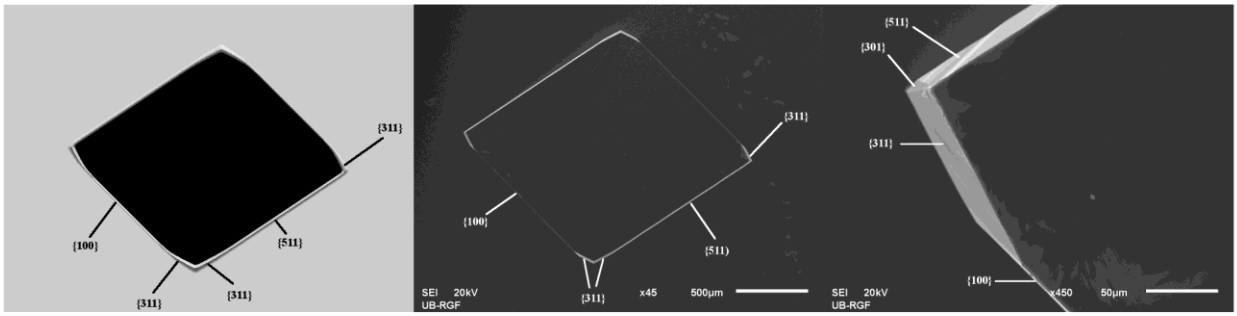


(b)

Fig. 1. Simulated etching profiles and SEM micrographs of the etched parallelogram patterns with sides along <100> and: (a) <210> directions; (b) <310> directions. Enlarged details of etched acute angles in the masking layer in both cases.



(a)



(b)

Fig. 2. Simulated etching profiles and SEM micrographs of the etched parallelogram patterns with sides along <100> and: (a) <410> directions; (b) <510> directions. Enlarged details of etched acute angles in the masking layer in both cases.

surfaces and 1-5 Ωcm resistivity have been used. Anisotropic etching has been done in pure TMAH 25 wt. % water solution (Merck). The etching temperature was 80 ℃. Wafers were standard cleaned and covered with SiO_2 thermally grown at 1100 ^{0}C in an oxygen ambient saturated with water vapour (at least 1 μm thick). SiO_2 was etched in BHF in a photolithographic process in order to define parallelogram patterns along determined crystallographic directions. Again, wafers were subjected to standard cleaning procedure and were dipped before etching for 30 s in HF (10 %) to remove native SiO_2 followed by rinsing in deionized water. Etching of whole 3" wafer was carried out in a thermostated glass vessel containing about 0.8 dm^3 of the solution with electronic temperature controller stabilizing temperature within ±0.5 ℃. The vessel was on the top of a hot plate and closed with a teflon lid that included a water-cooled condenser to minimize evaporation during etching. The wafer was oriented vertically in a teflon basket inside the glass vessel. Throughout the process, the solution was electromagnetically stirred with a velocity of 300 rpm. After reaching the desired depth, the wafer was rinsed in deionized water and dried with nitrogen.

III. RESULTS AND DISCUSSION

Parallelogram patterns in the masking layer are designed with sides that are along determined crystallographic directions <n10> (1<n<10) and <100>. The acute corners of islands in the masking layer formed by <n10> and <100> crystallographic directions are larger than 45^0 and smaller than 90^0. The values of acute corners are given in Table 1. Different 3D shapes are obtained during etching of silicon in 25 wt %. In the cases of n<7 during etching initial acute angles are split into two angles in the masking layer. In the case of n=2, they are acute and obtuse angles. In the case of n=3, they are right and obtuse angles. In other cases, we have two obtuse angles. The sidewalls of the first convex corners angle are defined only by {n11} and {311}-{301} (or {401}-{203}) families at the beginning of etching [25], as shown in Fig. 1-3. This is similar to the etching of square patterns in the masking layer with sides along <n10> crystallographic directions [25]. The sides of obtuse convex corners angles are defined by {100} and {311} families at the beginning of etching. In the cases of 6<n<10 during etching initial acute angles are split in three obtuse angles in the masking layer The sidewalls of the central convex corners angles are defined by two planes of {311} families at the beginning of etching [25], as shown in Fig. 3-4. This is similar to the etching of square pattern in the masking layer with sides along <100> crystallographic directions [25]. The sidewalls of two other obtuse convex corners angles are defined by {100} and {311} families or {n11} and {311}-{301} (or {401}-{203}) families. The etch depth in Fig. 1-4 is 55 μm. Appearance of {301} families is hard to observe in the simulated etching profiles. These planes have smaller surface areas than the dominant ones. The planes obtained in the simulation are more round and the edges of the convex corners tend to soften [12-13,25].

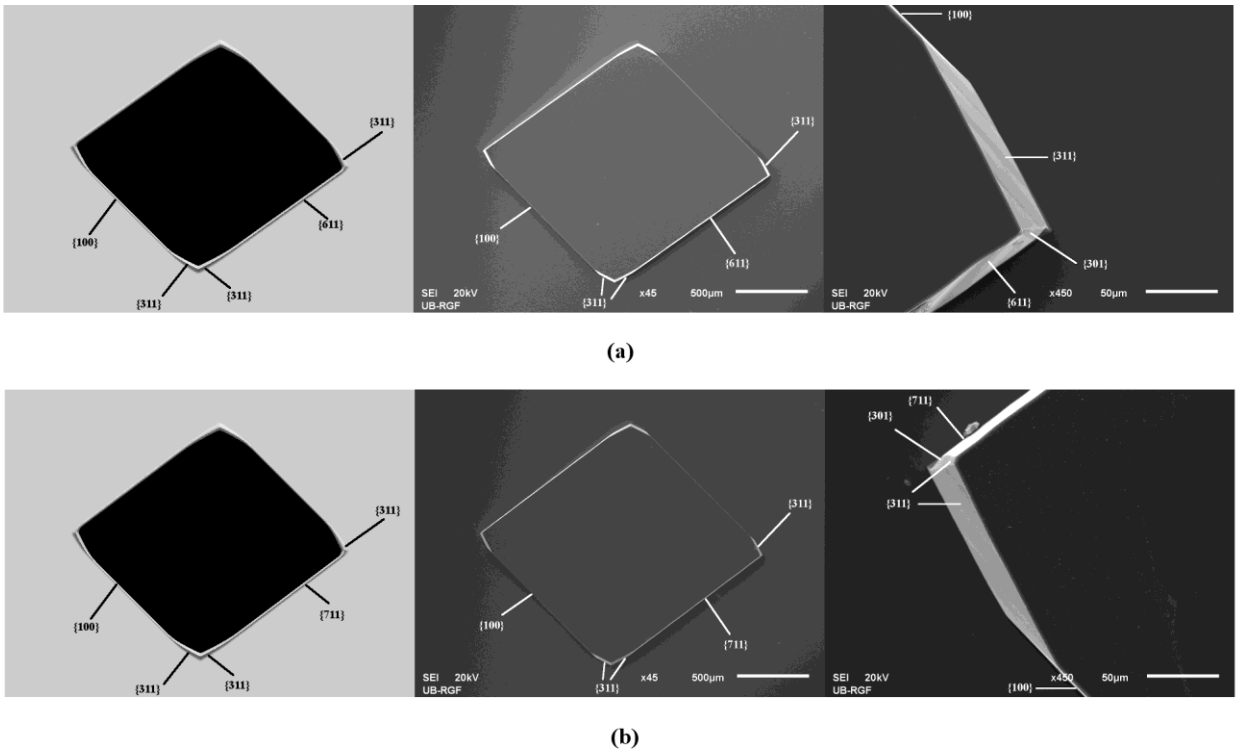


Fig. 3. Simulated etching profiles and SEM micrographs of the etched parallelogram patterns with sides along <100> and: (a) <610> directions; (b) <710> directions. Enlarged details of etched acute angles in the masking layer in both cases.

TABLE I
THE VALUES OF ACUTE ANGLES OF THE PARALLELOGRAMS

Crystallographic direction <n10>	Acute angle [^{o}]
<210>	63.4
<310>	71.6
<410>	76
<510>	78.7
<610>	80.5
<710>	81.9
<810>	82.9
<910>	83.7

The etching of island obtuse corners in the masking layer is the same for all cases except n=2, as shown in figures 1-4. During etching initial obtuse angles are split into three obtuse angles in the masking layer. The sidewalls of the first obtuse angle are defined by {311} and {n11} planes. The sidewalls of the second obtuse angle are defined by two planes of {311} family. This is similar to etching of square pattern in the masking layer with the sides along <100> crystallographic directions [25]. The sidewalls of the third obtuse angle are defined by {311} and {100} planes. For the case of n=2, the obtuse convex corner split only into two obtuse angles. The sidewalls of the first obtuse angle are defined by {311} and {211} planes. The second obtuse angle is defined by planes of {311} and {100} families.

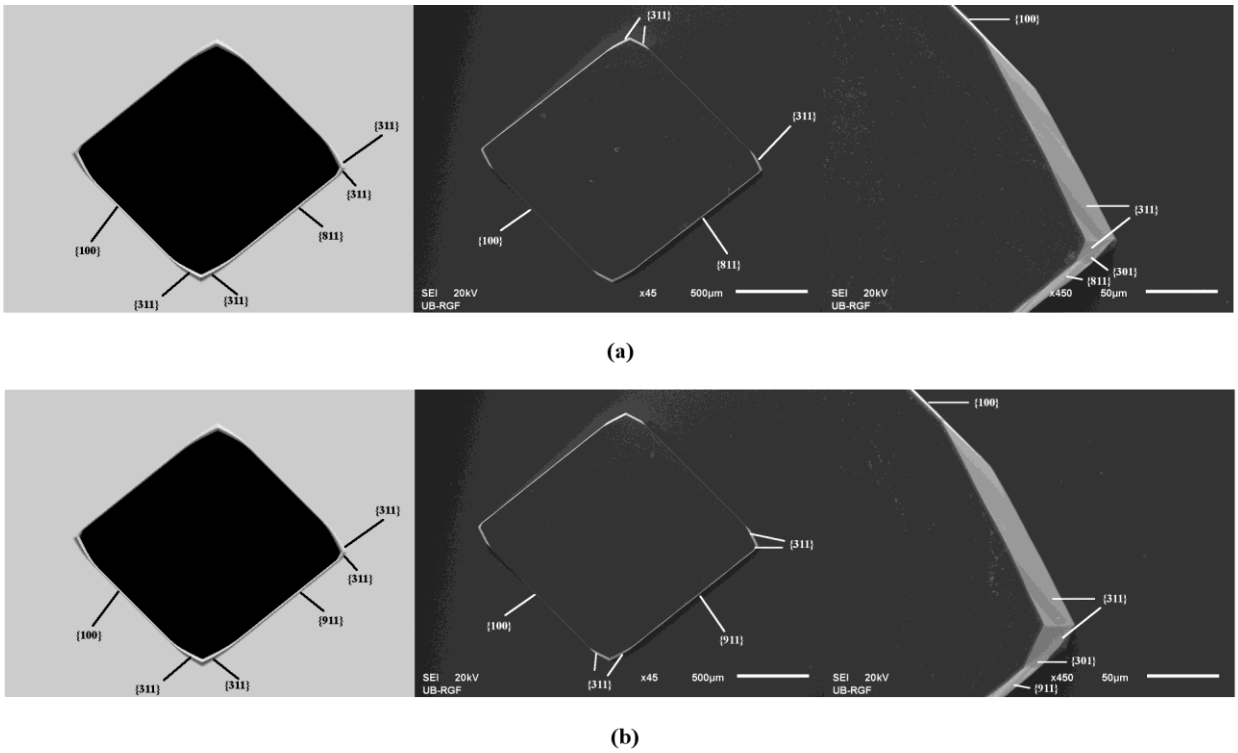


Fig. 4. Simulated etching profiles and SEM micrographs of the etched parallelogram patterns with sides along <100> and: (a) <810> directions; (b) <910> directions. Enlarged details of etched acute angles in the masking layer in both cases.

As etching continues sides planes of {311} family from the nearby convex corners become dominant and planes of {n11} (n>2) and {100} families disappear. The obtained shape is a truncated pyramid with the sidewalls defined by {311} and {301} families. The base of the pyramid in the masking layer is parallelogram with the sides along <310> crystallographic directions. In the case of n=2, the sidewalls are defined by {211}, {311} and {301} families. The base of the pyramid in the masking layer is parallelogram with sides along <310> and <210> crystallographic directions.

IV. CONCLUSION

In this paper, we studied silicon etching of parallelogram patterns in the masking layer in 25 wt % TMAH water solution at the temperature of 80 ℃. The sides of parallelogram islands were designed along <n10> and <100> crystallographic directions. We analyze the etching of islands in the masking layer using both the experiments and the simulations. All the crystallographic planes that appear during etching of silicon structures are determined. A good

agreement between dominant crystallographic planes through experiments and simulations is obtained. A comprehensive insight into the evolution of parallelogram patterns for different crystallographic directions can provide new ideas for the successful mask design of silicon microdevices.

ACKNOWLEDGMENT

This work has been partially funded by the Serbian Ministry of Education and Science within the framework of the project TR32008 and O171036.

REFERENCES

[1] V. Lindroos, M. Tilli, A. Lehto , T Motooka, "Wet Etching of Silicon" in *Handbook of Silicon Based MEMS Materials and Technologies*, William Andrew, Elsevier, 2010. https://www.elsevier.com/books/handbook-of-silicon-based-mems-materials-and-technologies/tilli/978-0-8155-1594-4

[2] J. Frühauf, *Shape and Functional Elements of the Bulk Silicon Microtechnique*, Springer-Verlag, Berlin, 2005. http://www.springer.com/gp/book/9783540221098

[3] M. Shikida, K. Sato, K. Tokoro, D. Uchikawa, "Differences in anisotropic etching properties of KOH and TMAH solutions", *Sensors and Actuators A* vol. 80 no. 2, pp. 179-188, Mar. 2000. https://doi.org/10.1016/S0924-4247(99)00264-2

[4] K. Sato, M. Shikida, T. Yamashiro, K. Asaumi, Y. Iriye, M. Yamamoto, "Anisotropic etching rates of single-crystal silicon for TMAH water solution as a function of crystallographic orientation", *Sensors and Actuators A* vol. 73 no. 1-2, pp. 131-137, Mar. 1999. https://doi.org/10.1016/S0924-4247(98)00271-4

[5] D. Resnik, D. Vrtacnik, U. Aljancic, S. Amon, "Wet etching of silicon structures bounded by (311) sidewalls", *Microelectronic Engineering* vol. 51-52 pp. 555-566, May 2000. https://doi.org/10.1016/S0167-9317(99)00519-5

[6] D. Resnik, D. Vrtacnik, S. Amon, "Morphological study of {311} crystal planes anisotropically etched in (100) silicon: role of etchants and etching parameters", *J. Micromech. Microeng.* vol. 10 no.3 pp. 430-439, Apr. 2000. https://doi.org/10.1088/0960-1317/10/3/319

[7] H. Yang, M. Bao, S. Shen, X. Li, D. Zhang, G. Wu, "A novel technique for measuring etch rate distribution of Si", *Sensors and Actuators A* vol. 79 no. 2 pp.136-140, Feb. 2000. https://doi.org/10.1016/S0924-4247(99)00270-8

[8] L.M. Landsberger, S. Naseh, M. Kahrizi, M. Paranjape, "On Hillocks Generated During Anisotropic Etching of Si in TMAH", *IEEE J. Microelectromech. Syst.* vol. 5 no. 2 pp. 106-116, Jul. 1996. DOI: 10.1109/84.506198

[9] I. Zubel, I. Barycka, K. Kotowska, M. Kramkowska, "Silicon anisotropic etching in alkaline solution IV: The effect of organic and inorganic agents on silicon anisotropic etching process", *Sensors and Actuators A* vol. 87 no. 3 pp. 163-171, Jan. 2001. https://doi.org/10.1016/S0924-4247(00)00481-7

[10] Trieu, H.K.;Mokwa, W. A generalized model describing corner undercutting by the experimental analysis of TMAH/IPA, *J. Micromech. Microeng.*, 1998, 8, 80-83. https://doi.org/10.1088/0960-1317/8/2/009

[11] Sarro, P.M.; Brida, D.; Vlist, W.v.d.; Brida, S. Effect of surfactant on surface quality of silicon microstructures etched in saturated TMAHW solutions, *Sensors and Actuators A*, 2000, 85, 340-345. https://www.sciencedirect.com/science/article/pii/S0924424700003174

[12] M.M. Smiljanić, V. Jović, Ž. Lazić, "Maskless convex corner compensation technique on a (100) silicon substrate in a 25 wt. % TMAH water solution", *J. Micromech. Microeng.* vol. 22 no.11 pp. 115011, Sep. 2012. https://doi.org/10.1088/0960-1317/22/11/115011

[13] M.M. Smiljanic, B. Radjenović, M. Radmilović-Radjenović , Ž. Lazić, V. Jović, "Simulation and experimental study of maskless convex corner compensation in TMAH water solution", *J. Micromech. Microeng.* vol. 24, no. 11 pp. 115003, Oct. 2014. https://doi.org/10.1088/0960-1317/24/11/115003

[14] R. Mukhiya, A. Bagolini, B. Margesin, M. Zen, S. Kal, "<100> bar corner compensation for CMOS compatible anisotropic TMAH etching", *J. Micromech. Microeng.* vol. 16 no. 11 pp. 2458-2462, Oct. 2006. https://doi.org/10.1088/0960-1317/16/11/029

[15] A. Bagolini, A. Faes, M. Decarli, "Influence of Etching Potential on Convex Corner Anisotropic Etching in TMAH Solution", *IEEE J. Microelectromech. Syst.* vol. 19 no. 5 pp. 1254-1259, Oct. 2010. **DOI:** 10.1109/JMEMS.2010.2067436

[16] R. Mukhiya, A. Bagolini, T.K. Bhattacharyya, L. Lorenzelli, M. Zen, "Experimental study and analysis of corner compensation structures for CMOS compatible bulk micromachining using 25 wt% TMAH", *Microelectronics Journal*, 2011, 42, 127-134. https://doi.org/10.1016/j.mejo.2010.08.018

[17] A. Merlos, M.C Acero, M.H. Bao, J. Bausells, J. Esteve, "A study of the undercutting characteristics in the TMAH-IPA system", *J. Micromech. Microeng.* 1992, 2, 181-183. https://doi.org/10.1088/0960-1317/2/3/014

[18] A. Merlos, M.C Acero, M.H. Bao, J. Bausells, J. Esteve, "TMAH/IPA anisotropic etching characteristics", *Sensors and Actuators A*, 1993, 37–38, 737-743 . https://doi.org/10.1016/0924-4247(93)80125-Z

[19] P. Pal, K. Sato, M. Shikida, M.A. Gosalvez, "Study of corner compensating structures and fabrication of various shape of MEMS structures in pure and surfactant added TMAH", *Sensors and Actuators A* vol. 154 no.2 pp. 192-203, Sep. 2009. https://doi.org/10.1016/j.sna.2008.09.002

[20] P. Pal, K. Sato, S. Chandra, "Fabrication techniques of convex corners in a (100)-silicon wafer using bulk micromachining: a review", *J. Micromech. Microeng.* vol. 17 no. 10 R111-R133, Sep. 2007. https://doi.org/10.1088/0960-1317/17/10/R01

[21] O. Powell, H.B. Harrison "Anisotropic etching of {100} and {110} planes in (100) silicon", *J Micromech. Microeng.* vol. 11 no. 3 pp. 217-220, Feb. 2001. http://iopscience.iop.org/article/10.1088/0960-1317/11/3/309

[22] P.Pal, K. Sato,"A comprehensive review on convex and concave corners in silicon bulk micromachining based on anisotropic wet chemical etching", *Micro Nano Syst. Lett.*, vol. 3 pp. 1-42, May 2015. https://mnsl-journal.springeropen.com/articles/10.1186/s40486-015-0012-4

[23] P. Pal, S. Haldar, S. S. Singh, A. Ashok, X. Yan, K. Sato, "A detailed investigation and explanation to the appearance of different undercut profiles in KOH and TMAH", J Micromech. Microeng., vol.24 no.9 pp. 095026 (1-9), Avg 2014. http://iopscience.iop.org/article/10.1088/0960-1317/24/9/095026/meta

[24] P. Pal, K. Sato, M.A. Gosalvez, M Shikida, "Study of rounded concave and sharp edge convex corners undercutting in CMOS compatible anisotropic etchants", J Micromech. Microeng., 2007, 17, 2299–2307. https://doi.org/10.1088/0960-1317/17/11/017

[25] M.M. Smiljanic, B. Radjenović, M. Radmilović-Radjenović, Ž. Lazić, V. Jović, "Evolution of Si crystallographic planes-etching of square and circle patterns in 25 wt % TMAH", *Micromachines*, *10*(2), 102, Jan. 2019. https://doi.org/10.3390/mi10020102

[26] P. Pal, M. A. Gosalvez, K. Sato5, H. Hida, Y. Xing, "Anisotropic etching on Si{1 1 0}: experiment and simulation for the formation of microstructures with convex corners", *J. Micromech. Microeng.* vol. 24, no. 12 pp. 125001, 2014. https://doi.org/10.1088/0960-1317/24/12/125001

[27] S. Osher, J.A. Sethian, "Fronts Propagating with Curvature Dependent Speed: Algorithms Based on Hamilton-Jacobi Formulations", *J. Comp. Phys.* vol. 79 no. 1 pp. 12-49, Nov. 1988. https://doi.org/10.1016/0021-9991(88)90002-2

[28] B. Radjenović, J.K. Lee, M. Radmilović-Radjenović, "Sparse field level set method for non-convex Hamiltonians in 3D plasma etching profile simulations", *Computer Physics Communications* vol. 174 no.2 pp. 127–132, Jan. 2006. https://doi.org/10.1016/j.cpc.2005.09.010

[29] B. Radjenović, M. Radmilović-Radjenović, M. Mitrić, "Non-convex Hamiltonians in 3D level set simulations of the wet etching of silicon", *Appl. Phys. Lett.* Vol. 89 no. 21 pp. 213102 (1-2), Oct. 2006. https://doi.org/10.1063/1.2388860

[30] B. Radjenović, M. Radmilović-Radjenović, "3D simulations of the profile evolution during anisotropic wet etching of silicon", *Thin Solid Films* vol. 517 no. 14 pp. 4233–4237, May 2009. https://doi.org/10.1016/j.tsf.2009.02.007

[31] B. Radjenović, M. Radmilović-Radjenović, M. Mitrić, "Level Set Approach to Anisotropic Wet Etching of Silicon", *Sensors* vol. 10 no.5 pp. 4950-4967, May 2010. doi:10.3390/s100504950

[32] C. Montoliu, N. Ferrando, M.A. Gosalvez, J. Cerda, R.J. Colom, “Level set implementation for the simulation of anisotropic etching: application to complex MEMS micromachining”, *J. Micromech. Microeng.* vol. 23 no.7 pp. 075017, Jun. 2013. https://doi.org/10.1088/0960-1317/23/7/075017

[33] C. Montoliu, N. Ferrando, M.A. Gosalvez, J. Cerda, R.J. Colom, “Implementation and evaluation of the Level Set method - Towards efficient and accurate simulation of wet etching for microengineering applications”, *Computer Physics Communications* vol. 184 no. 10 pp. 2299–2309, Oct. 2013. https://doi.org/10.1016/j.cpc.2013.05.016

[34] http://www.paraview.org/

Reversed ellipsoidal troughs sculpted in plasmonic multilayer nanomembranes

Marko Obradov, Zoran Jakšić, *Senior Member, IEEE* , Ivana Mladenović, Dragan Tanasković, Dana Vasiljević Radović

***Abstract*—Nanomembranes represent a versatile novel building block in micro- and nanoelectromechanical systems, inspired by biological cell membranes. If built as quasi-2D multilayer metal-dielectric nanocomposites, they represent a natural choice for the use in plasmonics and optical metamaterials, ensuring a new degree of design freedom and thus a number of new applications.**

In this contribution we consider surface sculpting of two types of three-layer nanomembranes – metal-insulator-metal (MIM) and insulator-metal-insulator (IMI) structures. The geometry we analyze represents troughs with ellipsoidal profiles sculpted in the multilayer nanomembranes, simultaneously acting as plasmonic waveguides and ensuring tailoring of frequency dispersion of the plasmonic nanomembranes.

We determine frequency dispersions of the scattering parameters for the simulated nanomembranes, as well as the spatial ditribution of the optical near fields in and around them, both evanescent (plasmon-polariton based) and propagating. To this purpose we utilize finite element method simulations.

The obtained results show that our IMI structures exhibit a behavior similar to that of the EOT arrays, resulting in optical transparency of the structure for resonant plasmonic modes. On the other hand, MIM structures offer an excellent confinement of electromagnetic radiation within the dielectric layer. We conclude that both MIM and IMI channels allow for additional degrees of freedom in customization the nanomembrane evanescent fields, making a way to numerous potential applications.

***Index Terms*—Nanomembranes; Plasmonics; Metamaterials; Optical Multilayers; Nanooptics**

Marko Obradov is with Centre of Microel. Technologies, Institute of Chemistry, Technology and Metallurgy, University of Belgrade, Njegoševa 12, 11000 Belgrade, Serbia (e-mail: marko.obradov@nanosys.ihtm.bg.ac.rs).

Zoran Jakšić is with Centre of Microelectronic Technologies, Institute of Chemistry, Technology and Metallurgy, University of Belgrade, Njegoševa 12, 11000 Belgrade, Serbia (e-mail: jaksa@nanosys.ihtm.bg.ac.rs).

Ivana Mladenović is with Centre of Microelectronic Technologies, Institute of Chemistry, Technology and Metallurgy, University of Belgrade, Njegoševa 12, 11000 Belgrade, Serbia (e-mail: ivana@nanosys.ihtm.bg.ac.rs).

Dragan Tanasković is with Centre of Microelectronic Technologies, Institute of Chemistry, Technology and Metallurgy, University of Belgrade, Njegoševa 12, 11000 Belgrade, Serbia (e-mail: dragant@nanosys.ihtm.bg.ac.rs).

Dana Vasiljević Radović is with Centre of Microelectronic Technologies, Institute of Chemistry, Technology and Metallurgy, University of Belgrade, Njegoševa 12, 11000 Belgrade, Serbia (e-mail: dana@nanosys.ihtm.bg.ac.rs).

I. Introduction

SYNTHETIC NANOMEMBRANES are artificial, biologically inspired, freestanding, self-supported structures resembling bilipid cell membranes but offering a much wider choice of materials, surface patterns and profiles [1]. They are quasi-2D structures, meaning that their aspect ratios can be huge – their widths and lengths can be even six to seven orders of magnitude larger than their thickness [2]. The possibility to functionalize nanomembranes [3] opens a pathway towards myriads of pratical applications. One of the possible approaches to such functionalization is 3D surface sculpting of nanomembranes, as proposed in [4] and described in more details in [3].

Plasmonic nanomembranes represent mono- or multilayer structures that can be described as 1D plasmonic crystals [5]. They consist of one or more of conductive materials whose electron dynamics is described by Drude model, with possible incorporation of dielectric(s). Freestanding plasmonic nanomembranes represent a new building block in photonics, subwavelength optics and plasmonics and they can be as simple as a nanometer-thin metallic sheath surrounded by dielectric (air). The subwavelength thickness of nanomembranes together with their electromagnetic symmetry ensures coupling of surface plasmons polaritons (SPP) from their both sides, resulting in the appearance of the SPP with extremely large propagation paths. Such SPPs are known as long range (LR) SPPs [6].

Similarly to metamaterials and nanoplasmonic structures, plasmonic nanomembranes exhibit peculiar phenomena, many of which are not met in natural materials, including extremely high values of refractive index, very low and zero values, as well as negative refractive index. It may be safely said that plasmonic nanomembranes enable one to manipulate evanescent near fields, practically making it possible to tailor light at will. A whole new field of electromagnetic optics developed from these properties – the transformation optics [7, 8]. This on the other hand ensures various applications like ultrasensitive chemical and biological sensors with single-molecule sensitivity, superlenses and hyperlenses, superabsorbers, superconcentrators, invisibility shields (cloaking devices), all-optical integrated circuits, and many more [9].

Since SPPs exist outside of light cone coupling of incident light to these modes requires some form of impedance matching between the two. The use of diffractive gratings for

this purpose ensures not only coupling between propagating modes and SPPs but also allows tailoring of dispersive properties by changing the geometrical properties of the grating. Plasmonic modes of these structures are characterized by high field localization within subwavelength openings.

The analysis of plasmonic nanomembranes integrated with diffractive couplers is of large practical interest since it gives us data on the optimum ways of coupling between evanescent modes of LR SPP and propagating light. More generally, it ensures tailoring of frequency dispersion including the determination of modes that simultaneously exist within the light cone an outside of it.

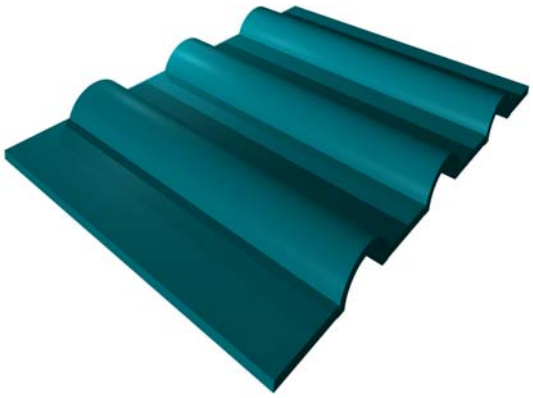


Fig. 1. General presentation of a sculpted nanomembrane with a diffractive grating formed by an array of reversed troughs with ellipsoidal profiles.

In this contribution we analyze a specific geometry of plasmonic nanomembrane integrated with a diffractive grating. The case we analyze is represented in Fig. 1. It shows troughs with ellipsoidal cross-section sculpted in the surface of a nanomembrane. Contrary to [10], here we analyze both IMI and MIM multilayer nanomembranes. To this purpose we utilize COMSOL Multiphysics RF module. We analyze a freestanding plasmonic nanomembrane with embedded diffractive grating. The unit cell of our proposed structure is shown in Fig. 2.

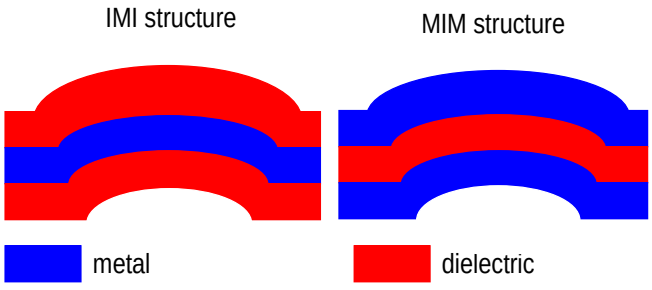


Fig. 2. Cross section view of unit cell of a corrugated multilayered freestanding nanomembrane. Insulator-metal-insulator (IMI) structure (left) and metal-insulator-metal (MIM) structure (right).

II. THEORY

For most conductors based on free electrons their electromagnetic properties in the optical range are well described by lossy extended Drude model. The frequency dispersion of ther complex relative dielectric permittivity $\varepsilon(\omega)$ is given by the following relation [11]:

$$\varepsilon(\omega) = \varepsilon_\infty - \frac{\omega_p^{\ 2}}{\omega^2 + i\gamma\omega}, \tag{1}$$

ε_∞ is the asymptotic dielectric permittivity and $\gamma = 1/\tau$ is the characteristic frequency related to the damping of electron oscillations due to collisions, where τ is the relaxation time of the electron gas and plasma frequency is determined by the concentration of free carriers

$$\omega_p = \frac{ne^2}{m^*\varepsilon_0} \tag{2}$$

where n is electron concentration, e is the free electron charge ($1.6\cdot10^{-19}$ C), ε_0 is the dielectric permittivity of the vacuum ($8.854\cdot10^{-12}$ F/m), and m^* is the effective mass of electrons.

Dispersion relation of SPP propagating on a metal-dielectric interface is given by :

$$k_{spp} = k_0\sqrt{\frac{\varepsilon_d\varepsilon_m}{(\varepsilon_d + \varepsilon_m)}} \tag{3}$$

where $k_0=2\pi/\lambda$ is the wavevector in vacuum, ε_d is the relative permittivity of dielectric and ε_m is the relative permittivity of metal described by Drude model (1).

Coupling between propagating waves and SPP bound on metal-dielectric interface can be achieved by different impedance-matching techniques. By embedding diffractive gratings in the metal-dielectric interface, the impedance matching between SPP and propagating waves is achieved through the diffracted modes of the grating. The wave vector of the diffracted mode is determined by the grating constant a:

$$k_d = \pm m\frac{2\pi}{a} \tag{4}$$

where m is an integer. Coupling of the propagating wave with the SPP occurs when the following condition is met:

$$\vec{k}_{spp} = \vec{k}_d + \vec{k}_p \tag{5}$$

where k_p is the wavevector of the wave propagating in-plane, parallel to the interface

$$k_p = \frac{\omega}{c}\sin\theta \tag{6}$$

where c is the speed of light in the medium above the plasmonic surface, ω is the angular frequency and θ is the incident angle of the propagating modes.

For multilayered metal-dielectric structures SPPs on multiple interfaces can couple, leading to the splitting of resonant states, starting with the even and odd states with only two interfaces and expanding into optical bands as the number of layers (interfaces) increases.

III. Results and Discussion

We examined optical properties of the freestanding corrugated IMI and MIM nanomembranes as shown in Fig. 1. using RF module of Comsol Multiphysics software package. The width of the entire unit cell is a = 1000 nm, which is also the periodicity of the embedded diffractive grating. Embedded curvatures of the reversed troughs are modeled as ellipses with 100 nm and 200 nm semi axes (bottom curvature) and 250 nm and 350 nm semi axes (top curvature). Individual metal and dielectric layers are 50 nm thick. The structure is surrounded by air. Metal is chosen to be nickel with Drude model parameters taken from the literature [12] and dielectric is polymer with a refractive index n=1.4.

Our finite element simulations determine the spatial field distributions as well as the frequency dispersion of the transmission and reflection coefficient for TM plane waves incident on the structure from various angles. Two parallel ports, one active and one passive were added above and below the structure to simulate the flow of optical radiation through the simulation domain, with light entering the domain from the top. Floquet boundary conditions are applied to the edges of the unit cell to simulate the periodicity of the structure. The parametric sweep of the wavelengths and incident angles was used to determine the dispersive properties of the scattering parameters and the spatial distributions of the electromagnetic field.

The dispersive properties of the IMI structure are shown in Fig. 3. It is observed in Fig. 3 that IMI structure behaves similarly to extraordinary optical transmission (EOT) arrays, exhibiting increased transparency due to surface plasmonic modes. For normal incidence the IMI structure roughly supports two narrow transparency bands as shown in Fig. 3a with sharp resonant peak at 610 nm. For oblique incidence shown in Fig. 3b and Fig. 3c for 30º and 60º incident angles the structure exhibits a rich modal behavior with a multitude of narrow resonant peaks, especially for the 30º incident angle.

Spatial field distributions for IMI structure at some of the resonant wavelengths are shown in Fig.4-7. Presented field distributions illustrate exceptional capabilities of IMI structure in tailoring spectral and spatial near field enhancement due to plasmonic resonance. Fig. 4 and 5 present two modes for normal incidence with complementary spatial distributions each stemming from different substructure of the complex IMI structure. The first mode at 580 nm stems from the core-shell substructure (reversed troughs of the multilayer nanomembrane) with enhanced scattering on both sides of the membrane.

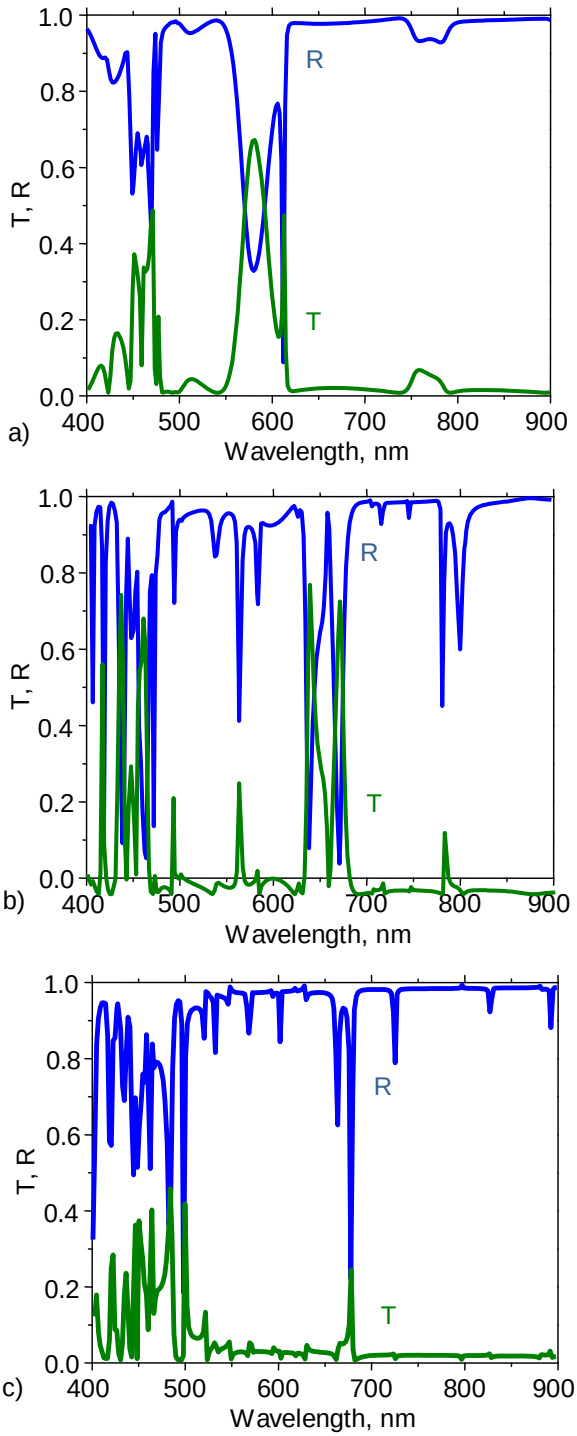


Fig. 3. Dispersive properties of IMI freestanding nanomembrane for different incident angles: a) normal incidence; b) 30º incident angle; c) 60º incident angle. Green: transmission coefficient; blue: reflection coefficient.

The second mode at 610 nm is bound to the surface at the planar part of the structure. Fig. 6 shows that near field enhancement can be localized on the back side of the structure by changing the incident angle. Fig. 7 shows hybridization of the surface modes from different parts of the structure resulting in high field localization spread across the entire structure.

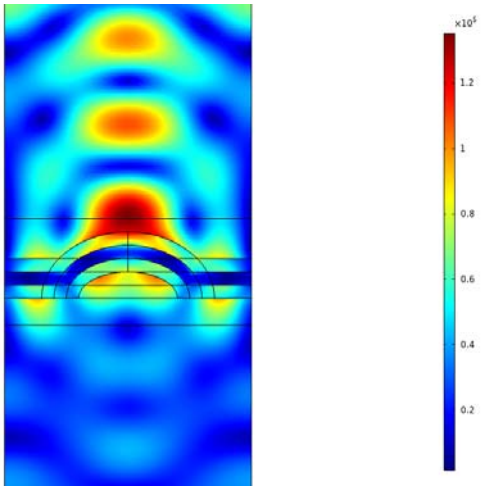


Fig. 4 IMI freestanding nanomembrane, electric field spatial distribution for normal incidence at 580 nm.

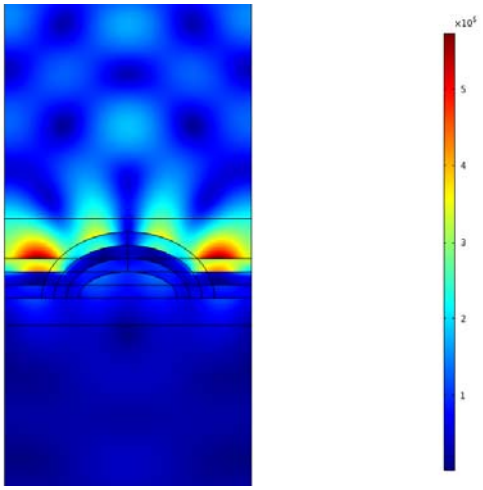


Fig. 5 IMI freestanding nanomembrane, electric field spatial distribution for normal incidence at 610 nm.

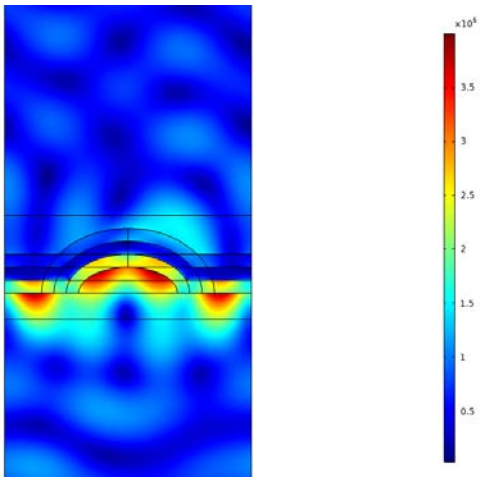


Fig. 6. IMI freestanding nanomembrane, electric field spatial distribution for 30º incident angle at 565 nm.

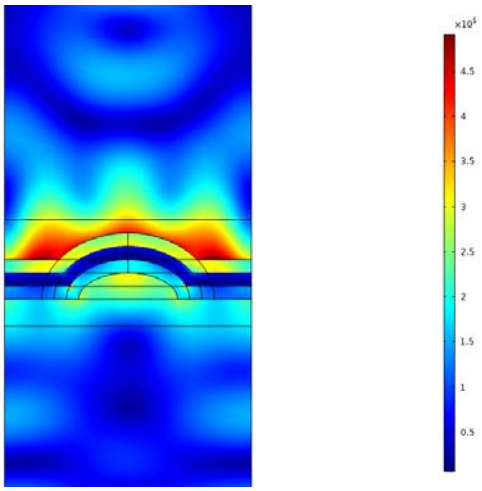
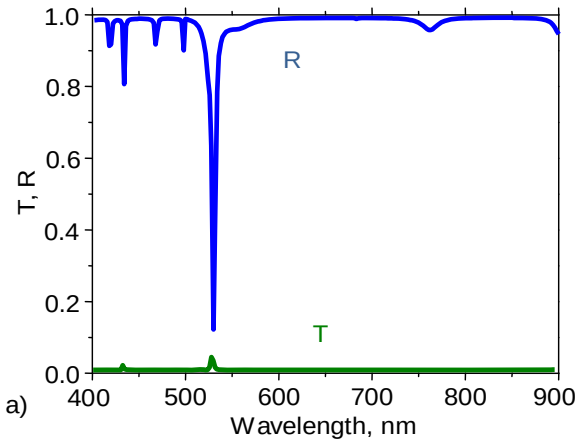


Fig. 7. IMI freestanding nanomembrane, electric field spatial distribution for 60º incident angle at 665 nm.

The dispersive properties of the MIM structure are shown in Fig. 8. Unlike the IMI structure, the MIM structure is highly opaque, denoting that the part of the energy of the propagating wave that isn't being reflected is being funneled into bound surface modes when plasmonic resonance occurs. For normal incidence (Fig. 8a) coupling between the propagating waves and the bound modes is possible due to the corrugated structure of the nanomembrane resulting in a sharp resonant reflection dip at 530 nm. Due to strong coupling between the propagating and the bound modes a high field localization is achieved within the dielectric layer of the MIM structure, as shown in Fig. 9. The number of the bound surface modes that the structure can support increases for oblique incidences, as shown in Fig. 8b and Fig. 8c. The field distribution for another mode with strong coupling for 30º incidence angle at 535 nm is shown in Fig.10. For this mode light localization is moved from the central dielectric layer to the surface of the membrane, allowing again for tailoring of field localization by changing the angle of incidence. Fig. 11 shows a situation similar to the one in Fig. 9 but with much weaker coupling between the propagating and the bound modes resulting in weaker near field enhancement, while retaining the spatial distribution.



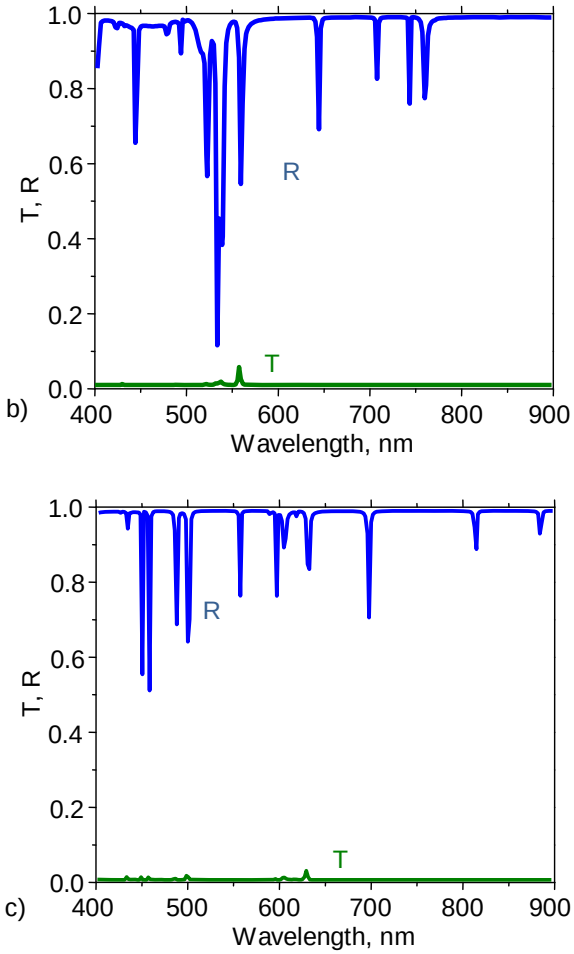


Fig. 8. Dispersive properties of MIM freestanding nanomembrane for different incident angles: a) normal incidence; b) 30º incident angle; c) 60º incident angle. Green: transmission coefficient; blue: reflection coefficient.

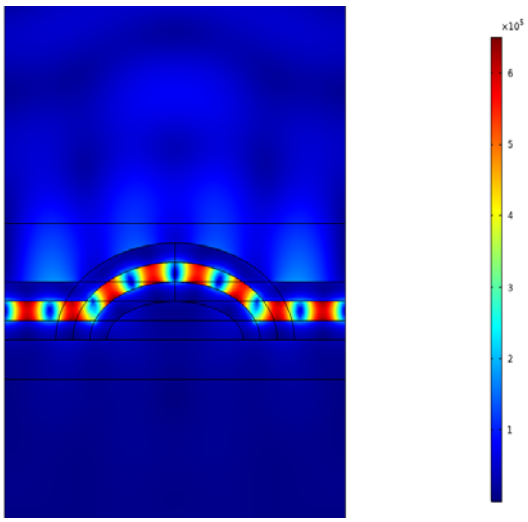


Fig. 9 MIM freestanding nanomembrane, electric field spatial distribution for normal incidence at 530 nm

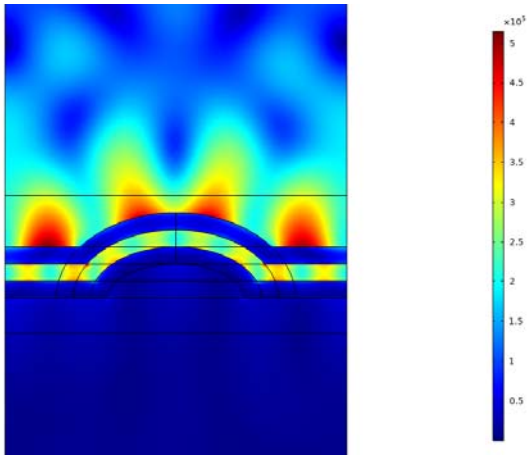


Fig. 10 MIM freestanding nanomembrane, electric field spatial distribution for 30º incident angle at 535 nm.

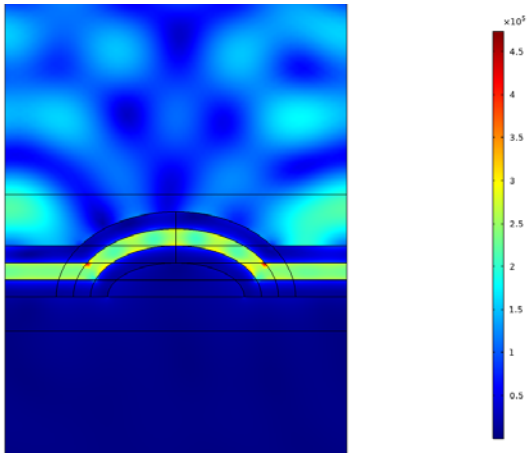


Fig. 11. MIM freestanding nanomembrane, electric field spatial distribution for 60º incident angle at 500 nm.

IV. Conclusion

We analyzed optical properties of corrugated IMI and MIM freestanding nanomembranes using FEM simulation. We have shown that our IMI structure exhibits a behavior similar to that of the EOT arrays, resulting in optical transparency of the structure for resonant plasmonic modes. Together with a rich modal behavior and an angular selectivity of the field spatial distributions, this allows for additional degrees of freedom in customizing the nanomembrane evanescent fields. The MIM structure offers excellent confinement of electromagnetic radiation within the dielectric layer and allows for fine tuning of both the near field enhancement and its spatial distribution by adjusting the angle of incidence.

Acknowledgment

This work was supported by the Serbian Ministry of Education, Science and Technological Development under Project TR32008.

REFERENCES

[1] C. Jiang, S. Markutsya, Y. Pikus, and V. V. Tsukruk, "Freely suspended nanocomposite membranes as highly sensitive sensors," *Nature Mater.,* vol. 3, no. 10, pp. 721-728, 2004.

[2] J. Matović, and Z. Jakšić, "Simple and reliable technology for manufacturing metal-composite nanomembranes with giant aspect ratio," *Microelectron. Eng.,* vol. 86, no. 4-6, pp. 906-909, 2009.

[3] Z. Jakšić, and J. Matovic, "Functionalization of Artificial Freestanding Composite Nanomembranes," *Materials,* vol. 3, no. 1, pp. 165-200, 2010.

[4] J. Matović, and Z. Jakšić, "Three-dimensional surface sculpting of freestanding metal-composite nanomembranes," *Microelectron. Eng.,* vol. 87, no. 5-8, pp. 1487-1490, 2010.

[5] S. M. Vuković, Z. Jakšić, I. V. Shadrivov, and Y. S. Kivshar, "Plasmonic crystal waveguides " *Appl. Phys. A,* vol. 103, no. 3, pp. 615-617, 2011.

[6] P. Berini, "Long-range surface plasmon polaritons," *Adv. Opt. Photon.,* vol. 1, no. 3, pp. 484-588, 2009.

[7] J. B. Pendry, D. Schurig, and D. R. Smith, "Controlling Electromagnetic Fields," *Science,* vol. 312, no. 5781, pp. 1780-1782, 2006.

[8] U. Leonhardt, and T. G. Philbin, "Transformation Optics and the Geometry of Light," *Progress in Optics*, E. Wolf, ed., pp. 69-152, Amsterdam, The Netherlands: Elsevier Science & Technology 2009.

[9] W. L. Barnes, A. Dereux, and T. W. Ebbesen, "Surface plasmon subwavelength optics," *Nature,* vol. 424, no. 6950, pp. 824-830, 2003.

[10] M. Obradov, Z. Jakšić, I. Mladenović, D. Tanasković, D. Vasiljević Radović, "Customization of evanescent near fields on freestanding plasmonic nanomembranes", *5th Internat. Conf. on Electrical, Electronic and Computing Engineering IcETRAN 2018*, Palić, June 11 – 14, pp. 957-960, 2018.

[11] S. A. Maier, *Plasmonics: Fundamentals and Applications*, Springer Science+Business Media, New York, NY, 2007.

[12] A. D. Rakić, A. B. Djurišić, J. M. Elazar, and M. L. Majewski, "Optical properties of metallic films for vertical-cavity optoelectronic devices," *Appl. Opt.,* vol. 37, no. 22, pp. 5271-5283, 1998.

Solution-processed Silver Nanowires as Transparent Electrodes in Solar Cells

Vuk Radmilović

Abstract— **As with all optoelectronic devices like displays, touch panels, or light emitting diodes (LED), solar cells require materials with high electrical conductivity and optical transparency for creating high performance transparent electrodes. Conventional materials used for this purpose have numerous drawbacks mostly regarding demanding and expensive processing methods which is why alternatives are explored. Such are silver nanowires, easily synthesized and processed very cost effective nanostructures. Our work consisted of utilizing silver nanowires as transparent electrodes in two types of solar cells - conventional silicon-based and new generation flexible organic-based. Results show that silver nanowire based transparent electrodes exhibit optoelectronic properties comparable to conventional transparent electrodes and hence, solar cells utilizing silver nanowire based transparent electrodes exhibit competitive power conversion efficiencies to ones utilizing conventional materials.**

Index Terms—**Silver Nanowires (AgNW), Transparent Electrodes, Solar Cells, Photovoltaics, Printed Electronics.**

I. Introduction

It is fairly obvious that, as we proceed into the 21st century one of the main concerns of the human race is providing sustainable energy for ~7.7 billion people on Earth. Keeping in mind the amount of reserves of finite energy sources like coal, petroleum, natural gas and uranium, along with their environmental drawbacks such as atmosphere pollution and radiation, drastic measures have to be taken in order to balance energy production and prevention of environmental disasters like nuclear meltdowns or the greenhouse effect. These drastic measures include completely switching to renewable and environmentally acceptable energy sources. Among these sources, the sun i.e. solar energy stands out as it has a potential of nearly 1.9×10^8 TWh/y [1], more than 200 times the amount of all renewable resources combined. Since the latest data shows that the total global energy consumption is around 20 TWh and is projected to be over 40 TWh by 2050 [1], it is clear that solar energy is predominant choice to make that is both large enough and environmentally acceptable for the planet which is why there is a growing global interest in research of solar cells or photovoltaics, through which solar energy is converted into electricity.

What is the core of the phenomenon of conversion of solar energy into electricity? Solar cells are comprised of semiconductors, materials that posses weakly bonded electrons which occupy the valence energy band. If an energy greater than the band gap energy (gap between valence and conduction bands) is applied to a valence electron, the bonds are broken and the electron is free to move to a higher energy band - the conduction band. Source of this energy is supplied by quanta of light - photons. When the electrons move to the conduction band, a selective contact (n-type electrode) collects such electrons and sends them to the external circuit, where electrons lose their energy (through work) and are restored to the solar cell by another contact (p-type electrode) which returns them to the valence band and their initial energy. This is called the photovoltaic effect. Three generations of solar cells have been developed so far. Unlike the silicon (Si) wafer-based first generation cells and thin film based second generation cells, third-generation technology utilizes far less demanding and less expensive processing methods. Organic compounds like polymers, used in third generation organic solar cells (OSCs) posses the potential of having their properties tailored through molecular design and synthesis yielding desirable intrinsic properties like low weight, flexibility and high optical transparency. Unfortunately, the disadvantages of this technology still include low efficiency (due to low dielectric constant of polymers, absence of a crystal lattice and spectral mismatch between solar spectrum and organic materials) and short lifetimes (conductive polymers degrade in the light)[2]. However, even with these disadvantages, this technology exhibits huge potential as third generation of solar cells can be processed on flexible substrates by various simple printing methods using novel materials in very small quantities. This is particularly attractive as the global market for printed electronics is expected to surpass 44 billion dollars in the coming years, 7.7 out of which will be from solar cell technologies [3].

In solar cell architecture, besides the active layers where the photoconversion actually occurs, electrode layers (selective contacts, as mentioned) serve a very important function – they form differences in potentials which help split charge carriers during photoconversion and guide them, thereby forming electron current. Indium tin oxide (ITO) is the most frequently used material for electrodes in solar cells due to its optimal high optical transparency and low electrical sheet resistance combination. Despite ITO's beneficial properties, needed for transparent electrodes (TE), alternative materials need to be explored due to numerous drawbacks such as high price of indium, it chemical instability, brittleness, demanding processing procedures etc. One of the most promising candidates to replace ITO are NW networks, specifically AgNW networks, due to their excellent optoelectronic properties, second only to ITO and their ease of synthesis and

Vuk Radmilović is with the Faculty of Technology and Metallurgy, University of Belgrade, 4 Karnegijeva, 11020 Belgrade, Serbia (e-mail: vukradmilovic@tmf.bg.ac.rs).

processability [2]. This was the motivation for the work done by our group and our goal was to utilize solution-processed silver nanowires in both, first generation solar cells i.e. crystalline silicon based solar cells, as well as third generation solar cells i.e. tandem bulk heterojunction (BHJ) OSCs, where multiple polymer nanocomposite active layers are stacked atop of each other in order to increase absorption bands and carrier mobility in the device as a whole.

II. The Method

A. AgNWs in Si-based Solar Cells

For application as TE in Si-based solar cells, AgNWs were synthesized in house. For AgNW synthesis, the polyol reduction method [4] was used where ethylene glycol (EG) was utilized as the reducing agent and solvent. By adding EG to the precursor silver nitrate ($AgNO_3$), silver ions (Ag^+) reduce to nuclei (Ag atoms), which are unstable. As these nuclei grow, they form larger clusters, fluctuations disappear and they adopt the role of seeds from which various structures can be grown. These seeds can be multiply twinned (as in the case of AgNWs), single twinned or single crystalline. Formation of aggregates is fuelled by mechanisms of surface diffusion and energy minimization, which is why controlling the kinetics of each step can lead to the control of nanocrystal morphology. In the synthesis of AgNWs, surfactant used was polyvinylpyrrolidone (PVP), a polymer which controls the shape of AgNWs during synthesis by preventing coalescence of nuclei during initial growth. It binds to {100} NW facets, passivating them, enabling growth of only {111} facets in the [110] direction resulting in a one-dimensional (1D) NW structure. By analyzing multiple scanning electron micrographs (SEM) of the NW network, it was deduced that diameter size distribution of AgNWs is characterized by relative monodispersity with an average diameter of ~130nm [2].

After synthesis, AgNWs were annealed at 250°C for 3min in order to reduce electrical resistance of the network as resistance at NWs junctions is much higher than in the NWs themselves. Thermal activation occurs at relatively low T (well below T_m) which leads to surface diffusion of Ag atoms, solid-state wetting and the formation of welded NW junctions [2]. Decrease of the total free energy of the junction is the driving force for the local enhanced diffusion [5,6]. This process is essentially sintering. After local sintering at the AgNW junctions, electrical resistance starts to decrease as PVP encapsulating organic layers starts to decompose [7]. After annealing, the sample was coated with a layer of aluminum doped zinc oxide (AZO) by atomic layer deposition (ALD), which produced a uniform thickness coating of ~100nm. AZO was chosen because it tightens junctions and fills empty space between NWs enhancing conductivity, although resulting in slightly poorer optical transmittance [2].

Solar cell device fabrication was prepared with a diffused p–n junction i.e. boron doped Si wafer pieces (p-type) were used which were doped by phosphorous (n-type) containing spin-on-glass. This was achieved by spin coating and subsequent annealing for dopant diffusion. Finally, after spin-on-glass removal, various top electrodes were deposited [4].

The aim of this work was to characterize the microstructure of AgNWs/AZO nanocomposite and subsequently to develop a new nanocomposite material superior to other electrode solutions such as pristine AZO, Ag grid, or various combinations of the four materials mentioned.

B. AgNWs in BHJ OSCs

For application as TE in tandem BHJ OSCs, AgNWs dispersion (ClearOhm ink) was received from Cambrios Technologies Corporation, with an average AgNW diameter of ~30 nm. Solar cell device fabrication was prepared as follows: for the fully printed cells, all layers have been deposited from ink (solution/suspension) on glass and PET substrates by doctor blading in ambient atmosphere. The layers were processed on substrate and dried for solvent evaporation after each deposition, in the following order: PEDOT:PSS, opaque Ag, PEDOT:PSS, GEN-2:PCBM, ZnO, PEDOT:PSS ,pDPP5T-2:PCBM, ZnO, AgNWs. The role of the first PEDOT:PSS layer is to enhance the adhesion of the reflective Ag to the substrate. The role of the opaque Ag layer is that of a highly reflective electrode for allowing multiple passages of photons into the active layers in order for photocurrent generation increase. The role of the second PEDOT:PSS layer is to act as an electron blocking layer. The GEN-2:PCBM is the first active nanocomposite layer or BHJ, in which actual photoconversion occurs, where GEN-2 (Merck) is the polymer matrix and electron donor (1.76 eV bandgap) while PCBM is [6,6]-phenyl-C61-butyric acid methyl ester, a the fullerene derivative of C60 acting as a nanofiller and electron acceptor. The next two layers, ZnO nanoparticle layer and PEDOT:PSS polymer act as charge carrier recombination layers. The second active layer, again a BHJ nanocomposite layer is pDPP5T-2:PCBM where PCBM is the nanofiller while diketopyrrolopyrrole-quinquethiophene alternating copolymer (pDPP5T-2) is a low bandgap (1.41 eV) polymer matrix. The next ZnO layers serves as a electron transporting layer while the subsequent layer of AgNWs serves as a transparent electrode. For the reference cell, all layers have been printed by means of doctor blading except for ITO and Ag, which were sputtered and thermally deposited under vacuum, respectively [8].

The aim of this work was characterize the microstructure of the solar cell and to develop a fully printed tandem BHJ organic solar cell utilizing materials in the form of inks, among which have been AgNWs used as a top transparent electrode.

III. Main Results

A. AgNWs in Si-based Solar Cells

When two NWs are touching, they are in contact with each other through a junction, extending the percolation of the network but increasing resistance because junctions, as previously mentioned, exhibit enormous resistances of up to 1GΩ [9], several orders higher than individual AgNWs. Nanojoining of AgNW through modification of junction morphology, also called welding or joining, decreases electrical resistance i.e. enhances conductivity along with

bonding strength and fracture resistance. This leads to higher mechanical durability [10] of NW networks, a very important property for possible application in flexible optoelectronic devices.

Examples of AgNW junctions where sintering or "welding" has occurred after annealing are presented in Fig. 1 along with welded AgNW junctions with subsequent AZO deposition.

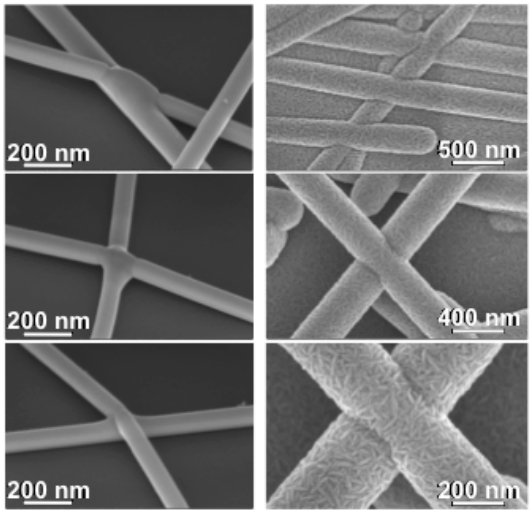


Fig. 1. SEM images of various welded AgNW junctions before (left) and after (right) AZO depositon [2].

In order to understand the morphology of this TE nanocomposite, a cross-section of the sample was made by focused ion beam (FIB) machining, a way of sample preparation for scanning transmission electron microscopy (STEM) analysis. Fig. 2 is a low magnification STEM annular dark field (ADF) image of AgNW cores encapsulated with AZO columnar nano-grain shells, displayed as white and grey regions, respectively. It is clear that the AZO film is very homogenous in thickness distribution and structure, and it seems to grow radially from the AgNWs and substrate outward. AgNWs are confirmed to have a mostly pentagonal twinned structure with fairly uniform diameter distribution [2]. Some AgNWs present in Figure 2, appear to be blurry, the result of their axes being tilted with respect to electron beam. Also shown are the electron and ion beam protective carbon deposited layers which assist against sample amorphization during FIB machining [11].

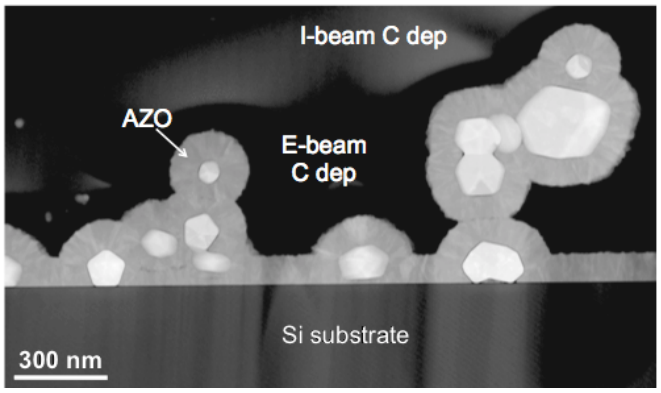


Fig. 2. ADF STEM image of AgNW/AZO cross-section [2].

HAADF STEM image of AgNW/AZO cross-section along with EDS maps noting elemental distribution of silver (turquoise), aluminum (yellow), zinc (green), oxygen (orange) are presented in Fig. 3. It can be seen that aluminum and zinc is present in the shell, while the presence of silver which is limited to the cores of the analyzed structure.

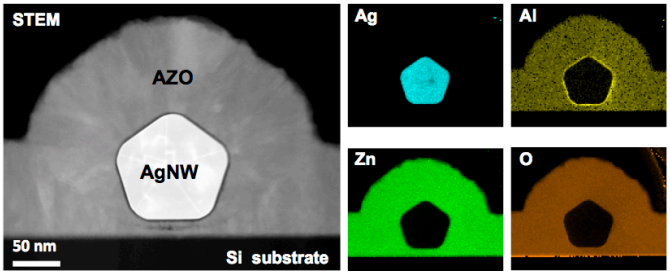


Fig. 3. HAADF STEM image of AgNW/AZO cross-section along with appropriate EDS maps noting elemental distribution of Ag, Al, Zn and O [4].

To put things in perspective of solar cells application, this nanocomposite electrode was pitted against several other electrodes including: thermally evaporated Ag grid, AZO, Ag grid/AZO, and AgNWs/AZO/Ag grid. Among all the material combinations utilized as TE in Si-based solar cell, AgNW/AZO with larger surface area of coverage (34%) exhibited the largest power conversion efficiency (PCE), as shown in Table I (noted in bold and italic), in which, besides PCE values for open circuit voltage (V_{OC}), short circuit current density (J_{SC}), fill factor (FF) and sheet resistance (R_S) are given for all electrodes [4].

TABLE I
SOLAR CELL PERFORMANCE UTILIZING VARIOUS TE [4].

Electrode	V_{OC} (mV)	J_{SC} (mA/cm^2)	FF (%)	R_S (Ω cm^2)	PCE (%)
Ag grid	553	16.3	68.4	1.8	6.1
AgNWs/AZO/ Ag grid	556	18.2	73.6	1.0	7.4
AZO/Ag grid	559	19.8	72.1	1.4	8.0
AZO	547	25.1	30.1	17.9	4.1
AgNW (16%)/AZO	559	27	31.3	23.6	4.7
AgNW (34%)/AZO	***559***	***28***	***60.1***	***4.4***	***9.4***

B. AgNWs in BHJ OSCs

For application as TE in tandem BHJ OSC, as mentioned, AgNWs have been printed from ink, by means of doctor blading. Fig. 4shows top view of the AgNW TE of our cell with an architecture (bottom to top): Glass substrate / PEDOT:PSS / Opaque Ag/ PEDOT:PSS / GEN-2 / PCBM / ZnO / PEDOT:PSS /pDPP5T-2:PCBM / ZnO / AgNW. As can be seen, the network consists of very long AgNWs,

beneficial to the percolation of electrical properties while still have enough of empty space for not hindering the optical transparency of the TE.

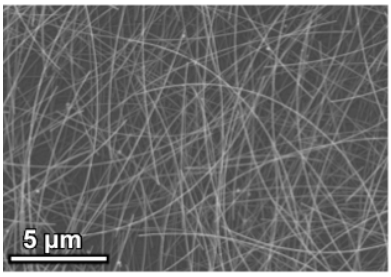


Fig. 4. SEM image of top view of cell i.e. AgNW transparent electrode [2].

Fig. 5 (left) represents the schematic of the solar cell while Fig. 5 (right) is a bright field conventional transmission electron micrograph (CTEM) obtained by lifting out a thin lamella from sample by means of FIB, just like in the previous section regarding the AgNW/AZO nanocomposite. The same architecture, but with polyethylene terephthalate (PET) polymer as a flexible substrate, was not possible to characterize by means of TEM as it was not possible to acquire a lamella by FIB. From the CTEM image it can be seen that fairly uniform thicknesses can be observed for all layers except the AgNWs, suggesting that charge carrier diffusion to their respective electrodes should be uniform throughout the cell. In the case of AgNWs, non-homogenous thickness as well as high surface roughness is expected as AgNWs are essentially a network, not a continuous film.

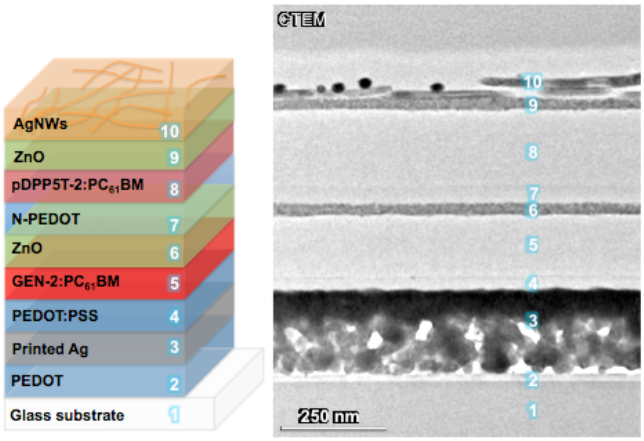


Fig. 5. Left: Schematic of tandem BHJ OSC architecture with layers numbered "1-10" in order of deposition (bottom to top); Right: CTEM image cell cross-section, numbers "1-10" note layers which correspond to image on the left [8].

Concerning AgNWs, a polymer capping agent PVP, introduced during synthesis in order to stabilize AgNWs [8], prevents direct contact of AgNWs and ZnO nanoparticles, which increases resistance and contributes to lower FF and PCE. This capping agent, noted as an amorphous region in Fig. 6, although negatively impacting the PCE (by increasing electrical sheet resistance), is detrimental to the formation and growth of AgNWs during synthesis. It can cleary be seen that the cross sections of the AgNWs are of pentagonal shape, due to them having five tetrahedral five twinned sub-segments inherent from the synthesis process [12], as well as the uniform thickness of the ZnO layer, as previously mentioned.

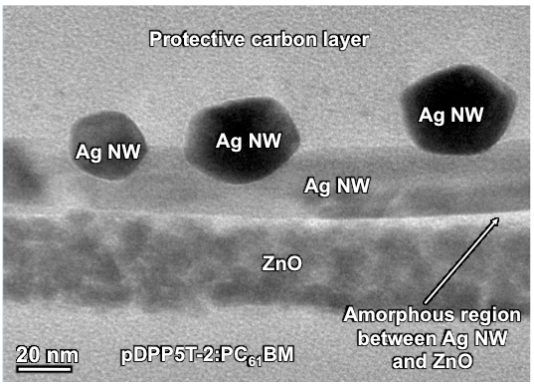


Fig. 6. HAADF STEM image of AgNW/AZO cross-section along with appropriate EDS maps noting elemental distribution of Ag, Al, Zn and O [2].

Exceptional properties of AgNWs such as high optical transparency and low sheet resistance (as well as high reflectance and low sheet resistance of the, also printed opaque Ag reflective electrode) ensured sufficient light absorption in the active layers and efficient charge carriers collection. With the incorporation into a fully printed tandem OSC where interface engineering and optimization of layer compatibility was performed, the resulting cells without the use of ITO and vacuum-deposition steps achieved high PCEs of 5.81% and 4.85% on glass and flexible PET substrate, respectively [8].

TABLE II
SOLAR CELL PERFORMANCE UTILIZING VARIOUS ELECTRODES [8].

SC Architecture	V_{OC} (V)	J_{SC} (mA/cm^2)	FF (%)	PCE (%)
Glass / ITO / GEN-2 / pDPP5T-2 / Ag	1.29	7.61	68.4	6.48
Glass/ P_Ag / GEN-2 / pDPP5T-2 / Ag	1.29	7.38	73.6	5.81
PET / P_Ag/ GEN-2 / pDPP5T-2 / AgNWs	1.28	7.02	72.1	4.85

IV. CONCLUSIONS

Novel materials or novel applications for existing materials have to be constantly invented as the needs of industries like the optoelectronic industry are ever-growing. Flexible, low cost materials as well as cost effective methods of synthesis and processing are something that is highly desired. In our work we have utilized AgNW based transparent electrodes in two different kinds of solar cells. In the case of first generation Si-based solar cells, the AgNW/AZO

nanocomposite has outperformed several materials. In the case of third generation tandem BHJ OSC, the results show that the solar cell architecture with AgNWs as the top TE exhibits a power conversion efficiency comparable to similar architecture using materials whose processing methods are very demanding and expensive. It has also been shown that the flexible version of the BHJ OSC with AgNW as the TE has also a very comparable power conversion efficiency, exhibiting its potential in flexible optoelectronics.

ACKNOWLEDGMENT

The author acknowledges support by the Ministry of Education, Science and Technological Development of the Republic of Serbia, grant no. III45019. The author would also like to thank prof. dr. Velimir R. Radmilović from the University of Belgrade, Serbia for useful discussions as well as prof. dr. Erdmann Spiecker from the Friedrich Alexander University, Erlangen, Germany where electron microscopy has been performed.

REFERENCES

[1] "International Energy Agency 2018 Report – Key World Energy Statistics", IEA, Paris, France, 2018.

[2] V.V. Radmilović, "Transparent Nanocomposite Films for Plastic Electronics Applications" Ph.D. dissertation, FTM, UB, Belgrade, Serbia, 2016.

[3] R. Das, P. Harrop, Printed and organic electronics: forecasts, players and opportunities 2011–2021", IDTechEx, Cambridge, UK, 2011.

[4] M. Goebelt, R. Keding, S. W. Schmitt, B. Hoffmann, S. Jaeckle, M. Latzel, V.V. Radmilovic, V.R. Radmilovic, E. Spiecker, S. Christiansen, "Encapsulation of silver nanowire networks by atomic layer deposition for indium-free transparent electrodes", *Nano Energy*, 16, 196-206, 2015.

[5] R.W. Messler Jr., Principles of welding: processes, physics, chemistry and metallurgy, New York, USA: John Wiley & Sons, 1999.

[6] M. Brochu, Microjoining and nanojoining, Cambridge, UK: Woodhead Publishing Ltd., 2008.

[7] D.P. Langley, M. Lagrange, G. Giusti, C. Jiménez, Y. Bréchet, N.D. Nguyen, D. Bellet, "Metallic nanowire networks: effects of thermal annealing on electrical resistance", *Nanoscale*, 6, 13535-13543, 2014.

[8] F. Guo, N. Li, V.V. Radmilovic, V.R. Radmilovic, M. Turbiez, E. Spiecker, K. Forberich and C. Brabec, "Fully printed organic tandem solar cells using solution-processed silver nanowires and opaque silver as charge collecting electrodes", *Energy Environ. Sci.*, 8, 1690-1697, 2015.

[9] L.B. Hu, H.S. Kim, J.Y. Lee, P. Peumans and Y. Cui, "Scalable coating and properties of transparent, flexible, silver nanowire electrodes", *ACS Nano*, 4, 2955– 2963, 2010.

[10] K.S. Siow, "Mechanical properties of nano-silver joints as die attach materials", *J. Alloys and Compd.*, 514, 6-19, 2012.

[11] V.V. Radmilovic, M. Goebelt, S. Christiansen, E. Spiecker, V.R. Radmilovic, "Low temperature solid-state wetting and formation of nanowelds in silver nanowires", *Nanotechnology*, 8, 385701, 2017.

[12] J. Monk, J.J. Hoyt, and D. Farkas, "Metastability of multitwinned Ag nanorods: Molecular dynamics study", *Phys. Rev. B*, 78, 024112, 2008.

Procedure merenja električnih karakteristika naprezanih p-kanalnih VDMOS tranzistora snage

Snežana Đorić-Veljković, *Member, IEEE*, Vojkan Davidović, *Member, IEEE*, Danijel Danković, *Member, IEEE*, Snežana Golubović, *Member, IEEE* i Ninoslav Stojadinović, *Life Fellow, IEEE*

***Sadržaj* — U ovom radu izvršena je komparativna analiza električnih karakteristika i napona praga p-kanalnih VDMOS tranzistora snage, podvrgnutih NBT (negative bias temperature) naprezanju, zavisno od primene različitih mernih procedura i konfigurisanih mernih uređaja. Pokazano je da iako tokom NBT naprezanja postoji odlično kvalitativno slaganje u promenama napona praga, rezultati dobijeni merenjem u proceduri pri kojoj su korišćeni uređaji u konfiguraciji sa jednom SMU (source measure unit) u izvesnoj meri odstupaju od onih sa dve SMU, a odstupanje može iznositi i do 20% nakon 168 h naprezanja.**

***Ključne reči* — VDMOS tranzistor snage; napon praga; NBT naprezanje; ozračivanje; merne procedure.**

I. Uvod

MOS komponente snage, među kojima posebno mesto zauzimaju VDMOS (*Vertical Double Diffused MOS*) tranzistori snage, zbog svojih performansi kao što su velika prekidačka brzina, visoki probojni napon, dobra termička stabilnost, visoka ulazna impedansa i mogućnost upravljanja visokim strujama, imaju veoma široku primenu. Koriste se u brojnim elektronskim uređajima i sistemima projektovanim za različite primene (u prekidačkim izvorima napajanja, u audio pojačavačima [1, 2]), u automobilskoj industriji u uređajima i sistemima koji predstavljaju i dodatnu i primarnu opremu [3]. U navedenim primenama, VDMOS tranzistori u radnom režimu trpe permanentni uticaj i jakog električnog polja i povišene temperature iz ambijenta ili kao posledica samozagrevanja. Međutim, VDMOS tranzistori snage imaju, takođe, široku primenu i u elektronskim uređajima i sistemima projektovanim za specijalne namene u radijacionom okruženju: u telekomunikacionim satelitima (kao prekidački elementi u izvorima napajanja), u vojnoj industriji, u sistemima automatike kod nuklearnih postrojenja. U ovim primenama VDMOS tranzistori snage su izloženi negativnom kompleksnom dejstvu BT (*Bias Temperature*) naprezanja i jonizujućeg zračenja, pa se od njih, osim visoke pouzdanosti, zahteva i velika otpornost na jonizujuće zračenje. [4-10]. To je osnovni razlog za višegodišnje proučavanje efekata jonizujućeg zračenja i BT naprezanja kod VDMOS tranzistora snage [11, 12]. Glavni rezultat tih proučavanja je činjenica da naprezanje tranzistora ima za posledicu nestabilnosti njihovih električnih parametara i karakteristika, u osnovi kojih su nestabilnosti gustina naelektrisanja u oksidu gejta i površinskih stanja prouzrokovanih naprezanjem. Iz tog razloga proučavanja su produbljena i usmerena na analizu formiranja i prirode naelektrisanja u oksidu gejta i površinskih stanja [13-16], u cilju praćenja i prognoze ponašanja električnih parametara ispitivanih tranzistora. Kako je za to neophodno praćenje električnih karakteristika tranzistora tokom naprezanja, imperativ je koristiti što efikasniju mernu proceduru koja podrazumeva optimalno korišćenje merne opreme. Merna oprema korišćena u ovim eksperimentima, sa hronološkog aspekta, sadržala je najpre dve SMU (*Source Measure Unit*). Međutim, iz praktičnih razloga, prešlo se na korišćenje samo jedne SMU, s obzirom da olakšava mobilnost merne konfiguracije, što je od posebnog značaja kada se merenja određenih sekvenci istog eksperimenta moraju vršiti u različitim laboratorijama, koje su međusobno fizički distancirane.

U ovom radu je izvršena komparativna analiza rezultata merenja električnih karakteristika p-kanalnih VDMOS tranzistora snage, podvrgnutih NBT naprezanju dobijenih primenom različito konfigurisanih mernih uređaja (prva konfiguracija uređaja sadržala je dve SMU dok je druga sadržala samo jednu).

II. Eksperiment

U ovim istraživanjima su kao eksperimentalni uzorci korišćeni komercijalni p-kanalni VDMOS tranzistori snage tipa IRF 9520, koji su proizvedeni u standardnoj Si-gejt tehnologiji, sa nominalnom debljinom oksida gejta od 100 nm. Tokom NBT naprezanja, pri temperaturi od 175°C, kod svih komponenata je bila primenjena polarizacija gejta od - 45 V, dok su sors i drejn bili uzemljeni. NBT naprezanje je obavljeno u Heraeus HEP2 komorama, koje su obezbeđivale stabilnu temperaturu.

U cilju praćenja degradacije komponenata NBT naprezanje je prekidano nakon prethodno definisanih vremenskih perioda

Snežana Đorić-Veljković – Univerzitet u Nišu, Građevinsko-arhitektonski fakultet, Aleksandra Medvedeva 14, 18000 Niš, Srbija (e-mail: snezana.djoric.veljkovic@elfak.ni.ac.rs).

Vojkan Davidović – Univerzitet u Nišu, Elektronski fakultet, Aleksandra Medvedeva 14, 18000 Niš, Srbija (e-mail: vojkan.davidovic@elfak.ni.ac.rs).

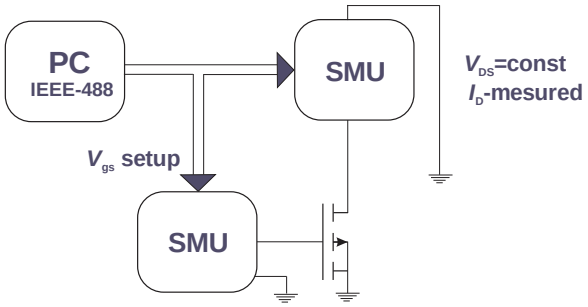
Danijel Danković – Univerzitet u Nišu, Elektronski fakultet, Aleksandra Medvedeva 14, 18000 Niš, Srbija (e-mail: danijel.dankovic@elfak.ni.ac.rs).

Snežana Golubović – Univerzitet u Nišu, Elektronski fakultet, Aleksandra Medvedeva 14, 18000 Niš, Srbija (e-mail: snezana.golubovic@elfak.ni.ac.rs).

Ninoslav Stojadinović – Univerzitet u Nišu, Elektronski fakultet, Aleksandra Medvedeva 14, 18000 Niš, Srbija; Srpska akademija nauka i umetnosti (SANU) - ogranak u Nišu, Univerzitetski trg 2, 18000 Niš, Srbija (e-mail: ninoslav.stojadinovic@elfak.ni.ac.rs).

da bi se snimale prenosne karakteristike I_D-V_{GS}, na osnovu kojih je određen napon praga. Vrednosti napona praga određivane su na osnovu nadpragovskog dela prenosnih karakteristika u oblasti zasićenja, ekstrapolacijom linearnog dela $\sqrt{I_D}$–V_G krivih do preseka sa V_G-osom. Sva merenja su obavljana na sobnoj temperaturi.

Na Sl. 1 je dat šematski prikaz konfiguracije opreme za snimanje prenosnih strujno-naponskih karakteristika p-kanalnih VDMOS tranzistora snage korišćenjem dve SMU. Mernim sistemom, koji se sastojao od uređaja Keithley 237 (za polarisanje drejna i merenje struje drejna) i Keithley 2400 (za variranje-setovanje napona na gejtu u zadatim koracima), upravljano je preko personalnog računara u koji je ugradjena interfejs kartica tipa IEEE 488.



Sl. 1. Šematski prikaz konfiguracije opreme za snimanje prenosnih strujno-naponskih karakteristika p-kanalnih VDMOS tranzistora snage korišćenjem dve SMU.

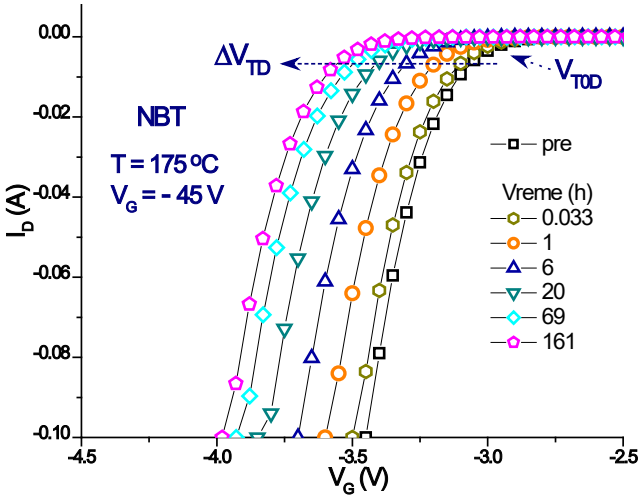
Na Sl. 2. je dat šematski prikaz i fotografija konfiguracije za snimanje prenosnih strujno-naponskih karakteristika p-kanalnih VDMOS tranzistora snage korišćenjem jedne SMU. Pri tome je korišćen visoko precizni Keysight Technologies B2901A, kontrolisan pomoću laptopa preko USB-a.

Sl. 2. Šematski prikaz i fotografija konfiguracije za snimanje prenosnih strujno-naponskih karakteristika p-kanalnih VDMOS tranzistora snage korišćenjem jedne SMU.

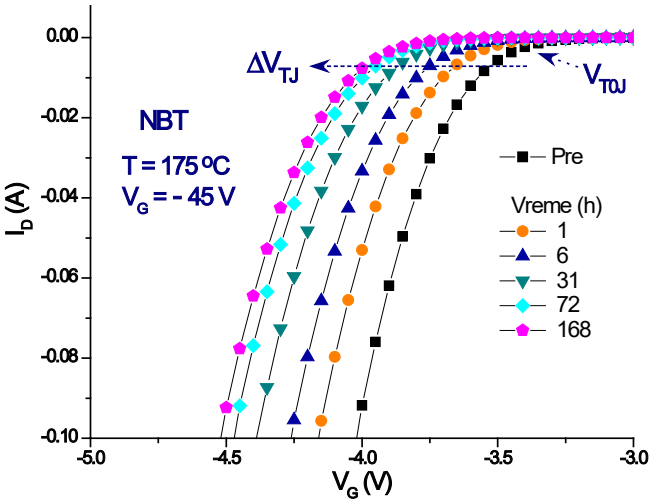
Konfiguracija prikazana na Sl. 2 (sa samo jednom SMU) je svakako jednostavnija za primenu i olakšava njenu mobilnost, što je od posebnog značaja kada se merenja moraju vršiti u različitim, međusobno udaljenim laboratorijama. To je bilo posebno značajno kada su sprovedeni uporedni eksperimenti NBT naprezanje-ozračivanje i ozračivanje-NBT naprezanje [11, 13, 14]. U ovim eksperimentima su uzorci podvrgnuti NBT naprezanju u Laboratoriji Katedre za mikroelektroniku na Elektronskom fakultetu u Nišu, a ozračivani γ-zračenjem na Co-60 izvoru u Laboratoriji za zaštitu od zračenja i zaštitu životne sredine, Instituta za nuklearne nauke “Vinča”.

III. Rezultati i diskusija

Na osnovi izvršenih merenja primenom dve i jedne SMU, promene prenosnih karakteristika p-kanalnog VDMOS tranzistora snage, tokom NBT naprezanja, prikazane su na Sl. 3 i Sl. 4, respektivno.



Sl. 3. Promene prenosnih karakteristika p-kanalnog VDMOS tranzistora snage, tokom NBT naprezanja, pri merenju konfiguracijom sa dve SMU.

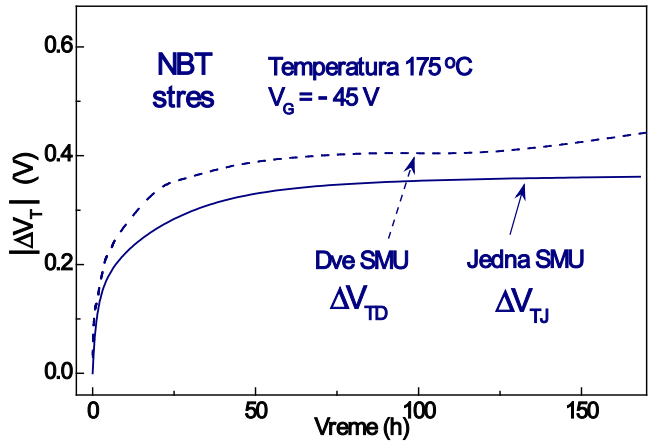


Sl. 4. Promene prenosnih karakteristika p-kanalnog VDMOS tranzistora snage, tokom NBT naprezanja, pri merenju konfiguracijom sa jednom SMU.

Na osnovu Sl. 3 i Sl. 4 se jasno uočava da tokom NBT naprezanja dolazi do pomeranja prenosnih karakteristika duž V_G ose ka negativnijim vrednostima, što ukazuje na porast napona praga ispitivanih tranzistora po apsolutnoj vrednosti. Pokazano je da do ovakvog ponašanja napona praga tokom NBT naprezanja dolazi usled porasta gustina pozitivnog naelektrisanja u oksidu gejta i na površinskim stanjima, koje su posledica odgovarajućih procesa u oksidu i na međupovršini [3, 7, 8, 13, 14, 16].

Na Sl. 5 su prikazane promene napona praga p-kanalnih VDMOS tranzistora snage određene na osnovu izlaznih karakteristika koje su merene sa dve (ΔV_{TD}) i sa jednom (ΔV_{TJ}) SMU tokom NBT naprezanja. Uočava se da su promene napona praga, dobijene merenjem u proceduri pri

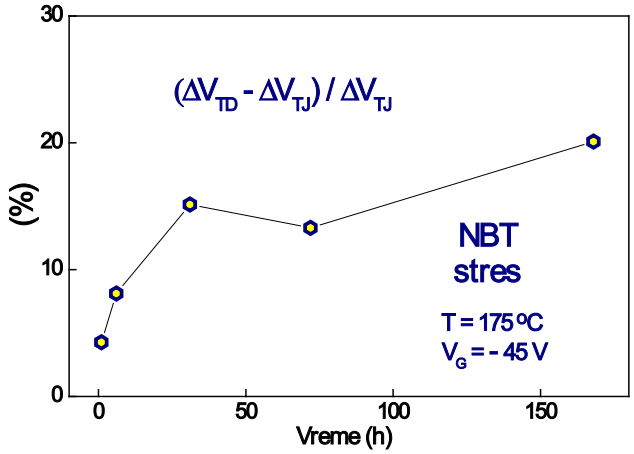
kojoj su korišćeni uređaji u konfiguraciji sa jednom SMU, manje od onih dobijenih primenom dve SMU.



Sl. 5. Promene napona praga p-kanalnih VDMOS tranzistora snage određene na osnovu izlaznih karakteristika merenih sa dve i sa jednom SMU tokom NBT naprezanja.

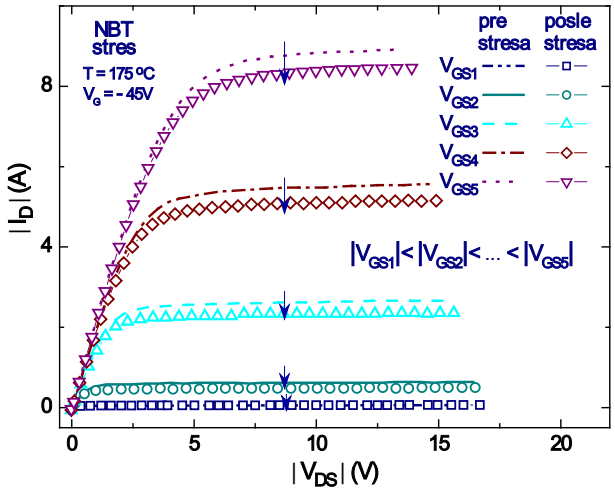
Sa Sl. 5 se primećuje da pri kraćim vremenima NBT naprezanja (do dva sata) gotovo da nema razlike u vrednostima promena napona praga, bez obzira da li se merenja vrše sa jednom ili sa dve SMU, ali da sa vremenom te razlike postaju sve izraženije do 30-ak sati, nakon čega se znatno sporije povećavaju. Na Sl. 6 su prikazane relativne promene napona praga p-kanalnih VDMOS tranzistora snage tokom NBT naprezanja, određene na osnovu izlaznih karakteristika koje su merene sa dve i sa jednom SMU. Može se videti da su te promene samo 4% nakon jednog sata NBT naprezanja, dok nakon 30-ak sati dostižu vrednost od 15% i povećavaju se do 20% nakon 168 h naprezanja.

U cilju dalje detaljnije analize uočenih razlika pri merenju karakteristika sa dve i sa jednom SMU, na Sl. 7 su prikazane izlazne karakteristike p-kanalnog VDMOS tranzistora snage pre i posle NBT stresa (naprezanja).

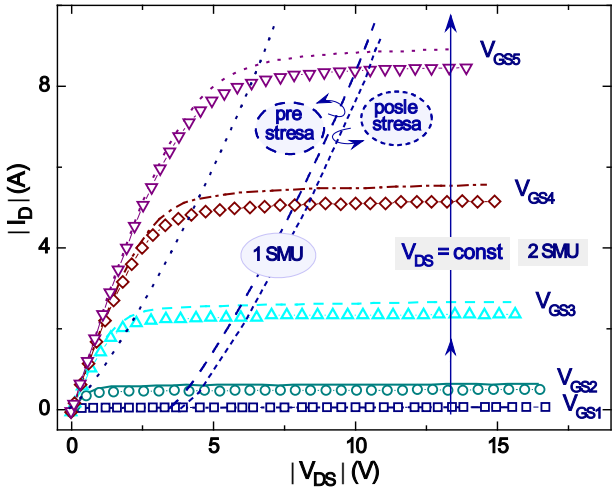


Sl. 6. Relativno odstupanje dobijenih promena napona praga p-kanalnih VDMOS tranzistora snage, pri merenju prenosnih I-V karakteristika sa dve i sa jednom SMU tokom NBT naprezanja.

Na Sl. 7a se može uočiti tendencija promena izlaznih karakteristika nakon NBT stresa, odnosno njihovo potiskivanje ka manjim (apsolutnim) vrednostima struja drejna, usled porasta napona praga, a sa Sl. 7b razlike pri merenju prenosnih karakteristika sa dve i sa jednom SMU.



(a)



(b)

Sl. 7. Izlazne karakteristike p-kanalnog VDMOS tranzistora snage pre i posle NBT stresa na kojima se zapaža: (a) tendencija pomeranja karakteristika i (b) razlike pri merenju prenosnih karakteristika sa dve i sa jednom SMU.

Pri merenju sa dve SMU napon na drejnu V_{DS} je konstantan (tipično 5 V ili više), napon koji se dovodi na gejt (V_{GS}) menja se u zadatim koracima, a meri se struja drejna (I_D). Pri tome, tranzistor je duboko u zasićenju, a merene tačke leže na vertikalnoj pravoj. Kao što je pomenuto, jednom SMU je variran napon na gejtu, dok je drugom vršeno polarisanje drejna i merenje struje drejna. U slučaju merenja sa jednom SMU drejn i gejt su kratkospojeni, tranzistor je konstantno u zasićenju, ali napon na drejnu klizi tokom merenja. Kako je parabola koja razdvaja triodnu od oblasti zasićenja definisana naponom $V_{DSsat}=V_{GS}-V_T$, to se merenjem sa kratkospojenim gejtom i drejnom ($V_{GS}=V_{DS}$) krećemo po paraboli koja je za vrednost V_T potisnuta u oblast zasićenja. Pored toga, sa povećanjem napona na drejnu smanjuje se efektivna dužina kanala (modulacija dužine kanala), te karakteristike u oblasti zasićenja nisu idealno paralelne naponskoj osi (V_{DS}), što je posebno izraženo kod p-kanalnih tranzistora. Sa Sl. 7 se može

primetiti da su razlike između karakteristika pre i posle stresa veće pri većim vrednostima V_{DS}, kada je tranzistor dublje u zasićenju. Kako se merenja karakteristika primenom dve SMU vrše kada je tranzistor duboko u zasićenju jasno je da će i dobijene vrednosti za promenu napona praga biti veće. Pri istoj vrednosti V_{GS} procedura sa dve SMU daće veću struju I_D, što prividno odgovara manjoj vrednosti napona praga, ali i njegovoj većoj promeni. Značajno odstupanje rezultata sa Sl. 3 i Sl. 4 posledica je ovog efekta, ali i same činjenice da su početne vrednosti napona praga bile različite, pri čemu su u slučaju dve SMU (Sl. 3) bile $V_{T0D} = 3$ V, a u slučaju jedne SMU (Sl. 4) $V_{T0J} = 3.4$ V.

IV. Zaključak

Nestabilnosti električnih parametara MOS tranzistora pri NBT naprezanju direktna su posledica nestabilnosti gustina naelektrisanja u oksidu gejta i površinskih stanja. U cilju rasvetljavanja odgovornih elektrofizičkih mehanizama, istraživanja su usmerena na analizu električnih parametara, pri čemu je za praćenje električnih karakteristika tranzistora tokom naprezanja neophodno koristiti što efikasniju mernu proceduru koja podrazumeva i optimalno korišćenje merne opreme. U ovom radu je izvršena komparativna analiza rezultata merenja električnih karakteristika p-kanalnih VDMOS tranzistora snage, podvrgnutih NBT naprezanju, dobijenih primenom različito konfigurisanih mernih uređaja - prva konfiguracija uređaja sadržala je dve SMU, dok je druga sadržala samo jednu SMU.

Pokazano je da iako postoji odlično kvalitativno slaganje u promenama napona praga tokom NBT naprezanja, rezultati dobijeni merenjem u proceduri pri kojoj su korišćeni uređaji u konfiguraciji sa jednom SMU u izvesnoj meri odstupaju od onih sa dve SMU, a odstupanje može iznositi i do 20% nakon 168 h naprezanja. Dobro kvalitativno slaganje i neznatno odstupanje ukazuje da se rezultati dobijeni u prethodnim, kao i u izmenjenim (unapređenim) mernim procedurama mogu u određenoj meri koristiti i u drugim eksperimentima različitih naprezanja p-kanalnih VDMOS tranzistora snage, a u cilju dobijanja pouzdanih podataka neophodnih za prognozu ponašanja njihovih električnih parametara.

Zahvalnica

Prikazani rezultati dobijeni su u okviru istraživanja na projektima OI171026 i TR32026 koje finansira Ministarstvo prosvete, nauke i tehnološkog razvoja Republike Srbije i na projektu "Osobine tankih i ultratankih oksidnih slojeva" (F-148), koji finansira Srpska akademija nauka i umetnosti-SANU.

Literatura

[1] V. Benda, J. Gowar, D. Grant, *Power Semiconductor Devices*, John Wiley & Sons, New York, 1999.

[2] B. Jayant Baliga, *Fundamentals of Semiconductor Power Devices*, Springer, New York, 2008.

[3] S. Gamerith, M. Polzl, "Negative bias temperature stress in low voltage p-channel DMOS transistors and role of nitrogen", *Microelectron. Reliab.*, vol. 42, pp. 1439-1443, 2002.

[4] N. Stojadinović, S. Golubović, V. Davidović, S. Djorić-Veljković and S. Dimitrijev, "Modeling Radiation-Induced Mobility Degradation in MOSFETs", *Phys. Stat. Sol. (a)*, vol. 169, no. 1, pp. 63-66, 1998.

[5] D. Danković, I. Manić, A. Prijić, S. Djorić-Veljković, V. Davidović, N. Stojadinović, Z. Prijić, S. Golubović, "Negative bias temperature instability in p-channel power VDMOSFETs: recoverable versus permanent degradation", *Semicond. Sci. Technol.*, vol. 30, pp. 105009 (9), 2015.

[6] A. Prijić, D. Danković, Lj. Vračar, I. Manić, Z. Prijić, N. Stojadinović, "A Method for Negative Bias Temperature Instability (NBTI) Measurements on Power VDMOS Transistors", *Meas. Sci. Technol.*, vol. 23, no. 8, art. no. 085003, 2012.

[7] D.K. Schroder and J.A. Babcock, "Negative bias temperature instability: Road to cross in deep submicron silicon semiconductor manufacturing", *J. Appl. Phys.*, vol. 94, pp. 1-18, 2003.

[8] S. Ogawa, M. Shimaza and N. Shimano, "Interface trap generation at ultrathin SiO_2 (4-6 nm) - Si interfaces during negative-bias temperature aging", *J. Appl. Phys.*, vol. 77, pp. 1137-1148, 1995.

[9] N. Stojadinović, D. Danković, S. Djorić-Veljković, V. Davidović, I. Manić, S. Golubović, "Negative bias temperature instability mechanisms in p-channel power VDMOSFETs", *Microelectron. Reliab.*, vol. 45, pp. 1343-1348, 2005.

[10] D. Danković, I. Manić, V. Davidović, S. Djorić-Veljković, S. Golubović, N. Stojadinović, "Negative bias temperature instabilities in sequentially stressed and annealed p-channel power VDMOSFETs", *Microelectron. Reliab.*, vol. 47, pp. 1400-1405, 2007.

[11] V. Davidović, D. Danković, A. Ilić, I. Manić, S. Golubović, S. Djorić-Veljković, Z. Prijić, N. Stojadinović, "NBTI and Irradiation Effects in P-Channel Power VDMOS Transistors", *IEEE Trans. Nucl. Sci.*, vol. 63, pp. 1268-1275, 2016.

[12] P. Picard, C. Brisset, A. Hoffmann, J.P. Charles, F. Joffre, L. Adams, and A. Holmes-Siedle, "Use of electrical stress and isochronal annealing on power MOSFETs in order to characterize the effects of ^{60}Co irradiation", *Microelectron. Reliab.*, vol. 40, pp. 1647-1652, 2000.

[13] N. Stojadinović, S. Djorić-Veljković, V. Davidović, S. Golubović, S. Stankovic, A. Prijić, Z. Prijić, I. Manić, D. Danković, "NBTI and irradiation related degradation mechanisms in power VDMOS transistors", *Microelectron. Reliab.*, vol. 88-90, pp. 135-141, 2018.

[14] V. Davidović, D. Danković, A. Ilić, I. Manić, S. Golubović, S. Djorić-Veljković, Z. Prijić, A. Prijić, and N. Stojadinović, "Effects of consecutive irradiation and bias temperature stress in p-channel power vertical double-diffused metal oxide semiconductor transistors", *Jpn. J. App. Phys.*, vol. 57, pp. 044101-1-044101-10, 2018.

[15] D. Danković, I. Manić, V. Davidović, A. Prijić, M. Marjanović, A. Ilić, Z. Prijić, N. Stojadinović, "On the Recoverable and Permanent Components of NBTI in P-Channel Power VDMOSFETs", *IEEE Trans. Dev. Mater. Reliab.*, vol. 16, no. 4, pp. 522-531, art. no. 7536114, 2016.

[16] V. Davidović, D. Danković, S. Golubović, S. Djorić-Veljković, I. Manić, Z. Prijić, A. Prijić, N. Stojadinović, "NBT Stress and Radiation Related Degradation and Underlying Mechanisms in Power VDMOSFETs", *Facta Univers., Ser.: Electron. Energ.*, vol. 31, no. 3, pp. 367-388, 2018.

Abstract

In this paper comparative analyzis of electrical characteristics and threshold voltage shift of NBT (negative bias temperature) stressed p-channel power VDMOS transistors, depending on applied measuring procedures, was performed. It was shown that although there is good qualitative behaviour of threshold voltage shift obtained by measuring configuration by one and by two SMU, relative deviation of threshold voltage shift might be almost 20% after 168 h of NBT stress.

Electrical characteristics measurement procedures of stressed p-channel power VDMOS transistors

Snežana Đorić-Veljković, Vojkan Davidović, Danijel Danković, Snežana Golubović and Ninoslav Stojadinović

Wideband Antenna Array for mm-Wave Radar Modules Characterization

Siniša Jovanović, *Member, IEEE*, Ivan Milosavljević, *Member, IEEE,* Veselin Branković

***Abstract*—This paper outlines the design and realization of one type of antenna array developed for testing of integrated FMCW transmitter modules operating in the unlicensed 60-GHz band.**

All important features such as the type and design of the radiating elements of the array, the configuration of the biasing network as well as the substrate stack-up are adjusted for achieving a very wide operational frequency range, with almost constant antenna gain, from 50 to 70 GHz. To accommodate various test scenarios, four antenna versions are fabricated with low or high gain and with a coaxial connector or waveguide input port.

The fabricated antennas are employed in various test setups for characterization of two versions of millimeter-wave FMCW chips realized in 0.13-µm SiGe BiCMOS technology: a transmitter module operating from 59.5 GHz to 70.5 GHz and a complete radar transceiver module operating from 54 GHz to 65 GHz.

***Index Terms*—Antenna array, millimeter-wave frequency range, FMCW radar module.**

I. Introduction

An unlicensed industrial, scientific and medical (ISM) band ranging from 57 to 64 GHz has recently been gaining increased attention due to its potential to meet the demand of high data-rate wireless communications as well as short-range radar and secure communication within a limited area [1]. All potential applications in this frequency range benefit from a high path loss caused by atmospheric oxygen absorption which allows massive usage of such devices in relative proximity without unwanted interference.

Significant potential for implementation in various fields of everyday human occupations belongs to miniature short-range contactless microwave radar systems due to their capability to operate in bad weather conditions and in harsh environments. The potential fields of application for these radar systems are healthcare, the automotive industry, surveillance, infrastructure maintenance, agriculture and many more areas yet to be acknowledged [2]–[4].

For wider and cost-effective commercial applications, mm-wave radar sensors should be single-chip integrated devices with a low unit cost and small physical size. Over the past few years, Novelic Microsystems developed and realized two modules for application in this frequency range: a SiGe highly integrated FMCW transmitter module that can generate 11 GHz frequency ramps (from 59.5 to 70.5 GHz) with a single continuous sweep [5]; and a fully integrated FMCW radar sensor containing both a transmitter and receiver chain and operating from 55 to 64 GHz [6]. For providing the required working environment as well as for the testing and characterization of these devices, several specific RF and microwave components were developed and integrated within various RF boards [7]. All RF components were required to cover a wider frequency range than the expected operational limits of the FMCW units in order to compensate for the fabrication tolerances as well as the uncertainty induced by the innovative antenna topology.

Various types of antenna arrays were designed and realized to support the development of these FMCW modules. Some antennas were integrated within RF boards and connected to Tx and Rx ports of the FMCW chips. Other antennas were required within the test setups during the testing and characterization of the FMCW modules with an operational frequency range even wider than the expected range of the modules. This paper describes the design and realization of these supporting antennas that were indispensable for the complete characterization of the newly developed FMCW radar modules.

II. Antenna Design

To provide the applicability in all possible testing scenarios and fabrication tolerances during the fabrication of integrated radar modules, the test antennas should cover a very broad frequency range - as wide as from 50 GHz up to 70 GHz. Since a bandwidth of classic microstrip patch antennas is up to 5%, typically, they were unsuitable for usage as the array's radiating elements. However, the required frequency bandwidth can be achieved by an antenna array consisting of customized bowtie-shaped dipoles illustrated in Fig. 1. The basic radiation element of all arrays presented in this paper is a two-sided printed dipole placed on the top and the bottom layer of a thin dielectric substrate with a reflecting plane placed at $\approx\lambda_0/4$ behind the antenna plane, with air between the reflector and the antenna. The dipole can be defined with a small set of parametrized dimensions and modeled in various program packages for EM analysis and simulation [8,9]. By fine-tuning the dipole's parameters, its input impedance can be adjusted to match approximately 100 Ω over a broad frequency range similar to pentagonal-shaped dipoles described in [10].

Siniša Jovanović is with IMTEL Komunikacije, Blvd M. Pupina 165B, 11070 Novi Beograd, Belgrade, Serbia (e-mail: siki@insimtel.com).

Ivan Milosavljević is with the Department of Electronics, School of Electrical Engineering, University of Belgrade, Blvd Kralja Aleksandra 73, 11020 Belgrade, Serbia; NovelIC Microsystems, V. Dugoševića 54, 11000 Belgrade, Serbia (e-mail: ivan.milosavljevic@novelic.com).

Veselin Branković is with NovelIC Microsystems, V. Dugoševića 54, 11000 Belgrade, Serbia (e-mail: veselin.brankovic@novelic.com).

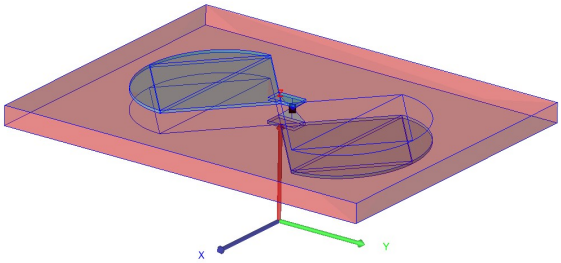


Fig. 1. 3D model in WIPL-D [8] of two-sided printed dipole which is a basic radiating element of antenna array

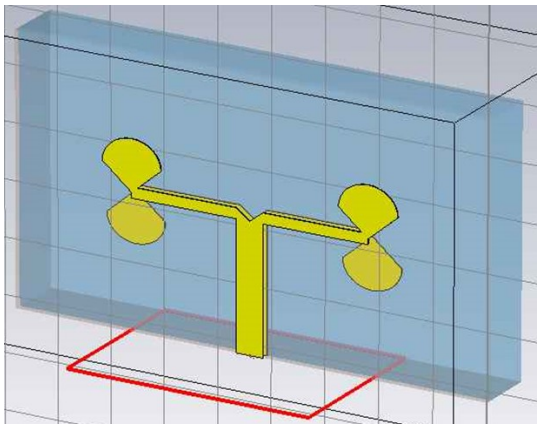


Fig. 2. 3D model of 2×1 antenna array.

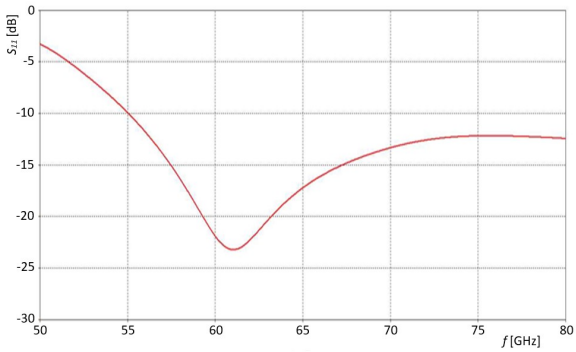


Fig. 3. Return Loss frequency characteristics of 2×1 antenna array obtained from 3D EM simulation.

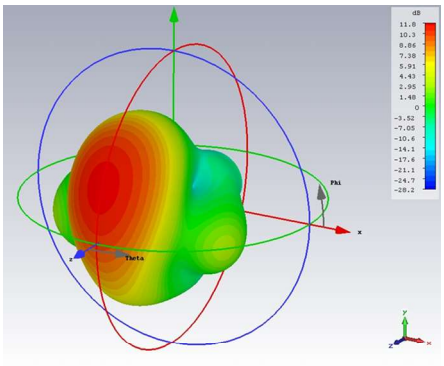


Fig. 4. 3D Radiation diagram of 2×1 antenna array obtained from 3D EM simulation.

Each dipole is fed by a symmetrical microstrip transmission line with a characteristic impedance of 100 Ω. Pairs of adjacent dipoles with associated feeding lines are connected over a simple T-junction forming a dipole pair with 50 Ω input impedance as shown in Fig. 2. In that manner, good matching can be achieved in a very wide frequency range as presented in Fig. 3 that shows an *RL* better than 10 dB in a frequency range from 55 GHz to above 80 GHz.

Fig. 4 shows the 3D radiation diagram of the 2×1 antenna array. It has a very wide 3 dB radiation beam of 60° in E plane (yz) and 35° in H plane (xz) with a maximum gain of 11.8 dBi which is steered by about 20° in E plane due to the influence and coupling with the main feeding line. For higher order arrays, having a bigger number of radiation elements, this influence is less pronounced because the feeding network and antenna array structure is more symmetrical.

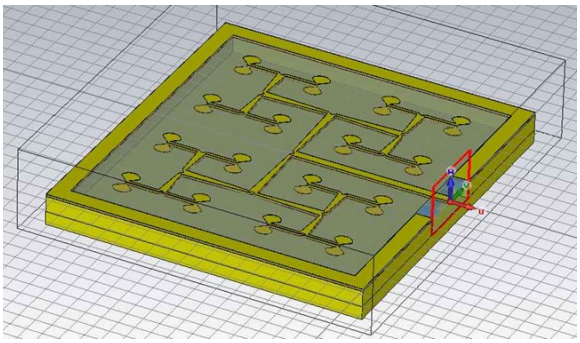


Fig. 5. 3D model of 4×4 antenna array

Fig. 5 shows a 3D model of the 4×4 antenna array. The distances between the radiating elements are $d_v = 4.4$ mm along v axis, and $d_u = 4.8$ mm along u axis. The model contains a metalized frame around the array for mechanical support of the array and preventing its bending. The frame is included into the 3D model to ensure that it will not have a negative influence on the overall characteristics.

Besides the main feeding line of $Z_c = 50\ \Omega$ and dipole feeding line of $Z_c = 100\ \Omega$, the feeding network contains a system of tapered 50 Ω to 100 Ω transformers. In that manner, good matching is achieved in a wide frequency range as shown in Fig. 6 with *RL* better than 10 dB from 54 GHz to above 80 GHz.

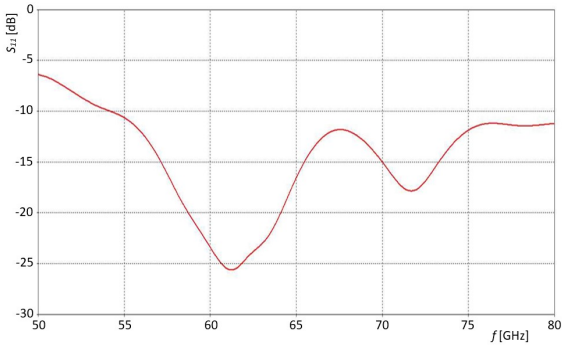


Fig. 6. Return Loss frequency characteristics of 4×4 antenna array obtained from 3D EM simulation.

Fig. 7 presents a 3D radiation diagram of the 4×4 antenna array showing a maximum gain of 20.8 dBi.

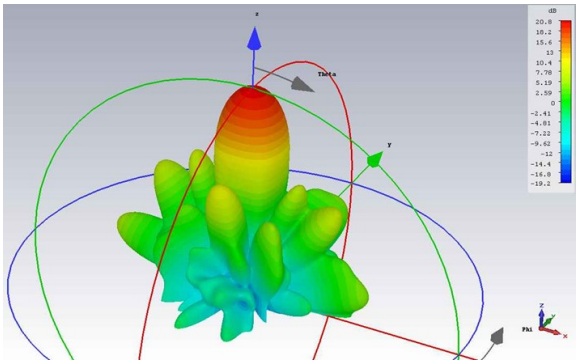


Fig. 7. 3D Radiation diagram of 4×4 antenna array obtained from 3D EM simulation.

Fig.8 presents the maximum gain of the 4×4 array from 50 to 80 GHz showing that the antenna array has a very wide usable frequency range with a gain better than 20 dBi from 57 to 72 GHz and a gain better than 18 dBi within the whole observed range.

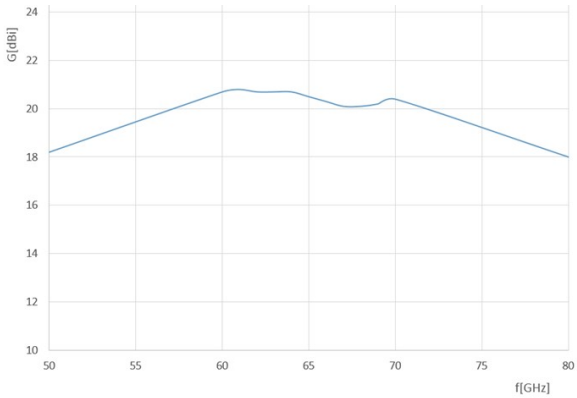


Fig. 8. Maximum Gain of 4×4 array vs. frequency obtained from 3D EM simulation.

All presented 3D models contain a multi pin waveguide port that provides the ideal balanced antiphase excitation for the analyzed arrays. For practical realization, it is necessary to provide a transition from such a balanced port to more practical ports such as either coaxial or waveguide. Fig. 9 shows a 3D model of CPWG to symmetrical microstrip (SMS) transition which consists of five sections. At the right end of the structure shown in Fig. 9, it is a CPWG transmission line with a central conductor surrounded with reference ground both at the bottom layer of the dielectric and at the top layer. The length of this section must be sufficient to match mechanical dimensions for the selected connector type. The next section is a transition from CPWG transmission line to standard unbalanced microstrip transmission line, followed by a very short unbalanced microstrip line. The next section is unbalanced to balanced microstrip line transition and finally a section of balanced microstrip line which belongs to an antenna array feeding network.

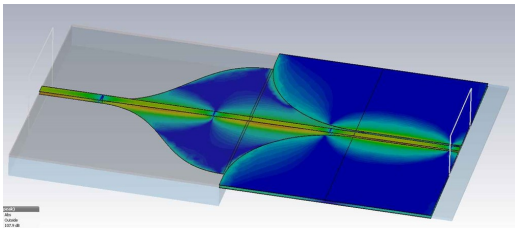


Fig. 9. E-field distribution at a transition from CPWG to symmetrical microstrip transmission line .

Fig. 10 shows the partial cross section of a 3D model of a symmetrical microstrip to WR-15 waveguide transition. The transition can be optimized by changing the length and the curvature of the fences. Both transitions were optimized to provide minimum insertion loss, less than 0.2 dB within the whole frequency range of interest.

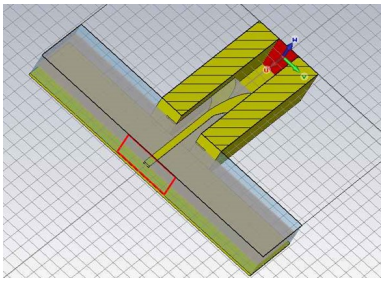


Fig. 10. 3D model of a transition from WR-15 type waveguide port to symmetrical microstrip transmission line (partial cross-section).

III. Realization

Based on the analysis performed in the previous section, four versions of antenna arrays are designed for fabrication of experimental models: 4×4 and 2×1 antenna arrays combined with both SMS to WG and SMS to coaxial connector transitions.

All printed antennas are fabricated on a teflon-fiberglass substrate having $\varepsilon_r = 2.17$, $\tan\delta = 0.0009$ and $h = 0.127$ mm. A standard photolithography process is used with tolerances better than ±10 μm. Fig. 11 shows a photo of all four versions of the antenna arrays with brown copper metallization of the top layer and black shadow of the bottom metallization.

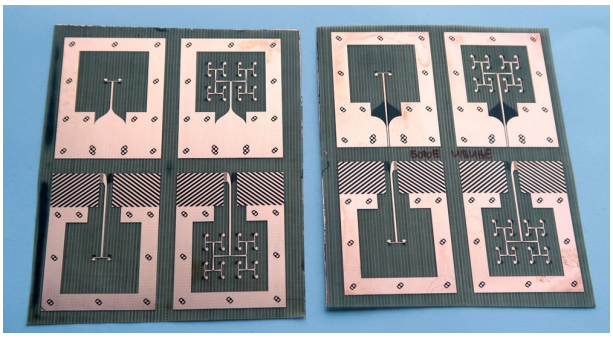


Fig. 11. Photo of two manufactured sets of printed antenna arrays: bottom side (left) and top side (right).

Fig. 12 shows a partially and a fully assembled 4×4 antenna array with WR-15 flange, while Fig. 13 shows assembled 4×4 and 2×1 antenna arrays with 2.4 mm SMA connectors.

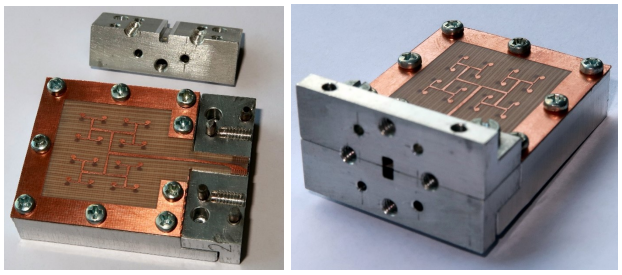


Fig. 12. Partially assembled (left) and assembled (right) 4×4 antenna array with WR 15 waveguide

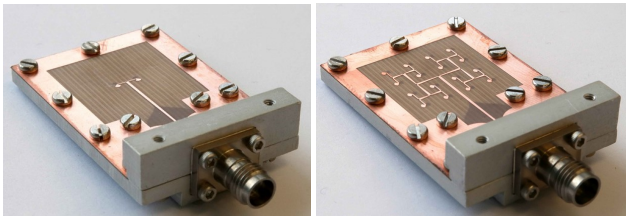


Fig. 12. Assembled antenna arrays with 2.4 mm connector

IV. Measurements

Fig. 13 shows Vector Network Analyzer ANRITSU MS4647A that is used for experimental verification of all realized antennas. The VNA, having a frequency range up to 70 GHz is used for Return Loss measurement as well as free space loss measurement. The free space loss data is post-processed for obtaining the maximum gain of the measured antennas.

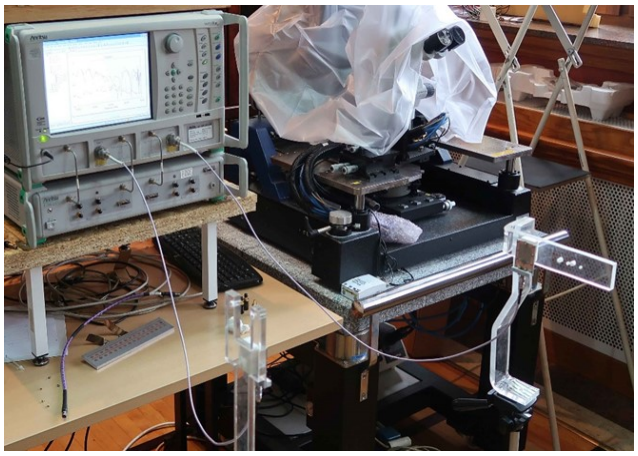


Fig. 13. Setup for free space loss measurement of two identical antennas

Fig. 14 shows measured results for the 2×1 antenna array with a 2.4 mm connector. The S_{11} and S_{22} are the return losses at the transmitting and the receiving antenna, respectively. A matching better than -10 dB is achieved in the wide frequency range starting from 53 GHz. The *RL* results include the influence of the 2.4 mm connector. The G trace shows the antenna gain reduced for the influence of the insertion loss of the connector, while Gc shows the corrected antenna gain.

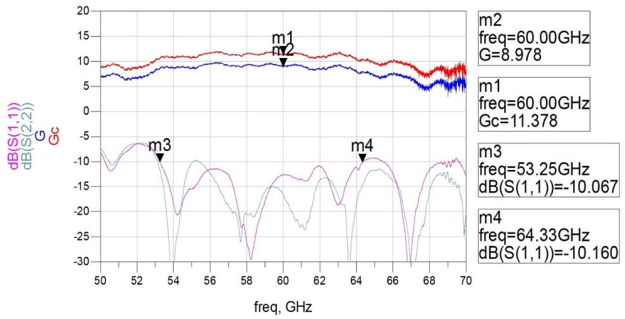


Fig. 14. Measured results for the gain of 2×1 antenna array: G (blue line) – gain with connector losses; Gc (red line) – gain excluding connector losses

Fig. 15 shows measured results for the 4×4 antenna array with the 2.4 mm connector. The S_{11} and S_{22} are the return losses at the transmitting and the receiving antenna, respectively. A matching better than -10 dB is achieved in a wide frequency range starting from 54.3 GHz. The *RL* results include the influence of the 2.4 mm connector. The G trace shows the maximum antenna gain of 17.9 dBi at 60 GHz, which is reduced due to the influence of the insertion loss of the connector. Gc shows the corrected antenna gain with a maximum value of 20.1 dBi at 60 GHz. A high antenna gain above 18 dBi is achieved in a very wide frequency range from 54 GHz to 66 GHz. The measurements of the antenna arrays with waveguide connectors show the same results for the maximum antenna gains.

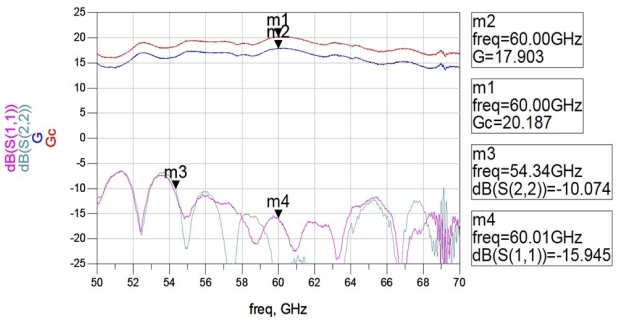


Fig. 15. Measured results for the gain of 4×4 antenna array: G (blue line) – gain with connector losses; Gc (red line) – gain excluding connector losses

V. Conclusion

The paper presents design realization and measuring results for a set of wideband antenna arrays operating at a mm-wave frequency range. The obtained characteristics of all versions of the realized antennas are in good accordance with the results obtained by 3D EM analysis. The antennas were employed within various test setups for the characterization of FMCW integrated radar modules described in papers [5,6].

ACKNOWLEDGMENT

The results shown in this paper are the outcome of a development project funded by Novelic Microsystems.

This paper was partially supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia under grant TR-32024.

REFERENCES

[1] T. Yilmaz, E. Fadel and O. B. Akan, "Employing 60 GHz ISM band for 5G wireless communications," *2014 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom)*, Odessa, 2014, pp. 77-82.

[2] C. Li, Z. Peng, T. Y. Huang, T. Fan, F. K. Wang, T. S. Horng, J. M. Muñoz-Ferreras, R. Gómez-García, L. Ran, and J. Lin, "A Review on Recent Progress of Portable Short-Range Noncontact Microwave Radar Systems," *IEEE Trans. Microw. Theory Techn.*, vol. 65, no. 5, pp. 1692– 1706, May 2017.

[3] M. Pauli, B. Göttel, S. Scherr, A. Bhutani, S. Ayhan, W. Winkler, and T. Zwick, "Miniaturized Millimeter-Wave Radar Sensor for High Accuracy Applications," *IEEE Trans. Microw. Theory Techn.*, vol. 65, no. 5, pp. 1707–1715, May 2017.

[4] J. Hasch, E. Topak, R. Schnabel, T. Zwick, R. Weigel, and C. Waldschmidt, "Millimeter-Wave Technology for Automotive Radar Sensors in the 77 GHz Frequency Band," *IEEE Trans. Microw. Theory Techn.*, vol. 60, no. 3, pp. 845–860, Mar. 2012.

[5] I. M. Milosavljević, D. P. Krčum, Đ. P. Glavonjić, S. P. Jovanović, V. R. Mihajlović, D. M. Tasovac, and V. M. Milovanović, "A SiGe Highly Integrated FMCW Transmitter Module With a 59.5-70.5 GHz Single Sweep Cover", *IEEE Trans. Microw. Theory Techn*. vol. 66, no. 9, pp. 4121–4133, Sep. 2018.

[6] I. M. Milosavljević, Đ. P. Glavonjić, D. P. Krčum, S. P. Jovanović, V. R. Mihajlović, and V. M. Milovanović, "A 55 to 64 GHz Fully Integrated Miniaturized FMCW Radar Sensor Module for Short-Range Applications", *IEEE Microwave and Wireless Components Letters*. (submitted for publication)

[7] S. Jovanovic, I. Milosavljevic, V. Brankovic, "Using Rat Race Balun Transition for Characterization of 60 GHz FMCW Transmitter Module" 5th International Conference on Electrical, Electronic and Computing Engineering, IcETRAN 2018, Palic, Serbia, June 2018, Proceedings of Papers, pp MTI1.2, 1-4.

[8] WIPL-D Pro v15, Belgrade:WIPL-D d.o.o, 2018.

[9] CST Studio. Computer Simulation Technology (CST) GmbH

[10] A. Nesic, I. Radnovic and V. Brankovic, "Ultrawideband printed antenna array for 60 GHz frequency range," IEEE Antennas and Propagation Society International Symposium 1997. Digest, Montreal, Quebec, Canada, 1997, pp. 1272-1275 vol.2.

Comparison of the Measured Characteristics of Schottky Diodes for Power Harvesting Applications

Branka Milosevic, Milos Radovanovic, Branka Jokanovic, *Senior Member, IEEE*

***Abstract*—This paper presents a new procedure for wide-band measurement of the Schottky diode impedance and rectified DC voltage. Two different Si Shottky diodes are measured, zero-bias SMS 7630 and low barrier MP 2005. Measurement is performed with Vector Network Analyzer (VNA) in the frequency range 0.5 - 5 GHz. The RF input power is varied from -28 to 2 dBm and several different values of load resistance are used. Since measurement required definite power levels at the diode ports, both S-parameter calibration and nonlinear power calibration of VNA were applied. Complex impedance and power conversion efficiency (PCE) of two diodes were compared. It was concluded that SMS 7630 is more sensitive in the almost whole power range. Also, it was noted that both diodes have similar complex impedance which means they can be used with the same antenna, that is designed to have an input impedance complexly conjugate in relation to the rectifying diode.**

***Index Terms*—Complex impedance measurement, power conversion efficiency, zero-bias Schottky diode, low barrier Shottky diode.**

I. Introduction

The main motivation for measuring the complex impedance and power conversion efficiency (PCE) of the Shottky diodes is their use for RF energy harvesting applications. RF energy harvesting is a process in which ambient electromagnetic radiation is collected, rectified, and then stored or used for powering low-power devices. Knowing that ambient electromagnetic radiation posseses very low power density, it is necessary to have rectification element that is able to work on lower input powers and at the frequencies assigned for wireless transmission. The basic elements of RF Energy Harvesting system are antenna, rectifier and power management circuitry. Combination of antenna, impedance matching circuit and rectifier, is often referred in literature with the common name rectenna [1]. One of the final goals of the diode measurement is to get data necessary for the design of rectenna with good performances, but without impedance matching circuit, because it brings additional losses to already low-level collected power. For that reason, it was useful to measure complex impedance of the rectifying diode, in order be able to design an antenna that would have acceptably matching impedance on the frequencies of interest.

In order to increase DC power collected from RF energy harvesting, multiple rectenna elements are used. Two approaches can be considered while designing a rectenna array. Block diagrams of both approaches are shown on Fig. 1. Rectenna array in Fig. 1. (a) combines collected RF power from antennas into a single rectifier. In rectenna array shown in Fig. 1. (b) output power of every antenna element is first rectified, and then combined later. According to results presented in [2], first approach leads to more harvested DC power near the main beam, while second topology results in more harvested power at angles away from broadside. Knowing power conversion efficiency of rectifying element can help when deciding which of these two topologies to use.

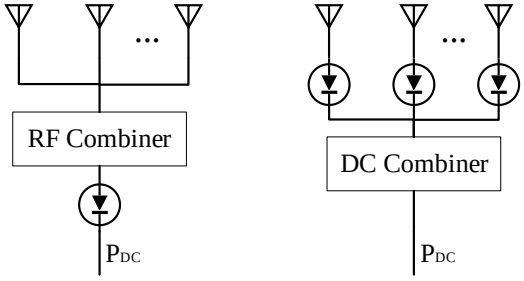


Figure 1. Two types of rectenna array: (a) with RF combiner and only one rectifying element, and (b) with rectifying element on every antenna and DC combiner

In this paper, two different Schottky diodes are measured, and results of the measurements are compared. Schottky diodes are chosen as the most frequently used diodes in RF energy harvesting application. Measured diodes are *flip-chip* Si Schottky diodes: SMS 7630 (zero-bias) and MP 2005 (low barrier). Parameters of these diodes, according to their data sheets [3, 4], are shown in Table 1.

TABLE I
PARAMETERS OF THE SHOTTKY DIODES

Parameter	**SMS 7630**	**MP 2005**
Breakdown voltage V_B	2 V (I_R=100 µA)	1 V (I_R=10 µA)
Forward voltage (max) V_F	240 mV (I_f=1 mA)	290 mV (I_f=1 mA)
Total capacitance C_T	0.2 pF (f=1 MHz)	0.18 pF (f=1 MHz)
Series resistance R_S	20 Ω	16 Ω

Branka Milosevic, Milos Radovanovic, Branka Jokanovic and are with the Institute of Physics, University of Belgrade, Pregrevica 118, 11080 Pregrevica, Serbia (e-mails: brankam@ipb.ac.rs, brankaj@ipb.ac.rs, rmilos@ipb.ac.rs).

II. Measurement and Results

For measurement of the diode's impedance, VNA was used that also served as RF source with frequency and power sweep. The output DC voltage of the diode, which was used to calculate PCE was measured by oscilloscope. Power calibration that provides accurate power levels for both impedance and PCE measurements was done by power meter.

Block diagram for both calibration and measurement is shown in Fig. 2. Three different pairs of reference planes can be identified at: AA', BB' and CC'. At reference planes AA', Vector Network Analyzer (VNA) can be considered calibrated, but measurements are supposed to be performed at reference planes CC'. For that reason, it is necessary to calibrate VNA, and complete calibration procedure will be explained in following subsection.

A. Calibration procedures

For the S-parameter calibration at the reference planes BB', full two-port SOLT (short-open-load-thru) calibration is performed. Coaxial standards needed for this calibration are provided in calibration kit.

In order to calibrate reference plane CC', custom LRL (line-reflect-line) calibration set was designed (see Fig. 3). The LRL is a calibration algorithm that uses two or more transmission lines and a reflect standard for each port. Difference in the transmission line lengths ΔL is important parameter for the LRL calibration, since it will not work if ΔL equals to integer multiplicity of the half wavelength. Theoretically, the electrical length of ΔL should be between 0 and 180 degrees for all frequencies. Due to line parasitics, spurious modes and other problems, it is recommended to keep ΔL between 20 and 160 degrees [5].

The reflect standard for LRL calibration can be open or short circuit, and it is only important that it's symmetric with not too high return loss. In this case, short circuit was used. All standards used for this calibration are displayed in Fig. 2. Standards are designed as 50 Ω-coplanar waveguide (CPW) transmission lines on the substrate RO 3010 (ε_r=11.2, h=0.635 mm, t=0.017 mm) since the device under test (DUT) also contains 50 Ω-CPW lines with series mounted flip-chip diode. CPW technology was chosen due to simple diode assemblies and almost constant line characteristics in a wide frequency range (low dispersion).

Measurement of PCE requires accurate power levels at reference plane CC'. For that reason the source power calibration was performed with external HP 438A Power Meter and HP 8481 power sensor. It wasn't possible to use power sensor at the reference planes CC', so the power calibration was performed at the reference planes BB'. The difference in power levels between reference planes BB' and CC' was recalculated later.

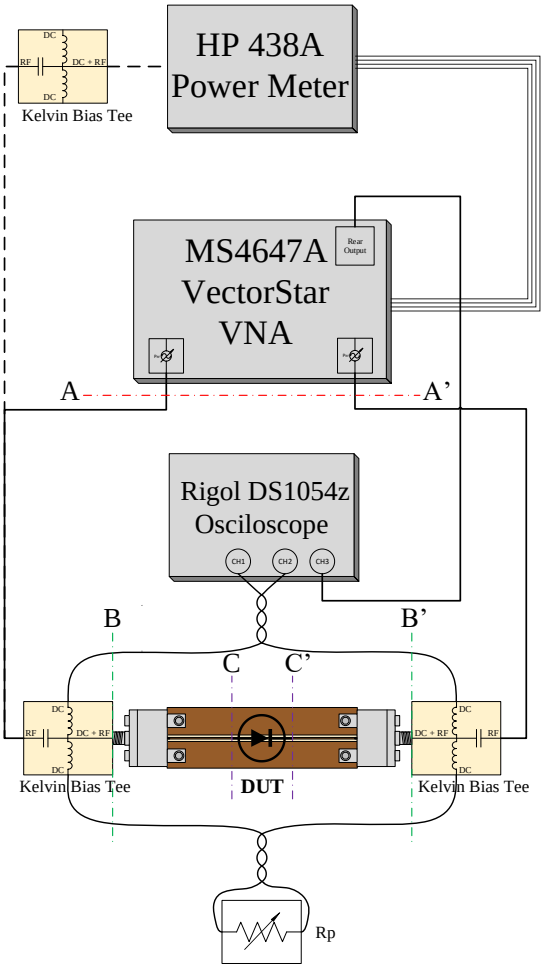


Fig. 2. Block diagram of the measurement set-up. DUT is placed between C-C' reference planes.

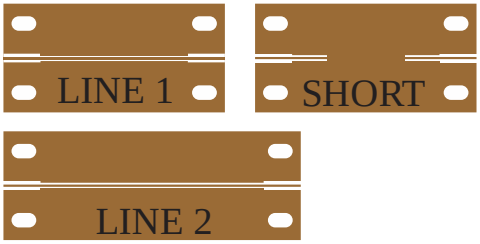


Fig. 3. Standards for the LRL calibration with CPW transmission lines: LINE1=20.2 mm, LINE2=30.6 mm, double SHORT circuit.

B. Results and comparison

Results of interest in our case were complex impedance and PCE of the two diodes. In order to calculate complex impedance from measured S parameters equivalent π-network was used, with diode's impedance in the middle and port's impedance on sides. According to [6], value of that impedance is:

$$Z_D = -\frac{1}{Y_{21}(S_{21}, S_{11})}, \tag{1}$$

where Y_{21} is obtained from following expression:

$$Y_{21} = \frac{-2S_{21}}{(1+S_{11})^2 - S_{21}^2} \cdot \frac{1}{Z_0}. \tag{2}$$

PCE was calculated from measured voltage as:

$$\eta = \frac{\frac{V_{OSC}^{\ 2}}{R_{LOAD}}}{P_{IN}}. \qquad (3)$$

In Fig. 4.-6. complex impedance of both diodes is shown. It can be noted that measured values of impedance are similar for input powers greater than -8 dBm, which is especially visible on Fig. 5. On Fig. 4., values for impedance are vastly different, because MP 2005 diode wouldn't work for low input power P_{IN} = -18 dBm.

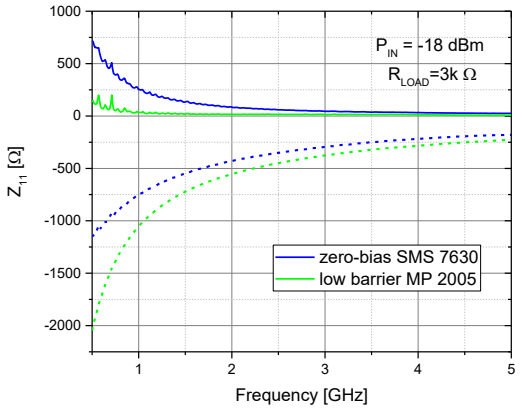


Fig. 4. Complex impedance of SMS 7630 and MP 2005 diodes for input power P_{IN} = -18 dBm and R_{LOAD} = 3 kΩ. Real part of the impedance is presented with solid, while imaginary part is presented with dashed line.

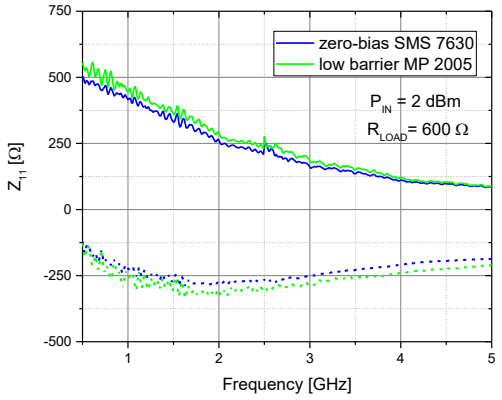


Fig. 5. Complex impedance of SMS 7630 and MP 2005 diodes for the input power P_{IN} = 2 dBm and R_{LOAD} = 600 Ω. Real part of impedance is presented with solid, while imaginary part is presented with dashed line.

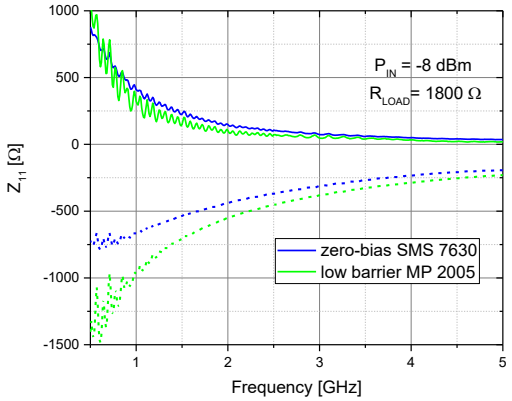


Fig. 6. Complex impedance of SMS 7630 and MP 2005 diodes for the input power P_{IN} = -8 dBm and R_{LOAD} = 1.8 kΩ. Real part of impedance is presented with solid, while imaginary part is presented with dashed line.

Complex impedance of both diodes, as function of input power, is shown in Fig. 7.-8. It can be noted that MP 2005 has significantly lower resistance for input power smaller than -8 dBm, probably because it barely works at power levels lower than that.

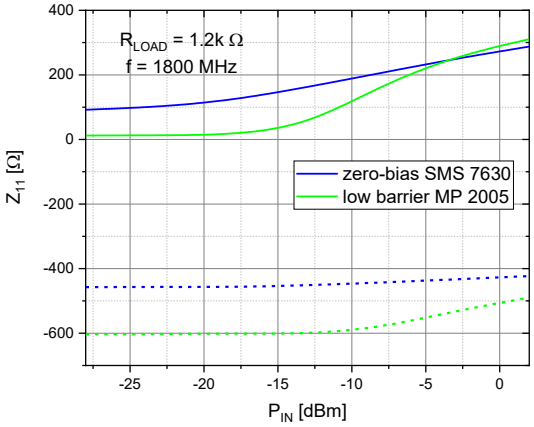


Fig. 7. Complex impedance of SMS 7630 and MP 2005 diodes as function of input power P_{IN}, on frequency f = 1800 MHz and load impedance R_{LOAD} = 1.8 kΩ. Real part of impedance is presented with solid, while imaginary part is presented with dashed line.

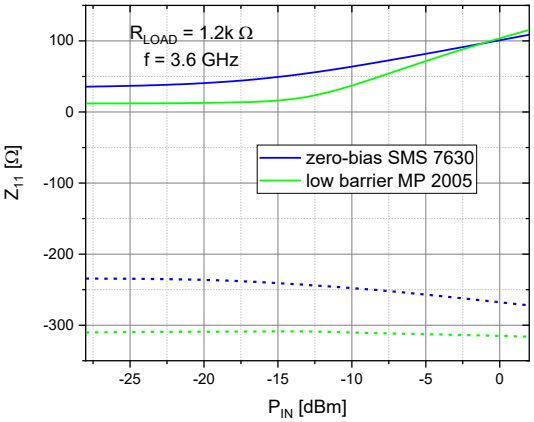


Fig. 8. Complex impedance of SMS 7630 and MP 2005 diodes as function of input power P_{IN}, on frequency f = 3.6 GHz and load impedance R_{LOAD} = 1.8 kΩ. Real part of impedance is presented with solid, while imaginary part is presented with dashed line.

Comparison of PCE of two diodes is shown in Fig. 9.-10. It can be seen that the diode SMS 7630 always perform better for lower input power, which was expected. However, MP 2005 diode becomes more efficient for the input powers larger than 1 dBm, at working frequency f = 900 MHz (Fig. 9.). If working frequency is f = 2.4 GHz, then SMS 7630 diode would be more efficient for all values of the input power that is used in this measurement (Fig. 10.). If we could choose between two topologies shown in Fig. 1., first approach probably would be preferable for SMS 7630 diode, but MP 2005 would work significantly better with second topology.

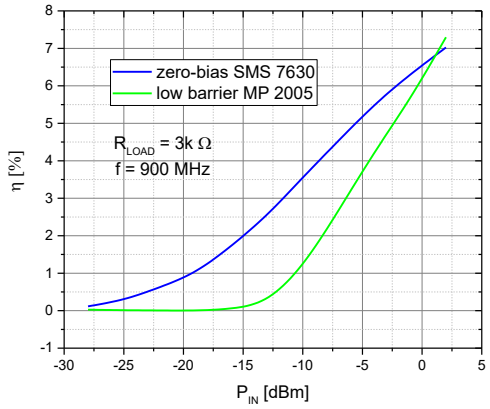


Fig. 9. Comparison of efficiency of SMS 7630 and MP 2005 diodes. Efficiency was measured for load resistance R_{LOAD} = 3 kΩ, and at 900 MHz frequency.

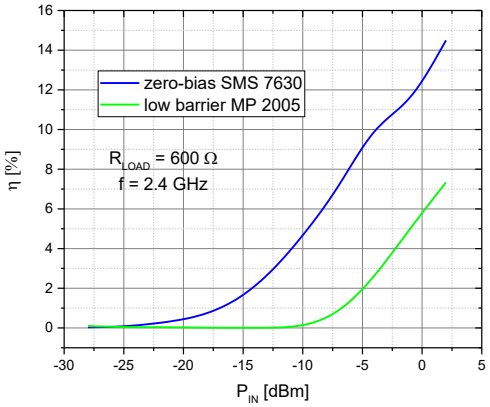


Fig. 10. Comparison of efficiency of SMS 7630 and MP 2005 diodes. Efficiency was measured for load resistance R_{LOAD} = 600 kΩ, and at 2.4 GHz frequency.

III. Conclusion

This paper presents the measurement procedure and obtained results of rectifying Si Schottky barrier diodes that are planned to be used for power harvesting application. As expected, measured results confirm that the low barrier diode MP 2005 is less sensitive than zero-bias diode SMS 7630 in the almost whole applied power range (from -28 to 2 dBm). At a higher frequency range, above 0.9 GHz, the low-barrier diode exhibits even much lower efficiency in respect to the zero bias one. Measured impedance of the diodes are quite similar, which means they can be used with the same antenna, that is designed to have an input impedance complexly conjugate in relation to the rectifying diode.

Acknowledgment

This work was supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia, as a part of the project TR 32024.

References

[1] J. A. Hagerty, F. B. Helmbrecht, W. H. McCalpin, R. Zane and Z. B. Popovic, "Recycling ambient microwave energy with broad-band rectenna arrays," in IEEE Transactions on Microwave Theory and Techniques, vol. 52, no. 3, pp. 1014-1024, March 2004.

[2] U. Olgun, C. Chen and J. L. Volakis, "Investigation of Rectenna Array Configurations for Enhanced RF Power Harvesting," in *IEEE Antennas and Wireless Propagation Letters*, vol. 10, pp. 262-265, 2011.

[3] http://www.skyworksinc.com

[4] https://www.mpulsemw.com

[5] Understanding VNA calibration, Anritsu.

[6] Pozar, David M. (2005); *Microwave Engineering*.

VHF Gysel 3 dB Power Divider/Combiner in Microstrip Technology

Veljko Crnadak, *Member, IEEE,* and Siniša Tasić

***Abstract*—In this paper, the theory and the design of the VHF Gysel 3 dB power divider/combiner are presented. The Gysel 3 dB power divider/combiner is designed as a microstrip circuit, for the frequency range from 150 MHz to 200 MHz, with the central frequency at 175 MHz. The Gysel divider/combiner is realized as a microstrip circuit on Rogers 4003 substrate, with the copper traces. Knowledge of the conditions, which must be satisfied by the Gysel divider/combiner, enables the determination of the dimensions of the Gysel divider/combiner, through the process of computer simulation. In order to verify the project, values of the S-parameters of the simulated Gysel divider/combiner are compared to the measured values of the S-parameters of the produced Gysel divider/combiner.**

***Index Terms*—combiner; divider; even mode; Gysel; microstrip; odd mode; Rogers 4003; VHF; Wilkinson.**

I. Introduction

A two-way power divider/combiner can split an RF signal into two paths. In reverse direction, it can combine two signals onto a single path. There are two types of power dividers/combiners, resistive ones, and reactive ones. A resistive two-way power divider/combiner has a nominal insertion loss of 6 dB, while a reactive one has a nominal insertion loss of 3 dB. Isolation between input/output ports in a resistive divider/combiner is 6 dB, while in a reactive one is typically 20 dB. If we want to combine outputs of two or more power amplifiers we must use reactive power combiners. The two most famous members of a reactive power divider/combiner group are Wilkinson and Gysel power dividers/combiners. The Wilkinson 3 dB power divider/combiner, Fig. 1, has one crucial drawback and that is its 100 Ω lumped resistor. The lumped resistor has to be small compared to the wavelength, so it is difficult to produce it at high frequencies, and it is problematic to heat-sink it because of its balanced nature [1]. In his design, Gysel swapped the lumped resistor with a network of transmission lines and two 50 Ω ballast resistors. The Gysel 3 dB combiner is used to combine the outputs of two amplifiers in order to obtain a high-power amplifier because proper heat sinking of the external ballast resistors is achievable [2]. The purpose of the paper is to show the differences that arise when realizing the Gysel divider/combiner in microstrip technology, compared to strip-line technology that was implemented in the previous work [3].

Veljko Crnadak is with the School of Electrical Engineering, University of Belgrade, Bulevar kralja Aleksandra 73, 11120 Belgrade, Serbia, and with the Company for Microwave and Millimeter-Wave Techniques and Electronics IMTEL-Komunikacije Joint-Stock Company Belgrade, Bulevar Mihajla Pupina 165b, 11070 Novi Beograd, Serbia (e-mail: veljko@insimtel.com).

Siniša Tasić is with the Company for Microwave and Millimeter-Wave Techniques and Electronics IMTEL-Komunikacije Joint-Stock Company Belgrade, Bulevar Mihajla Pupina 165b, 11070 Novi Beograd, Serbia (e-mail: tasa@insimtel.com).

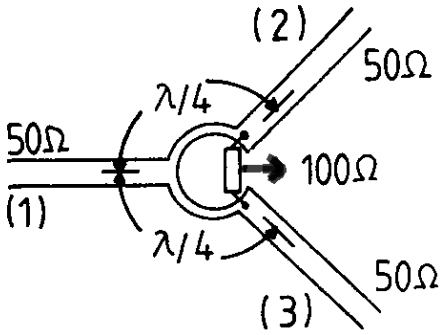


Fig. 1. The Wilkinson 3 dB power divider/combiner [1].

II. Design Theory

The Gysel 3 dB power divider/combiner circuit is shown in Fig. 2.

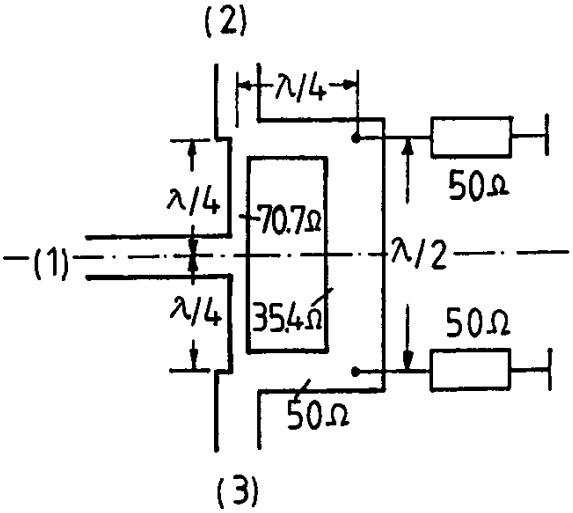


Fig. 2. The Gysel 3 dB power divider/combiner [1].

Port 1 is connected via transmission-line of arbitrary length and the characteristic impedance of 50 Ω, with the rest of the

divider/combiner. Lengths of all transmission-lines in Fig. 2, are determined at the center frequency of the divider/combiner's bandwidth. Two transmission-lines, each a $\lambda/4$ long and with the characteristic impedance of 70.7 Ω, lead from the input port to the two output ports. Two transmission-lines, each a $\lambda/4$ long and with the characteristic impedance of 50 Ω, connect each ballast resistor with its corresponding output port [4]. The two ballast resistors are connected via transmission-line, a $\lambda/2$ long and with the characteristic impedance of 35.4 Ω. Ports 2 and 3 are connected, each via transmission-line of arbitrary length and the characteristic impedance of 50 Ω, with the rest of the divider/combiner. The ballast resistors, 50 Ω each, can be replaced by a transmission-line of the characteristic impedance of 50 Ω, arbitrary in length and terminated in a load of value 50 Ω [4]. Therefore, each ballast resistor has been substituted with an external load, that can sustain high-power. The loads are no longer the power-limiting factor, while the breakdown voltage of the divider/combiner's microstrip lines is the limiting factor [4]. External loads also provide high-isolation between input/output ports. The function of the Gysel 3 dB divider is to divide an input signal into two equal outputs (equal by phase and amplitude). In the reverse direction, the Gysel 3 dB divider works as a combiner, that is it combines two in-phase signals into an output signal. The scattering matrix of an ideal Gysel 3 dB divider/combiner is [3],

$$S = -\frac{j}{\sqrt{2}}\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}. \quad (1)$$

The total power that is dissipated on the external loads of the Gysel 3 dB combiner is,

$$P_{tot} = \frac{1}{2}(P_2 + P_3) - \sqrt{P_2 P_3}\cos(\Delta\varphi_{23}), \quad (2)$$

where P_2 and P_3 are the input powers, while $\Delta\varphi_{23}$ is the phase difference between the signals at the input ports.

Because the Gysel power divider/combiner is symmetrical with respect to port 1, we can apply the bisection theorem, that is we can analyze the circuit in the even and the odd mode. The equivalent circuit for the even mode is shown in Fig. 3.

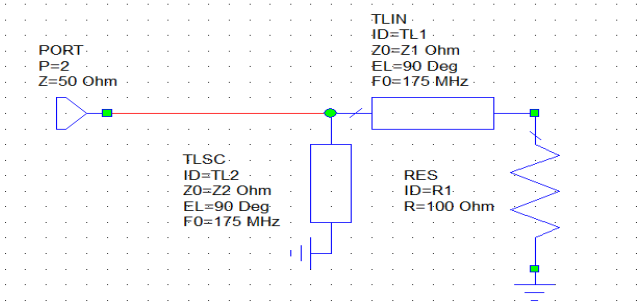


Fig. 3. The equivalent circuit for the even mode.

As Fig. 3 shows, the transmission-line with the characteristic impedance Z_1, behaves as a $\lambda/4$-impedance transformer, while the transmission-line with the characteristic impedance Z_2, behaves as a short-circuited $\lambda/4$-stub. The equivalent circuit for the odd mode is shown in Fig. 4.

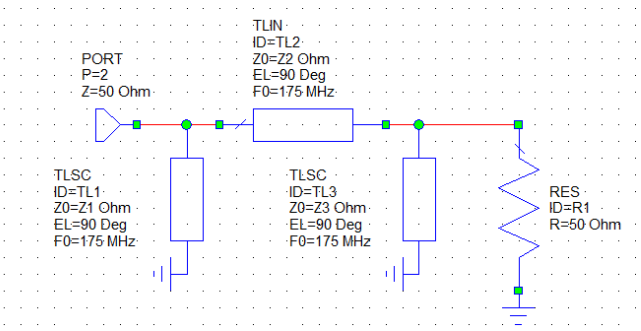


Fig. 4. The equivalent circuit for the odd mode.

As Fig. 4 shows, the transmission-lines with the characteristic impedance Z_1 and Z_3 behave as short-circuited $\lambda/4$-stubs, while the $\lambda/4$-transmission-line that connects them, with the characteristic impedance of Z_2, behaves as an impedance inverter. The circuit in Fig. 4, acts as a band-pass filter of the second order. The basic coupling between the circuits in Figs. 3 and 4 is done through the transmission-lines Z_1 and Z_2 [1]. The Gysel 3 dB power divider/combiner is realized as a microstrip circuit on 1.52 mm thick Rogers 4003 substrate with a relative dielectric constant of ε_r =3.55, and with copper traces 35 μm thick. The width of every input/output line is 3.36 mm. The length of the line that connects port 1 with the rest of the circuit is 5 mm. Both ports 2 and 3 are connected with the rest of the circuit, via the transmission-line of the length 29.5 mm. The width of both lines with the impedance Z_1, 66.87 Ω, is 2.02 mm and the length is 282.63 mm. The width of both lines with the impedance Z_2, 49.76 Ω, is 3.39 mm and the length is 290.6 mm. The width of the line with the impedance Z_3, 23.76 Ω, is 9.62 mm and the length is 565.52 mm. The width of both lines that are connected to the external loads is 3.36 mm and the length is 31.18 mm. The three SMA connectors are mounted on the circuit. The picture of the fabricated Gysel 3 dB power divider/combiner is shown in Fig. 5. The dimensions of the circuit, in Fig. 5, are 190 mm x 102 mm.

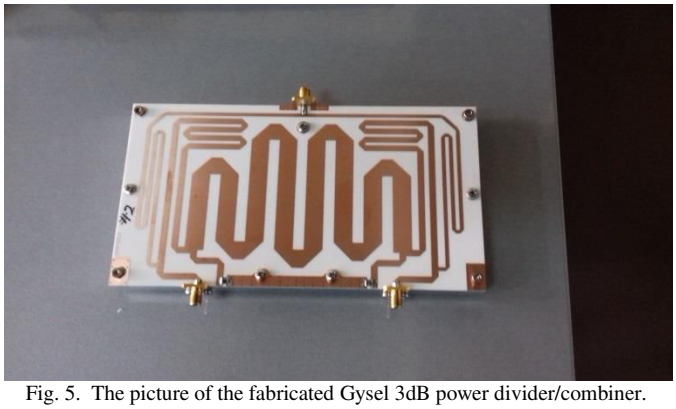


Fig. 5. The picture of the fabricated Gysel 3dB power divider/combiner.

III. Simulation And Measurement Results

A. Simulation

The circuit in Fig. 3 behaves as a resonator with two degrees of freedom, Z_1 and Z_2. In this case, by tuning the impedances Z_1 and Z_2, we can only achieve one zero of the reflection coefficient S_{22} at the center frequency, as it is shown in Fig. 6.

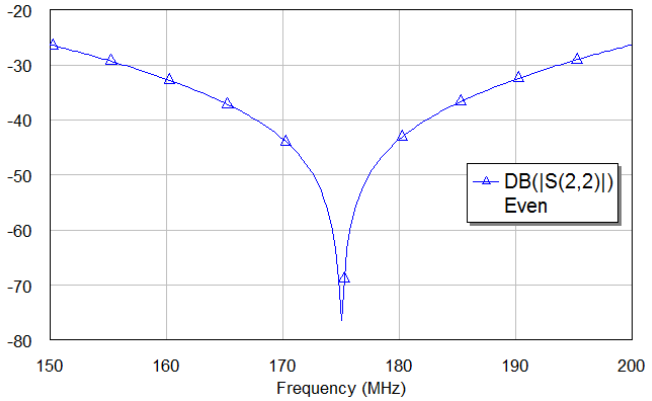


Fig. 6. The reflection coefficient S_{22} in the even mode.

On the other hand the circuit in Fig. 4, behaves as a resonator with three degrees of freedom, Z_1, Z_2 and Z_3. By tuning the impedances Z_1, Z_2 and Z_3, we can achieve two zeros of the reflection coefficient S_{22}, Fig. 7, which increases the bandwidth of the Gysel divider/combiner.

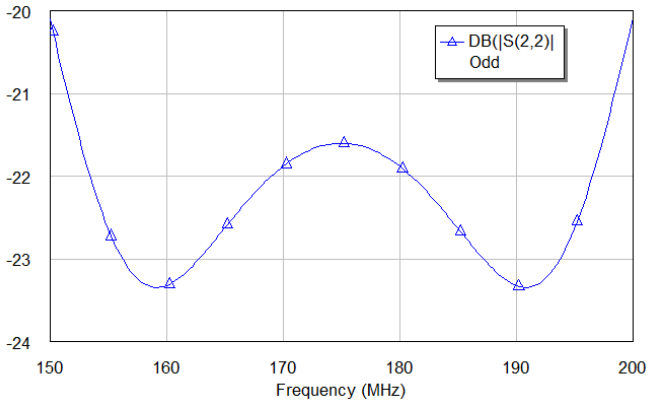


Fig. 7. The reflection coefficient S_{22} in the odd mode.

Having been aware of the conditions, which had to be met by the Gysel power divider/combiner (S_{11}, S_{22}, S_{33} and S_{23}, lower than -20 dB, and S_{12}, S_{13}, lower or equal than -3 dB), we were in the position to direct the optimization process towards satisfying those same conditions, and obtaining the dimensions of the Gysel divider/combiner. The top view of the model of the Gysel 3 dB divider/combiner in MWO Axiem simulator, is given in Fig. 8.

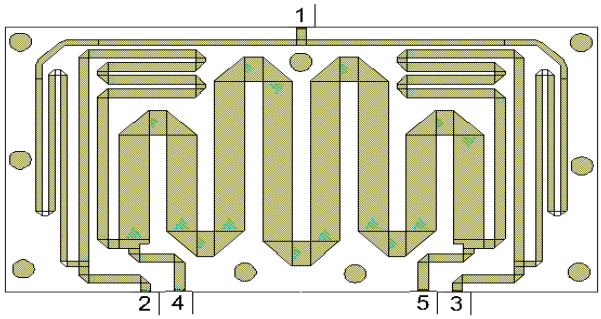


Fig. 8. The top view of the model of the Gysel divider/combiner in MWO Axiem simulator.

B. Simulated and measured results

Figure 9 shows that the measured reflection coefficient S_{11} is lower than -21.1 dB, over the frequency band of interest. The measured reflection coefficient S_{22} is lower than -24.78 dB, over the frequency band of interest, as can be seen in Fig. 10. Figure 11 shows that the measured reflection coefficient S_{33} is lower than -24.88 dB, over the frequency band of interest. The measured isolation S_{23} is lower than -17.65 dB, over the frequency band of interest, as can be seen in Fig. 12. Insertion loss is equal or lower than 3.191 dB and the amplitude balance is ±0.014 dB, over the frequency band of interest, as can be seen in Fig. 13 and Fig. 14. In Fig. 15 we can see that the phase angles of S_{12} and S_{13} are practically the same, over the frequency band of interest.

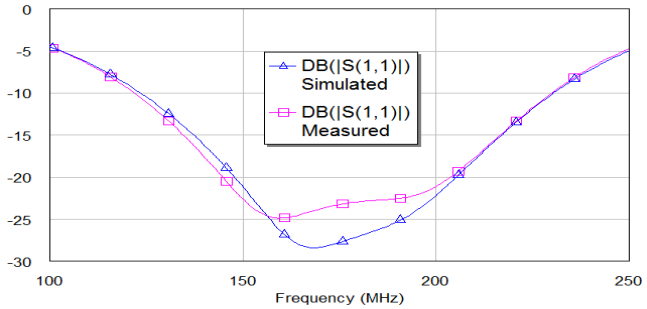


Fig. 9. Comparison of the simulated and measured S_{11}.

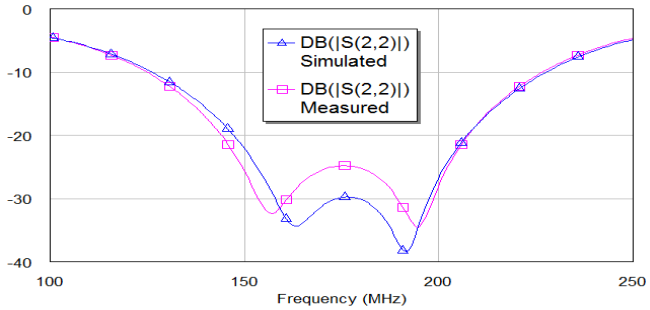


Fig. 10. Comparison of the simulated and measured S_{22}.

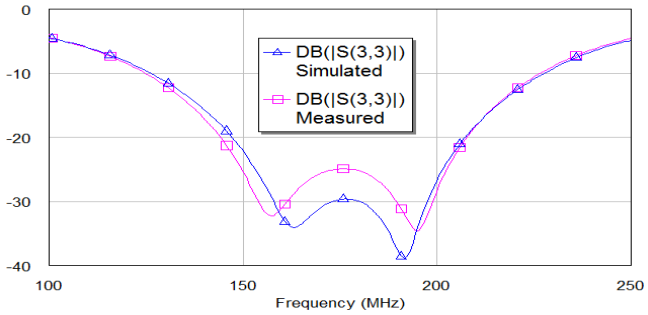


Fig. 11. Comparison of the simulated and measured S_{33}.

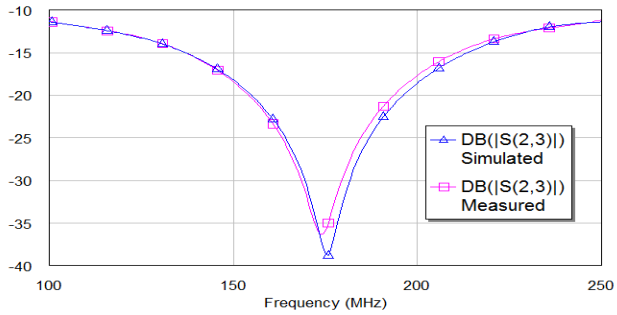


Fig. 12. Comparison of the simulated and measured S_{23}.

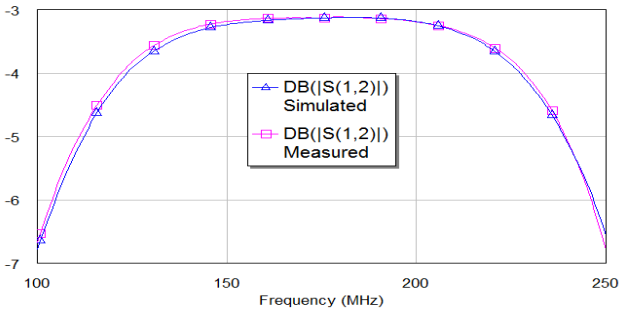


Fig. 13. Comparison of the simulated and measured magnitude of S_{12}.

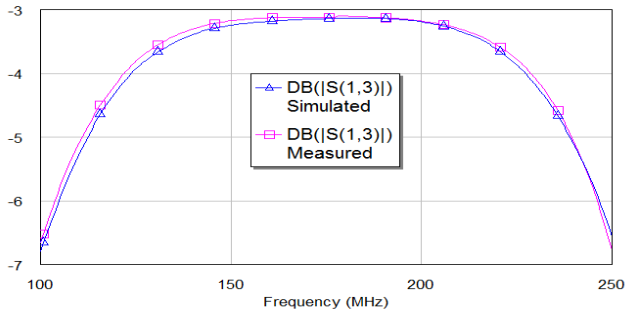


Fig. 14. Comparison of the simulated and measured magnitude of S_{13}.

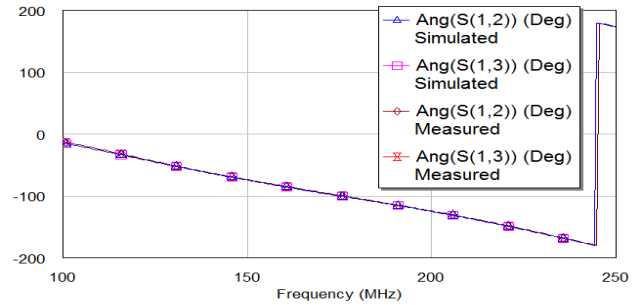


Fig. 15. Comparison of the simulated and measured phases of S_{12} and S_{13}.

IV. Conclusion

In this article, we have described the theory and the design of the VHF Gysel 3 dB power divider/combiner. The objectives regarding the level of the return loss on all three ports, the level of the insertion loss and the level of isolation between the input/output ports were met. The bandwidth of the realized Gysel divider/combiner is 28.57 %. The main advantages of the Gysel 3dB divider/combiner are external isolation loads (permitting high-power loads), easily realizable geometry and monitoring capability for imbalances at the output ports [4]. Microstrip technology favors substantially lower dimensions of the circuit, but higher losses, while strip-line technology provides the capability of handling high-power signals, in the continuous wave mode, at the cost of larger dimensions of the circuit.

Acknowledgment

This paper was partially supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia under grant TR-32024.

References

[1] R. Knochel and B. Mayer, “Broadband Printed Circuit 0°/180° Couplers and High Power In phase Power Dividers,” 1990 IEEE MTT-S Int. Microwave Symp. Dig., pp. 471-474.

[2] A. Grebennikov, “RF and Microwave Power Amplifier Design”, 2nd ed., McGraw-Hill Education, 2015, pp. 353-355.

[3] Veljko Crnadak, Siniša Tasić, “VHF Gysel 3 dB Power Divider/Combiner”, The 4th International Conference on Electrical, Electronics and Computing Engineering, IcETRAN 2017, June 05-08, Kladovo, Serbia.

[4] U. H. Gysel, “A New N-Way Power Divider/Combiner Suitable for High-Power Applications”, 1975 IEEE MTT-S Int. Microwave Symposium. Dig., pp. 116-118.

EM Modelling of Microstrip T-Junction with an Open Stub Printed over a Dielectric Cylinder

Tomislav Milošević, Dusan Nešić, *Member, IEEE*

***Abstract*—This paper presents EM modelling of microstrip T-junction with open stub printed over a dielectric cylinder. Software tool used for modelling and simulations is a full 3D EM Method-of-Moments, Surface Integral Equation solver applied to quadrilateral mesh elements. The paper investigates the stability of simulation results with respect to various parameters. This includes settings of the numerical kernel, quality of cylindrical surface approximation using segments of bilinear surfaces, different modelling of microstrip edge effects, and comparison between various excitation configurations with and without de-embedding. The purpose of this comprehensive investigation was to establish the optimum calculation parameters for the case of the particular resonator structure. With the optimum parameter settings, the high numerical efficiency of the calculations has been confirmed.**

***Index Terms*—EM modelling, geometrical modelling, microstrip T-junction, open stub, simulation.**

I. Introduction

IN the last couple of decades, electromagnetic (EM) modelling has become a standard process in research and development of various microwave devices. The analytical methods for analysis of such structures are limited to several simple cases. For realistic, complex devices only numerical solutions are adequate. Tremendous efforts have been made to develop efficient tools for versatile, yet intuitive geometrical modelling and efficient numerical analysis of various classes of real-world EM devices. The structures of interest include antennas, scatterers, passive microwave circuits etc. The main requirement for a numerical method suitable for the analysis of such structures is to accurately determine the distribution of the electromagnetic field, currents and charges. Such an analysis is often referred to as the electromagnetic modeling.

Within electromagnetic modeling, various structures can be treated following similar guidelines. For example, passive microwave circuits, such as microstrip transmission lines, usually represent structures consisting of metallic and dielectric parts collectively referred to as composite metallic and dielectric structures. If all the parts are made of linear materials, analysis can be facilitated in the frequency domain [1]. In this paper, we exploited WIPL-D Pro, a full wave, 3D EM frequency-domain Method-of-Moments (MoM) based code for geometrical modelling and simulation which applies higher order basis functions (HOBFs) on quadrilateral mesh elements using Surface Integral Equations (SIE) [4].

Dušan Nešić is with Centre of Microelectronic Technologies, Institute of Chemistry, Technology and Metallurgy, University of Belgrade, Njegoševa 12, 11000 Belgrade, Serbia (e-mail: nesicad@nanosys.ihtm.bg.ac.rs).

Tomislav Milošević is with WIPL-D d.o.o., Gandijeva 7, 11073 Belgrade, Serbia (e-mail: tomislav.milosevic@wipl-d.com).

The stability of output results is usually checked by manually varying numerical kernel parameters. In addition, the influence of some other parameters of interest for EM modelling can be investigated for optimum numerical efficiency.

The geometrical modeling assumes drawing a composite metallic and dielectric structure combining available building elements, bilinear surfaces – plate entities, and wire entities. The modeling can in general be exact or approximate. In the particular case of geometry modeling considered here, the modeling of curved\cylindrical structures is approximate and a search for an acceptable approximation basically involves finding a minimum number of straight segments to approximate curved\cylindrical geometry [1].

The microstrip structures are widely used in microwave engineering. Usually, standard microstrip devices are planar and easily modeled. However, in some applications, the basic microstrip structure can be modified and the properties of a modified structure exploited for a particular advantage, as a structure described in [3]. This paper focuses on EM modelling of a modified, not strictly planar microstrip structure which represents fluid sensors and can be used in measurements of a fluid characteristics [2,3].

The model considered here is a non-strictly planar microstrip structure operating between 0.4 GHz and 4.0 GHz. It is a T-junction which is suitable to be used as a fluid sensor. The results of various simulations are presented and explained.

II. Geometrical Model

The particular T-junction structure with an open stub over the dielectric cylinder is shown in Fig. 1 The dimensions of the parts are indicated in the same figure. The structure has been modeled using one symmetry plane named (A)Symmetry. (A)Symmetry represents a software feature where a symmetry of the structure is exploited to reduce an original number of unknowns. The original number of unknowns is approximately halved. With the feature made active, the numerical kernel is automatically invoked two times and the results are automatically combined.

In order to efficiently control meshing of cylindrical areas, software built-in objects Body-of-Rotation (BoR) were used (Fig. 2). Finite metallization thickness (Fig. 2) is also added using software built-in manipulation. All dimensions of the structure are parametrized through the symbolic variables and can be changed easily.

To optimize numerical accuracy, the resonator structure was modeled by introducing ‘imaging’ where necessary (Fig. 2). This means that closely located ‘upper’ and ‘lower’ surfaces from the model are meshed in the same way. This improves the stability of MoM matrix.

The dielectric substrate is with the following parameters: dielectric constant is 3, while loss tangent is 0.01. Dielectric constant of a lossless cylinder is 10. The cylindrical dielectric in the interior of the structure and the neighboring conductors are shown in Fig. 3.

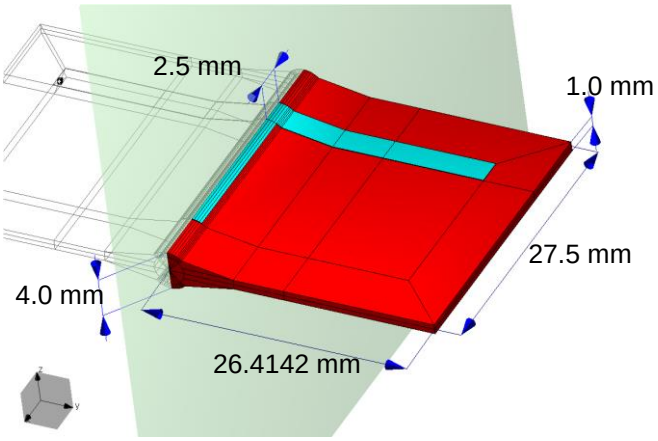


Fig. 1. Simulated structure with overall dimensions and one symmetry plane.

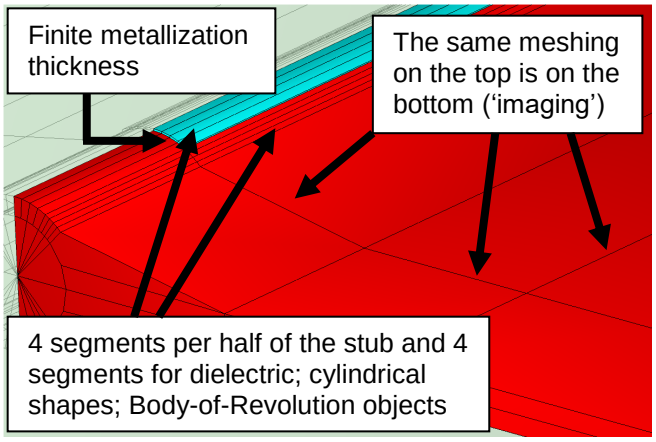


Fig. 2. Simulated structure-the detail. Body-of-Revolution objects modelling curved cylindrical surfaces, finite metallization thickness, and meshing lines on the 'upper' surface can be recognized.

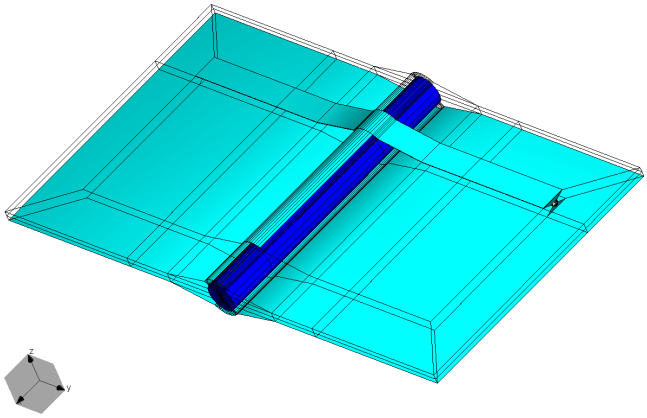


Fig. 3. The cylindrical dielectric and neighboring conductors in the full model without the symmetry applied.

Improved handling of edge effects in particular MoM SIE based software was achieved by automatic subdividion of the plates having a common edge which in fact separates areas belonging to different materials. The automatic subdivision follows after applying, so called, Edge-ing manipulation. Details of modified structure are in Fig. 4.

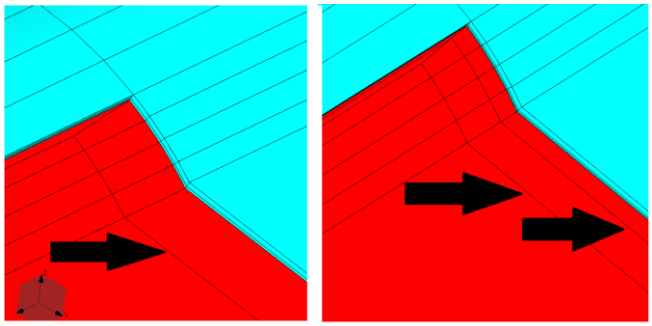


Fig. 4. Handling of edge effects – the pre-simulated structure modification after application of single Edge-ing and double Edge-ing manipulations.

The location of the reference plane required for the de-embedding and position of generator used for feeding are shown in Fig. 5.

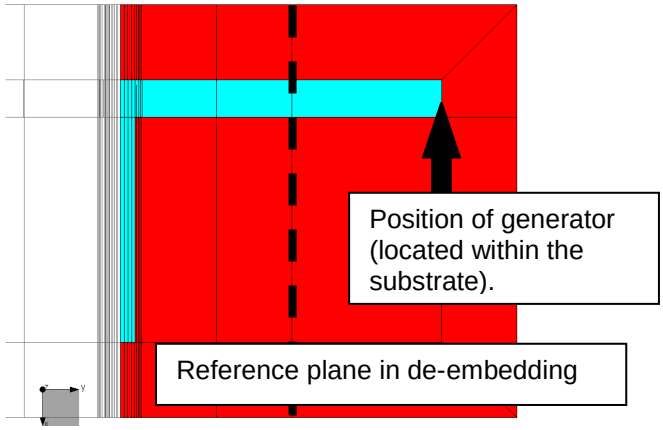


Fig. 5. Reference plane defined in the de-embedding process and position of the generator used for feeding.

III. Simulation Results

In order to demonstrate EM modelling of the structure, to highlight the most important steps in the investigation of this, and to discuss some actions required in usual EM modeling, we present here S-parameters from several simulation scenarios. At first instance, we investigate the convergence of the simulations by varying standard calculation parameters. After that we investigated influence of a number of bilinear surface segments used in cylindrical objects modelling. The quality of edge effects modelling was investigated next. Finally, the results obtained with and the results obtained without the de-embedding are compared.

All of the models have been simulated at 21 frequency points. The smooth curves are resulting from excellent graphical fitter used. All of the models were simulated on standard Intel® Core™ i7-7700 CPU @ 3.60 GHz with 64 GB RAM. The most time-consuming simulation is the one where a model of the structure includes Edge-ing manipulation applied two times (total simulation time is 900 seconds). The most of the other simulations finish after

500 seconds - 850 seconds.

A. *Obtaining Optimally Stable Results*

Convergence check was performed by manually varying numerical kernel parameters. The first step was increasing parameter Integral accuracy (IA) from *normal* to *Enhanced 1* and graphically comparing the responses obtained (Fig. 6). Parameter IA influences evaluation of integrals related with accurate calculation of MoM matrix elements. In general, higher IA enables more accurate results, exploiting longer simulation time. Results obtained after IA set to *normal* and set to *Enhanced 1* are denoted as IA = 0 and IA = 1, respectively.

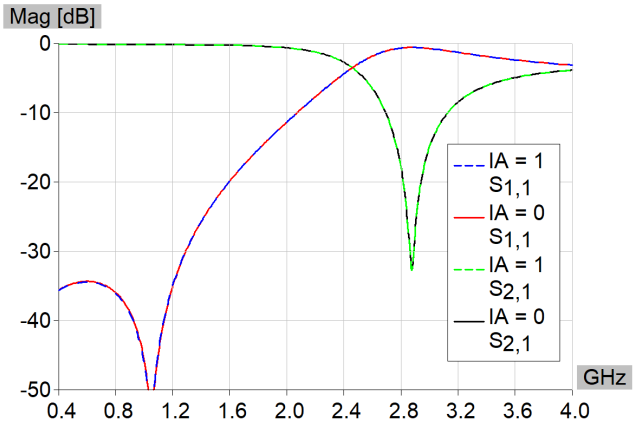


Fig. 6. S-parameters with Integral accuracy parameter varied.

The effect of increasing a number of unknowns has been investigated next. The number of unknowns represents the number of unknown coefficients used for current distribution approximation. Comparison of the results is shown in Fig. 7 and Fig. 8.

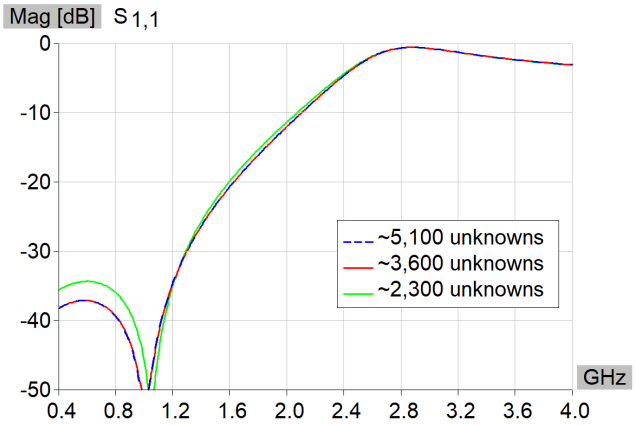


Fig. 7. S_{11}-parameters obtained for varied number of unknowns.

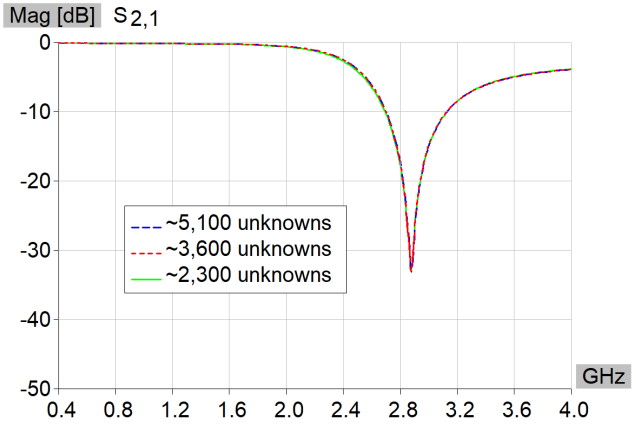


Fig. 8. S_{21}-parameters obtained for varied number of unknowns.

A tradeoff between sufficient accuracy and minimum computer resources required suggests that a value for the integral accuracy should be set to *normal* with a number of unknowns set to 3,600.

B. *Bilinear Segments Used in Cylindrical Surfaces Modelling*

Cylindrical objects are approximated using planar bilinear surfaces, as in BoR objects shown in Fig. 2. As a next step towards establishing the optimum calculation parameters, a number of segments is varied and the influence of changing number of segments to the results investigated. Results with various numbers of segments determining stub curvature and dielectric curvature are shown in the Fig. 9 and Fig. 10, respectively. A quality of approximation using a total of 12 and 8 segments was investigated for the case of a stub curvature, while a total of 6 and 4 segments were used for modelling the dielectric curvature.

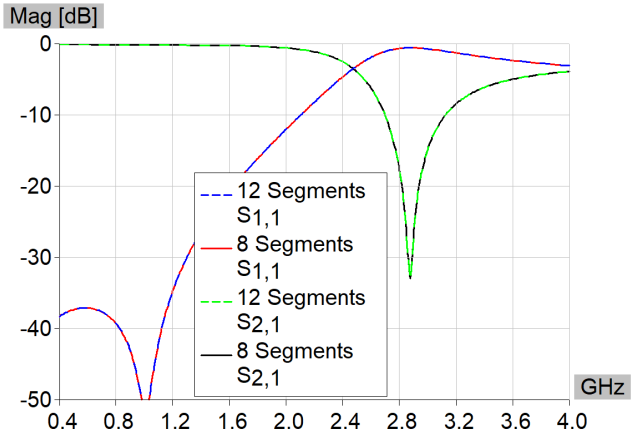


Fig. 9. S-parameters for varied number of segments of the Body-of-Revolution object modelling stub curvature.

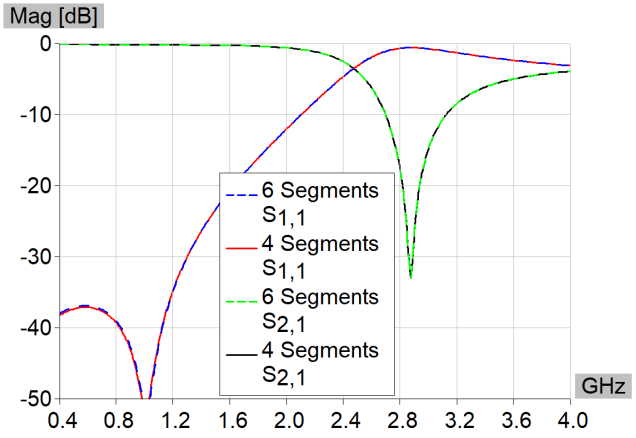


Fig. 10. S-parameters for varied number of segments of the Body-of-Revolution object modelling dielectric curvature near stub.

C. *Investigating Edge Effects*

According to the recommendations for the particular software used for the simulations, at least one Edge-ing manipulation (Fig. 4) should be used for microstrip and microstrip-like structures. Following the recommendations, in the analysis of the model of the structure with IA set to *normal* requiring 3,600 unknowns, which S-parameters are shown in Fig. 7 and Fig. 8 one Edge-ing manipulation has been used already. However, the influence of Edge-ing manipulation to the accuracy of the simulations is investigated as follows. The S-parameters obtained without Edge-ing are compared with results with one Edge-ing manipulation applied and with two Edge-ing manipulations applied. The results are shown in Fig. 11.

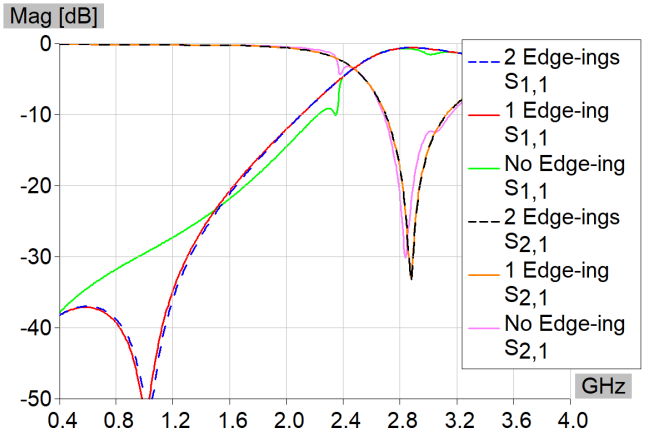


Fig. 11. S-parameters for models where Edge-ing manipulation is not applied with models where one and two manipulations Edge-ing are applied.

D. *The Influence of De-Embedding Procedure*

In the simulations presented so far, S-parameters are calculated with respect to the reference plane coinciding with the location of the generators. In order to calculate S-parameters in reference planes presented in Fig. 5, the De-embedding technique is applied. The comparison between S-parameters obtained after De-embedding procedure applied and the simulation without de-embedding procedure are shown in Fig. 12.

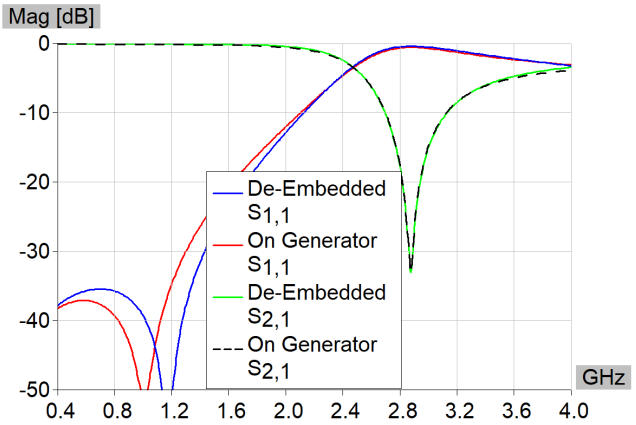


Fig. 12. De-embedded S-parameters and S-parameters calculated without -embedding, on the generators.

IV. Conclusion

This paper presented EM modelling of microstrip T-junction with open stub printed over a dielectric cylinder. Software tool used for modelling and simulations was a full 3D EM Method-of-Moments, Surface Integral Equation solver applied to quadrilateral mesh elements.

The paper investigates the stability of EM simulation results with respect to various parameters: settings of the numerical kernel through changing parameters of a numerical integration and a number of unknowns required, quality of approximation of cylindrical surfaces, different modelling of microstrip edge effects, and comparison between various excitation configurations with and without de-embedding.

The investigation of EM modeling of the structure was established in order to reveal optimum simulation parameters for the case of the particular resonator structure which modification can act as a real-life sensor for fluid measurement. With the optimum parameter settings, the high numerical efficiency of the calculations has been confirmed. It has been shown that, with proper EM model, even with the standard settings applied to control the operation of the numerical kernel, the utilized software produced very stable S-parameters. The stability has been confirmed as changes in several numerical kernel parameters haven't introduced noticeable changes in S-parameter values. It was shown that quality of approximation of cylindrical structures and stability of S-parameters is high even with relatively small number of linear segments utilized. Also, it has been shown that edge effect has to be properly taken into the account if a non-strictly planar microstrip-like structure is simulated. The edge effect has been conveniently modeled with application of a single Edge-ing manipulation. Finally, we showed that De-embedding procedure is not mandatory if the proper feeding structure is used for excitation of the EM model.

The high efficiency can be confirmed in a different sense - through the simulation times as they are all relatively short. From a practical side this is very significant result as the simulations were carried out on a standard desktop platform.

The further investigation of this structure will include test

sample fabrication and utilization in measurement of various fluid characteristics.

ACKNOWLEDGMENT

This is part of the Serbian Ministry of Science project 32005, realized by the Department of General Electrical Engineering, School of Electrical Engineering, Belgrade, Serbia.

REFERENCES

[1] B. M. Kolundzija, A. R. Djordjevic, *Electromagnetic Modeling of Composite Metallic and Dielectric Structures, 1st* ed. Norwood, Massachusetts, USA: Artech House, 2002.

[2] Dusan Nesic, "A New Type of Microstrip Resonator for Permittivity Measurement," IcETRAN, Zlatibor, Serbia, 2016

[3] Muhammad Akram Karimi, Muhammad Arsalan, Atif Shamim, "Low Cost and Pipe Conformable Microwave-Based Water Cut Sensor," *IEEE Sensors Journal*, vol. 16, *Issue 21*, pp. *7636-7645*, 2016.

[4] WIPL-D Pro v15, WIPL-D d.o.o, Belgrade 2019. www.wipl-d.com

A New Type of Microwave Coaxial Resonant Permittivity Sensor

Dušan Nešić, *Member, IEEE*

***Abstract*—A new type of microwave coaxial resonant permittivity sensor is introduced. It is constructed using only commercial SMA connectors: one T-junction and one jack-to-jack adapter. The complete open stub is formed of a needle (the central conductor) and the hollow jack-to-jack adapter (the outer conductor). The hollow jack-to-jack adapter is container to be filled with the material under test. The sensor is simulated and preliminary measured in a wide dielectric constant range, from 1 to 4 and from 30 to 80.**

***Index Terms*—Microwave sensor, Permittivity measurement, Coaxial resonator.**

I. Introduction

MICROWAVE sensors are becoming a widespread type of sensors [1-3]. Among them permittivity sensors are important in characterization of microwave substrates and other material (gasses, fluids and solids) [2-4]. Resonant method has some advantage as a type of sensing. The resonant method has the lowest uncertainties for the permittivity [2] and the permittivity is determined from measurements of the resonance frequency.

Coaxial structure is the one of the most convenient and frequently used technique to measure lossy materials at high frequencies (i.e RF and microwave). The most common is coaxial probe method which detects reflected signal (phase and magnitude) from the tested material [3]. It is also applicable for fluids but has some disadvantages. The disadvantages are repetitive calibrations, some deflection for the low permittivity materials and cable stability [3,4]. Another coaxial method to measure the dielectric properties of fluids are coaxial open stub resonators [5-8]. This technique is less sensitive to errors especially for outer electromagnetic influence. In all cases in [5-8] systems are for relatively higher amount of fluids, mainly for water monitoring (diameter of open stub part is around 25 mm and length over 250 mm).

Coaxial resonant permittivity sensor presented in this paper is for a small amount of material. It is constructed using only commercial SMA connectors. The applied SMA connectors are one T-junction and one jack-to-jack adapter, Fig. 1. The sensor is fabricated, simulated and preliminary measured in a wide dielectric constant range.

Fig. 1. SMA: a) Jack-to-plug-to-jack Tee adapter; b) Jack-to-jack adapter

Dušan Nešić is with the Centre of Microelectronic Technologies, Institute of Chemistry, Technology and Metallurgy, University of Belgrade, Njegoseva 12, Belgrade, Serbia, (e-mail: nesicad@nanosys.ihtm.bg.ac.rs).

II. Sensor Construction

The presented sensor is constructed of one SMA jack-to-plug-to-jack Tee adapter and one SMA jack-to-jack adapter, Fig. 1a,b. A needle 11mm long is mounted with silver-epoxy on the Tee adapter as a central conductor of the open stub, Fig. 2a. The inner content of the jack-to-jack adapter, Fig. 1b, is removed. Then, the hollow metal cylinder is mounted on the Tee adapter with the needle, Fig. 2a, and forms final structure in Fig. 2b. Now, the complete open stub is formed of the needle (the central conductor) and the hollow jack-to-jack adapter (the outer conductor). The hollow jack-to-jack adapter is filled with fluid to perform measurement.

a) b)

Fig. 2. a) 11 mm long needle mounted with silver-epoxy on the SMA Tee adapter presented in Fig. 1a ; b) Hollow SMA jack-to-jack adapter, Fig. 1b, mounted on the SMA Tee adapter with a needle presented in a).

III. Simulation and Measurement

Simulation is performed in Program Package WIPL-D Microwave Pro v5.1 for the coaxial structures [9]. The scheme is presented in Fig. 3. Sensing part, needle (the central conductor) and the hollow jack-to-jack adapter (the outer conductor) are presented as the Coaxial open end in Fig. 3.

Simulation results are presented in Fig. 4 for lower real dielectric constant ε_R (1 to 4) and in Fig. 5 for higher ε_R (30 to 80). The lower ε_R includes, for example, gasoline and oils. The higher ε_R includes ethanol, water and their mixture.

Preliminary measurement results are presented in Fig. 6. to Fig. 9. for: air, gasoline, 70% ethanol-water mixture and tap water. According to the simulation results ε_R of gasoline is 1.95 against 2.0 in literature [10]. For 70% ethanol-water mixture is 45 against 39.5 [11]. For the tap water it is 80 against 76 [12]. Obviously, higher ε_R produce higher step in characteristic impedance and high fringing effect.

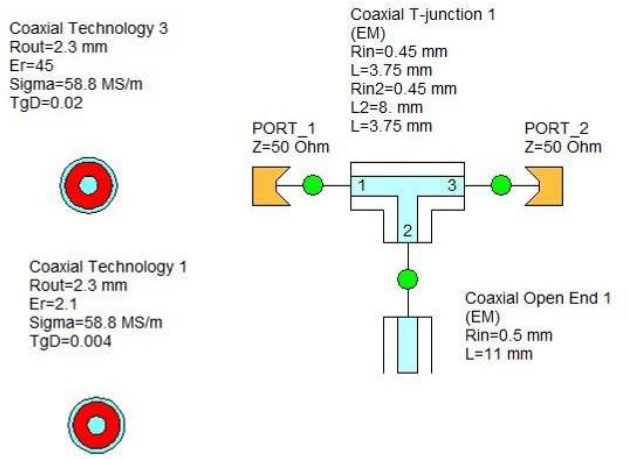


Fig. 3. Scheme of the sensor modeled in Program Package WIPL-D Microwave Prov5.1 for the coaxial structures. Sensing part, needle (the central conductor) and the hollow jack-to-jack adapter (the outer conductor) are presented as the Coaxial open end.

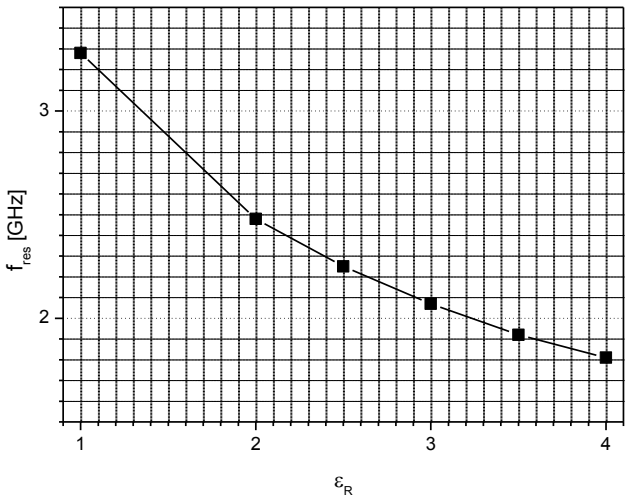


Fig. 4. Resonant frequency f_{res} vs. relative dielectric constant ε_R simulated on model in Fig. 1 for lower dielectric constants.

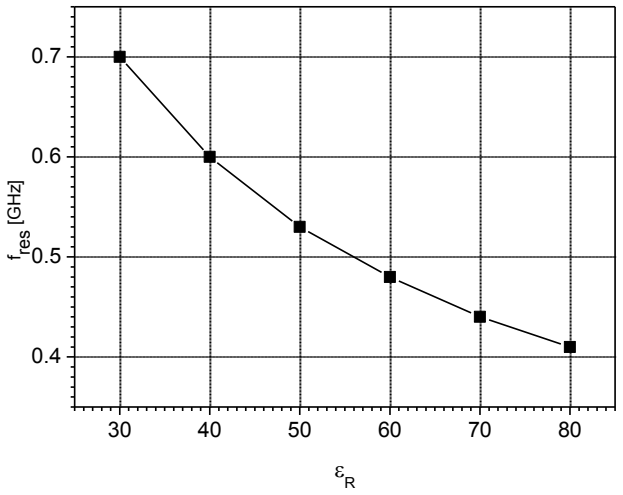


Fig. 5. Resonant frequency f_{res} vs relative dielectric constant ε_R simulated on model in Fig. 1 for higher dielectric constants.

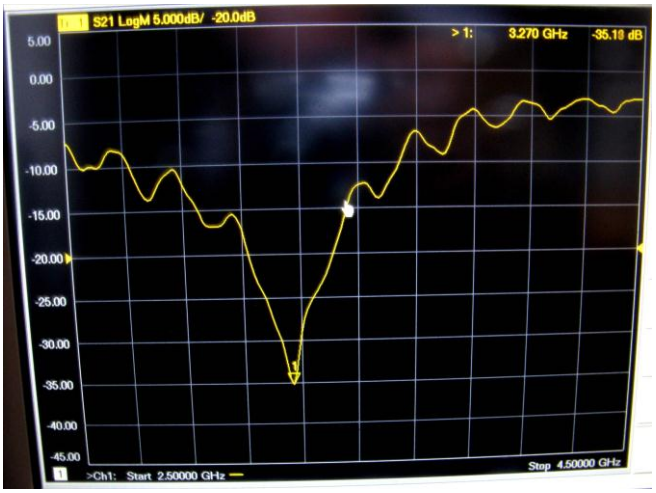


Fig. 6. S_{21}-parameters without tested material (only air).

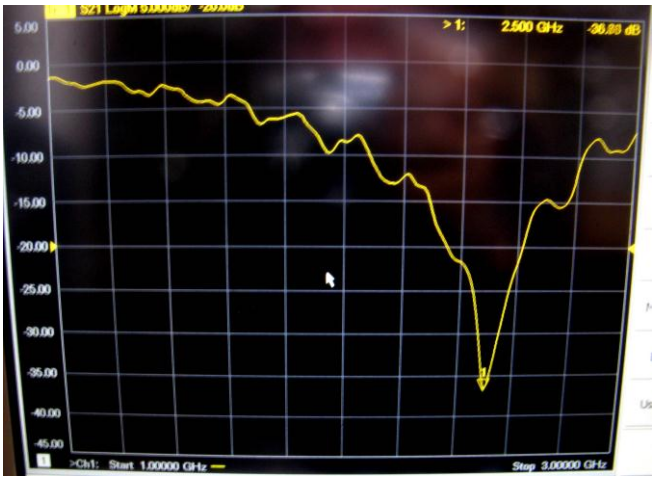


Fig. 7. S_{21}-parameters for gasoline.

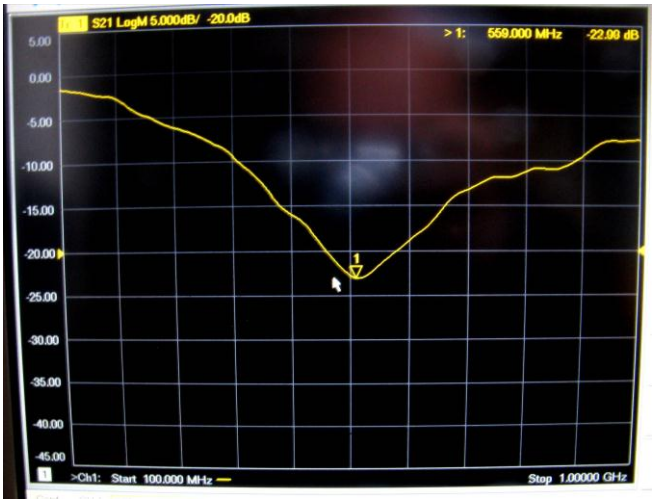


Fig. 8. S_{21}-parameters for 70% ethanol.

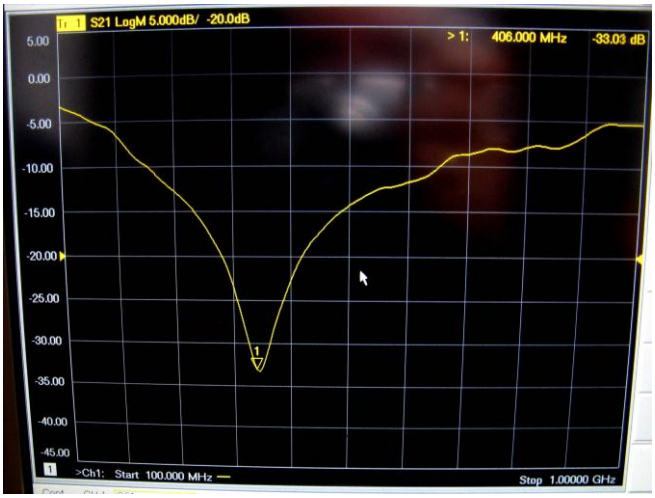


Fig. 9. S_{21}-parameters for tap water.

IV. Option of a Differential Sensor

A general drawback of resonance-based sensors is that permittivity depends on environmental conditions (temperature, moisture) and thus the resonance frequency can shifted by spurious effects. Typical solution to deal with changing environmental factors is through differential measurements e.g. differential sensor [13]. Model of the presented sensor applied as a differential sensor is presented in Fig. 10. The graphs in Fig. 11. present an example of simulated S_{21}-parameters for the differential sensor for two pairs of dielectric constants: 2.1 against 2.1 and 2.1 against 2.2.

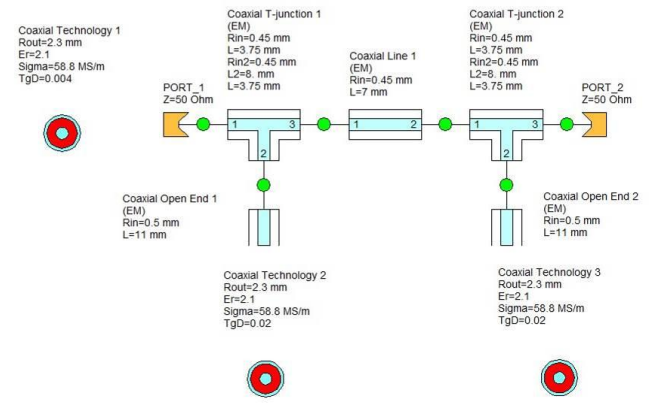


Fig. 10. Scheme of the differential sensor structure simulated in Program Package WIPL-D Microwave Pro v5.1 for the coaxial structures.

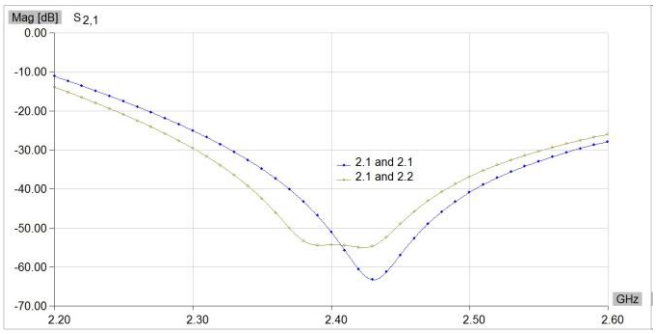


Fig. 11. An example of S_{21}-parameters in simulation of differential sensor for two pairs of dielectric constants: 2.1 against 2.1 and 2.1 against 2.2.

V. Conclusion

This work is introducing a new type of coaxial resonant sensor. It is constructed using only two types of SMA connectors: one SMA jack-to-plug-to-jack Tee adapter and one SMA jack-to-jack adapter. It is easy for construction and useful for smaller amount of fluid under test (cylinder diameter 4.4 mm and to 15 mm high). The sensor is simulated and tested for the wide range of the real dielectric constant, from 1 to 4 and from 30 to 80. High ε_R produces high impedance step and high fringing effect. In future work it needs complete 3D simulation in WIPL-D.

Simulation is also experimentally done for a differential sensor and gives very good results.

Acknowledgment

Author would like to thank L. Novaković, Dr. M. Frantlović and R. Đorđević for help in fabrication and to professor M. Potrebić from the University of Belgrade, School of Electrical Engineering, for her assistance in performing the measurements.

This work was funded by the Serbian Ministry of Education and Science within the project TR 32008.

References

[1] S. Dey, J.K. Saha, and N.C. Karmakar, "Smart Sensing", *IEEE Microwave Magazine*, November 2015, pp. 26-39

[2] J. Baker-Jarvis, M. D. Janezic and D. C. DeGroot, High-Frequency Dielectric Measurements, *IEEE Instrumentation & Measurement Magazine*, 2010, pp. 24 – 31

[3] M. T. Jilani, M. Z. ur Rehman, A. M. Khan, M. T. Khan, S. M. Ali, A Brief Review of Measuring Techniques for Characterization of Dielectric Materials, *International Journal of Information Technology and Electrical Engineering*, Volume 1, Issue 1, 2012, pp. 1-5

[4] *Agilent Basics of Measuring the Dielectric Properties of Materials*, Application Note, www.agilent.com

[5] N.A. Hoog-Antonyuk, W. Olthuis, M.J.J. Mayer, D. Yntema, H. Miedema, A. van den Berg, On-line fingerprinting of fluids using coaxial stub resonator technology, *Sensors and Actuators B*, 163 (2012) pp.90– 96

[6] N.A. Hoog, M.J.J. Mayerb, H. Miedemac, W. Olthuisa, F.B.J. Leferinkd, A. van den Berg, Modeling and simulations of the amplitude–frequency response of transmission line type resonators filled with lossy dielectric fluids, *Sensors and Actuators A*, 216 (2014) pp.147–157

[7] N.A. Hoog, M.J.J. Mayer, H. Miedema, W. Olthuis, A.A. Tomaszewska, A.H. Paulitsch-Fuchs, A. van den Berg, Online monitoring of biofouling using coaxial stub resonator technique, *Sensing and Bio-Sensing Research*, 3 (2015) pp.79–91

[8] N. A. Hoog, M. J.J. Mayer, H. Miedema, W. Olthuis and A. van den Berg, Coaxial Stub Resonator for Online Monitoring Early Stages of Corrosion, *Key Engineering Materials*, Vol. 605 (2014) pp 111-114

[9] *Program Package WIPL-D Microwave Pro v5.1 (coaxial)*, WIPL-D d.o.o, Belgrade 2019. www.wipl-d.com

[10] F. S. Jafari, J. Ahmadi-Shokouh, Reconfigurable microwave SIW sensor based on PBG structure for high accuracy permittivity characterization of industrial liquids, *Sensors and Actuators A* 283 (2018) 386–395

[11] A. Megriche1, A. Belhadj and A. Mgaidi, "Microwave Dielectric Properties of Binary Solvent Water-Alcohol, Alcohol-Alcohol Mixtures at Temperatures Between -35°C and +35°C and Dielectric Relaxation Studies", *Mediterranean Journal of Chemistry* 2012, 1(4), 200-209

[12] Water and Microwaves, *Water Structure and Science*, http://www1.lsbu.ac.uk/water/microwave_water.htm

[13] J. Naqui, Member, C. Damm, A. Wiens, R. Jakoby, L. Su, IEEE, J. Mata-Contreras, and F. Martín, Transmission Lines Loaded With Pairs of Stepped Impedance Resonators: Modeling and Application to Differential Permittivity Measurements, *IEEE Transactions on Microwave Theory and Techniques*, vol. 64, no. 11, 2016, pp. 3864-3877

A Simple Analog Control System for Electromagnetic Levitation Small Object

Nenad Popović[1], Predrag Manojlović[1], and Bojan Virijević[2]

***Abstract*—This paper describes a design of an electromagnetic levitation system for small iron sphere on a small distance from the electromagnet. The control circuit for controlling the distance between the metal ball and the electromagnet consists of a switching electronics with electromagnet and Hall or opto-coupler sensors.**

***Index Terms*—Electromagnetic Levitation, Feedback Control circuit, Opto-Coupler Sensor.**

I. Introduction

A magnetic levitation system is one of the most popular systems in levitation control of small objects.

Electromagnetic levitation is a method by which a metal object (ferromagnetic) floats with the aid of a magnetic field. It uses the electromagnetic force to neutralize the effect of the gravity on the object. In other words, a stable leap of the object in the field of gravity. In real-time, the system responds predictably to unpredictable external influences [1] - [9].

The control of the floating object is usually carried out either via an optical sensor or through a Hall sensor. The Hall sensor solution is mechanically simpler [4].

The first practical application in the field of public transport was applied at the MAGLEV's Shanghai-run railroad in length of 30 kilometers, connecting the Shanghai center and the new Shanghai Pudong airport [7].

II. Phisical Model

The mathematical model of this system is based on the physical model encompasses actuator, electromagnet and metal ball. Electromagnet acts with its electromagnetic attraction force on a metal ball, attracting it against the opposite effect of its weight.

The electromagnetic levitation model can be viewed as a feedback control system [1], the basic block diagram which is shown in Figure 1.

Nenad Popović[1] is with IMTEL Komunikacije a.d., Belgrade, Bulevar Mihaila Pupina 165b, 11070 N. Belgrade, Serbia (e-mail: nenad@ insimtel.com).

Predrag Manojlović[1] is with IMTEL Komunikacije a.d., Belgrade, Bulevar Mihaila Pupina 165b, 11070 N. Belgrade, Serbia (e-mail: pedja@ insimtel.com).

Bojan Virijević[2] is with Military Technical Institute (VTI) Žarkovo, Ratka Resanovica 1, 11030 Belgrade, Serbia

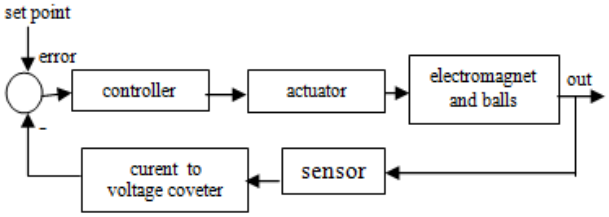


Figure 1. Functional block diagram of the levitation device.

III. Mathematical Dynamic Model

Figure 2 shows the graphic model of the electromagnetic actuator and the metal ball with the forces acting on it [1], [2].

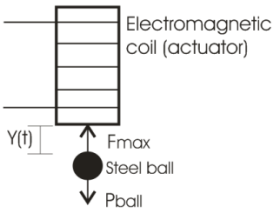


Figure 2. The ball weight is equilibrated by the magnetic force.

The system of mathematical equations for the modelthat is shown in Figure 2, can be written in the following way.
In static equilibrium, it is given by the equation (1) [1],[2].

$$Fmag = Pball = mg \tag{1}$$

m - The steel ball weight
g - 9.81 $[\frac{m}{s^2}]$

Dynamic behavior of the model can be described with equation (2).

$$F_R = F_{mag}^{din} - P_{ball} = m\frac{d^2y(t)}{dt^2} \tag{2}$$

the acceleration is: $a = \frac{d^2y(t)}{dt^2}$ (3)

The dynamic electromagnetic force is given by sum of the static component that compensates for the weight of the steel ball and the variation of the magnetic force, as shown by equation (4).

$$F_R = F_{mag} + \Delta F_{mag} - P_{ball} = m\frac{d^2y(t)}{dt^2} \quad (4)$$

In accordance with equation (1), static equilibrium, the electromagnetic force equals the weight of metal balls ($F_{mag} = F_{ball}$). Thus, these two forces in equation (4) can be mutually canceled so that the result is now given by the following expression (5):

$$F_R = \Delta F_{mag} = m\frac{d^2y(t)}{dt^2} \quad (5)$$

y - distance from electromagnet to ball

We can also calculate the results and observe the energy equilibrium, which in this case includes electrical influences:

$$dW_e = dW_{meh} + dW_t + dW_M \quad (6)$$

where are:
dW_e - total energy
dW_{meh} - mechanical energy
dW_t - heat energy
dW_M - magnetic energy

Through electromagnetic energy we can achieve an attractive electromagnetic force F_M [6]:

$$F_M = -\left[\frac{\partial W_M}{\partial y}\right] i - cost \quad (7)$$

$$W_M = \frac{L(y)\, i^2}{2} \quad (8)$$

where inductance L (y) is given by the following expression [3].

$$L(y) = Lo + L_1 e^{-(\frac{y}{y_0})^2} \quad (9)$$

L_0 - electromagnetic inductance without the presence of balls,
L_1 - electromagnetic inductance with ball presence,

now equation is:

$$-\left[\frac{\partial W_M}{\partial x}\right]_{i-const} = -\frac{\partial L(y)}{\partial y}\frac{i^2}{2} =$$
$$-\frac{\partial}{\partial y}\left(L_0 + L_1 e^{-\left(\frac{y}{yo}\right)^2}\right)\frac{i^2}{2} =$$
$$-L_1\frac{\partial}{\partial y}\left(e^{-\left(\frac{y}{yo}\right)^2}\right)\frac{i^2}{2} =$$
$$-L_1\frac{i^2}{2}\left[-2\left(\frac{y}{yo^2}\right)e^{-\left(\frac{y}{yo}\right)^2}\right] =$$
$$L_1 i^2\left[\left(\frac{y}{yo^2}\right)e^{-\left(\frac{y}{yo}\right)^2}\right] \quad (10)$$

So F_M is the same:

$$F_M = -\left[\frac{\partial W_M}{\partial x}\right] i - cost = L_1 i^2\left[\left(\frac{y}{yo^2}\right)e^{-\left(\frac{y}{yo}\right)^2}\right] \quad (11)$$

The inductance of the electromagnet windings depends on the position of the metal ball. The balloon can have two end positions, one when it touches the electromagnet core and the other when it is away from it.

In the expression (11), y_0 has a constant value so that the variable is part of the expression of the scaling can be written in the form of function $f(y) = ye^{-y^2}$, whose graphic representation [10] is given in Figure 3. Note that value of y = 0 to y = 3 is an interesting part for us.

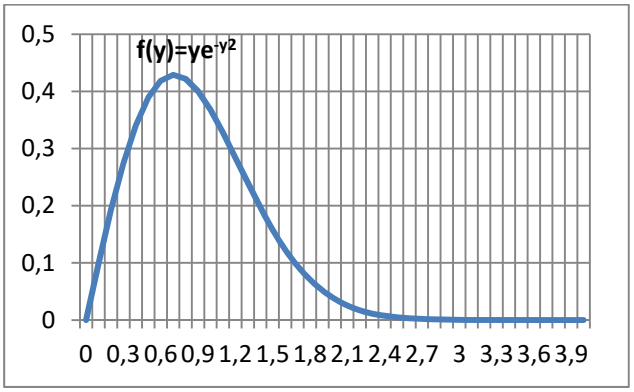


Figure 3. Graphic function $f(y) = ye^{-y^2}$.

By applying Laplace transform to the expression (10) in the time domain, we can calculate its corresponding shape in the frequency domain [1-3].

$$\mathcal{L}\left[L_1 i^2\left(\frac{y}{yo^2}\right)e^{-\left(\frac{y}{yo}\right)^2}\right] \quad (12)$$

For equation,

$$L_1 i^2 \mathcal{L}\left[\left(\frac{y}{yo^2}\right)e^{-\left(\frac{y}{yo}\right)^2}\right] \quad (13)$$

we get the following expression [10]:

$$L_1 i^2 \frac{1}{4}[2 - e^{\frac{s^2}{4}}\sqrt{\pi}\, s\, \mathrm{Erfc}(\tfrac{s}{2})] \quad (14)$$

where,

Erfc[z] – is a complementary error function

IV. REALIZATION

An electromagnet with a corrugated core is attached to an aluminum yoke that is attached to a fiberglass stand, which also has a circular circuit breaker. Electromagnet has the following characteristics:

$$L = 73.2 \text{ mH}, N = 1300 \text{ winds}, r = 17.5\ \Omega.$$

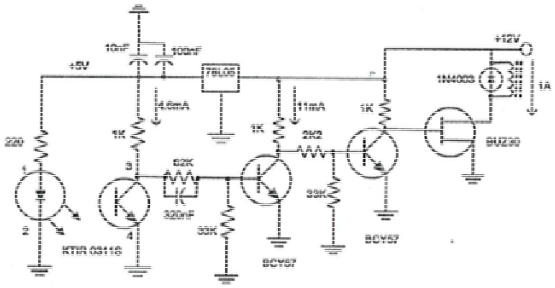


Figure 4. Shematic diagram of realized control circuit for electro-magnetic levitation.

Intesity of coil current is controled by signal of opto-coupler placed nearby sensor plane. Switching electronics operate very efficiently and allow stabile position of the iron rod. Design of controler unit is very simple because we have only two possible states of the opto-coulper output.

Figure 5 shows a photo of a realized model for electromagnetic levitation with switching electronics.

Figure 5. Photo of a realized yoke for electro-magnetic levitation.

V. CONCLUSION

The device function was tested in laboratory conditions and the operating area was established. An area where exist the stable behaviors of the levitating object with assigned weight is guaranteed. The system gain was experimentally checked for defined weights. Levitation of an object is the basic form of electromagnetic levitation creating an opportunity to be familiarized with all the fundamental aspects of control system and electromagnet theory. Implemented design only imposes control in the vertical direction. Electromagnetic levitation system is difficult to implement in 3 dimensions because of its intrinsic nonlinearity.

The force actuator and the sensor can be controled in linearized model by using panoply of coils and sensors instead of single opto-coupler element.

ACKNOWLEDGMENT

This paper is co-financed by the Ministry of Education, Science and Technological Development of the Republic of Serbia within the Technological Development Project TR32052.

REFERENCES

[1] R. R. Gomes, Daniel C.B.V. da Silva, Jose Luiz da Silva Neto, “Electromagnetic Levitation Using MATLAB Real Time Control Toolbox,” 2003 IEEE International Symposium on Industrial Electronics, 9-11 June 2003, Rio de Jeneiro, Brazil.

[2] M.Gon Yoon and Jung Ho Moon, "A Simple Analog Controller for a Magnetic Levitation Kit," IJERT, Vol.5, pp.94-97, Issue 03, March 2016.

[3] R. A. Jose and A S. Gandi, “Development and Control of a Non Linear Magnetic Levitation System,” IJTARME, Vol. 2, Issue-1, pp.62-65, 2013, India.

[4] M. Deshpande and B.L. Mathur, “Modeling of Actuator and Position Sensors for Attraction Typ Magnetic Levitation System,” IJCA, Vol.9, No.2, pp.36-49, Nov. 2010.

[5] L.Baghli and A. Rezzoug, “Levitation magnetique, une approche obet-projet, CETSIS 2008, pp.1-5, 27-29.oct. 2008, Bruxelles, Belgium.

[6] V.Dolga and L. Dolga, “Modeling And Simulation of A Magnetic Levitation System,” Fascicle of Management and Technological Engineering, Vol.VI (XVI), pp.1118-1124, 2007, Romania.

[7] T. Hron, “Model of the Electromagnetic Levitation Device,” Czech Technical University in Prague, Faculty of Electrical Engineering.

[8] H.M. Aguilar,"Magnetic levitation and Newton`s third law," The Physics Teacher, Vol.45, May. 2oo7, pp.278-279.

[9] J. Surutka, "Elektromagnetika," Građevinska knjiga, pp.399-400, Beograd, 1971.

[10] *www.mathworks.com/products/matlab*.

Modelovanje pojačavača snage za LTE sisteme primenom RVTDNN mreže

Jelena Mišić, Milan Čabarkapa, Vera Marković, Đurađ Budimir

***Apstrakt*—U ovom radu predstavljeno je modelovanje izrazito nelinearnih multifrekvencijskih pojačavača snage za širokopojasne LTE sisteme primenom veštačkih neuronskih mreža. Za modelovanje pojačavča korišćena je dinamička neuronska mreža sa kašnjenjem koja ima realne ulazne vrednosti (eng. *Real-Valued Time Delay Neural Network* - RVTDNN). Prilikom modelovanja, ravijen RVTDNN model je optimizovan po broju skrivenih slojeva, neurona u skrivenim slojevima, i dubini memorije ulaznog i izlaznog signala. Metrike koje su korišćene prilikom optimizacije su NMSE (*Normalized Mean-Square Error*) i NAMSE (*Normalized Absolute Mean-Square Error*). Optimalni model ima 12 i 15 neurona u skrivenim slojeva, i dubinu memorije izlaznog signala 3. AM/AM, AM/PM i spektralna karakteristika razvijenog modela na test skupu podataka, koji se razlikovao od trening skupa, imaju visok stepen podudaranja sa odgovarajućim merenim karakteristikama pojačavača.**

***Ključne reči*—ANN; pojačavač snage; LTE; modelovanje komponenti.**

I. UVOD

BROJ mobilnih korisnika je u stalnom porastu poslednjih godina jer je mobilnost postala osobina koja se podrazumeva na polju govorne komunikacije. Takođe, od mobilnih mreža se zahteva da prenesu velike količine podataka za relativno kratko vreme. Trenutna tehnologija koja se koristi u mobilnim mrežama je 4G tehnologija, odnosno LTE (*Long Term Evolution*), koja je dizajnirana da zadovolji zahteve kosinika u pogledu brzine prenosa i količine podataka.

LTE tehnologija podržava fleksibilne opsege učestanosti zahvaljujući primeni OFDMA (*Orthogonal Frequency Division Mulitple Access)* i SC- FDMA (*Single Carrier Frequency Division Multiple Access)* pristupnim šemama. LTE na *downlink*-u koristi OFDMA tehniku višestrukog pristupa, dok na *uplink*-u koristi SC-FDMA.

Jedna od najbitnih komponenti na predajoj i prijemnoj strani LTE sistema je pojačavač snage (eng. *Power amplifier* - PA) osnosno malo-šumni pojačavač (eng. *Low noise amplifier* - LNA). U modernim mobilnim sistemima, kao što su sistemi bazirani na LTE standardu, koriste se napredne modulacione tehnike, kod kojih pojačavač snage radi blizu oblasti saturacije kako bi se postigla maksimalna izlazna snaga. Međutim, u ovoj oblasti prenosna karakteristika pojačavača više nije linearna, što u slučaju primene signala kao što je OFDM signal može dovesti do distorzije.

Jedna od najrasprostranjenihih metoda za linearizaciju pojačavača je DPD (*Digital pre-distortion*) koja se zasniva na uvođenju bloka čija je prenosna krakteristika jednaka inverznoj prenosnoj karakteristici pojačavača koji se koristi. Kao što se može zaključiti, za implementaciju DPD modela neophodno je razviti kako model samog pojačavača tako i njegov inverzni model. Postoji mnogo tehnika za modelovanje mikrotalasnih komponenti. Jedna od trenutno najzastupljnijih tehnika su veštačke neuronske mreže (eng. *Artifical Neural Networks* - *ANNs*) [1]. Veštačke neuronske mreže se primenjuju za modelovanje različitih RF komponenti među kojima su i pojačavači snage [2-5]. Postoji veliki broj tipova ovih mreža i kod modelovanja komponenti je najbitnije odabrati najpogodniji model mreže i prilagoditi njegove parametre komponenti koja se modeluje.

Zbog značaja izbora najpogodnijeg ANN modela, u ovom radu predstavljena je detaljna analiza modelovanja jednog pojačavača, kao i razvijeni ANN model.

Ostatak rada je organizovan na sledeći način. U Sekciji II, predstavljene su veštačke neuronske mreže. U Sekciji III, data je analiza modelovanja i konačni model pojačavača, kao i njegove karakteristike. U Sekciji IV, dati su zaključci i smernice za budući rad.

II. VEŠTAČKE NEURONSKE MREŽE

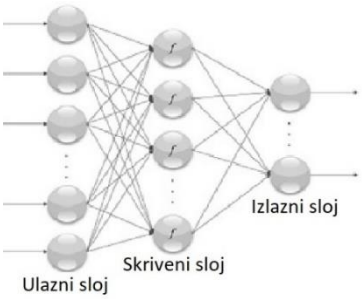
ANN mreža se može definisati kao paralelna matematička struktura sastavljena od određenog broja jednostavnih, međusobno povezanih elemenata – neurona. Neuroni u ANN su tipično organizovani u slojevima. Jedna ANN mreža može imati različit broj slojeva. Prvi sloj u ANN mreži se naziva ulazni sloj (eng. *input layer*), posle njega sledi jedan ili više skrivenih slojeva (eng. *hidden layers*), nakon kojih je izlazni sloj (eng. *output layer*). Broj neurona u ulaznom sloju jednak je broju nezavisnih ulaznih parametara, dok je broj izlaznih neurona jednak broju parametara koje treba modelovati. Najčešće korišćeni tip ANN mreža je *feedforward* ANN mreža kod koje se signal kreće u smeru od ulaznog sloja ka izlaznom sloju bez ikakvih povratnih signala kroz samu mrežu. Uprošćeni prikaz *feedforward* ANN mreže koja se sastoji od 3 sloja dat je na Sl. 1.

Jelena Mišić – Elektronski fakultet, Univerzitet u Nišu, Aleksandra Medvedeva 14, 18106 Niš, Srbija (e-mail: ms.jelena.misic@gmai.com).

Milan Čabarkapa – Elektrotehnički fakultet, Univerzitet u Beogradu, Bulevar Kralja Aleksandra 73, 11020 Beograd, Srbija (e-mail: cabmilan@etf.bg.ac.rs).

Vera Marković – Elektronski fakultet, Univerzitet u Nišu, Aleksandra Medvedeva 14, 18106 Niš, Srbija (vera.markovic@elfak.ni.ac.rs).

Đurađ Budimir – Wireless Communications Research Group, University of Westminster, London, UK (d.budimir@wmin.ac.uk).



Sl. 1. Uprošćeni prikaz structure troslojne ANN mreže.

Kao što se sa Sl. 1 može videti, neuroni jednog sloja su povezani sa svim neuronima u narednom sloju. Svaki od neurona ANN mreže se karakteriše svojom prenosnom funkcijom i vrednošću bijasa, dok se svaka veza između neurona karakteriše težinskim faktorom. Neuroni u jednom sloju imaju istu prenosnu funkciju. Prenosna (aktivaciona) funkcija predstavlja funkciju koja povezuje ulazni i izlazni signal neurona. Da bi se mreža trenirala da simulira izlaze sa potrebnom preciznošću, pragovi funkcija aktiviranja neurona i težine neuronske veze se optimizuju tokom procesa obuke mreže.

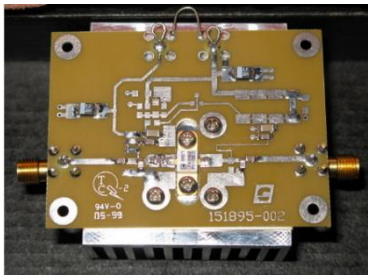
ANN mreže se zbog svoje sposobnosti visoke aproksimacije nelinearnih funkcija, uspešno primenjuju kod modelovanja RF i mikrotalasnih uređaja/kola [6-9]. Teorema uopštene aproksimacije (eng. *Universal Approximation Theorem* - UAT) [10] pokazuje da *feedforward* ANN mreža sa jednim skrivenim slojem i promenljivim, monotono rastućim, kontinualnim aktivacionim funkcijama može da aproksimira bilo koju nelinearnu funkciju sa željenom verovatnoćom greške. Dakle, prema UAT teoremi, struktura sa jednim skrivenim slojem (eng. *Single Hidden Layer* - SHL) je dovoljna za aproksimaciju uniformnog nelinearnog sistema.

III. MODELOVANJE POJAČAVAČA PRIMENOM ANN

Ono što karakteriše pojačavače osim njihove osnovne funckije pojačavanja signala, je i tzv. memorijski efekat koji se oglada u činjenici da izlaz pojačavača zavisi kako od trenutnog ulaza u pojačavač tako i od jedne ili više prethodnih vrednosti ulaznog i izlaznog signala pojačavača. Memorijski efekat je karakteristika koja je vrlo bitna prilikom modelovanja pojačavača. Ukoliko se prilikom modelovanja ne uzme u obzir i memorijski efekat može doći do greške u predikciji izlaza modela, što u krajnjoj liniji dovodi do greške i u DPD modelu, imajući kao krajnji cilj loš efekat linearizacije.

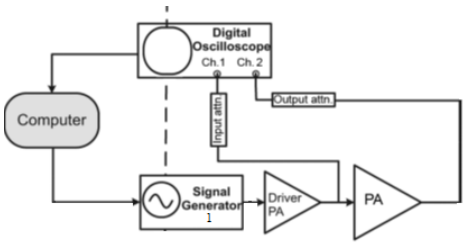
A. Prikupljanje podataka za modelovanje

U okviru ovog rada, mi smo izvršili analizu i modelovanje pojačavača sa oznakom CFH-2163-P3 (*datasheet* je u [8]), koji je prikazan na Sl. 2.



Sl. 2. Pojačavač CFH-2163 P3.

Realni pojačavač je u ovom radu predstavljen svojim matematičkim modelom. Za dobijanje funkcije koja predstavlja matematički model koristili smo laboratorijsku postavku prikazanu na Sl. 3.

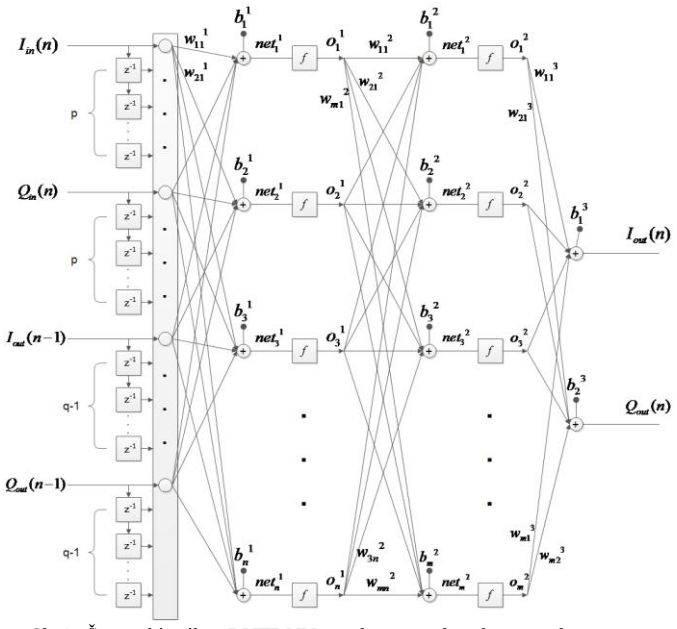


Sl. 3. Postavka za merenje gde PA predstavlja DUT (eng. *Device Under Test*).

Kao što se vidi na prethodnoj slici, u laboratorijskim uslovima, korišćen je generator signala. Signal generator generiše WCDMA signal sa frekvencijom nosioca na 2.14 GHz, a demodulacija je podešena da prihvata WCDMA signal na 2.14 GHz. Frekvencija i vrsta signala su odabrane tako da odgovaraju LTE specifikacijama. Signal WCDMA je korišćen kao *proof of concept* jer u trenutku razvijanja modela OFDM signal nismo imali u Matlab formatu.

B. Odabir ANN modela

Kako bi se uzeo u obzir prethodno pomenuti memorijski efekat, za modelovanje pojačavača korišćena je dinamička neuronska mreža sa kašnjenjem koja ima realne ulazne vrednosti (eng. *Real-Valued Time Delay Neural Network* - RVTDNN). Ustanovljeno je da je ova ANN topologija efikasna u modelovanju nelinearnog PA [4]. Nelinearni efekti praćeni izraženim memorijskim efektima su mnogo izraženiji u odnosu na rad [4], jer se koristi tri puta širi spektar da bi se video potencijal primene ovo metodologije u izrazito nelinearnim širokopojasnim multifrekvencijskim sistemima. Ovde se pod ulazom sa realnim vrednostima podrazumeva kompleksni ulazni signal (širokopojasni signal) koji se sastoji od dve komponente, od kojih jedna predstavlja vrednost realnog dela signala, a druga vrednost imaginarnog dela signala. Na Sl. 4 dat je šematski prikaz RVTDNN topologije sa dva skrivena sloja, sa dubinom memorije p i q. Na Sl. 4, RVTDNN model ima četiri ulaza: dve komponente ulaznog signala (I_{in}, Q_{in}) i dve komponente zakašnjenog izlaznog signala (I_{out}, Q_{out}). Predložena RVTDNN mreža se zasniva na *feedforward* strukturi ANN mreže [11], sa dodatkom četiri linije za kašnjenje (eng. *Tapped Delay Line* - TDL) u ulaznom nivou, na svakom ulazu po jedan TDL.



Sl. 4. Šematski prikaz RVTDNN topologije sa dva skrivena sloja.

Na Sl. 4, p označava dubinu memorije ulaznog signala, a q dubinu memorije izlaznog signala; broj TDL linija za kašnjenje kod izlaznog signala na ulazu mreže jednak je (q-1), jer je prvi zakašnjeni signal prikazan kao direktno povezan na ulazni sloj. Shodno tome, TDL struktura ugrađena u RVTDNN obezbeđuje modelovanje kratkoročnih memorijskih efekata koje pokazuje PA. Što se dugoročnih memorijskih efekata tiče, one su ugrađene u samu topologiju RVTDNN, i to kroz učenje pod nadzorom (eng. *supervised learning*), koje predstavlja jedan od načina obučavanja ANN mreže. Ovakav način modelovanja dugoročne memorije može da pomogne kod praćenja sporih dinamičkih promena nelinearnih karakteristika PA tokom vremena.

Prikazana RVTDNN može biti predstavljena ulaznim vektorom $2(p + q + 1)$, gde p i q predstavljaju dubinu memorije u vidu prethodnih odbiraka ulaznih komponenata (I_{in}, Q_{in}) i izlaznih komponenata (I_{out}, Q_{out}) u datom trenutku. Ulazni vektor u datom trenutku obuhvata realne vrednosti trenutnog i prethodnih odbiraka obe komponente ulaznog i izlaznog signala, i može biti pretstavljen kao:

$$X_{in} = [\ I_{in}(n), I_{in}(n-1), \dots, I_{in}(n-p), \quad (1)$$
$$Q_{in}(n), Q_{in}(n-1), \dots, Q_{in}(n-p),$$
$$I_{out}(n-1), \dots, I_{out}(n-q),$$
$$Q_{out}(n-1), \dots, Q_{out}(n-q)].$$

C. Metrike korišćene u analizi

Prilikom modelovanja pojačavača, pored analize dubine memorije, vršena je analiza i u pogledu broja skrivenih slojeva i broja neurona u skrivenim slojevima. Metrike koje su korišćene prilikom analize su NMSE (*Normalized Mean-Square Error*) i NAMSE (*Normalized Absolute Mean-Square Error*). Naime, NMSE metrika daje dobar uvid u *in-band* odziv, ali ne pokazuje performanse u susednim kanalima, gde je lociran najveći deo nelinearne distorzije koji uzrokuje smetnje u drugim komunikacionim kanalima. NMSE metrika se definiše na sledeći način:

$$NMSE_{dB} = 10\log_{10}\left(\frac{\sum_{n=1}^{K}|y_{meas}(n)-y_{est}(n)|^2}{\sum_{n=1}^{K}|y_{meas}(n)|^2}\right), \quad (1)$$

gde su $y_{meas}(n)$ i $y_{est}(n)$ izmereni i predviđeni (modelovani) izlazni signali, respektivno, K je broj odbiraka izlaznog signala, BW predstavlja normalizovan opseg, a f je normalizovana frekvencija.

NAMSE metika služi za robusniju procenu performansi i definiše se kao:

$$NAMSE_{dB} = 10\log_{10}\left(mean_{f\epsilon[f\ min, f\ max]}\left(\frac{|Y_{meas}(f)-Y_{est}(f)|}{|Y_{meas}(f)|}\right)\right), \quad (2)$$

gde su $Y_{meas}(f)$ i $Y_{est}(f)$ spektralne gustine snage signala na izlazu DUT-a i izlazu modela, respektivno. Frekvencijski opseg $[f\ min, f\ max]$ se može prilagoditi tako da uzima u obzir celokupan frekvencijski opseg, ili samo specifične opsege koji odgovaraju intermodulacionim produktima trećeg i petog reda.

NAMSE razmatra grešku u frekvencijskom domenu između izmerenog i modelovanog spektra. Izbegavanjem integracije, NAMSE predstavlja više mikroskopsku metriku koja daje jednak značaj svim spektralnim regionima, nezavisno od nivoa snage. Prema tome, očekuje se da će ova metrika da da vrlo preciznu kvantifikaciju performansi modela ponašanja.

D. Obučavanje mreže

Tokom obučavanja (treniranja) ANN mreže koristili smo postupak kros-validacije (eng. *cross-validation*). U sklopu ovog postupka postoje dve faze, faza treniranja (eng. *training phase*) i faza testiranja (eng. *test phase*). Treniranje predložene mreže izvrešeno je korišćenjem različitih segmenata izmerenog ulaznog i izlaznog signala, kako bi se obezbedila dobra sposobnost generalizacije ANN modela. Trening skup sastojao se od 9.000 odbiraka, a test skup od 26.000 odbiraka. Broj odabiraka u test skupu je s namerom izabran skoro tri puta veći od onog koji je korišćen tokom treniranja, kako bi se procenila sposobnost generalizacije ANN modela. Sposobnost generalizacije ANN mreže definiše se kao sposobnost mreže da predvidi tačan izlaz za ulazni signal koji se razlikuje od ulaznih signala koji su korišćeni prilikom treninga ANN mreže. Bitno je napomenuti, da su odabirci u trening i test skupovima bili različiti. Test faza ima za cilj proveravanje sposobnosti ANN modela da predvidi izlaz za ulazne signale koji nisu korišćeni prilikom obučavanja mreže; na ovaj način se proverava sposobnost generalizacije ANN modela. Maksimalno broj epoha prilikom treniranja je bio postavljen na 10,000, ali je pri obuci svih razvijenih RVTDNN modela broj epoha bio znatno manji od maksimalnog, odnosno zadate performanse modela su bile postignute sa mnogo manje epoha. Nivo učenja je bio postavljen na 0.1. Ova vrednost izabrana je na osnovu naših

prethodnih istraživanja, jer se pokazalo da veće vrednosti novoa učenja dovode do smanjenja generalizacije mreže.

E. Optimalna ANN struktura

Kao što je već napomenuto, kako bi se odredila optimalna ANN struktura za izabrani PA izvršena je analiza nekoliko ANN parametara, tačnije, optimizacija ANN strukture je vršena u pogledu četiri parametra, broja skrivenih slojeva, broja neurona u skrivenim slojevima, dubine memorije ulaznog signala, i dubine memorije izlaznog signala. Metrike koje su korišćene prilikom odabira optimalnih vrednosti pomenutih četiri parametara, su gore definisane metrike, NMSE i NAMSE. Prilikom odabira optimalnog modela, pored ovih metrika upoređivane su i sledeće karakteristike: amplitudska karakteristika, fazna karakteristika, i *Error Spectrum* karakteristika, koje se uobičajno koriste prilikom procene podobnosti ANN modela pojačavača.

Prilikom modelovanja razvijeno je više RVTDNN modela, koji se mogu svrstati u dve kategorije: RVTDNN modeli sa jednim skrivenim slojem, i RVTDNN modeli sa dva skrivena sloja. Broj skivenih slojeva nije povećavan preko 2 jer shodno UAT teoremi, ANN sa jednim skrivenim slojem i promenljivim, monotono rastućim, kontinualnim aktivacionim funkcijama može da aproksimira bilo koju nelinearnu funkciju sa željenom verovatnoćom greške, tako da smo povećanje broja skrivenih slojeva preko 2 smatrali neoptimalnim. Takođe, srodne studije pokazuju da je u najvećem broju slučajeva za modelovanje pojačavača potrebno realizovati ANN mrežu sa maksimalno dva skrivena sloja. Prema rezultatima, mreža sa dva skrivena sloja je u našem slučaju dala bolje rezultate, tako da je RVTDNN struktura sa dva skrivena sloja uzeta kao optimalna. U pogledu broja neurona u skivenim slojevima, najbolje performanse su dobijene kada je broj neurona u prvom skrivenom sloju bio 12, a u drugom 15, kao što je prikazano u Tabeli I.

Analiza je pokazala da je optimalna dubina memorije ulaznog signala jednaka nuli. Odnosno, dovođenje zakasnelih ulaznih signala zajedno sa trenutnim ulaznim signalom nije rezultiralo poboljšanjem performansi ANN modela.

Što se tiče dubine memorije izlaznog signala, rezultati su pokazali da je optimalno rešenje dubina memorije 3, odnosno dovođenjem tri zakasnela izlazna signala na ulaz mreže dobijeni su najbolji rezultati u pogledu NMSE i NAMSE performansi, i AM/AM, AM/PM i *Error Spectrum* karakteristika. Rezultati za različite dubine memorije izlaznog signala prikazani su u Tabeli II.

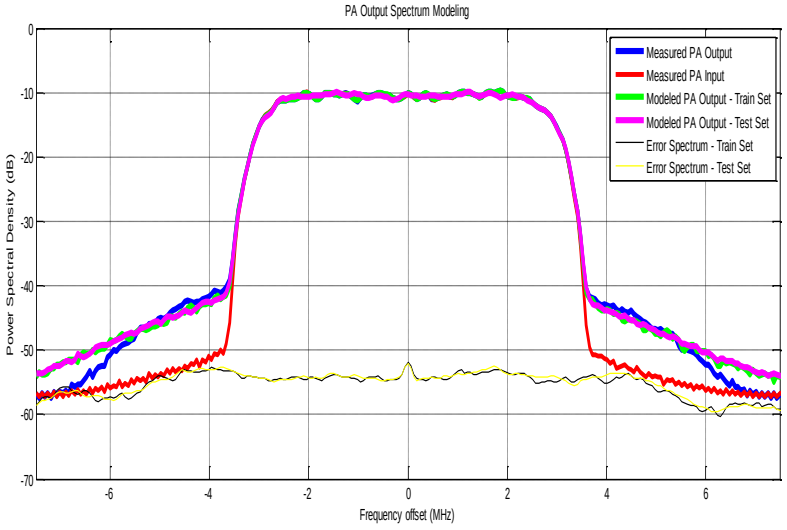
TABELA I
BROJ NEURONA U SKRIVENIM SLOJEVIMA

Broj neurona	*NMSE Train* (dB)	*NMSE Test* (dB)	*NAMSE Train* (dB)	*NAMSE Test* (dB)
I_HL=3, II_HL=3	-34.7318	-34.6660	-20.0209	-19.6848
I_HL=5, II_HL=3	-33.8056	-33.9593	-20.0906	20.1189
I_HL=7, II_HL=3	-37.7937	-37.7562	-21.7221	-21.3805
I_HL=5, II_HL=12	-38.1693	-38.1740	-22.0861	-21.9001
I_HL=12, II_HL=15	-40.5347	-40.3897	-22.6781	-22.2835

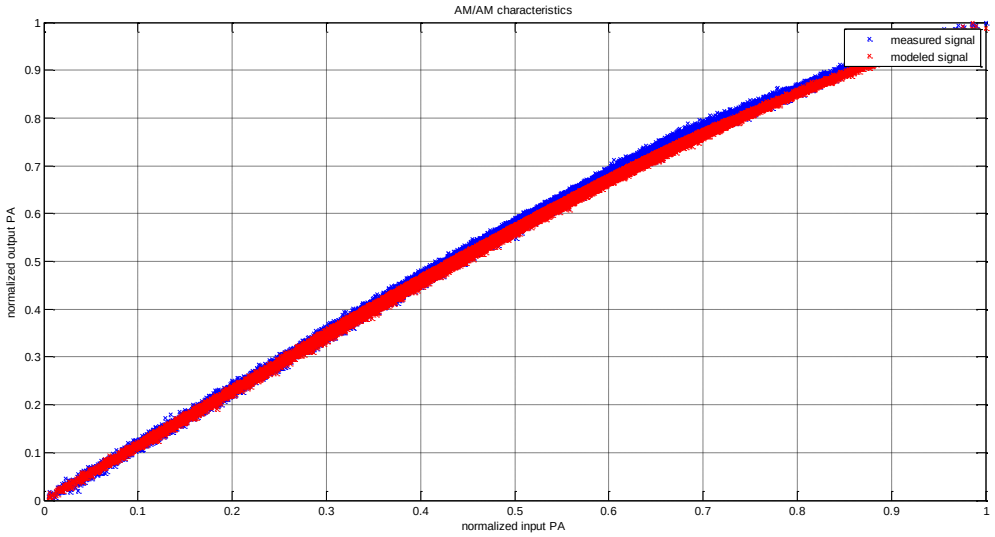
TABELA II
DUBINA MEMERIJE IZLAZNOG SIGNALA

Dubina memorije	*NMSE Train* (dB)	*NMSE Test* (dB)	*NAMSE Train* (dB)	*NAMSE Test* (dB)
$q = 0$	-37.4279	-37.4267	-20.5748	-20.5171
$q = 2$	-39.9849	-39.8766	-22.5726	-22.2998
$q = 3$	-40.1665	-40.0677	-22.4296	-22.2632
$q = 5$	-38.6017	-38.4715	-22.0837	-21.7430

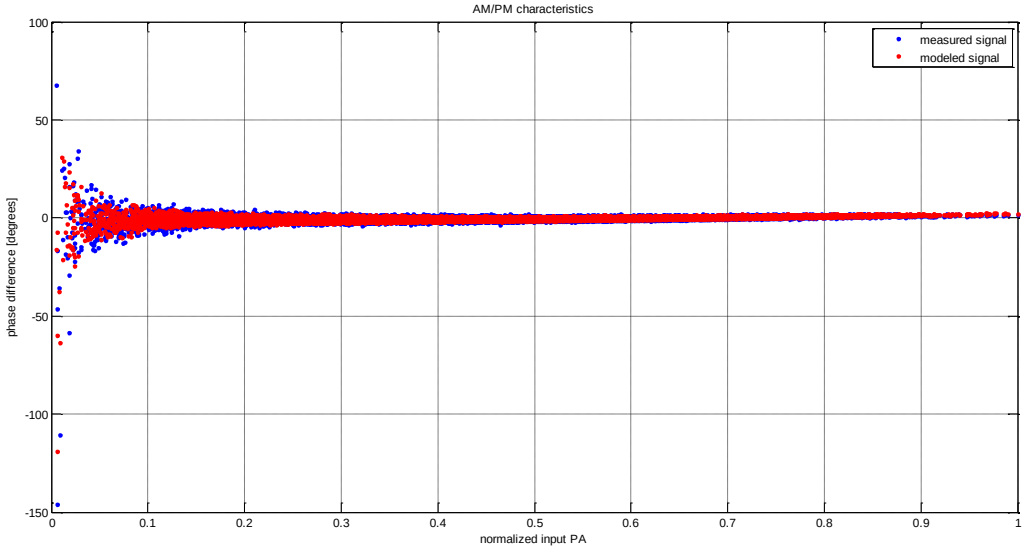
Uporedni izgled spektra signala za mereni i modelovani izlaz za trening i test skup za optimalni RVTDNN model prikazan je na Sl. 5. Na Sl. 6-9 date su AM/AM, AM/PM i *Error Spectrum* (razlika između izmerene i modelovane spektralne karakteristike) karakteristike.



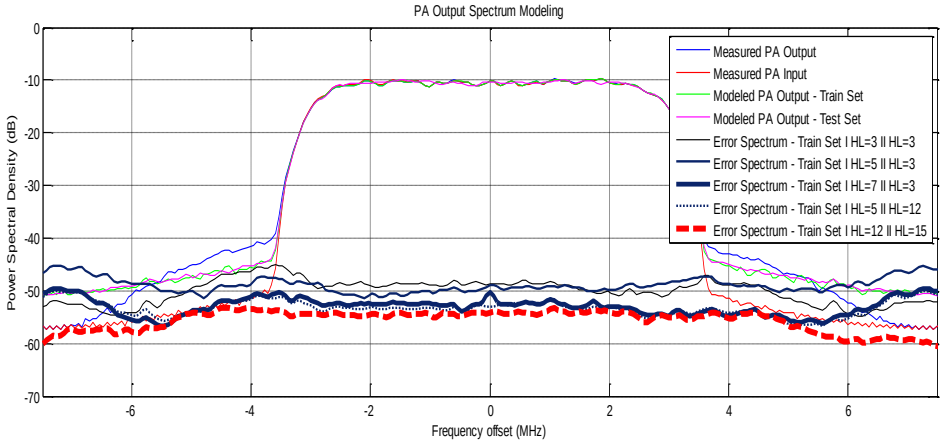
Sl. 5. Izgled spektra signala na izlazu iz RVTDNN mreže za q=3 (izmereni spektri su nornalizovani).



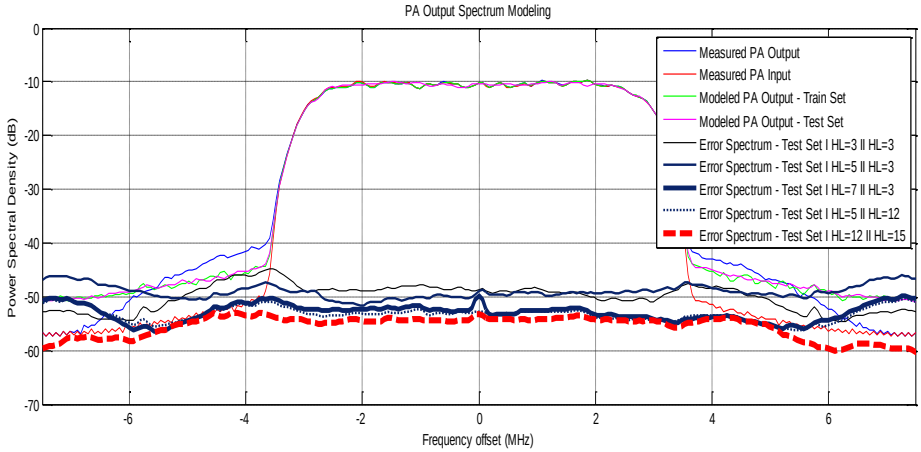
Sl. 6. AM/AM karakteristika za q=3.



Sl. 7. AM/PM karakteristika za q=3.



Sl. 8. Normalizovani spektri signala i *Error Spectrum* karakteristika za trening skup za različite brojeve neurona u skrivenim slojevima.



(b)

Sl. 9. Normalizovani spektri signala i *Error Spectrum* karakteristika za test skup za različite brojeve neurona u skrivenim slojevima.

Kao što se na prethodnim slikama može videti, razvijen RVTDNN model ima dobre performanse po svim razmatranim kriterijumina.

IV. Zaključak

U ovom radu prikazan je postupak modelovanja pojačavača snage CFH-2163-P3 primenom veštačkih neuronskih mreža i sa signalom koji je tri puta širi u odnosu na prethodna istraživanja. Za modelovanje pojačavača korišćena je RVTDNN mreža sa dva skrivena sloja i sigmoidalnom prenosnom funkcijom. Razvijeni model je optimizovan po broju neurona u skrivenim slojevima i dubini memorije ulaznog i izlaznog signala; optimlan broj neurona u skrivenim slojevima je bio 12 i 15, dubina memorije ulaznog signal je jednaka nuli, a dubina memorije izlaznog signal je jednaka 3. Performanse razvijenog modela u pogledu NMSE i NAMSE su ispod -40 dB i -20 dB, respektivno, što predstavlja dobar rezultat. Poklapanje AM/AM, AM/PM i spektara modelovanog i merenog signala je dobro, što potvrđuje podobnost razvijenog modela pojačavača.

U okviru našeg daljeg rada biće razvijem inverzni model za pojačavač snage CFH-2163-P3 primenom veštačkih neuronskih mreža, zarad razvoja DPD modela za dati izrazito nelinearni pojačavač i izrazito širokopojasne sisteme.

Literatura

[1] Q. J. Zhang, K. C. Gupta, "Neural Networks for RF and Microwave Design," MA: Artech House, Boston, 2000.

[2] H. Kabir, L. Zhang, M. Yu, P.H. Aaen, J. Wood, Q.J. Zhang, "Smart modelling of microwave devices," Microwave Magazine, IEEE 11, pp. 105–108, 2010

[3] F. Ortega-Zamorano, J.M. Jerez, D. Urda Munoz, R.M. Luque-Baena, L. Franco, "Efficient Implementation of the Backpropagation Algorithm in FPGAs and Microcontrollers, Neural Networks and Learning Systems," IEEE Transactions on 99, 2015

[4] M. Vaskovic, D. Budimir, "Compensation of nonlinear distortion in RF power amplifiers for LTE applications," Microwave and Optical Technology Letters 56, pp. 1910–1913, 2014

[5] N. Gunavathi, D. Sriramkumar, "Estimation of resonant frequency and bandwidth of compact unilateral coplanar waveguide-fed flag shaped monopole antennas using artificial neural network," Microwave and Optical Technology Letters 57, pp. 337-342, 2015

[6] M. Čabarkapa, N. Nešković, A. Nešković, Đ. Budimir, Adaptive nonlinearity compensation technique for 4G wireless transmitters, *Electron. Lett.*, Vol. 48, No. 20, pp. 1308 - 1309, Dec, 2012.

[7] M. Čabarkapa, N. Nešković, Đ. Budimir, A Generalized 2-D Linearity Enhancement Architecture for Concurrent Dual-Band Wireless Transmitters, *IEEE Trans. Microw. Theory and Tech*., Vol. 61, No. 12, pp. 4579 - 4590, Dec, 2013

[8] M. Cabarkapa, *Digital predistortion of RF amplifiers using baseband injection for mobile broadband communications*, PhD Thesis, University of Westminster, London, UK, 2014.

[9] M. Čabarkapa, N. Nešković, M. Prokin, Đ. Budimir, Modelovanje ponašanja pojačavača snage i digitalna predistorzija za 4G bežične komunikacione sisteme, *ETRAN 2016, Jun, 2016*

[10] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Math. Control, Signals, Syst.*, vol. 2, no. 4, pp. 303–314, Feb. 1989.

[11] I. W. Sandberg, Nonlinear Dynamical Systems: Feedforward Neural Network Perspectives. New York: Wiley, 2001.

Abstract

In this paper, a power amplifier intended for LTE systems is modeled using artificial neural networks. The Real-Valued Time Delay Neural Network – RVTDNN is used in modeling. The RVTDNN model is optimized regarding the number of hidden layers, the number of neurons in the hidden layers, memory depth of the input signal, and memory depth of the output signal. The measures used in model evaluations are NMSE (Normalized Mean-Square Error) and NAMSE (Normalized Absolute Mean-Square Error). The optimized RVTDNN model has 12 and 15 neurons in the first and second hidden layers, respectively, and memory depth of the input and output signals are equal to 0 and 3, respectively. The AM/AM, AM/PM and error spectrum characteristics of the optimized model show excellent agreement with the measured characteristics on the test dataset, which differs from the training dataset.

Modeling of PA for LTE Systems Based on RVTDNN Neural Networks

Jelena Misic, Milan Cabarkapa, Vera Markovic, Djuradj Budimir

Influence of Mechanical Activation on Electrical Properties of Ceramic Materials in VHF Band

Nina Obradović and Antonije Đorđević

***Abstract*—Mechanical activation is commonly used as a pre-sintering process in order to enhance the reactivity of materials, reduce the particle size, increase diffusion rates, accelerate the reaction, and lower the sintering temperature. The mechanical activation can affect the final electrical and mechanical characteristics. In this paper we consider the influence of the mechanical activation on the permittivity and the loss tangent. We outline methods for evaluation of these parameters, with emphasis on our coaxial-chamber technique for measurements in the VHF band.**

***Index Terms*—Ceramic materials; mechanical activation; sintering; electrical properties; numerical modeling; measurements of dielectric parameters.**

I. Introduction

Ceramic materials are solid materials obtained by sintering of appropriate mixtures of powders. "Ceramic" is derived from the Greek word "keramos", which means potter's clay or pottery [1]. Ceramic materials have many technical applications. Advanced materials, such as alumina, aluminum nitride, borides, zirconia, silicon carbide, silicon nitride, silicon- and titania-based materials offer a high-performance and economic alternative to conventional materials, such as glass, metals, and plastics.

The sintering is one of the most important phases in the production process of ceramic materials. Specific predetermined properties of ceramics are required along with cost-effectiveness of the manufacturing process [2]. In order to obtain materials with the desired properties during the sintering process, one should simultaneously take into account several factors: the size, shape, and structure of the particles, particle-size distribution, particle packing, the purity of the starting mixtures, the density of compacts (samples obtained after consolidation, prior to the sintering process), the appropriate sintering regime (temperature and sintering time, heating/cooling rate), the atmosphere in which the process is performed, the existence of impurities and/or additives, etc. It has been empirically determined that a decrease in the particle size of reactants leads to an increase of the isothermal process constant rate without changes in the activation energy of the process [3]. The increase of the specific surface area that is induced by the mechanical activation plays an important role. Extremely small particles require application of a higher pressure during the consolidation process. Due to the non-homogeneous package and gas retention, samples get stratified and the sintering leads to formation of cracks. The presence of large particles causes an excessive grain growth during the sintering. By adjusting the conditions of the pre-treatment, we can prepare the starting powder with the predefined optimal particle-size distribution.

Basically, the sintering processes can be divided into two types: the solid-state sintering and the liquid-phase sintering. The solid-state sintering occurs when the powder compact is densified wholly in the solid state at the sintering temperature. The liquid-phase sintering occurs when a liquid phase is present in the powder compact during the sintering [4].

Ceramic materials are used in electrical engineering as insulators, chip carriers, dielectrics for capacitors, in fabrication of multilayer ceramic (MLC) devices, in particular LTCC (low-temperature co-fired ceramics), for various optical devices, etc. Various properties of these materials are important, depending on the application. The relevant electromagnetic properties of the sintered materials include dielectric properties (permittivity, losses, and breakdown), insulating/conductive properties, magnetic properties (permeability, losses, nonlinearities, and Curie temperature), and possible piezoelectric properties. The thermal conductivity is important for heat transfer in semiconductor devices. Adherence of metals and semiconductors, including dilatation and compatibility with adhered materials, is essential for fabrication processes, etc.

In this paper, we consider only nonmagnetic materials and are focused on the electrical properties of the ceramics. We assume that the material is linear. We consider the complex relative permittivity and quantities related to it (the loss tangent and the conductivity), predominantly in the VHF (Very High Frequency) band, which extends from 30 MHz to 300 MHz. Fig. 1 presents the position of the VHF band in the spectrum. The data include free-space wavelength (λ), the frequency (f), the energy of a quantum, and the designation.

We present our measurement technique for the relative permittivity of ceramic samples in the VHF band, which can also be applied in higher frequency bands, in the microwave region, up to around 10 GHz.

This paper is organized as follows. In Section II, we briefly present methods of mechanical activation and the sintering process. In Section III, the complex permittivity is introduced. In Section IV, we provide a description of our measurement technique. In section V, we give examples of measured permittivities of ceramic materials. Finally, Section VI concludes the paper.

Nina Obradović – Institute of Technical Sciences of SASA, Knez Mihailova 35/IV, 11000 Belgrade, Serbia (e-mail: nina.obradovic@itn.sanu.ac.rs).

Antonije Đorđević – School of Electrical Engineering, University of Belgrade, Bulevar kralja Aleksandra 73, 11120 Belgrade, Serbia, and Serbian Academy of Sciences and Arts, Knez Mihailova 35, 11000 Belgrade, Serbia (e-mail: edjordja@etf.bg.ac.rs).

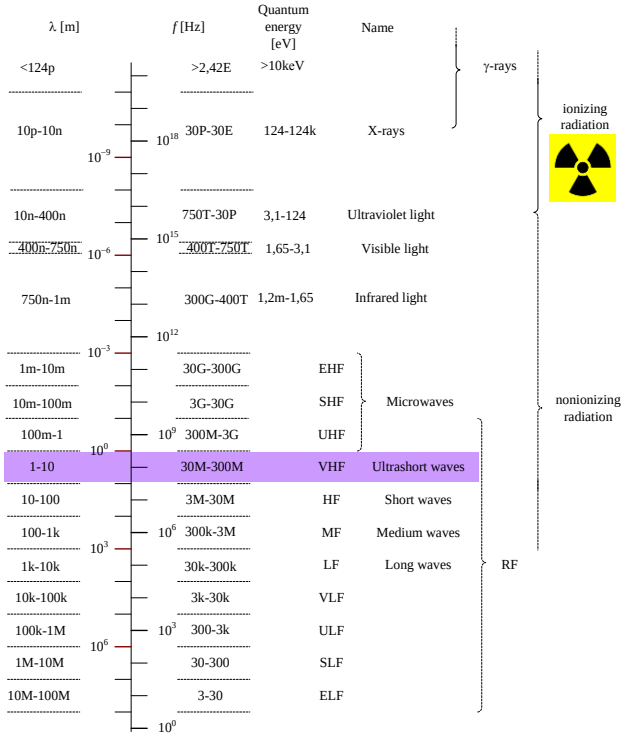


Fig. 1. Position of the VHF band in the electromagnetic spectrum.

II. Mechanical Activation and Sintering

The mechanical activation represents a complex physico-chemical process, which leads to an increase of chemical activity and potential energy of the treated material. Such a process can cause changes in the specific surface area and internal energy within the system. The free energy of the system is also increased [5]. During the mechanical activation of an inorganic material, attrition of the material takes place and reduces the crystallite size, which further leads to deformation or changes in the crystal structure, accompanied by generation of defects [6–9]. Grinding of the material is carried out by successive fracturing of particles (Fig. 2). The method of obtaining the finest powder is called dispersion milling [10]. Equipment used for grinding is a mill. There are several types of high-energy mills: attrition mill, vibro mill, horizontal ball mill, and planetary mill.

In terms of the grinding process, the influence of the mechanical forces on the material leads to formation of elastic and plastic deformations. During the elastic collision, the whole energy of the system remains unchained, as well as the material structure. During the plastic collision, the energy is partially transformed into the energy of deformation. Physico-chemical properties of materials are changed by these deformations [11]. The reaction capability of solid materials increases as a direct result of structural changes within the material by applying the method of mechanical activation. Changes of the structure of the crystal lattice are reflected trough the generation of point defects (vacancies, interstitial atoms, and impurity atoms), line defects (atom aggregations on the crystal surface), volume defects (pores and impurities), and electron defects (electrons and holes) [12, 13]. The occurrence of defects during mechanical activation represents the general tendency of crystals to transform the mechanical energy into the energy of the crystal lattice defects. During the mechanical treatment, the condition of the treated system is constantly changing, from the stage of the initial powder, where particles are poorly connected, to the stage of a heterogeneous system, where larger agglomerates form out of smaller powder particles with clearly defined boundaries.

Grinding of the material is carried out by successive fracturing of the powder particles in the mills. The mechanical activation leads to fragmentation, decrease of the particle size, and change of the free surface and the physico-chemical properties of crushed material. It is performed in order to improve the reactivity of the system or for the purpose of its amorphization. This treatment increases the reactivity of the material and lowers the temperature and time of sintering [15, 16].

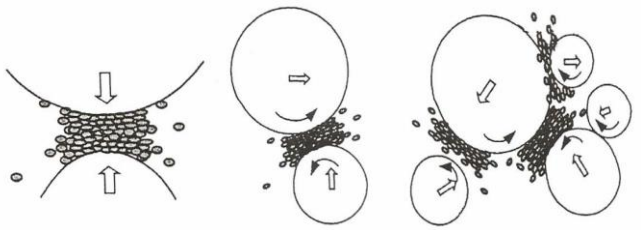


Fig. 2. Collisions that occur during milling in a high-energy planetary ball mill [14].

The method of mechanical activation has a variety of advantages compared to other methods of powder preparation, such as relatively cheap equipment (mills), inexpensive production, possible appliance to all classes of materials, and obtaining large quantities of material for further investigation. The main disadvantage of this method is the formation of agglomerates in the powder after milling. The powder agglomeration can be controlled and avoided by choosing adequate synthesis parameters. Firstly, the time of activation. According to the literature data, authors apply times in the range of two to five minutes, adequate for mixing and homogenization of powders [17, 18], up to dozens of hours, in cases of very hard materials or difficult and complex reactions occurring in the system [19–21]. Secondly, the ball-to-powder mass ratio. Authors choose the ball-to-powder mass ratio depending on how much energy they wish system to get during milling. Authors usually choose 20:1 or 40:1 ball-to-powder mass ratio, in order to obtain great energy transfer and easier and faster mechanochemical reaction [22–24]. Thirdly, selection of the material that the mill is made of. There is a variety of materials in usage, such as ZrO_2, WC-Co or hardened steel [25–27]. There is a possibility that during prolonged activation some particles from balls and vessels drop off and incorporate into the powder particles. Hence, prolonged milling time is not very suitable or milling equipment needs to be new and unused. Finally, the choice of the milling medium is also very important. It could be dry milling, conducted in air, or wet milling, e.g., in ethanol [28, 29]. Usually, milling is performed in air, but if the objective is to avoid the agglomerates, wet milling can be performed. To protect the powders from oxidation, the milling can take place in an inert atmosphere (at atmospheric pressure), either Ar or N_2 [30]. The method of mechanical activation attracts great attention due to its simplicity and because many parameters can be changed and optimized, which makes the work in a laboratory interesting, challenging, and novel each time.

G. V. Samsonov gave the most complete definition of the sintering process: “Sintering is a set of complex and interconnected mass transport processes that occur between the particles and within the particles of dispersed system during its consolidation” [31]. The necessary energy needs to be transferred into the system, so the sintering process can take place. It can be thermal (heating), mechanical (sintering under high pressure), or of some other form of energy (e.g., microwave sintering). The free energy of the system, which is in the non-equilibrium state, represents a driving force for the sintering process. The non-equilibrium state is a consequence of a developed specific surface area of powder particles and structural defects within them.

The sintering process of real materials can be conditionally divided into three stages [15], as illustrated in Fig. 3:

(1) In the initial stage, a contact is made between individual particles, but particles retain their structural individuality. During this stage, the density changes slightly and at the end of this stage, its value is 60–75 % of the theoretical density. At this sintering stage, the establishment of a better contact between particles and neck growth are a consequence of simultaneous acting of several mechanisms (viscous flow, surface diffusion, and volume diffusion). Open porosity is dominant, with pores of irregular shape and fracture trough grains.

(2) In the middle stage, particles accrue to one another. This leads to fast growth of contact necks, space between particles loses its shape, and particles lose their individuality. Pores get a more regular shape; their closure begins along with grain growth. Densification is most intensive in this intermediate phase. Fractures between grains and fractures trough grains are present at this sintering stage.

(3) The final stage is characterized by further formation of closed pores and grain growth. Due to shrinkage, open pores are too narrow to be stable and they are transformed into closed pores; then sintering enters the final stage. Density values are 90–93 % of the theoretical density. The final sintering stage is characterized by elimination of closed pores and approaching the theoretical density. The volume diffusion and diffusion along grain boundaries are dominant in the middle and final sintering stages. There are no sharp boundaries between these three stages of sintering.

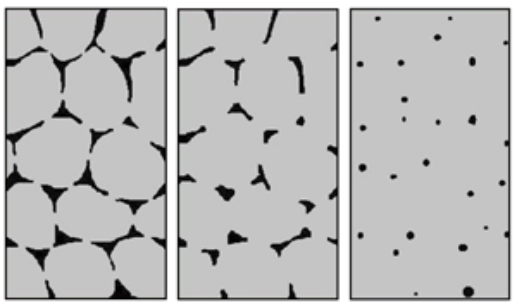


Fig. 3. The three stages of solid-state sintering [32].

In practice, technological factors have a great influence on the sintering process. Most important are: particle shape and size within the initial powder, the purity of the starting powders, the density of compacts (compaction pressure), sintering temperature, sintering regime, sintering atmosphere, etc.

There are a number of equations describing the process of sintering in the literature, which certainly indicate the great complexity of the process. Each of these equations refers to either the individual stages of sintering, or to certain mechanisms. Therefore, the basic questions concerning the kinetics and mechanisms are solved by usage of models (two-sphere model, sphere and flat plain model, etc.). A two-sphere model is most commonly used (Fig. 4). An increase of the contact surface may be caused by several mechanisms. Different types of diffusion in a solid state allow transport of the mass:

(1) The viscous flow consists of a deformation of the crystal structure under the influence of the surface tension. The mass transport occurs owing to the directional movement of atoms from the volume of the particles to the contact neck. During this process, the work that is carried out by the forces of the surface tension is equated with the work of forces of internal friction. An increase of the contact neck occurs along with reducing the distance between the particles [33].

(2) The surface diffusion is the most general mechanism which occurs during the sintering of crystalline materials. The spontaneous transport of atoms on the particle surface is dominant. During the growth of the contact neck, the distance between particle centers remains the same, without an increase in density of the material. The surface diffusion is the most common mechanism for lower temperatures of sintering and small particles.

(3) The volume diffusion is a mechanism where the mass transfer occurs by the diffusion flow from the particle surface to the contact neck. During this type of sintering, the vacancy concentration has a great impact on the diffusion velocity. This is crucial for materials with additives, where vacancies depend directly on the concentration of dopants. In this way, the concentration of the added material has an impact on the diffusion coefficient, and consequently on the mass transfer.

(4) The evaporation-condensation is a mechanism where the mass transfer occurs due to the difference of the vapor pressure of the solid phase in various parts of the system. Owing to the existence of a higher vapor pressure of the solid phase on the convex surface, and a lower vapor pressure of the solid phase on the concave surface, the material evaporates from the convex surface and via the gas phase gets transferred to the concave surface, where it condenses. The distance between the particles does not change; hence, the evaporation-condensation mechanism does not contribute to the densification of the system.

(5) The diffusion along grain boundaries is a mechanism in which the mass transport occurs from the border area between the particles to the surface of the contact neck, by diffusion along the edge. The contribution of this mechanism to the total mass transport is not large because the extent of the grain boundaries is limited. The diffusion of substances from the grain border area to the surface of the contact neck ensures particle approaching, and produces the effect of a compact shrinkage during sintering. It also causes a change in the shape and size of pores, reducing the porosity.

One of the most important issues that need to be understood in the context of studying the process of

sintering is the way of pore elimination.

Based on experimental data, it was concluded that materials which contain large pores are sintered slower, while the smaller pores can be eliminated from the system during heating at significantly lower temperatures. The system will enter the final sintering stage when the temperature is sufficient to ensure an increase in the grain size by that level where the critical pore size reaches the size of the largest pore in the system. When the system reaches the phase where all pores are smaller than the critical pore size, it may result in a full densification of the ceramic material. That explains the poor sinterability of agglomerated powders [34].

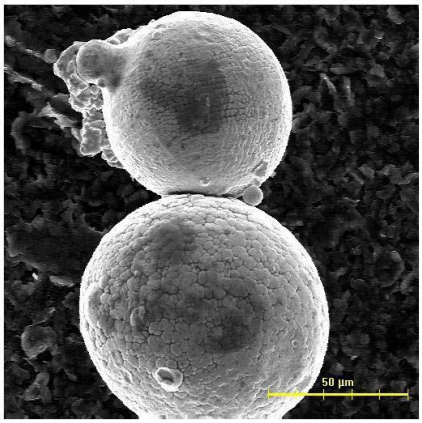


Fig. 4. An accretion of two real particles obtained by scanning electron microscopy [14].

Agglomerates are sets of small related particles that form large pores in the compact during the sintering process. As the process progresses, pores between agglomerates become larger, while pores within the agglomerates are reduced. The agglomeration or other processes, whose action makes a different pore size, suppress the sintering as long as there are two types of pores in the structure—larger ones between the agglomerates and smaller ones within the agglomerates. As the process develops, the large pores grow, while the smaller ones decrease. In the case of samples with a high initial density (the density after the compaction process, prior to the sintering), the pores between the agglomerates are reduced, and the negative effect of the existence of the two types of pores is decreased [35, 36].

Various techniques have been developed to obtain dense ceramics with a desired microstructure and phase composition. In general, these methods involve a combination of a heating regime and applied pressure. Heating regimes can be simple, as in the isothermal sintering, or can have a complex temperature-time relationship, as in the rate-controlled sintering, while the pressure may be applied either uniaxally, with or without a die, or by a surrounding gas [37]. The control of the sintering atmosphere is also important, and a precise control of oxygen or nitrogen partial pressures as a function of temperature may in some cases be beneficial or essential. Insoluble gas trapped in closed pores may obstruct the final stages of densification; hence, a change in the sintering atmosphere or a vacuum sintering is required. There are various heating regimes, such as: isothermal sintering, constant-heating-rate sintering, multi-stage sintering, rate-controlled sintering, microwave sintering, spark-plasma sintering, hot pressing, and hot-isostatic pressing.

Before we can chose the appropriate pre-sintering treatment and adjust its parameters, it is very important to have in mind that the resulting structure of the obtained compound is a direct result of structural changes at all hierarchical levels, caused by the applied treatment. It is possible to influence the development of the microstructure and to achieve the optimum and desired properties of the final material by a proper choice of the pre-sintering treatment and the type of the sintering process.

III. Complex Permittivity

In this paper, we consider dielectric properties of ceramic materials. These properties can be described by various parameters, such as polarizability, permittivity, dissipation factor (loss tangent), etc. In engineering applications, the complex relative permittivity is often used to characterize a dielectric material.

From the engineering standpoint, we live in the time domain: various physical quantities are functions of time, such as voltages, currents, electric fields, etc. In reality, the time dependence is known only for the past. However, in order to enable a mathematical description, we often assume that the time dependence is given analytically, by a certain mathematical function.

Furthermore, our world is causal: the response to a given excitation cannot occur before the excitation starts. If we consider electromagnetic waves, we even have a delay due to the propagation, so that the response is delayed after the excitation. The fastest known propagation is in a vacuum, where the velocity is $c_0 = 2{,}99792458 \cdot 10^8$ m/s .

In the engineering analysis, time-harmonic (sinusoidal) functions are often used to describe various periodic physical quantities [38]. Such functions also occur in the Fourier analysis. For example, a time-harmonic current, shown in Fig. 5, is given in the standard (canonical) form as

$$i(t) = I_\mathrm{m} \cos(\omega t + \psi), \qquad (1)$$

where t is the time ($-\infty < t < +\infty$), I_m is the amplitude ($I_\mathrm{m} \geq 0$), $\omega = 2\pi f$ is the angular frequency and f is the frequency, whereas ψ is the initial phase (usually reduced to the interval $-\pi < \psi \leq \pi$). The period (T) and the frequency are related as $fT = 1$.

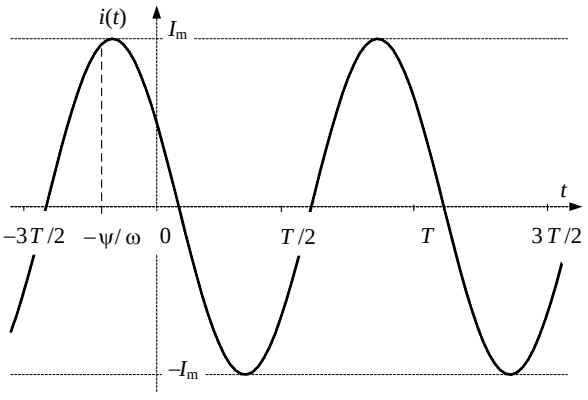


Fig. 5. Time-harmonic current.

Time-harmonic functions cannot occur in reality, because they span infinite time ($-\infty < t < +\infty$), but they can

conveniently describe various quantities in electrical engineering. In particular, if we consider a linear system and apply a time-harmonic excitation, the response will, in most practical cases, also be a time-harmonic function.

However, there are several difficulties in the analysis of such systems. The main one occurs when evaluating a sum or a difference of two time-harmonic quantities. The result is also a time-harmonic quantity of the same frequency, but expressions for the amplitude and phase of the result are rather cumbersome.

This difficulty is bypassed by introducing phasor (complex) representatives of the time-domain quantities and performing the analysis in the frequency (complex) domain. For example, the complex representative $\underline{I}$ of the time-domain current (1) is introduced by:

$$i(t) = \mathrm{Re}\left\{ \underline{I}\sqrt{2}\, \mathrm{e}^{-\mathrm{j}\omega t} \right\}, \tag{2}$$

where Re denotes the real part, $\mathrm{e} = 2.718281\ldots$ is the base of the natural logarithm, and $\mathrm{j} = \sqrt{-1}$ is the imaginary unit. In the exponential form, $\underline{I} = I\mathrm{e}^{\mathrm{j}\psi}$. The modulus (magnitude) of $\underline{I}$ is the root-mean-square (rms) value of $i(t)$, $I = |\underline{I}| = \frac{I_\mathrm{m}}{\sqrt{2}}$, and the argument (angle) of $\underline{I}$ is the initial phase of $i(t)$. Note that (2) is related to the Fourier and Laplace transforms.

Once we have phasors, addition and subtraction in the time domain is replaced by addition and subtraction of phasors, which is easily performed because we merely have to add or subtract complex numbers.

Additionally, the phasors facilitate solutions of linear differential equations: the differentiation in the time domain is mapped into the phasor domain as a multiplication by $\mathrm{j}\omega$. Hence, a linear differential equation in the time domain is mapped into an algebraic equation in the complex domain.

Equation (2) is valid for scalar quantities. A similar mapping can be applied to time-harmonic vectors as well [39]. Let us first define a time-harmonic vector, e.g., the electric-field vector, $\mathbf{E}(t)$. In Cartesian coordinates,

$$\mathbf{E}(t) = \mathbf{u}_x E_x(t) + \mathbf{u}_y E_y(t) + \mathbf{u}_z E_z(t), \tag{3}$$

where $\mathbf{u}_x$, $\mathbf{u}_y$, and $\mathbf{u}_z$ are the unit vectors of the Cartesian system, and

$$E_x(t) = E_x\sqrt{2}\cos(\omega t + \theta_x), \tag{4}$$

$$E_y(t) = E_y\sqrt{2}\cos(\omega t + \theta_y), \tag{5}$$

$$E_z(t) = E_z\sqrt{2}\cos(\omega t + \theta_z) \tag{6}$$

are time-harmonic functions of the same angular frequency ω, but can have different rms values (E_x, E_y, E_z) and initial phases ($\theta_x, \theta_y, \theta_z$). The complex representative of $\mathbf{E}(t)$ is now defined as

$$\underline{\mathbf{E}} = \mathbf{u}_x \underline{E}_x + \mathbf{u}_y \underline{E}_y + \mathbf{u}_z \underline{E}_z, \tag{7}$$

where $\underline{E}_x$, $\underline{E}_y$, $\underline{E}_z$ are the complex representatives of the components $E_x(t)$, $E_y(t)$, and $E_z(t)$, defined in accordance with (2).

Note that the complex domain is only a mathematically introduced "parallel world". Phasors (complex representatives) cannot be physically related to time-domain quantities.

The state in a polarized dielectric is macroscopically described by the polarization vector, **P** [40]. In the analysis of electromagnetic fields, in order to avoid dealing with the polarization vector, the electric displacement vector is introduced as defined by:

$$\mathbf{D} = \varepsilon_0 \mathbf{E} + \mathbf{P}, \tag{8}$$

where:

$$\varepsilon_0 = \frac{1}{\mu_0 c_0^2} = \frac{25\cdot 10^5}{299792458^2\,\pi}\ \mathrm{F/m} \tag{9}$$

is the permittivity (dielectric constant) of a vacuum. If the dielectric is linear, then the polarization vector is linearly proportional to the electric field (**E**) and, consequently, the vector **D** is also linearly proportional to **E**:

$$\mathbf{D} = \varepsilon \mathbf{E}, \tag{10}$$

where ε is the absolute permittivity (dielectric constant). Further, we can set $\varepsilon = \varepsilon_\mathrm{r}\varepsilon_0$, where ε_r is the relative permittivity (relative dielectric constant) of the material. Generally, a linear dielectric may be anisotropic (e.g., quartz), when the vectors **D** and **E** are not collinear, and ε_r is a tensor. In this paper, however, we assume the dielectric to be isotropic, so that **D** and **E** are collinear, and ε_r is a scalar. Note that $|\mathbf{D}| = |\varepsilon||\mathbf{E}|$. For a vacuum, $\varepsilon_\mathrm{r} = 1$. For static fields in linear, isotropic dielectrics, $\varepsilon_\mathrm{r} > 1$.

If the dielectric is located in a time-varying electric field $\mathbf{E}(t)$, the polarization vector is a function of time, $\mathbf{P}(t)$. Hence, the displacement vector is also a function of time, $\mathbf{D}(t) = \varepsilon_0\mathbf{E}(t) + \mathbf{P}(t)$. As a particular case, in the time-harmonic regime, all three vectors are time-harmonic quantities. However, in the general case, they are not in phase. If we switch to the complex domain, we can write

$$\underline{\mathbf{D}} = \underline{\varepsilon}\,\underline{\mathbf{E}}, \tag{11}$$

where $\underline{\varepsilon} = \underline{\varepsilon}_\mathrm{r}\varepsilon_0$ is the complex absolute permittivity and $\underline{\varepsilon}_\mathrm{r}$ is the complex relative permittivity.

The complex relative permittivity can be written as:

$$\underline{\varepsilon}_\mathrm{r} = |\underline{\varepsilon}_\mathrm{r}|\,\mathrm{e}^{-\mathrm{j}\delta} = \varepsilon_\mathrm{r}' - \mathrm{j}\,\varepsilon_\mathrm{r}'' = \varepsilon_\mathrm{r}'(1 - \mathrm{j}\tan\delta), \tag{12}$$

where $|\underline{\varepsilon}_\mathrm{r}|$ is the modulus, $-\delta$ is the argument, ε_r' is the real part, $-\varepsilon_\mathrm{r}''$ is the imaginary part, and $\tan\delta = \varepsilon_\mathrm{r}''/\varepsilon_\mathrm{r}'$ is the loss tangent. The minus sign is introduced because the vector **D** lags behind **E**; the argument of $\underline{\varepsilon}_\mathrm{r}$ is negative, but $\delta > 0$. Obviously, $-\delta$ is the phase difference between **D** and **E**. Further, $|\underline{\mathbf{D}}| = |\underline{\varepsilon}_\mathrm{r}|\,\varepsilon_0\,|\underline{\mathbf{E}}|$ so that $|\underline{\varepsilon}_\mathrm{r}|$ plays a role similar to ε_r in a static field.

Since $\mathbf{D}(t)$ and $\mathbf{E}(t)$ are collinear, we can introduce an axis collinear with these two vectors and consider the projections $D(t)$ and $E(t)$ on this axis. These projections are time-harmonic quantities. We consider the plot shown in Fig. 6, where the abscissa is $E(t)$ and the ordinate is $D(t)$. If $D(t)$ and $E(t)$ are in phase, the point in this plane periodically moves along a straight line (the red dashed line in Fig. 6). If they are not in phase, the point moves along an

ellipse in the counter-clockwise direction, creating a hysteresis loop (Rayleigh loop). Such a loop indicates that the dielectric is lossy. The volume density of the energy loss during one cycle is proportional to the surface area bounded by the ellipse, i.e.,

$$\frac{\mathrm{d}w_\mathrm{h}}{\mathrm{d}v} = \varepsilon_\mathrm{r}''\varepsilon_0 |\underline{\mathbf{E}}|^2 = \varepsilon_\mathrm{r}'\varepsilon_0 |\underline{\mathbf{E}}|^2 \tan\delta, \qquad (13)$$

where $|\underline{\mathbf{E}}|$ is the rms of the electric field. Obviously, the larger is $\tan\delta$, the larger are the losses. Hence, $\tan\delta$ is referred to as the dissipation factor.

Note that in our model $-\varepsilon_\mathrm{r}''$ describes all losses in the dielectric: both polarization and ohmic. If needed, the losses may be described by the equivalent conductivity of the medium (σ) and $-\mathrm{j}\varepsilon_\mathrm{r}''$ replaced in (12) by $-\mathrm{j}\dfrac{\sigma}{\omega\varepsilon_0}$.

For high-quality (low-loss) dielectrics, δ is small, $\tan\delta \ll 1$, and $\varepsilon_\mathrm{r}' \approx |\underline{\varepsilon}_\mathrm{r}|$. What can be considered as a low-loss dielectric, depends on applications. For example, in microwave engineering, for frequencies about 1 GHz, $\tan\delta$ of the order of 0.001 can usually considered to be low.

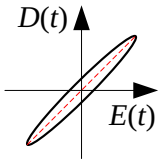


Fig. 6. Hysteresis loop.

The real and the imaginary parts of the complex permittivity are functions of frequency. Due to the causality conditions [41], these functions must satisfy the Hilbert transform (or, equivalently, the Kramers-Kronig relations). Hence, under certain conditions and limitations, if we know ε_r', we can reconstruct $-\varepsilon_\mathrm{r}''$, and vice versa. An important consequence of these relations is that if ε_r' changes with frequency, $-\varepsilon_\mathrm{r}''$ must be nonzero, or, practically speaking, the material must be lossy.

If we need to extract the function $\underline{\varepsilon}_\mathrm{r}(f)$ from measured data, it is convenient to approximate it by an analytic function of the complex frequency $\underline{s}$ ($\underline{s} = \mathrm{j}\omega$ on the imaginary axis of the complex plane), as done, for example, in [42] for FR-4, a material that has an almost constant loss tangent in a very broad frequency range.

IV. Measurement Techniques

Various methods and equipment can be used for measurement of the dielectric parameters [43], depending on the frequency range, aggregate state of samples, shape and size of available samples, etc.

For lower frequencies, up to about 100 MHz, the most commonly used method is based on inserting the material sample between two parallel electrodes, thus forming a parallel-plate capacitor. The capacitance is measured and the dielectric parameters are identified from the result. This technique is simple and convenient for solid samples in many cases. However, for barium titanate, magnesium titanate, and similar high-permittivity materials, air gaps between the sample and the electrodes are formed, which reduce the accuracy. Also, the measurement structure cannot be assumed to be quasistatic already at about 100 MHz. At higher frequencies, internal resonances may occur and jeopardize the measurements.

As the frequency increases, the parallel-plate capacitor has a strong electromagnetic coupling with the environment and it must be shielded. However, the shield creates a resonant cavity, which further aggravates the measurements. In order to push these resonances towards higher frequencies, the size of the cavity should be small.

Another problem with the parallel-plate structure is that it requires samples of an appropriate shape and size. For example, commercially available meters (such as Agilent E4991A meter with Agilent 16451B probe) require a large sample diameter (at least 15 mm), whereas our samples of ceramic materials are much smaller (often up to around 8 mm in diameter). Examples are pill-shaped sintered samples [44], which practically cannot be machined or otherwise adapted to the measurement system.

The parallel-plate capacitor measurement technique is convenient for liquid dielectrics. The main limitation is the relatively low upper frequency limit.

For frequencies above about 1 GHz, open coaxial lines can be used to characterize various materials [45]. This method is convenient because it usually does not require special forming of the sample. However, in order to provide accurate measurements, the sample should be big enough and it tightly positioned on the coaxial-line opening.

It is also possible to characterize a material by placing a sample between two antennas and measure the transfer between the antennas [46]. However, relatively large samples are required (e.g., in the form of a large sheet). Similar techniques exist for measurements in coaxial and waveguide systems [47], but they usually require that the samples have an appropriate shape and dimensions.

Dielectric substrates used in microwave engineering can be characterized by various resonant methods [48]. However, these are narrowband techniques. In order to provide data for a wide frequency band, other techniques can be used, such as manufacturing and measuring various transmission lines [42], which also require a special shape of the dielectric or metallization.

In [49], a method is described that was designed specifically for measurements of samples of ceramic materials that have a cylindrical shape (i.e., a pill-shape) and whose dimensions correspond to the majority of our samples. We have developed a small coaxial chamber, which is, naturally, shielded from the environment. We measure the reflection coefficient of the chamber using a network analyzer and then apply numerical techniques to extract the complex relative permittivity of the sample. However, this chamber requires that the bases of the cylinder are metallized. This can be done using silver-based conductive ink, which does not adhere well to ceramic materials and is messy for application.

We have further improved the measurement method by machining a new coaxial chamber (Fig. 7) and developing the corresponding software for measurements at frequencies in the VHF band [50]. In the meantime, we have developed

new numerical models that enable us to perform measurements up to around 10 GHz. We briefly describe this method in the following paragraphs.

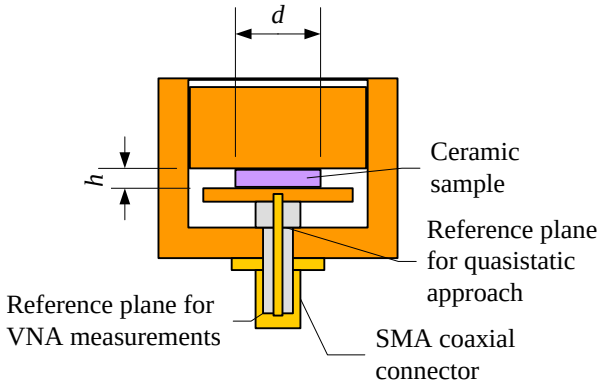


Fig. 7. Sketch of the test fixture for measurement of dielectric parameters of sintered samples.

The test fixture is a rotationally-symmetrical structure. It comprises a standard SMA coaxial connector at the bottom. The inner conductor of the connector carries a thin metallic disc, which acts as the lower electrode of a parallel-plate capacitor. The upper electrode is a thick metallic disc, which has a screw thread whose pitch is 1 mm. A mating thread is cut in the outer wall of the chamber. The upper electrode can be removed from the chamber so that a ceramic sample can be inserted into the chamber. The upper electrode is placed back so that it lightly presses the sample.

The ceramic sample is cylindrical, its diameter is *d*, and the height is *h*.

All metallic parts of the chamber are made of brass, except for the coaxial SMA connector, which is made of gold-plated stainless steel. The dielectric of the connector is Teflon. A Teflon disk is placed under the lower electrode in order to restrict the pressure on the inner conductor of the SMA connector.

We use a vector network analyzer (VNA) Agilent E5061A to measure the reflection coefficient at the SMA connector. The VNA is calibrated using an SMA calibration kit. The reference plane is at the lower end of the SMA connector.

At lower frequencies, up to around several hundred MHz, the chamber is assumed to be small compared to the wavelength and the quasistatic approximation for the electromagnetic fields is applied. The chamber is numerically modeled using a similar approach as in [51, 52]. A set of integral equations for the total charges (free plus bound) is formulated based on the boundary conditions. The first subset of boundary conditions is for the potential at conductor surfaces. The second subset of boundary conditions is for the normal component of the electric field at dielectric-to-dielectric interfaces. This system of integral equations is solved using the method of moments [53] with a piecewise-constant approximation for the charge distribution. The Galerkin technique is applied for testing.

The integrals for the potential are double. The first integration is carried out analytically, using the complete elliptic integral of the first kind. For the second integration, an adaptive Gauss-Legendre formula. The Galerkin integration is implemented using a Gauss-Legendre formula applied to equal-length subsegments along the generatrix of the structure.

The normal component of the electric field is obtained by numerical differentiation of the potential. Thereby, the field segment is incrementally moved away, perpendicularly from its original position, and the potential is evaluated for two positions of this segment.

In the numerical model, the conductors are assumed to have a finite thickness. Hence, the surface density of the free charges on the conductors is extracted from the density of the total charges by multiplication by the relative permittivity of the surrounding dielectric. In comparison with the method used in [51, 52], this approach significantly improves the quality of the solution.

The dielectric losses of the ceramic sample are taken into account by using the complex relative permittivity. The complex capacitance of the chamber is extracted from the measured complex admittance and the complex permittivity of the sample is adjusted in the numerical model so that the computed capacitance matches the measured one. This adjustment involves several interpolations.

For smaller ceramic samples, a simpler procedure is possible. The complex capacitance is measured when the sample is present and when it is removed. The complex relative permittivity of the sample is thereafter evaluated from the difference of these two capacitances, assuming that the electric field within the sample is homogeneous.

The accuracy of the quasistatic approach is enhanced by shifting the VNA reference plane to the beginning of the chamber (Fig. 7). Additionally, a first-order correction is implemented for the parasitic inductance of the chamber. As the result, the measurements can be performed in the whole VHF frequency band (30–300 MHz) and even extended to 500 MHz. The measurement uncertainty is 2% for ε_r' and 0.003 for $\tan\delta$. At lower frequencies, the VNA noise becomes a problem, which jeopardizes measurements particularly below around 10 MHz.

For higher frequencies, we have developed full-wave models of the chamber in software WIPL-D [54] and HFSS [55]. The models include the feeding coaxial line, up to the reference plane for VNA measurements. In each model, we supply the geometrical dimensions of the chamber and the sample, as well as data for the conductors (brass, gold) and dielectrics (Teflon, air) of the chamber. We assume several values for the relative permittivity of the sample and sweep over the frequency to compute the reflection coefficient at the reference plane for VNA measurements. Finally, we interpolate the relative permittivity to obtain the best agreement with the measurements.

V. Examples of Measured Permittivities

We present two examples in order to illustrate how the dielectric parameters of ceramic materials are affected by the mechanical activation (Section II) and also demonstrate the results obtained by the measurement technique described in Section IV.

The first example is spinel ceramic ($MgAl_2O_4$) [56]. We examined two different powders. One was not activated and the other one was activated in a high-energy planetary ball mill for 60 min. The results of the particle-size analysis (PSA) are shown in Fig. 8. The average particle size d(0.5)

was 2.1 µm for the non-activated powder AM–0. After mechanical activation for 60 minutes, the average particle size decreased to about 1.3 µm. Likewise, d(0.1) decreased from 0.69 µm for AM–0 to 0.43 µm for AM–60. Interestingly, d(0.9) increased from 4.74 µm for AM–0 to 8.04 µm for AM–60. Milling also produced a bimodal particle-size distribution in the AM–60 powder with peaks in the distribution at about 0.8 µm and 5 µm. While the milling significantly reduced the average particle size, it also produced some agglomerates that were larger than in the starting powder.

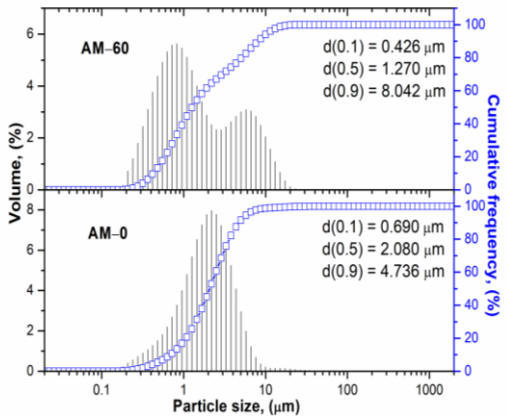


Fig. 8. PSA of the non-activated powder (AM–0) and powder activated for 60 minutes (AM–60).

Fig. 9a shows the relative permittivity and loss tangent of the material, sintered at 1400 ºC for 2h, when the starting powders (MgO and Al_2O_3) were not mechanically activated. These dielectric parameters are shown as a function of frequency in the range from 10 MHz to 500 MHz. The material is lossy; the loss tangent exceeds 0.3 at 10 MHz. The relative permittivity decreases with frequency, which is in accordance with the causality conditions.

Fig. 9b shows the results for spinel, also sintered at 1400 ºC for 2h, when the starting powders were activated. The relative permittivity is increased compared to the non-activated material and the loss tangent is smaller. The increased relative permittivity is in agreement with the increased density: the density of the non-activated material is 2.05 g/cm^3, whereas the density of the activated material is higher, 2.22 g/cm^3.

The second example is cordierite [57], which was mechanically activated. The milling times were up to 160 minutes. Also in this case, a strong correlation was found to exist between ε'_r and the sample density (ρ): the relative permittivity is higher for denser samples (i.e., for samples which have lower porosity), as shown in Fig. 10.

The data points in Fig. 10 closely follow the Lichtenecker-Rother logarithmic law of mixing [58]. The density of cordierite without pores is $\rho = 2.6\,\mathrm{g/cm^3}$, for which this fitting formula yields $\varepsilon'_r = 6.1$.

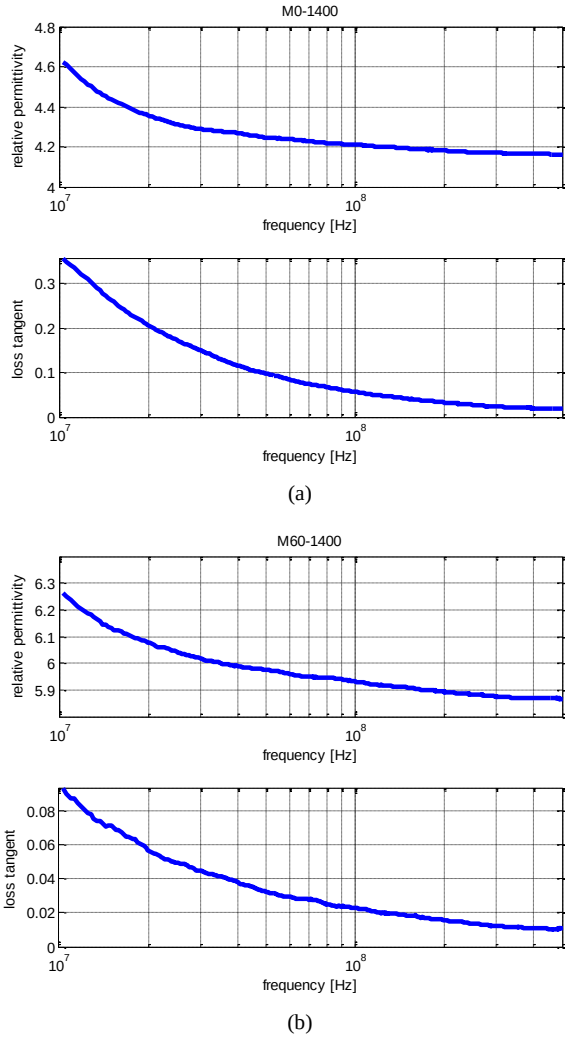


Fig. 9. Relative permittivity and loss tangent, as a function of frequency, for spinel when the starting powders were (a) not activated and (b) activated by milling for 60 minutes.

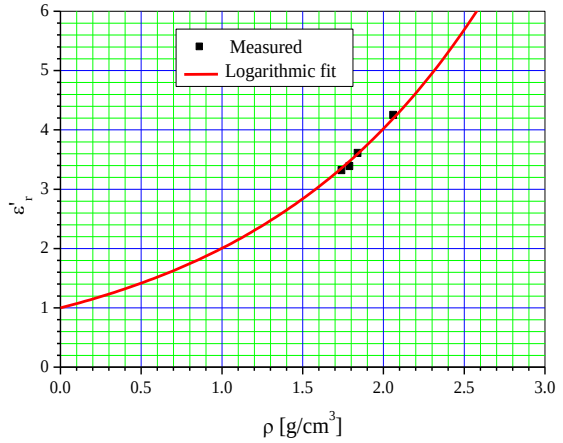
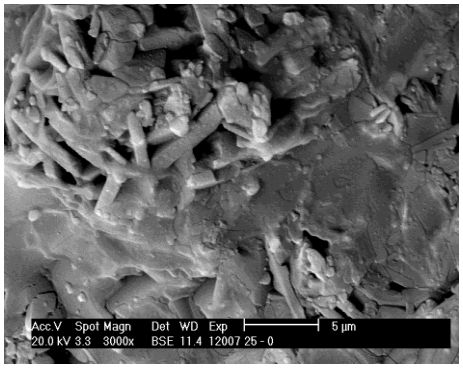
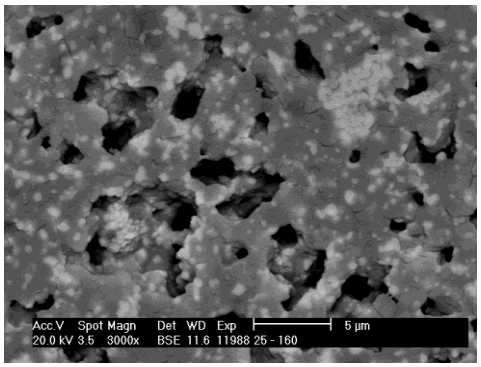


Fig. 10. Real part of the relative complex permittivity versus density for cordierite [57].



(a)



(b)

Fig. 11. SEM micrographs of two-step sintered powder mixtures: (a) MAS-0-S2 and (b) MAS-160-S2.

The average sample densities after the two-step sintering regime ranged from 60.3 to 74.8 % of the theoretical density. The scanning electron micrographs of the two-step regime sintered samples are shown in Fig. 11. The MAS-0-S2 sample (non-activated) consists of two different areas. One is a well sintered matrix and the other one is made of rod-like parts where the sintering process had not finished. This microstructure is a consequence of non-homogeneous mixture sintering. The sample MAS-160-S2 (activated for 160 minutes) has the highest relative density and the lowest relative open porosity. The obtained microstructures and densities after sintering are in a good accordance with the results of electrical measurements.

VI. Conclusions

Preparation conditions influence significantly the final structure and electrical properties of ceramic materials. An important goal is to obtain a ceramic material with nearly the full density and fine grains with a homogeneous distribution. The mechanical activation, as a pretreatment for the shaping and sintering process, is a very convenient way of powder preparation with a uniform particle-size distribution [59]. Processes such as increase in the number of contact and massive neck formation occur during the early stage of sintering. The neck growth is controlled by numerous diffusion mechanisms whose rates are determined by the total flux of atoms coming to the neck, suggesting that the dominant processes occur at grain boundaries.

Many functional properties of ceramics are also strongly influenced by the grain boundaries. From that point of view, electrical properties must be measured using AC techniques, so that the effects of grains and grain boundaries can be assessed separately.

Furthermore, sintering conditions can drastically change electrical properties. We considered the influence of the mechanical activation on the permittivity and the loss tangent of ceramic materials. We presented several examples demonstrating this influence. The results for the permittivity and the loss tangent were obtained using a newly developed measurement technique, which is revealed in the paper. The technique was designed to characterize relatively small samples of ceramic materials, without the need to metallize the surfaces of the samples. The electrical parameters in the VHF frequency band are computed from measured data using a quasi-static model. For higher frequencies, in the microwave region up to 10 GHz, we developed full-wave models, which are yet to be fully verified.

Acknowledgment

This research was performed within Projects OI 172057 and TR 32005 funded by the Ministry of Education, Science and Technological Development of the Republic of Serbia, and within Project F133 funded by the Serbian Academy of Sciences and Arts.

References

[1] http://www.linkinternationalkl.com/did-you-know-the-word-ceramics-comes-from-the-greek/
[2] P. L. Wise et al., Structure microwave property relations of Ca and Sr titanates, J. Eur. Ceram. Soc., 21 (2001) 2689-2632.
[3] С. Я. Гордеев, Физико-хемические основы керамической технологии, ИХТИ, Иваново, 1979.
[4] S. J. L. Kang, Sintering, densification, grain growth & microstructure, Elsevier, 2005.
[5] A. M. Maričić, S. M. Radić, M. M. Ristić, Fizički i fizičkohemijski principi tehnologije keramičkih materijala, Monografije nauke o materijalima, 37 CMS BU, Beograd, 1998.
[6] M. M. Ristić, S. Đ. Milošević, Mechanical activation of inorganic materials, SASA Monographs 38, Beograd, 1998.
[7] Е. Г. Аввакумов, Механические методы активации хумических процессоб, Новосибирск: Изд-во Наука, Сибирское отд-ние, 1986.
[8] Lj. D. Andrić, Mehanohemijska aktivacija glinice i njen uticaj na promenu kristalne strukture, doktorska disertacija, BU, Beograd, 1999.
[9] A. Branković et al., Mechanochemical Activation of (SeO_2+Na_2CO_3) Mixture and Sodium Selenite Synthesis in Vibrational Mill, J. Sol. St. Chem., 135 (1998) 256–259.
[10] N. Obradović, Uticaj aditiva na sinterovanje sistema ZnO-TiO_2 saglasno trijadi "sinteza-struktura-svojstva", doktorska disertacija, BU, Beograd, 2007.
[11] А. С. Баланкин, Самоорганизация и диссипативные структуры в деформируемом теле, Письма в Журнал Технической Физики, 16 (1990) 14–20.
[12] G. Heinicke, Tribochemistry, Academie-Verlag, Berlin, 1984.
[13] П. Ю. Бутягин, Кинетика и природа механохемических реакций, Успехи хемии, 40 (1971) 1935–1959.
[14] S. Filipović, Uticaj mehaničke aktivacije na svojstva MgO-TiO_2 elektrokeramike, doktorska disertacija, Tehnički fakultet u Čačku, Univerzitet u Kragujevcu, 2014.
[15] M. M. Ristić, Principi nauke o materijalima, Posebno izdanje knjiga DCXVII, SANU, Beograd, 1993.
[16] E. Kostić et al., Activation of solid state processes in Sintering and Materials, International Academic Publishers, Beijing, (1995) 142–147.
[17] N. Đorđević, N. Obradović, D. Kosanović, M. Mitrić, V. B. Pavlović, Sintering of Cordierite in the Presence of MoO_3 and Crystallization Analysis, Sci. Sinter., 46 (2014) 307–313.
[18] A. Peleš, V. P. Pavlović, S. Filipović, N. Obradović, L. Mančić, J. Krstić, M. Mitrić, B. Vlahović, G. Rašić, D. Kosanović, V. B. Pavlović, Structural investigation of mechanically activated ZnO powder, J. Alloys and Comp., 648 (2015) 971–979.
[19] N. Obradović, N. Labus, T. Srećković, S. Stevanović, Reaction Sintering of the 2ZnO-TiO_2 System, Sci. Sinter., 39 (2007) 127–132.

[20] G. Chen et al., Effects of mechanical activation on structural and microwave absorbing characteristics of high titanium slag, Powd. Technol., 286 (2015) 218–222.
[21] B. S. Zlatkov, M. V. Nikolić, V. Zeljković, N. Obradović, V. B. Pavlović, O. Aleksić, Analysis and modeling of sintering of Sr-hexaferrite produced by PIM technology, Sci. Sinter., 43 (2011) 9–20.
[22] T. Tunç, A. Ş. Demirkıran, The effects of mechanical activation on the sintering and microstructural properties of cordierite produced from natural zeolite, Powd. Technol., 260 (2014) 7–14.
[23] D. Kosanović, J. Živojinović, N. Obradović, V. P. Pavlović, V. B. Pavlović, A. Peleš, M. M. Ristić, The influence of mechanical activation on the electrical properties of $Ba_{0.77}Sr_{0.23}TiO_3$ ceramics, Ceram. Inter., 40 (2014) 11883–11888.
[24] S. Filipović, N. Obradović, V. Petrović, Influence of Mechanical Activation on Structural and Electrical Properties of Sintered $MgTiO_3$ Ceramics, Sci. Sinter., 41 (2009) 117–123.
[25] N. Obradović, N. Mitrović, V. Pavlović, Structural and electrical properties of sintered zinc-titanate ceramics, Ceram. Inter., 35 (2009) 35–37.
[26] M. M. Ristić, N. Obradović, S. Filipović, A. Bykov, M. A. Vasilkovskaya, L. A. Klochkov, I. I. Timofeeva, Formation of magnesium titanates, Powd. Metall. Metal. Ceram., 48 (2009) 371–374.
[27] A. N. Maratkanova et al., Structural characterization and microwave properties of chemically functionalized iron particles obtained by high-energy ball milling in paraffin-containing organic environment, Powd. Technol. 274 (2015) 349–361.
[28] N. Obradović, S. Filipović, N. Đorđević, D. Kosanović, S. Marković, V. Pavlović, D. Olćan, A. Đorđević, M. Kachlik, K. Maca, Effects of mechanical activation and two-step sintering on the structure and electrical properties of cordierite-based ceramics, Ceram. Inter., 42 (2016) 13909–13918.
[29] M. V. Nikolić, N. Obradović, K. M. Paraskevopoulos, T. T. Zorba, S. M. Savić, M. M. Ristić, Far infrared reflectance of sintered Zn_2TiO_4, J. Mater. Sci., 43 (2008) 5564–5568.
[30] H. B. Jin et al., Influence of mechanical activation on combustion synthesis of fine silicon carbide (SiC) powder, Powd. Technol., 196 (2009) 229–232.
[31] Г. В. Самсонов, Конфигурационние представления електроного строения в физическом материаловедении, Наукова Думка, Киев, 1977.
[32] https://www.google.com/search?q=three+stages+of+sintering&client=firefox-b-d&source=lnms&tbm=isch&sa=X&ved=0ahUKEwjzouiAvbjhAhVBaVAKHbf5DJoQ_AUIDigB&biw=1366&bih=654#imgrc=sIJp7eWFmCmlwM:
[33] Ya. I. Frenkel, Viscous flow of crystalline bodies under action of surface tension, J. Phys., 9 (1945) 385.
[34] V. V. Srdić, Presovanje novih keramičkih materijala, Tehnološki fakultet, Novi Sad, 2004.
[35] R. M. German, Sintering-Theory and Practice, John Wilez & Sons, Inc., New York, 1996.
[36] M. N. Rahaman, Ceramic processing and sintering, Marcel Dekker, Inc. New York, 2003.
[37] L. C. De Jonghe, M. N. Rahaman, Sintering of ceramics, Handbook of applied ceramics, Elsevier, 2003.
[38] A. R. Djordjević, Fundamentals of Electrical Engineering, Part IV, AC Currents, Belgrade: Academic Mind, 2016.
[39] A. R. Djordjević, Electromagnetics, Belgrade: Academic Mind, 2008.
[40] A .R. Djordjević, Fundamentals of Electrical Engineering, Part I, Electrostatics, Belgrade: Academic Mind, 2016.
[41] A. R. Đorđević, D. V. Tošić, "Causality of circuit and electromagnetic-field models", Proc. of 5th European Conference on Circuits and Systems for Communications (ECCSC'10), pp. 12–21, Belgrade, Serbia, 2010.
[42] A. R. Djordjević, R. M. Biljić, V. D. Likar-Smiljanić, T. K. Sarkar, "Wideband Frequency-Domain Characterization of FR-4 and Time-Domain Causality", IEEE Trans. Electromagn. Compat., vol. 43, no. 4, pp. 662–667, 2001.
[43] O. V. Tereshchenko ; F. J. K. Buesink ; F. B. J. Leferink, "An overview of the techniques for measuring the dielectric properties of materials", General Assembly and Scientific Symposium, vol. 1320, pp. 1–4, Istanbul, Turkey, 2011.
[44] N. Obradovic, M. V. Nikolic, N. Nikolic, S. Filipovic, M. Mitric, V. Pavlovic, P. M. Nikolic, A. R. Đordevic, M. M. Ristic, "Synthesis of barium-zinc-titanate ceramics", Science of Sintering, vol. 44, no. 1, pp. 65–71, 2012.
[45] T. P. Marsland, S. Evans, "Dielectric Measurements with an Open-ended Coaxial Probe", IEE Proc. Microw., Antennas Propag., vol. 134, pp. 341–349, 1987.
[46] D. K. Ghodgaonkar, V. V. Varadan, "A Free-space Method for Measurement of Dielectric Constants and Loss Tangents at Microwave Frequencies", IEEE Trans. Instrum. Meas., vol. 37, no. 3, pp. 789–793, 1989.
[47] A. M. Nicolson, G. F. Ross, "Measurement of the Intrinsic Properties of Materials by Time Domain Techniques", IEEE Trans. Instrum. Meas., vol. IM–19, no. 4, pp. 377–382, 1970.
[48] http://literature.cdn.keysight.com/litweb/pdf/5989-5384EN.pdf
[49] A. Djordjević, J. Dinkić, M. Stevanović, D. Olćan, S. Filipović, and N. Obradović, "Measurement of permittivity of solid and liquid dielectrics in coaxial chambers", Microwave Review, Vol. 22, No. 2, December 2016, pp. 3–9.
[50] A. Đorđević et al., Chamber for Measurement of Relative Permittivity and Loss Tangent of Dielectrics, School of Electrical Engineering, University of Belgrade, Project TR32005, 2018.
[51] A. R. Djordjević, M. B. Baždar, R. F. Harrington, and T. K. Sarkar, LINPAR for Windows: Matrix Parameters for Multiconductor Transmission Lines, Artech House, Norwood, MA, 1999.
[52] M. M. Nikolić, A. R. Djordjević, and M. M. Nikolić, ES3D: Electrostatic Field Solver for Multilayer Circuits, Artech House, Norwood, MA, 2007.
[53] R. F. Harrington, Field Computation by Moment Methods, Wiley-IEEE Press, Hoboken, NJ, 1993.
[54] WIPL-D. Wipl-d pro 11.0. WIPL-D, 2013, http://www.wipld.com
[55] ANSYS. Ansys hfss 15.0.0. ANSYS, 2014, http://www.ansys.com
[56] N. Obradović, W. G. Fahrenholtz, S. Filipović, D. Kosanović, A. Dapčević, A. Đorđević, I. Balać, V. Pavlović, "The effect of mechanical activation on synthesis and properties of $MgAl_2O_4$ ceramics", Ceramics International 45 (2019) 12015–12021.
[57] N. Obradović, Synthesis of Cordierite-Based Ceramics, Belgrade: Academic Mind, 2016.
[58] K. Lichtenecker and K. Rother, Die Herleitung des logarithmischen Mischungsgesetz es aus allgemeinen Prinzipien der stationären Strömung: Physikalische Zeitschrift, 32 (1931) 255–260.
[59] S. Filipović, V. P. Pavlović, N. Obradović, V. Paunović, K. Maca, V. B. Pavlović, "The impedance analysis of sintered $MgTiO_3$ ceramics" Journal of Alloys and Compounds 701 (2017) 107–115.

Electrical characteristics and phase transformation of Ho doped $BaTiO_3$ ceramics

Miloš Đorđević, *Student Member, IEEE*, Vesna Paunović, *Member, IEEE,* Vojislav Mitić, and Zoran Prijić, *Member, IEEE*

***Abstract*- The dielectric characteristics and phase transformation of Ho doped $BaTiO_3$ ceramics is investigated in this article. The concentrations of Ho_2O_3 in doped samples were ranged from 0.05 to 1.0 at% Ho. The investigated samples were prepared by a conventional solid state sintering procedure and sintered at 1320°C for 4 hours. For low dopants concentration (0.05 at% Ho), SEM analysis shows abnormal grain growth with the average size range between 10 μm - 30 μm. With the increase of dopant amount in samples causes decrease of average grain size, and for samples doped with 1.0 at% Ho, grain size range from less than 1 μm - 2 μm. The dielectric characteristics was measured in temperature range from 30°C to 180°C at different frequencies, from 100 Hz to 1 MHz. The dielectric constant has higher values for samples with lower concentration of additives (ε_r=4250 for 0.05 at% Ho/$BaTiO_3$, while ε_r=990 for 1.0 at% Ho/$BaTiO_3$ at Curie temperature). After initial high values at lower frequencies, ε_r decreases with frequency increase and reaches a constant value for f>20kHz. The Curie temperature at which the transition from the ferroelectric to the paraelectric region occurs ranges from 126°C to 130°C. For all investigated samples, it is characteristic that as the temperature increases, the tangent angle of losses increases. Curie-Weiss's law and modified Curie-Weiss law were used to calculate parameters such as Curie constant C and Curie temperature Tc, parameter γ which describing the diffusion and degree of nonlinearity of the change ε_r of the temperature above the Curie temperature and parameter δ which describing change εr of the temperature and frequency. In all the samples examined, a sharp transition from the ferroelectric to the paraelectric region at the Curie temperature is characteristic, which shows the value of the critical exponent of the nonlinearity γ from 1.01 to 1.07.**

Index Terms - $BaTiO_3$ ceramics; microstructure; electrical resistivity.

INTRODUCTION

Ferroelectric ceramics, most commonly tested for their practical applications, are $BaTiO_3$ ceramics. The reasons for such wide use of barium titanate are many. Barium titanate has a relatively low Curie temperature of T_C (120-130°C).

Miloš Đorđević - University of Niš, Faculty of Electronic Engineering, Aleksandra Medvedeva 14, 18000 Niš, Serbia (email: milos.djordjevic@elfak.ni.ac.rs).

Vesna Paunović – University of Niš, Faculty of Electronic Engineering, Aleksandra Medvedeva 14, 18000 Niš, Serbia (email: vesna.paunovic@elfak.ni.ac.rs).

Vojislav Mitić – University of Niš, Faculty of Electronic Engineering, Aleksandra Medvedeva 14, 18000 Niš, Serbia (email: vojislav.mitic@elfak.ni.ac.rs).

Zoran Prjić – University of Niš, Faculty of Electronic Engineering, Aleksandra Medvedeva 14, 18000 Niš, Serbia (e-mail: zoran.prijic@elfak.ni.ac.rs).

This allows the maximum dielectric permittivity in a temperature range in which it can be efficiently used (e.g. use in high-power converters) [1]. Based on this, the $BaTiO_3$ ceramic application area is very large and some of the examples are for electronic components such as multilayer ceramic capacitors, thermistors, piezoelectric sensors and PTC resistors [2], [3].

The dielectric properties of polycrystalline $BaTiO_3$ strongly dependent on the microstructure development, which depends on the type, concentration and the distribution of dopants.

In order to obtain $BaTiO_3$ ceramics with a high value of dielectric constant, it is necessary to establish high density, homogeneous and fine-grained microstructure, as well as uniform distribution of dopants and additives [4]-[6]. Two types of dopants can be introduced into $BaTiO_3$ lattice.

For the rare-earth ion incorporation into the $BaTiO_3$ lattice, the $BaTiO_3$ defect chemistry mainly depends on the lattice site where the ion is incorporated [7]. Depending on the size of the rare earth radius ions, which is in magnitude between the ionic radii Ba^{2+} or Ti^{4+} ions, rare earth cations such as Er^{3+}, Yb^{3+} Ho^{3+} and Dy^{3+} can occupy Ba or Ti positions in the perovskite $BaTiO_3$ structure [8], [9].

During cooling, $BaTiO_3$ ceramic passes through a series of phase transformations, from cubic to tetragonal ferroelectric phase at 1280°C-1300°C, to orthorhombic ferroelectric at 0°C, and further cooling to the rhombohedral ferroelectric phase at -80°C [10].

Additives have the effect of moving the Curie temperature or moving the maximum dielectric permeability value into a temperature range of doped $BaTiO_3$ ceramics, and that can be efficiently used. As the applied dopants influence the temperature of the phase transformation and on the Curie constant, it is best seen through the dependence of the dielectric constant on the temperature. For $BaTiO_3$ doped ceramics, phase transformation can have a very sharp transition from the ferroelectric to the paraelectric region, but also the diffusion phase transition ("relaxor" ceramics) can also occur [11].

To investigate the behavior of ferroelectrics in the paraelectric phase, in addition to the Curie-Weiss law, the modified Curie-Weiss relations that describe the deviations from the linearity $\varepsilon r = f(T)$ due to the diffuse phase transformation and the dependence of the dielectric constant on the frequency are also used.

The results of the influence of the additive concentration and the obtained microstructure on the dielectric properties of Ho doped $BaTiO_3$ ceramics are given in this paper. The

microstructure of samples was observed by scanning electron microscope (SEM). The variation of dielectric permittivity with temperature were measured in the temperature range from 30°C to 180°C and the frequency range from 100Hz to 1MHz.

EXPERIMENTAL PROCEDURE

Modified $BaTiO_3$ ceramics doped with 0.05, 0.1, 0.5 and 1.0 at% of Ho_2O_3 was obtained by conventional solid-state sintering method starting from pure oxide powders $BaTiO_3$ (Rhone Poulenc) and Ho_2O_3 (Fluka chemika). Starting powders are ball milled in ethyl alcohol for 24 hours. After milling the slurries were dried in an oven at 200°C for several hours until constant weigh and PVA was added as a binder. The milling powders were drying for several hours, and pressed into pellets 2 mm thick and 7 mm in diameter under 120 MPa. The pellets were sintered in air at 1320°C for 4 hours.

The microstructure of the sintered samples were observed by scanning electron microscope JOEL-JSM 5300 equipped with EDS (QX 2000S) system. Before samples were observing, electrical contacts were prepared by silver paste. The variation of electrical resistance with temperature were measured in temperature interval from 30°C to 180°C by using LCR meter Agilent 4248A at different frequencies, from 100Hz to 1MHz.

RESULTS AND DISCUSSION

Microstructure characteristics

The samples of $BaTiO_3$ ceramics doped with Ho_2O_3 are characterized by spherical and irregular polygonal grains.

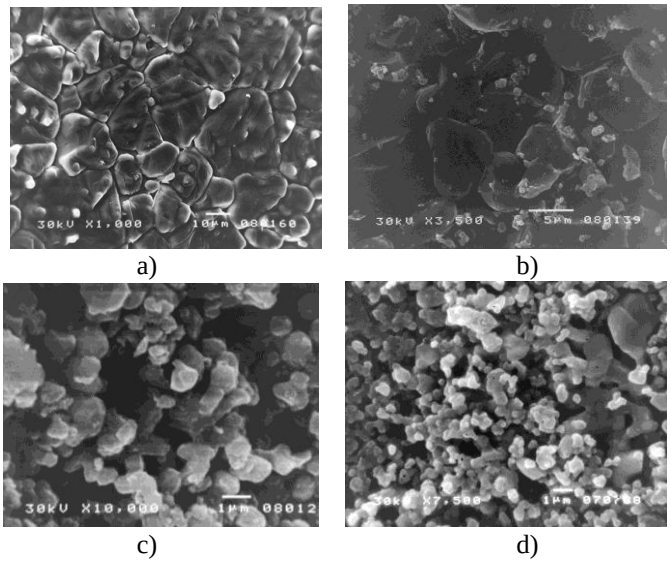


Fig. 1. SEM images of Ho doped $BaTiO_3$ ceramics sintered at Tsin=1320°C, a) 0.05 at%, b) 0.1 at%, c 0.5 at% and d) 1.0 at%.

The average grain size for samples doped with low content of Ho_2O_3 (0.05 at% Er) ranged from 10 μm to 30 μm (Fig. 1a)) for all measured samples. By increasing dopant concentration the grain size decreases. As a result, for 0.1 at% Ho of dopant the average grain size is from 5 μm to 15 μm (Fig. 1b)). With further increase of dopant concentration the average grain size decreases and for samples doped with 0.5 at% Ho is from 2 μm to 10 μm (Fig. 1c)), and for samples doped with higher dopant concentration (1.0 at% Ho), the average grain size is from the less than 1 μm to 2 μm (Fig. 1d)).

Such a microstructure is in accordance with the change of density of the investigated samples, which ranged from 82% to 91% of the theoretical density (TD). With the increase of dopant amount the increase of porosity is evident and density value decrease, so the highest density was for 0.05 at% doped samples (91% TD). The lowest value of density (82%) was measured for samples doped with 1.0 at% Ho.

EDS analysis of samples doped with 0.05 at% Ho_2O_3 did not reveal any Ho-rich regions, which indicated a uniform incorporation of dopants within the samples (Fig. 2a)).

EDS analysis cannot detect the additive content less than 1.0 at% Ho unless an inhomogeneous distribution or segregation of additive is present. The increase of dopant concentration leads to the appearance of Ho-rich regions (Fig. 2b)). These areas, rich of additives, are also characteristic for fine-grained microstructure.

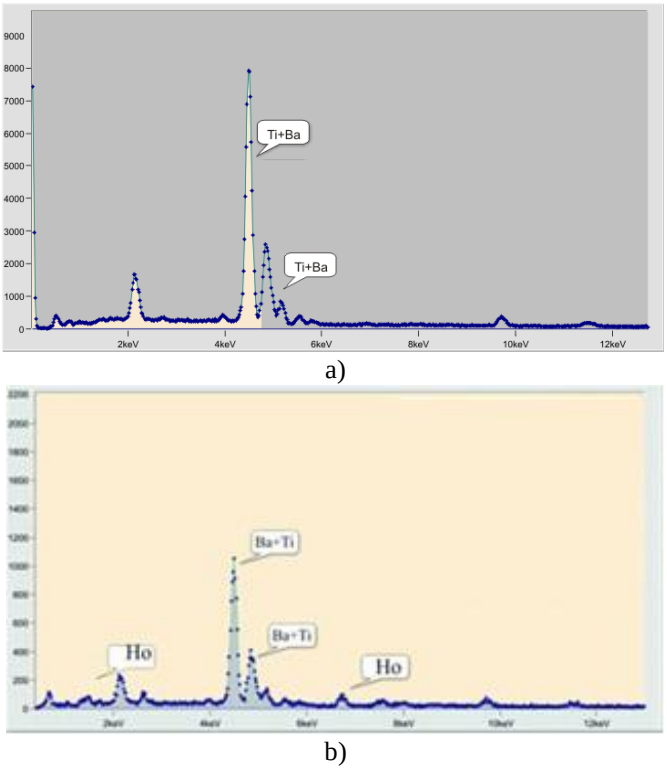


Fig. 2. EDS analysis of Ho doped $BaTiO_3$ sintered at 1320°C: a) 0.05 at% Ho-$BaTiO_3$, where no Ho is detected and b) Ho-rich regions in 1.0 at% Ho-$BaTiO_3$.

Electrical characteristics

The influence of the additive concentration and obtained microstructure on dielectric permittivity of the samples doped with Ho_2O_3 can also be examined through dependence of the dielectric permittivity on the temperature and frequency (Fig. 3 and Fig. 4). In this manuscript were investigated on

temperature in temperature range from 30°C to 180°C and frequency range from 100 Hz to 1 MHz. Based on the curves of dielectric constant dependence on temperature, it can be seen that the highest dielectric constant values at the Curie temperature ε_r = 3530 show samples with a concentration of 0.05 at% Ho. The lowest dielectric constant values at the Curie temperature have samples with the highest concentration of additives (1.0 at% Ho) and it is ε_r = 990.

Curie temperature (*Tc*) in which the transition from ferroelectric to paraelectric phase occur were in the range from 128°C to 130°C and it is lower relative to *Tc* undoped $BaTiO_3$ ceramic (*Tc* = 132°C).

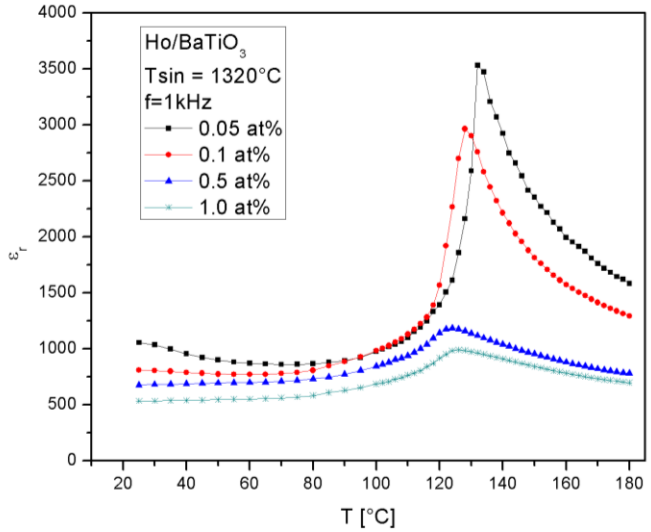


Fig. 3. The dependence of dielectric permittivity on temperature for the different doping concentration (0.05 – 1.0) at% Ho.

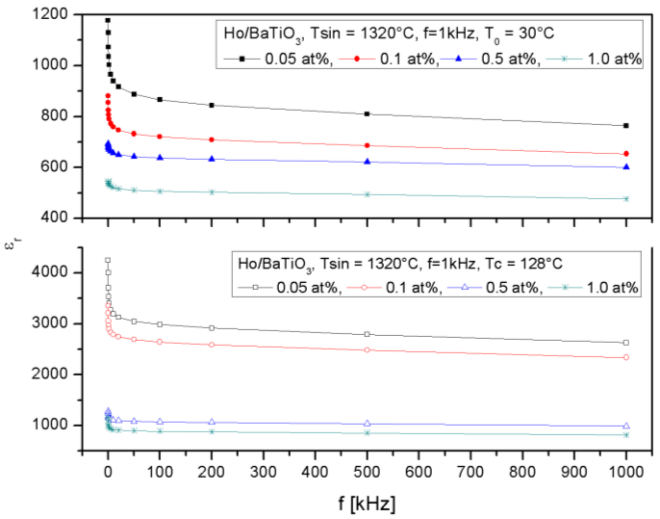


Fig. 4. The dependence of dielectric permittivity on frequency at room and Curie temperature.

The dielectric permittivity dependence on frequency at room and Curie temperature is shown in Fig. 4. As can be seen, the highest values of dielectric permittivity were for measured samples doped with lower content of dopant (0.05 at%). It has also been observed that the measured samples have higher values of dielectric permittivity at lower frequencies. Highest value of dielectric permittivity at room temperature (T = 30°C) has ε_r = 1177 for samples doped with 0.05 at% Ho, and at Curie temperature value of dielectric permittivity has ε_r=4250 for the same content of dopant. For the all measured samples it was noticed that after the initial high values dielectric permittivity values decreases increasing of values of frequency and achieves constant value for f > 20kHz.

On Fig. 5. it can be seen dependence of tangent of losses on temperature for all measured samples at frequency of 1 kHz. The dielectric losses of the measured samples (*tg δ*) with temperature have values ranging from 0.017 to 0.57. For the all measured samples is characteristic that as the temperature increases, the tangent loss increases. The smallest increase in the tangent losses is for samples with a concentration of additives of 1.0 at%, and the highest increase in samples with a concentration of additives of 0.05 at% Ho. Similar dependencies have been obtained for other frequencies.

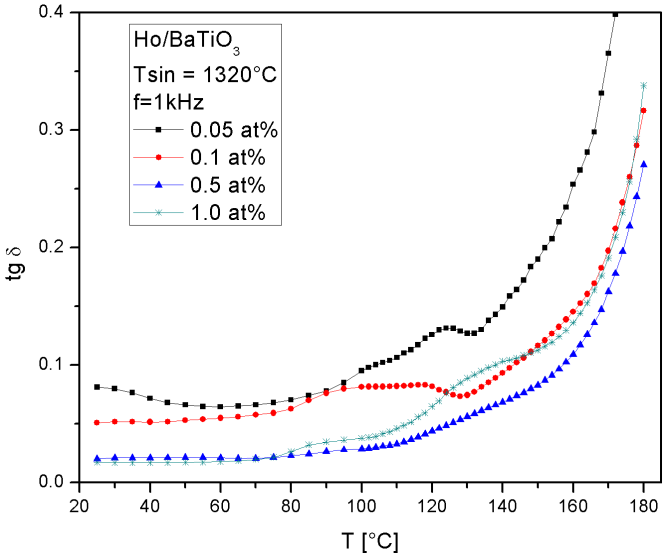


Fig. 5. The dependence of tangent (*tg δ*) losses on temperature for the different doping concentration (0.05 – 1.0) at% Ho at 1kHz.

Dielectric permittivity in ferroelectrics changes with temperature, reaches maximum value at Curie temperature and decreases with further increase of temperature. The dependence of dielectric permittivity on the temperature in the paraelectric region, i.e. in the area above Curie temperature, can be described by Currie-Weiss law:

$$\varepsilon_r = \frac{C}{T - T_0}. \qquad (1)$$

where is *C* – Currie constant, *T* – temperature, and T_0 – Currie-Weiss temperature.

By fitting the dependence of the reciprocal value of dielectric permittivity on the temperature, as shown in Fig. 6, the values of Curie-Weiss temperature T_0 are obtained. The Curie-Weiss temperature T_0 has a lower value than Currie temperature (*Tc* = 126°C - 130°C) for all measured samples. The highest value for T_0 was obtained for samples doped with 0.05 at% (T_0 = 93.05°C), and the lowest T_0 value for samples

was doped with 1.0 at% (T_0 = 10.22°C). Table I gives the values of Currie-Weiss temperature for all measured samples.

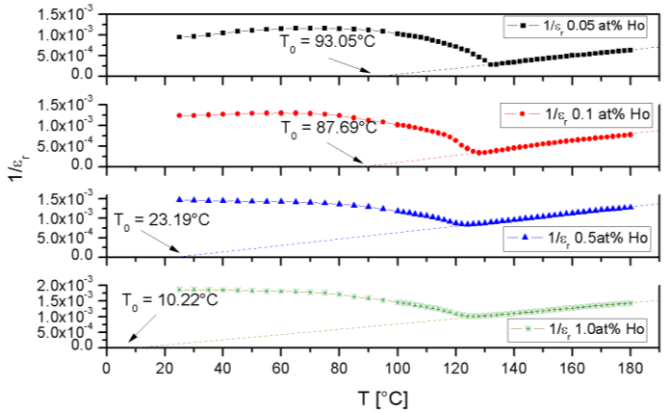


Fig. 6. The reciprocal value of the dielectric permittivity in the function of the temperature.

Based on the Currie-Weiss law, the values of the Curie constant for all measured samples are calculated. The value of Curie constant decreases with increasing additive concentration (Fig. 7) so that the highest value of the Curie constant is calculated for samples with additive concentration of 0.05 at% Ho ($C = 2.15 \cdot 10^5$ K), and the lowest for samples with an additive concentration of 1.0 at% Ho ($C = 5.9 \cdot 10^4$ K). The samples with highest value of C are characterized by fine-grained microstructure and higher sample density. The values for the Curie constant are in agreement with the change in the density of the tested samples as well as with the microstructural characteristics.

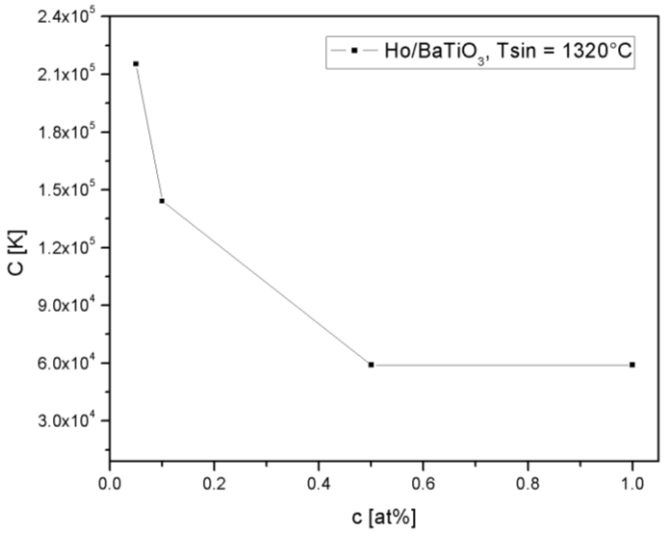


Fig. 6. Curie constant in the function of additive concentration.

For all the measured samples, a sharp transition from the ferroelectric to the paraelectric region is characteristic, as is the sudden increase of the dielectric permittivity at the Curie temperature. This fact can be confirmed by the ratio of the dielectric constant at the Curie temperature (ε_{rmax}) and at room temperature (ε_{rmin}), i.e. ($\varepsilon_{rmax}/\varepsilon_{rmin}$) (Table 1). The highest dielectric constant ratio was calculated for samples doped with 0.1 at%, $\varepsilon_{rmax}/\varepsilon_{rmin}$ = 3.66, and the lowest ratio of samples doped with 0.5 at%, 1.75.

TABLE I
DIELECTRIC PARAMETERS FOR HO/$BaTiO_3$ CERAMICS

	ε_r at T=300K f=1kHz	ε_r at T_C f=1kHz	Tc [°C]	$\varepsilon_{rmax}/\varepsilon_{rmin}$	T_0 [°C]	C [K] $\cdot 10^5$
0.05 at% Ho	1177	4250	130	3.35	93.05	2.153
0.1 at% Ho	880	3356	128	3.66	87.69	1.442
0.5 at% Ho	693	1285	124	1.75	23.19	0.591
1.0 at% Ho	545	1162	126	1.98	10.22	0.590

In order to explain the deviations from linearity of the Curie-Veiss law (1), in the investigated samples, Kirillov and Isupov proposed the equation [12,13]:

$$\frac{1}{\varepsilon_r} - \frac{1}{\varepsilon_{r\max}} = \frac{(T - T_{\max})^2}{2\varepsilon_{r\max}\delta^2}. \quad (2)$$

where ε_r is the dielectric constant, ε_{rmax} dielectric constant on Tc and δ parameter describing the change of ε_r from temperature and frequency. In addition to the Curie-Weiss law for testing the behavior of ferroelectrics in the paraelectric phase are also used modified Curie-Weiss relations that describe the deviations from linearity $\varepsilon_r = f(T)$ due to the diffuse phase transformation and the dependence of the dielectric permittivity on the frequency.

Using modified Curie-Weiss law:

$$\frac{1}{\varepsilon_r} = \frac{1}{\varepsilon_{r\max}} + \frac{(T - T_{\max})^{\gamma}}{C'}. \quad (3)$$

where the C' constant is similar to the Curie constant, a critical exponent of nonlinearity γ is determined, which shows a deviation from the linear dependence of the dielectric permittivity ε_r on the temperature in the paraelectric region. Based on linear fitting of curves $ln(1/\varepsilon r\text{-}1/\varepsilon r_{max})$ in function of $ln(T\text{-}T_{max})$ was obtained as a slope of the right, and the graphical representation for all samples is illustrated in Fig. 7.

The value of the critical exponent of the nonlinearity γ for measured samples ranges from 1.001 for 0.05at% Ho doped samples to 1.07 for samples doped with 1.0at% Ho (table 2). These values are in accordance with the experimental data, because for these samples there is a characteristic sharp transition from the ferroelectric to the paraelectric region indicating a structural phase change.

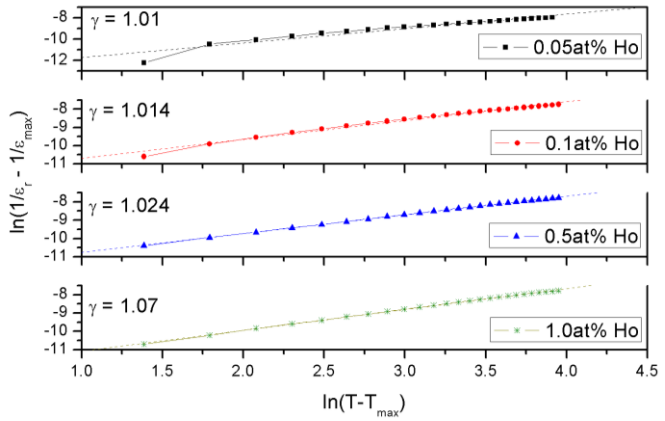


Fig. 7. Dependence $ln(1/\varepsilon_r - 1/\varepsilon_{rmax})$ of $ln(T-T_{max})$ for measured samples.

On the basis of equations (2) and (3), C' and δ are calculated which is a measure of the diffusion phase transition. As can be seen from Table 2 with increasing concentration of Ho_2O_3, the C' constant increases, so that the highest values are calculated for 1.0 at% Ho.

TABLE II
DIELECTRIC PARAMETERS FOR HO/BATIO$_3$ CERAMICS

	C' [K] $\cdot 10^5$	γ	δ
0.05at% Ho	1.32	1.01	22.53
0.1at% Ho	1.11	1.014	24.07
0.5at% Ho	1.29	1.024	43.38
1.0at% Ho	1.33	1.07	46.89

On Fig. 8 is shown the dependence of specific electrical resistance on temperature for all measured samples at a frequency of 1 kHz. The values of specific electrical resistance (ρ) with temperature have at low temperatures a slight increase in all measured samples up to Curie temperature, and above this temperature (128-130°C), starting with samples with the lowest concentration of additives (0.05 at% Ho), up to samples with the highest concentration (1.0 at% Ho) there is a sudden increase in specific electrical resistivity. For other frequencies on which samples were measured, similar addictions were obtained.

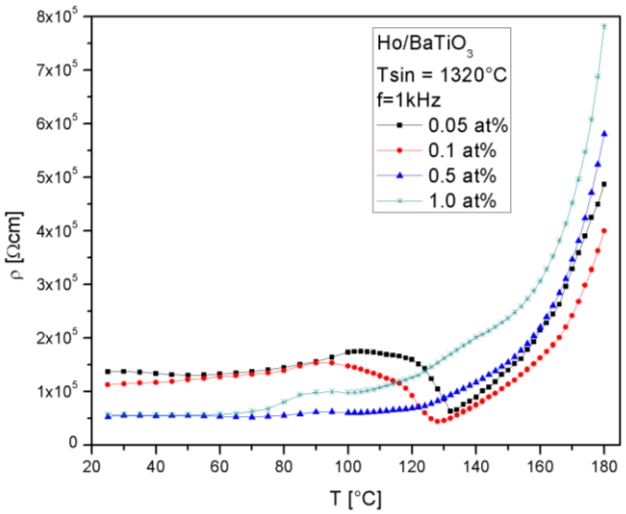


Fig. 8. The specific electrical resistivity on temperature for the different doping concentration (0.05 – 1.0) at% Ho at 1kHz.

IV. CONCLUSION

In this article the dielectric characteristics and phase transformation of Ho_2O_3 doped of $BaTiO_3$ ceramics has been investigated. Microstructural studies have shown that for lower concentration of dopant (0.05 at% Ho) characteristic spherical and irregular polygonal grain growth with the average size range between 10-30 μm for samples sintered at 1320°C. The increase of dopant concentration in samples, leads to decrease of average grain size and for samples doped with 1.0 at% Ho, the average grain size range from less than 1 μm - 2 μm. The highest value of dielectric permittivity at Curie temperature has measured samples ε_r = 3530 with a concentration of additives of 0.05 at% Ho. With increasing of dopant content, value of ε_r decreasing, and lowest value has measured for high dopant content (1.0 at Ho), εr=990. Value of Curie temperature was ranged from 128° to 130°C. Tangent angle of losses was in range from 0.017 to 0.57. Based on the Curie-Weiss law, the parameters such as the Curie-Weiss temperature (T_0) and the Curie constant (C) are determined. The highest value of the Curie constant is calculated for Ho doped BaTiO3 ceramics with additive concentration of 0.05 at% Ho (C = $2.95 \cdot 10^5$ K), and the lowest for samples with a concentration of additives 1.0 at% Ho ($C = 0.590 \cdot 10^5$ K). The critical exponent of non-linearity γ was in the range from 1.01 to 1.07, which is in accordance with the experimental data, because for all the samples a sharp transition from the ferroelectric to the paraelectric region is characteristic.

ACKNOWLEDGMENT

The authors gratefully acknowledge the financial support of Serbian Ministry of Education, Science and Technological Development. This research is a part of the Projects OI-172057, TR-32026 and TR-33035.

REFERENCES

[1] D. Mančić, V. Paunović: Application of impedance spectroscopy for electrical characterization of La doped $BaTiO_3$ ceramics, from the Monograph, Faculty of Electronic Engineering in Niš, Niš, 2012. [In Serbian]

[2] S.F. Wang, G.O. Dayton: Dielectric Properties of Fine-grained Barium Titanate Based X7R Materials, Journal of the American Ceramic Society, Vol. 82, No. 10, Oct. 1999, pp. 2677 – 2682.

[3] C. Pithan, D. Hennings, R. Waser: Progress in the Synthesis of Nanocrystalline $BaTiO_3$ Powders for MLCC, International Journal of Applied Ceramic Technology, Vol. 2, No. 1, Jan. 2005, pp. 1 – 14.

[4] P. Kumar, S. Singh, J.K. Juneja, C. Prakash, K.K. Raina: Influence of Calcium on Structural and Electrical Properties of Substituted Barium Titanate, Ceramics International, Vol. 37, No. 5, July 2011, pp. 1697 – 1700.

[5] Z.C. Li, B. Bergman: Electrical Properties and Ageing Characteristics of $BaTiO_3$ Ceramics Doped by by Single Dopants, Journal of the European Ceramic Society, Vol. 25, No. 4, April 2005, pp. 441– 445.

[5] V. Mitić, V. Paunović, D. Mančić, Lj. Kocić, Lj. Živković, V.B. Pavlović: Dielectric Properties of $BaTiO_3$ Doped with Er_2O_3, Yb_2O_3 based on Intergranular Contacts Model, Advances in Electroceramic Materials: Ceramic Transactions, Vol. 204, July 2009, pp. 137 – 144.

[6] M. Đorđević, M. Marjanović, V. Paunović, V. Mitić, Z. Prijić: Electrical Resistivity of Er/Yb doped $BaTiO_3$ ceramics, IcETRAN, Kladovo, Serbia, Proceedings 4th Conference IcETRAN, pp. NM1.2, 5-8, jun, 2017.

[7] D.Makovec, Z.Samardzija, M.Drofenik: Solid Solubility of Holmium, Ytrium and Dysprosium in $BaTiO_3$, J.Am.Ceram.Soc. 87, pp. 1324-1329, 2004.

[8] D. Lu, X. Sun, M. Toda Electron Spin Resonance Investigations and Compensation Mechanism of Europium-Doped Barium Titanate Ceramics Japanese Journal of Applied Physics Vol. 45, No. 11, 2006, pp. 8782-8788.

[9] S. M. Park, Y. H. Han, "Dielectric Relaxation of Oxygen Vacancies in Dy-doped BaTiO3", Journal of the Korean Physical Society, 2010, 57, No. 3, pp. 458- 463.

[10] M. Đorđević, M. Marjanović, V. Paunović, V. Mitić, Z. Prijić: Specific Electrical Resistivity of Er doped $BaTiO_3$ ceramics, ETRAN, Zlatibor, Serbia, Proceedings 60th Conference ETRAN, pp. NM1.1, 13-16, jun, 2016.

[11] K.J. Park, C.H. Kim,Y.J. Yoon, S.M. Song, "Doping Behaviors of Dysprosium, Yttrium and Holmium in BaTiO3 ceramics", Journal of the European Ceramic Society, 2009, vol. 29, pp. 1735-1741

[12] R.Zhang, J.F.Li, D.Viehland, "Effect of aliovalent substituents on the ferroelectric properties of modified barium titanate ceramics: relaxor ferroelectric behaviour" *J. Am. Ceram. Soc.*, 87 [5], pp.864-870, 2004.

[13] M. Đorđević, M. Marjanović, V. Paunović, V. Mitić, Z. Prijić: The Electrical Characteristics and Phase Transformation of Yb doped $BaTiO_3$ ceramic, ETRAN, Srebrno jezero, Serbia, Proceedings 59th Conference ETRAN, pp. NM1.1, 8-11, jun, 2015. [In Serbian]

Surface properties of polycrystalline diamonds for advanced applications

Sandra Veljković, *Student Member, IEEE*, Vojislav Mitić, Vesna Paunović, *Member, IEEE*, Goran Lazović, Markus Mohr and Hans Fecht

***Abstract* – The development of new materials as well as improvement of already known materials characteristics, can significantly contribute to the progress in the development of different areas. In that sense, polycrystalline diamonds are becoming more and more interesting because of their wide application. Considering that this material has an extreme potential, the research in this area is intense, and in this paper are presented the most important applications related to engineering. Intensive research of surface structure can contribute to the better insight of polycrystalline diamonds properties. An analysis of surface structure of nanocrystalline diamonds obtained by chemical vapor deposition method is presented in this paper.**

Index Terms – polycrystalline diamonds; application; MEMS; surface properties

I. Introduction

Diamond is one of allotropic modifications of carbon and due to its specific structure, has a very wide application [1]. Probably, its most known use is in jewelry, but it also has applications in medicine, in various industries for cutting, grinding, drilling and polishing [2]. Besides that, there is a very important use of diamonds in microelectromechanical systems (MEMS), microelectronics and in other areas.

However, natural diamonds are rear, expensive and it is hard to find them in proper sizes and shapes, which represents a limiting factor in their application. For that reason, the discovery that diamonds can be artificially synthesized was revolutionary. The first successful synthesis was achieved in the middle of the last century. The conditions in which diamonds are created in nature are artificially simulated: under high pressures (~GPa) and at high temperatures, which are about 1200°C. These parameters are used for the oldest technique for diamond synthesis, named high pressure high temperature (HPHT). Diamonds created by HPHT are very similar to natural diamonds. Although this technique is very effective, diamonds which are similar to stone are not applicable in engineering. For that reason, a better solution is chemical vapor deposition (CVD) method, by which diamonds can be created as thin films for covering different surfaces and shapes.

Sandra Veljković - University of Niš, Faculty of Electronic Engineering, Aleksandra Medvedeva 14, 18000 Niš, Serbia (email: sandra.veljkovic@elfak.rs).

Vojislav Mitić – Scientific advisor ITN SANU, University of Niš, Faculty of Electronic Engineering, Aleksandra Medvedeva 14, 18000 Niš, Serbia (email: vmitic.d2480@gmail.com).

Vesna Paunović – University of Niš, Faculty of Electronic Engineering, Aleksandra Medvedeva 14, 18000 Niš, Serbia (email: vesna.paunovic@elfak.ni.ac.rs).

Goran Lazović – University of Belgrade, Faculty of Mechanical Engineering, (email: goran.lazovic@gmail.com).

Markus Mohr – University of Ulm, Institute of Functional Nanosystems FNS, (email: markus.mohr@uni-ulm.de).

Hans Fecht – University of Ulm, Institute of Functional Nanosystems FNS, (email: hans.fecht@uni-ulm.de).

II. Natural Diamonds

Characteristics and properties of natural diamonds are the result of their specific structure, which is presented in Fig. 1a. It shows a unit diamond cell where each carbon atom is linked to the four nearest neighbors. The bond is made by sp^3 hybridized orbitals with angles of 109.5° between all neighbor bonds (Fig. 1b). Two intermediate surface-centered cubic lattices, which are shifted one of other one-quarter of its length can be seen in the diamond crystal structure. Each unit cell has eight atoms.

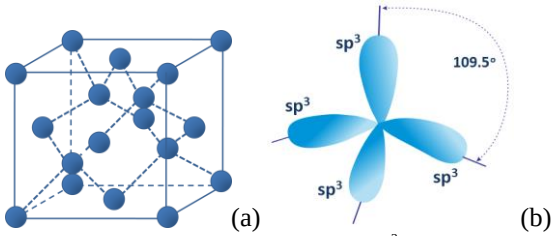


Fig. 1. (a) Cristal structure of diamond and (b) sp^3- hybrid orbitals.

Due to its structure, diamond is the hardest material on Mohs scale with the value of 10; it has the highest melting temperature which is 3547°C and the lowest molar entropy of 2.4 J mol^{-1} K^{-1}. Similarly, mechanical properties of diamonds are very interesting. Symmetrical tetrahedral structure and relatively short links between carbon atoms cause a very high value of Young's modulus – 1050 GPa. Also, monocrystalline diamond has a high fracture strength, about 2.8 GPa [3], which is by far the highest fracture strength value of all materials which are used for micro-mechanical purposes.

Despite great qualities of natural diamonds, it is not profitable to use them in mass production due to the price and complex processing. For that reason, diamond synthesis was a significant discovery, as it became possible to apply diamond layers on different materials and shapes.

III. Application of Polycrystalline Diamonds

A standard procedure for creating polycrystalline diamonds is the CVD technique. In this way, diamonds which are obtained are microcrystalline diamonds (MCD) and their grain size order of magnitude is micrometric. The main

characteristics of these layers are that they are very rough and have a very low resistance to fracture which limits their application. Further development and improving of this technique, made it possible to control the size of grain. Different forms of diamonds have been developed (with smaller grains) like nanocrystalline diamond (NCD) and ultra-nanocrystalline diamond (UNCD). These diamond materials have different structures and uses, but, what is the most important, they have significantly better characteristics in terms of mechanical strength and tribological properties in relation to MCD.

An efficient method for applying a diamond film of very high quality on large areas is to use hot filaments for the CVD method. There are variations of this method, but only a few variations are commercialized and applied for coating of tools for machine processing, cutting or surface protection.

The progress in development of the CVD method enabled control of the grain size, from micrometric to nanometric size. Grains, despite of their size, are always separated by non-diamond material such as graphite and C-H bonds with sp^2 and sp^3 hybridized carbon [1]. For that reason, any variation in the grain size of crystal can lead to a major impact on the total amount of atoms which are in contact with grain boundaries. In that sense, depending on the microstructure, the properties can be determined, and thus, the application of this materials.

Unlike natural diamond, which is an excellent electrical insulator with electrical resistivity of 10^{12}-10^{16} Ωcm, polycrystalline diamonds are very good electrical conductors and their conductivity varies between 10^{-4} and 300 S/cm [4]. For that reason, many intensive researches of electrical conductivity of polycrystalline diamonds can be found in literature. These novel materials are very interesting for application in MEMS, because of their mechanical and electrical properties. There is a requirement for the application of MEMS in sensitive and specific conditions [5].

Considering that polycrystalline diamonds have very good characteristics, their application in MEMS components is more frequent. The newest researches are related to finding the way to replace MEMS components in sensors with nanoelectromechanical systems (NEMS). With that replacement, there is a possibility to increase the resonant frequency and thus, the sensitivity of the sensor structure. NEMS structures can be made of different materials and because of that, there is a possibility to optimize individual properties such as hardness, low dissipation, compatibility with the work area, easy way for producing and better integration [6].

Structural properties of the high conductivity UNCD were especially investigated in order to determine the origin of electrical conductivity. During researching, it was observed that a small amount of nano-graphite exists between diamond grains. The mechanism which is responsible for conductivity in UNCD is attributed to the effect of πsp^2 bounds at the grain boundaries. The connection between structural properties of sp^2 carbon bounds and specific conductivity is observed. It can be said that the specific conductivity is firstly affected by structural properties of grain boundaries. Conductivity of thin films at room temperature can be explained by very low activation energies (meV), of sp^2 carbon bounds.

One of possible uses of polycrystalline diamonds is in radiofrequency (RF) MEMS resonators, and these materials were examined for this application [7]. The scanning electron microscopy (SEM) images of polycrystalline diamonds structures are shown in Fig. 2.

Also, the interest in the application of polycrystalline diamonds in sensors (piezo-resistant, temperature, gas and biosensors) has been increased due to their unique combination of properties [8]. Among other properties, they are chemically inert, resistant to corrosion, they have flexible modification of the surface, but the most important is that they are not toxic for humans and animals.

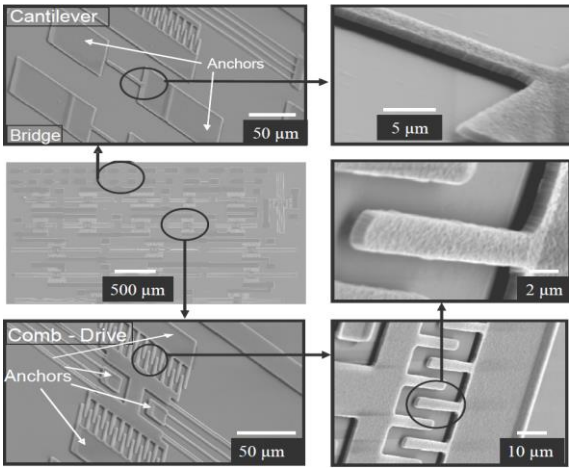


Fig .2. SEM image of polycrystalline diamond structures [7].

Fig. 3. presents the process of diamond probes creation, which have an undoped diamond layer (10^7 Ω-cm) and it serves as an insulator. The diamond layer which is doped with boron (10^3 Ω-cm) is used for bonds and electrodes.

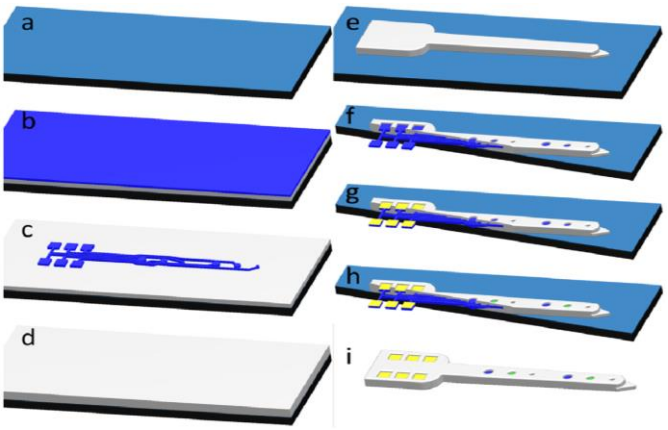


Fig. 3. Process of forming diamonds probes: (a) Si/SiO_2 substrate (b) undoped and doped diamond growth; (c) etching of doped diamond layer; (d) growing of undoped diamond layer; (e) etching undoped diamond layer for shaping probe; (f) undoped diamond etch to expose pads and electrodes; (g) gold bonding; (h) optional diamond functionalization; (i) release of probes in HF [8].

Fig. 3a illustrates Si/SiO_2 substrate, in 3b grow of a doped diamond layer can be seen and 3c shows etching of a doped diamond layer in order to define pads and electrodes. Next, 3d to 3f represent the growing and etching of an undoped diamond layer because probes and electrodes should be formed. Gold bonding and release of probes in HF are shown in 3g to 3i parts of the figure.

The usage of polycrystalline diamonds in the mechanical industry is very often, and especially for cutting, drilling or treating materials which are very complicated for processing such as carbon fiber reinforced plastics, plastic foils with metal additives. Also, one very interesting application of NCD is application in industry. For example, the tool of carbide tungsten is coated with NCD or NCD cutter blade for plastic films. In both cases NCD layer significantly extends the lifetime of the tool. The optical and thermal properties of microcrystalline diamonds with large grains are very useful for optical windows (for high–power laser window, vacuum windows, microwave windows) or head spreaders. Thick free-standing diamond disks up to several hundreds of micrometers can be grown and mechanically polished in order to achieve smooth surfaces.

In terms of tribological properties, the high wear-resistance and the low friction coefficient of diamond have very important role. NCD diamonds with smooth surfaces and high mechanical stability are excellent as the surface protective layer. The high shock and wear-resistance becomes highly effective when it finds application in bearings, pump seals or NCD-coated Si_3N_4 spheres. These spheres are used because of tribological properties, to improve their reliability and lifetime. Furthermore, the precise grow and microfabrication of NCD enabled the production of structures which are completely made of diamonds. Examples of these components are atomic force microscopy (AFM) probes which have a performance comparable to standard probes but with over 100 times lower wear rate [1].

The combination of the high wear-resistance and the low friction coefficient of smooth diamond surface is ideal for micromechanical systems. Very interesting and novel application of these materials is found in mechanical watch movements where important parameters are reliability, lifetime and accuracy. The production of mechanical watches represents one of few traditional crafts standing opposite to ever-increasing automation and mass production, which are increasingly present. However, traditional materials such as steel, nickel and ruby which are used in traditional manufacturing, now can be replaced with silicon and synthetic diamonds. The main aim of watch manufacturers is to produce the oil-free watch mechanism, which explains the demand for materials with very good tribological properties [9].

IV. Analysis of Polycrystalline Diamonds Surface

Fig. 4. represents a SEM image of polycrystalline diamonds obtained in the Institute of Functional Nanosystems in Ulm University, Germany. Wafers with a diameter of 6 inches were used for substrate silicon. The crystallographic orientation of these wafers was <100>. This sample was grown using the CVD method, but before growing the substrate was pre-treated. The solution which was obtained with container-shock synthesis, contained particles of diamonds. The average value of particles was about 30 nm. Clusters which were formed in this way had dimensions less than 100 nm. The substrate was ultrasonicated for almost 30 seconds and afterwards washed in isopropanol and deionized water. Diamond seeds were used for further initiation of grow nanocrystalline diamond films.

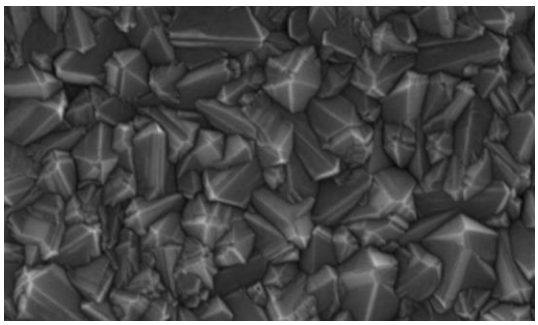


Fig. 4. SEM image of polycrystalline diamond surface.

In the production process of nanocrystalline diamond films CameCon CC800/Dia a reactor with wolfram filaments was used. The setup used for the CVD method contains a gas-cabinet in which there were CH_4, H_2 and several other gases. The source of gases was connected via mass flow controllers to the vacuum chamber. In the chamber, six rows for hot filaments and five positions for substrate could be located, which leads to 3D deposition. The pressure in the chamber can be reduced up to 1 Pa. In the vacuum-chamber there was a gas-shower which shall distribute the gas flow over the whole chamber cross-section. On the gas shower, were also mounted the filament holders which hold the wolfram filaments. Those were heated by an electric current to around 2000°C. The sample was held by a sample holder and the pressure in the chamber was controlled by a vacuum pump and a butterfly valve. The gas pressure during the process was 2000 Pa, while the hydrogen flow rate was set to 1500 sccm (standard cubic centimeters per minute) and the CH_4 flow to 20 sccm (1.3% CH_4/H_2). The process duration was 5 hours.

In order to visualize and analyze data, a modular program for scanning probe microscopy was used. In Fig. 5 is presented 3D view of the polycrystalline diamond surface of a

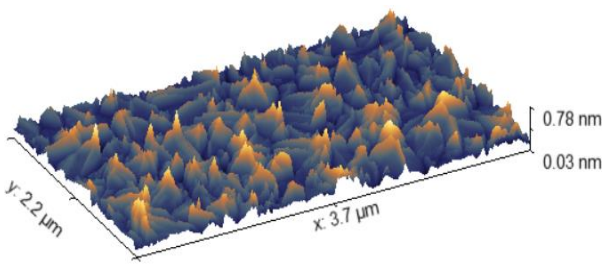


Fig. 5. 3D view of polycrystalline diamond surface.

grown film. Dimensions of the analyzed part of the surface are 2.2 μm and 3.7 μm, while the height of grains is in the range from 0.03 nm to 0.78 nm.

For further analysis standardized one-dimensional texture parameters in x and in y dimension were evaluated separately. In Fig. 6 are denoted specific A and A' (horizontal) and B and B' (vertical) directions for which the evaluation was performed. In the same picture, on the right side, can be observed a scale on which the shades of gray show the height of grains. The darkest gray represents the smallest height and the brightest gray represents the heights value of the grain. Obtained results are presented in figures 7 and 8.

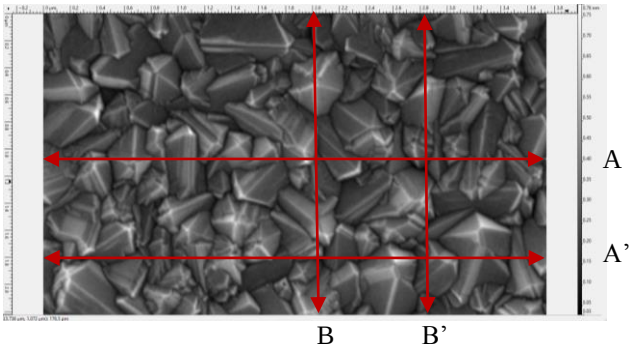


Fig. 6. Specific A and B directions for which evaluation was performed.

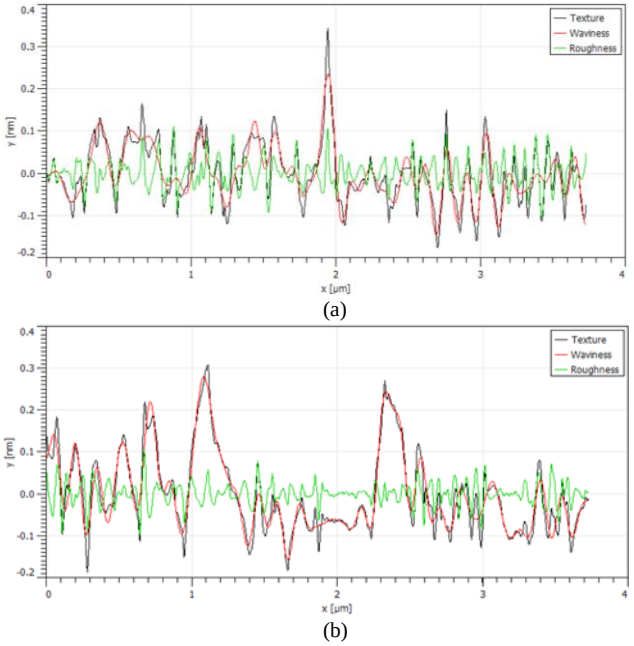
The one-dimensional texture is split to waviness (the low-frequency components defining the overall shape) and roughness (the high-frequency components) at the cut-off frequency. This frequency is specified in the units of the Nyquist frequency, that is the value of 1.0 corresponds to the Nyquist frequency. It is assumed that the mean value of r_j is zero, i.e.

$$r_j = z_j - \bar{z}. \quad (1)$$

where r_j is roughness, z_j is pixel-centered vertices and $\bar{z}$ is the mean value of z.

In Fig. 7. and Fig. 8. it can be seen that the texture and waviness are practically the same for all directions. The most pronounced peaks correspond to the highest parts of the grains for the chosen direction. Although, the waviness and texture are unique for any chosen direction, their values vary approximately from -0.2 nm to almost 0.4 nm for the presented directions. Maximal values of texture and waviness correspond to the highest parts of grains. It can be noticed that in direction A only one grain is dominant, while in direction A' several gains stand out and that caused the appearance of several peaks in Fig. 7b. Although, in direction B and B' there is almost the same number of prominent grains, in direction B' one grain (at position on 0.5 nm) is dominant.

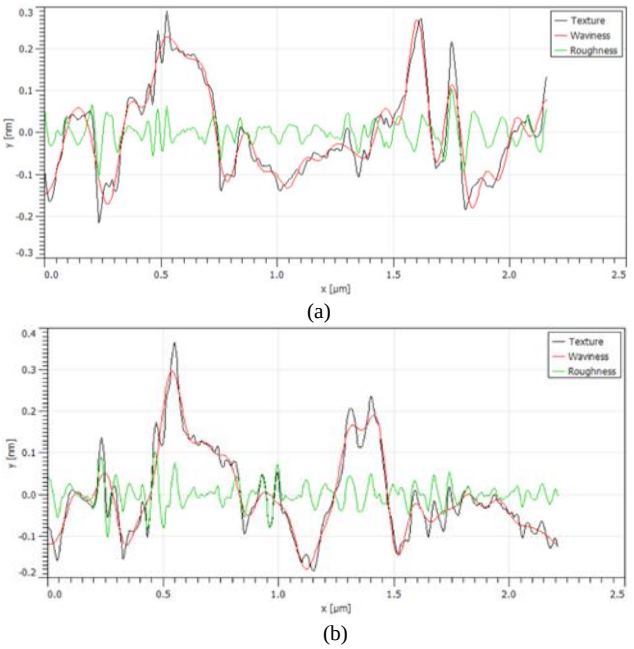
Also, Fig. 7. and Fig. 8. display that roughness as a component of surface texture, quantified by the deviations in the direction of the normal vector of a real surface from its ideal form, vary in all directions. In analyzing of surface roughness is typically considered to be the high-frequency component of a measured surface. However, in practice it is often necessary to know both the amplitude and frequency to ensure that a surface is fit for a purpose. Roughness is often a good predictor of the performance of a mechanical component, since irregularities may affect the properties of diamond films. Mechanical properties like Young's modulus and hardness strongly depend on grains size and structure. A detailed analysis of surface characteristics is also very important considering that pressure and gas mixture strongly influence the grain size grow. By varying these growth parameters, it is possible to influence properties of polycrystalline diamond films.



(a)

(b)

Fig. 7. Texture parameters of polycrystalline surface in A (a) and A' (b) direction.



(a)

(b)

Fig. 8. Texture parameters of polycrystalline surface in B (a) and B' (b) direction.

Additional parameters which can be useful for analysis are Root Mean Square Roughness (R_q,), the Amplitude Distribution Function (ADF), the Bearing Ratio Curve (BRC) and Skewness (R_{sk}).

The Root Mean Square Roughness represents the average of the measured height deviations taken within the evaluation length and measured from the mean line:

$$R_q = \sqrt{\frac{1}{N}\sum_{j=1}^{N} r_j^2} \tag{2}$$

where N is the number of samples along the assessment length. The value of this parameter R_q for the A direction is 33.79 pm.

The amplitude distribution function is a probability function that gives the probability that a profile of the surface has a certain height z at any position x. Obtained results for this parameter for the A direction are presented in Fig. 9.

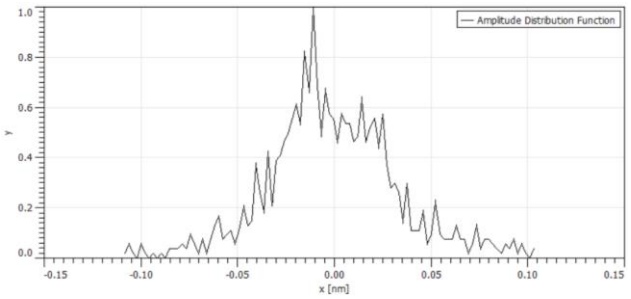


Fig. 9. The amplitude distribution function of polycrystalline surface in A direction.

The Bearing Ratio Curve is related to ADF, it is the corresponding cumulative probability distribution and it sees much greater use in surface finish. The bearing ratio curve is the integral (from the top down) of ADF.

Skewness is a parameter that describes the shape of ADF. Skewness is a simple measure of the asymmetry of ADF, or, equivalently, it measures the symmetry of the variation of a profile about its mean line which can be calculated by:

$$R_{sk} = \frac{1}{NR_q^3}\sum_{j=1}^{N} r_j^3 \tag{3}$$

The value of this parameter R_{sk} for the A direction is 0.1597.

V. Conclusion

Excellent electrical, thermal, mechanical and tribological properties are the reason why polycrystalline diamonds can be applied in many areas, in industry, medicine and especially for advanced electronics. Polycrystalline diamonds have a very important role in microelectronic and microelectromechanical systems as material used for sensors and probes. It is necessary to be familiar with characteristics of polycrystalline diamonds in order to improve current applications and also, to develop new ones. Considering that mechanical properties like the Young's modulus and hardness strongly depend on grains size and structure, a detailed analysis of surface characteristics, which includes various parameters, is also very important. Having in mind that pressure and gas mixture have a strong influence on the grain size grow, it is necessary to carefully investigate variations in these parameters.

In this paper an analysis of the polycrystalline diamond surface structure was performed. It was observed that for chosen directions, the texture feature is similar to the waviness. Although, waviness and texture are unique for any direction, their values are almost the same for the chosen directions and vary approximately from -0.2 nm to almost 0.4 nm. By varying growth parameters, it is possible to influence the properties of polycrystalline diamond films for specific applications.

Acknowledgment

These researches represent result which are a part of project OI 172057, which is financially supported of Serbian Ministry of Education, Science and Technological Development. Also, this paper is realized in cooperation with the Institute of Functional Nanosystems, University of Ulm, in Germany.

References

[1] Matthias Wiora, "Characterization of nanocrystalline diamond coatings for micro-mechanical applications", dissertation, University Ulm, 2013

[2] J. Asmussen and D.K. Reinhard, "Diamond Films Handbook", CRC Press, 2002.

[3] J.E. Field, "Properties of natural and synthetic diamond", Academic Press: London, 1992.

[4] M. Mertens, I.-N. Lin, D. Manoharan, A. Moeinian, K. Brühne, and H. J. Fecht, "Structural properties of highly conductive ultra-nanocrystalline diamond films grown by hot-filament CVD", AIP Advances, Vol. 7, arct. no. 015312, 2017.

[5] H.-J. Fecht, K. Br¨uhne, and P. Gluche, "Carbon-Based Nanomaterials and Hybrid Synthesis, Properties, and Commercial Applications", Pan Stanford, Singapore, 2014.

[6] L. Sekaric, J. M. Parpia, and H. G. Craighead, "Nanomechanical resonant structures in nanocrystalline diamond", Applied Physics Letters, Vol. 81, no. 23, pp. 4455-4457, 2002.

[7] Nelson Sepúlveda-Alancastro, Dissertation: Polycrystalline diamond RF MEMS resonator technology and characterization, Michigan State University, 2005.

[8] M.W. Varney, D.M. Aslam, A. Janoudi, H.Y Chan, D.H. Wang, "Polycrystalline-Diamond MEMS Biosensors Including Neural Microelectrode-Arrays", Biosensors, Vol. 1, no. 3, pp. 118-133, 2011.

[9] Sandra Veljković, Vojislav V. Mitić, Vesna Paunović, Goran Lazović, Markus Mohr, Hans Fecht, Karakteristike i primene polikristalnih dijamanata, IEEESTEC 2018, 11th International students projects conference.

Transport parameters of Ar^+ in Ar/BF_3 mixtures

Željka Nikitović

Abstract— **In this paper we present a cross section set for Ar^+ in Ar/BF_3 mixtures where existing experimentally obtained data are selected and extrapolated. A Monte Carlo simulation method is applied to accurately calculate transport parameters in hydrodynamic regime. We discuss new data for Ar^+ ions in Ar/BF_3 mixtures where mean energy, flux and bulk values of reduced mobility and other transport coefficients are given as a function of low and moderate reduced electric fields E/N (E-electric field, N-gas density).**

Key words— Ar/BF_3 mixtures, positive ions Ar^+, Monte Carlo simulations.

I. INTRODUCTION

Cold plasmas are frequently used in new technologies where they open up possibilities of non-intrusive production or modification of various substances [1]. These plasmas have high electron temperature and low gas temperature so non-equilibrium behavior of a large number of species becomes important [2]. Current computer resources allow studies of complex global models [3] which describe the behavior of such plasmas by taking into account a very large number of particles. The knowledge of ion-neutral reactions is generally accessible [4] although effects of reactions on transport parameters of particular ions are much less studied due to inability of instrumentation to detect rapidly vanishing ionic fluxes. This especially holds for ions whose transport is affected by fast reactions [5, 6].

Transport of Ar^+ plays significant role in various etching and deposition processes [7,8], in dark matter detection [9] and many more applications. It is known that transport parameters of Ar^+ in Ar are affected by resonant charge transfer reactions [10] leaving slow ions as a result. Charge transfer reactions are also a main collisional events in BF_3 gas where they introduce neutralization of Ar^+ ions [11]. The large rate coefficient for exothermic reactions (recombination energy of the ion is higher than the ionization potential of the gas particles) limits number of ions necessary to determine ion mobility. Boron dopant penetration in silicon is technologically achieved by DC pulsed plasma system (PLAD) most widely applying BF_3 gas [12, 13]. Uniform plasma and implantation with normal ion incidence are the main goals in this technological process. By using technique of Monte Carlo simulations one may calculate transport parameters for the cases that are out of the reach of experimental efforts provided complete cross section set is known. Reduced mobility data of both flux and bulk values as a function of *E/N*, are expected to be significantly affected by the presence of endothermic and exothermic reactions in the case of Ar^+ transport in the Ar/BF_3 mixtures.

Željka Nikitović – Institute of Physics, University of Belgrade, Pregrevica 118, 11080 Belgrade,Serbia (e-mail: zeljka@ipb.ac.rs).

II. CROSS SECTION SETS

A complete cross section sets for ion transport is scarce in spite of a broad range of specific methods relevant for quantification of particular cross sections. The main problem in heavy particle scattering, easily and precisely selecting the state of the projectile and target before the collision, is still very complicated for range of conditions, so databases for ion scattering [13, 4] are still devoid of such data. Phelps established the first worldwide accessible database with cross section sets [14] tested for each particular case either for swarm conditions of spatially resolved measurements of emission or ion mobility values. Another range of cross section sets was established by measurements of ionic transport coefficients [4] and this work is ongoing. In all cases only the most important cross sections may be established from the transport data. In the following section we will establish a complete cross section set for Ar^+ scattering on Ar and BF_3 from 0.1 meV to 1000 eV which will be used to calculate transport properties. Generally one may distinguish three characteristic energy ranges: the low energy regime where polarization scattering is dominant, the medium energy regime where polarization scattering is gradually replaced by hard sphere repulsion, and the high energy approximation regime.

Extensive discussion about transport properties of Ar^+ ions scattering in BF_3 gas applied to plasma physics problems was presented by Phelps [15] and Petrović and Stojanović [16]. Analytical expressions were offered in [11] to express apparently isotropic and anisotropic components of the cross section set (see Fig. 1). In order to focus on effects of reactive processes introduced by BF_3 we neglected all but these two components of the Ar^+ + Ar cross section set.

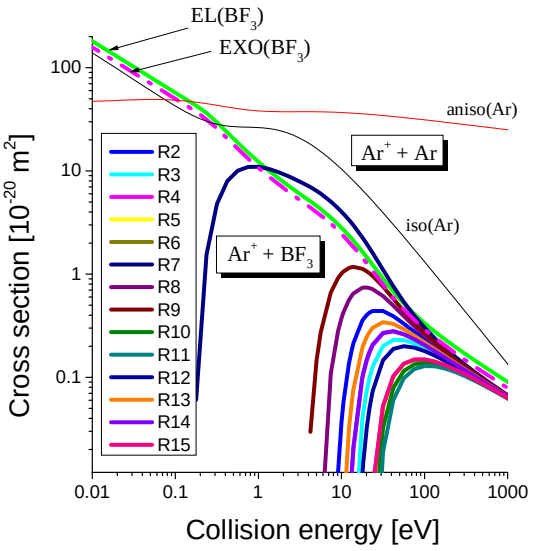


Fig.1. Cross sections for Ar^+ in BF_3 gas.

No	Reactions	Products	Threshold (eV)
1	EL	$Ar^+ + BF_3$ (el)	0.0000
2	R2	$Ar^+ + BF_2 + F$	7.4373
3	R3	$Ar^+ + BF + 2F$	12.2940
4	R4	$Ar^+ + BF + F2$	10.6461
5	R5	$Ar^+ + B + 3F$	20.1367
6	R6	$Ar^+ + B + F + F_2$	18.4889
7	R7	$BF_2+ + Ar + F$	0.1544
8	R8	$BF^+ + 2F + Ar$	5.0115
9	R9	$BF^+ + F_2 + Ar$	3.3637
10	R10	$F^+ + B + F_2 + Ar$	20.1488
11	R11	$F^+ + B + 2F + Ar$	21.7966
12	R12	$F^+ + BF + F + Ar$	13.9539
13	R13	$F^+ + BF_2 + Ar$	9.0972
14	R14	$F_2^+ + BF + Ar$	10.5846
15	R15	$F_2^+ + B + F + Ar$	18.4273
16	EXO	$BF_3^+ + Ar$	-0.2125

Table 1. Reaction products and thermodynamic thresholds for Ar+ + BF_3.

Complete cross section sets used in this work are shown in Fig. 1.

In this paper there are 15 endothermic and one exothermic processes. Therefore we have assumed in this work that exothermic processes are non-resonant, and neglected their effect on transport properties.

III. TRANSPORT PARAMETERS

Transport properties of species in gas plasmas are of great importance in understanding the nature of molecular and ionic interactions in gas mixtures [17, 18]. These properties include the mean energy, drift velocity, diffusion coefficients, ionization and chemical reaction coefficients, chemical reaction coefficients for ions and rarely. A Monte Carlo simulation code appropriate to calculate transport parameters [16, 17] of Ar^+ ions in Ar/BF_3 mixtures at non-zero temperature [19] has been used. In Monte Carlo simulations exothermic reactive collisions are followed in a similar way as all non-conservative collisions i.e. the swarm particle dissappears from the ensemble after exothermic collisions. This results in changes of the swarm particle number in the entire energy range introducing nonconserve effects in kinetic equations and thus division of transport parameters to flux and bulk ones [17].

In Fig. 2 we show results for mean energy as a function of *E/N* (*E*- electric field and *N*-gas density). Significant reduction and uniform control of mean energy of Ar^+ is obtained for argon content below 90 %. For Ar content below 10 % largest variations of mean energy are obtained for *E/N*>100 Td (1Td=10^{-21}Vm^2). These variations are the consequence of a reduced momentum transfer cross section for Ar^+ scattering with BF_3 (see Fig. 1) as compared to the scattering with Ar. At a high content of Ar charge transfer collisions dominate and make variation of mean energy with Ar content more uniform. Note, that for transport coefficients of Ar^+ in pure Ar one may find benchmark data presented in tabular form by Ristivojević and Petrović [19]. In Fig.3 we show the characteristic energies (diffusion coefficient normalized to mobility eD/K in units of eV) based on transversal (D_T) diffusion coefficients. Bulk and flux values of the characteristic energies are well matched.

In Fig. 4 variations of reduced mobility as a function of *E/N* are shown. The mobility *K* of an ion is a quantity defined as the velocity attained by an ion moving through a gas under the unit electric field. One often exploits the reduced or standard mobility defined as:

$$K_0 = \frac{v_d}{N_0 E} N \,, \qquad (1)$$

where v_d is the drift velocity of the ion, *N* is the gas density at elevated temperature *T*, N_0= $2.69\cdot10^{25}$ m^{-3} and *E* is the electric field. Behaviour of reduced mobility significantly changed with *E/N* with small additions of Ar, up to about 10 %. Especially intriguing are variations of flux reduced mobility which points to particle flux variations of Ar^+ ions as a function of their mean energy. On one side, small additions of Ar cause significant variations of particle flux at *E/N* close to 200 Td and on the other in that region we obtain significant difference of the flux and bulk drift velocities due to the BF_3 reactive processes [see also 10] which also significantly reduce Ar^+ density in favour of fast Ar [11]. Control of fast Ar flux at the surface thus can be easily achieved by small variations of Ar contents in Ar/BF_3 mixture. Due to reactive collisions bulk and flux values of reduced mobility are separated.

Longitudinal diffusion coefficients for Ar^+ ions in Ar/BF_3 mixtures as a function of *E/N* are shown in Fig. 5. Note that the difference between the flux and bulk values of diffusion coefficients, which have the same origin, have the same initial value as drift velocities. There are no published experimental data for the longitudinal and transverse diffusion coefficients of Ar^+ in Ar/BF_3.

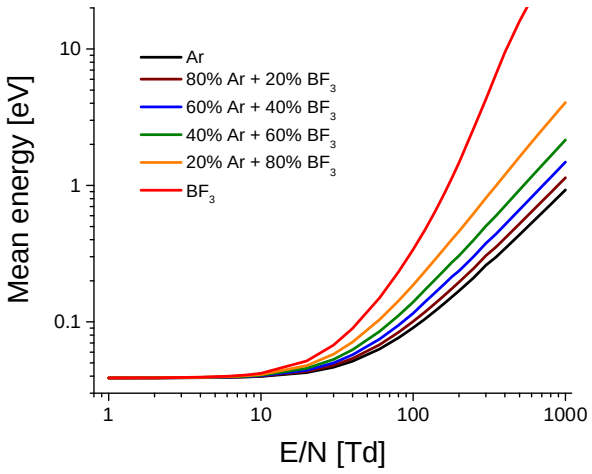


Fig.2. Mean energy as a function *E/N* for Ar^+ in Ar/BF_3 gas.

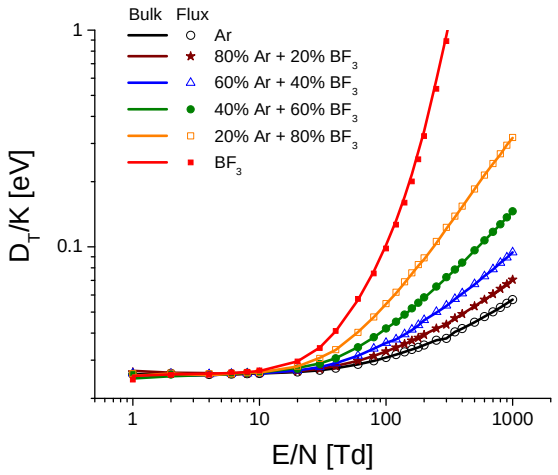


Fig.3. Characteristic energy as a function *E/N* for Ar^+ in Ar/BF_3 gas.

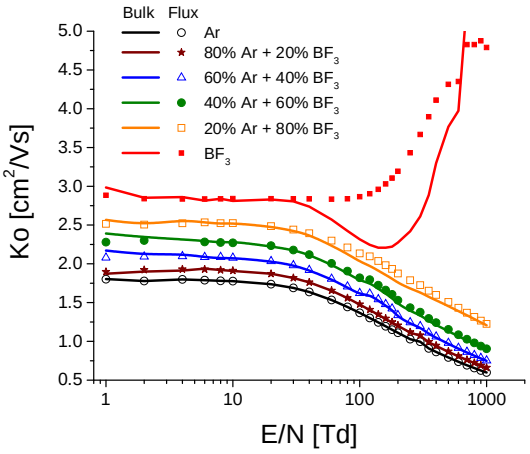


Fig.4. Reduced mobility as a function *E/N* for Ar^+ in Ar/BF_3 gas.

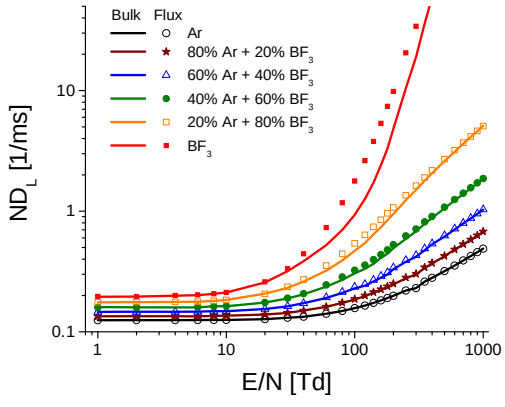


Fig.5. Longitudinal and diffusion coefficients as a function *E/N* for Ar^+ in Ar/BF_3 gas.

IV. CONCLUSION

Data for swarm parameters for ions are needed for hybrid and fluid codes and the current focus on liquids or liquids in the mixtures with rare gases dictates the need to produce data compatible with those models. In this paper we show transport parameters for the Ar^+ in Ar/BF_3 mixtures which do not exist in the literature. In addition to presenting the data we show here the effects of non-conservative collisions to ion transport. Due to exothermic cross sections that are dominant at low energies, for small abundances of Ar (<10 %) exothermic process may be larger than the elastic scattering cross section.

The Monte Carlo technique was applied to carry out calculations of the mean energy, reduced mobility, diffusion coefficients as a function of reduced DC electric field. The results are believed to be a good base for modeling, which could be further improved when measured values of transport coefficients become available and then we could perform this analysis again.

ACKNOWLEDGMENTS

Results obtained in the Institute of Physics University of Belgrade under the auspices of the Ministry of Education, Science and Technology, Projects OI 171037and III 45016.

Author is also grateful to Vladimir Stojanović and Zoran Raspopović.

REFERENCES

[1] T. Makabe, Z. Petrović, Plasma Electronics: Applications in Microelectronic Device Fabrication Taylor and Francis, CRC Press, New York, 2006.

[2] R. E. Robson, R. D. White and Z. Lj. Petrović., "Colloquium: Physically based fluid modeling of collisionally dominated low-temperature plasmas", Rev. Mod. Phys. **77,** pp.1303, 2005.

[3] T. Murakami, K. Niemi, T. Gans, D. O'Connell and W. G. Graham, "Interacting kinetics of neutral and ionic species in an atmospheric-pressure helium–oxygen plasma with humid air impurities" Plasma Sources Sci. Technol. **22,** 015003, 2013.

[4] https://nl.lxcat.net/data/set_type.php

[5] Ž. Nikitović, Z., Raspopović, V. Stojanović and J. Jovanović, "Transport parameters of F^- ions in Ar/BF_3 mixtures", EPL **108** , 35004, 2014.

[6] Ž. Nikitović, M. Gilić, Z. Raspopović and V. Stojanović, "Comparison between transport parameters for K^+ and Li^+ in 1, 2-dimethoxy ethane (DXE) gas", EPL **116,** 15002, 2016.

[7] V. Stojanović, Z. Raspopović, J. Jovanović, Ž. Nikitović and Z. Lj. Petrović, "Transport of F^- ions in F_2", EPL 101, 45003, 2013.

[8] M. A. Lieberman and A. J. Lichtenberg, Principles of Plasma Discharges and Materials Processing, Wiley, New York, 1994.

[9] A. Kaboth, J. Monroe, S. Ahlen , D. Dujmic, S. Henderson, G. Kohse, R. Lanza, M. Lewandowska, A. Roccaro, G. Sciolla, N. Skvorodnev, H. Tomita., R.Vanderspek, H. Wellenstein, R.Yamamoto, P. Fisher, "Measurement of Photon Production in Electron Avalanches in CF_4",Nucl. Instrum. and Meth. in Phys. Res. **A 592** pp. 63-92, 2008.

[10] E. Parker and F. S. M El-Ashhab., "Charge-transfer reactions of carbon tetrafluoride", Int. J. Mass Spectrom. Ion. Phys. **47,** pp. 159-162, 1983.

[11] A.V. Phelps, "The application of scattering cross sections to ion flux models in discharge sheaths", J. Appl. Phys. **76,** pp. 747, 1994.

[12] B.-W Koo., Z. Fang, L. Godet, S. B. Radovanov, C. Cardinaud, G. Cartry, A. Grouillet and D. Lenoble, IEEE Trans. Plasma Sci., **32,** pp. 456-463, 2004.

[13] Ž. Nikitović, S. Radovanov, L. Godet, Z. Raspopović, O. Šašić, V. Stojanović, Z. Lj. Petrović, "Measurements and modeling of electron energy distributions in the afterglow of a pulsed discharge in BF_3", EPL, **95,** 45003 (2011).
[14] www.ruf.rice.edu/~atmol
[15] A. Phelps database, private communication, www.lxcat.net, retrieved on February 4, 2019.
[16] Z. Lj. Petrović, V. D. Stojanović "The role of heavy particles in kinetics of low current discharges in argon at high electric field to gas number density ratio,"J. Vac. Sci. Technol A **16** (1), pp. 329-336, 1998.
[17]] R. E. Robson, "Transport Phenomena in the Presence of Reactions: Definition and measurement of transport coefficients", Aust. J. Phys.44, pp. 685-692, 1991.
[18]Z. Raspopović, S. Sakadžić, Z. Lj. Petrović and T. Makabe," Diffusion of electrons in time-dependent E(**t**)xB(**t**) fields", J. Phys. D 33, pp.1298-1302, 2000.
[19] Z. Ristivojević and Z. Lj. Petrović, "A Monte Carlo simulation of ion transport at finite temperatures", Plasma Sources Sci. Technol. Vol. 21, 035001, 2012.

Synthesis and characterization of Ti_3C_2 MXene film

Ivan Pešić, Daniel Mijailović, Vukašin Ugrinović, Miodrag Mitrić, Petar Uskoković, Vesna Radojević

***Abstract* — Two-dimensional (2D) transition metal carbides are known as MXenes can offer large surface area, excellent electrical conductivity and chemical stability. MXenes have shown great potential in a broad spectrum of applications such as photothermal cancer therapy, antibacterial effect, the improved electrical conductivity of polymers, hydrogen evolution reaction, energy storage, etc. Herein we report a successful synthesis and characterization of Ti_3C_2 MXene material. Properties of the material are tracked via scanning electron microscopy (SEM), X-ray diffraction (XRD), cyclic voltammetry (CV) and galvanostatic charge/discharge (GCD) experiments. SEM and XRD analysis revealed delaminated MXene structure, while XRD patterns also showed the presence of Mxenes structure, CV and GCD showed as electrode material for aqueous supercapacitors.**

***Index Terms* — Nanomaterials; MXenes; Ti_3C_2; Supercapactitor**

I. Introduction

The last few decades, the exponential development of technology reveals a large number of novel materials with specific physical and mechanical properties. MXenes represent class of 2D materials with the structure, analogue to graphene, *i.e.* 2D monolayers with very weak bonds between the layers They cover carbides and nitrides of transition metals with a unique combination of properties: excellent electrical conductivity, good mechanical properties, high thermal conductivity, large specific surface area and high resistance to thermal shock [1]. Not long after the discovery of the graphene [2], more materials could be found that can be classified as 2D materials. MXenes are the tertiary layered structure of carbide or nitrides of transition metals (most often from the third and fourth groups of the periodic table of elements). So far, more than 70 possible structures have been presumed, but only some have been successfully synthesized. Due to the synthesis process itself, the structure of surface is enriched with certain ions (depending on the synthesis method) such as (F^-, OH^- and / or O^-) which significantly affect the properties of the obtained MXens [3].

Typical commercial batteries require prolonged charging and therefore are limiting mobility of users. Pseudocapacitors show excellent ability to store a large amount of energy per unit mass.They are a sub-class of supercapacitors that are differentiated from electrical double-layer capacitors on the basis of charge storage mechanism store charge via formation of the electrical double layer at the electrode/electrolyte interface and, naturally, their capacitance is proportional to the electrode's specific surface area available for electrosorption of ions [4]. MXenes have the outstanding electrochemical performances which makes them perfect candidates for energy storage applications because: a conductive inner transition metal carbide layer enables fast electron supply to electrochemically active sites.

Ivan Pešić is with the faculty of Technology and Metalrgy, University of Belgrade, 4 Karnegijeva, 11000 Belgrade, Serbia (email: pesicivan2@gmail.com)

Daniel Mijailovic is with the innovation center of Faculty of Technology and Metalurgy, University of Belgrade, 4 Karnegijeva, 11000 Belgrade, Serbia (email: danielmijailovic@gmail.com)

Vukašin Ugrinović is with the innovation center of Faculty of Technology and Metalurgy, University of Belgrade, 4 Karnegijeva, 11000 Belgrade, Serbia (email: vugrinovic@gmail.com)

Miodrag Mitrić is with the Vinča institute of nuclear sciences, Mike Petrovića Alasa 12-14, 11351 Vinča, Belgrade, Serbia (email: mmitric@vin.bg.ac.rs)

Petar Uskoković is with the Faculty of Technology and Metallurgy, University of Belgrade, Karnegijeva 4, 11070 Belgrade, Serbia (email: puskokovic@tmf.bg.ac.rs)

Vesna Radojević is with the Faculty of Technology and Metallurgy, University of Belgrade, Karnegijeva 4, 11070 Belgrade, Serbia (e-mail: vesnar@tmf.bg.ac.rs).

II. Experimental section

1 g of Ti_3AlC_2 MAX phase is etched with HCl/LiF method [4,5]. 10 ml of concentrated hydrochloric acid (HCl) was mixed with 1g of LiF in teflon vial under magnetic stirring. After a few seconds, the 1 g of Ti_3AlC_2 MAX phase is added in portions. After 24h on a magnetic stirrer on 35 °C, the mixture is poured in the centrifuge tube and teflon vial is washed with deionized water (DI) and added to the tube. The tube is filled to 45 ml and put onto centrifugation (3500 rpm during 90 sec.). After several washing procedures, the sediment is redispersed in DI water and put into the ultrasonic treatment for 1h in an inert atmosphere. The supernatant is collected and used for further experiments.

The concentration of obtained supernatant was determined via a simple mass difference method. A certain amount is poured into a small vial and then the mass is measured before and after drying. Concentration was 2.82 mg/ml.

A. Electrochemical measurements

40 µl of suspension of active material was drop cast on a glassy carbon (GC) electrode and left to dry under the lamp. After 2h, the commercial Nafion solution was dropped above the dried sample and left under the lamp overnight. All the measurements were performed on Gamry potentiostat in 1 M KOH aqueous solution with three electrode (3E) setup: reference (saturated calomel electrode, SCE), counter (Pt mesh) and the working electrode [5].

B. SEM

The rest of the supernatant is vacuum filtered on DURAPORE MEMBRANE FILTERS (0.22μm). After filtering overnight, we obtained thin "cake" of MXenes which is used for SEM. Small pieces were cut and put into SEM without sputtering.

C. XRD

The above mentioned 'cake' is also used for this examination. Small piece of the sample was cut and put into the device.

III. Results and Discusion

Fig 1. shows cyclic voltammetry (CV) profiles collected at various potential sweep rates *i.e.*, 5, 10, 50 and 100 mV s^{-1}.. Rectangular-like shape of CV curves indicates typical capacitive behaviour in the potential range from -0.8 to -0.2V vs. SCE. By decreasing the sweep rate, the voltammetric charge increased which is usual for electric double-layer capacitors (EDLC). [6-9]. Namely, this kind of sweep rate dependence is due to limited electrolyte accessibility of internal structure of Ti_3C_2.

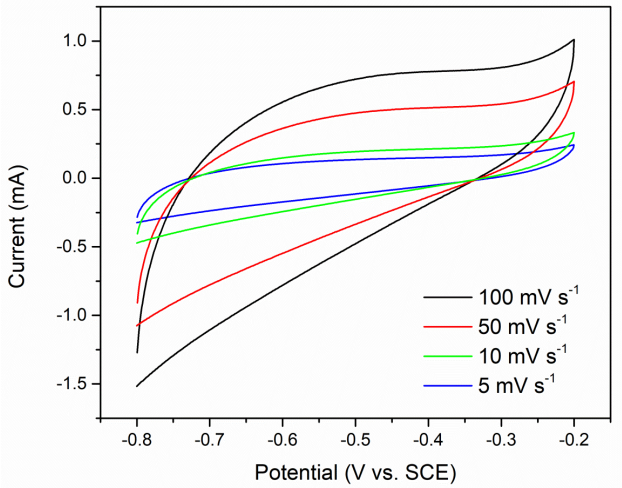


Fig. 1 - Electrochemical performance of $Ti_3C_2T_x$ MXene in 1M KOH electrolyte at different scanning speeds

The galvanostatic charge/discharge (GCD) curves at various current densities, *i.e.* 1, 5 and 10 Ag^{-1}, are shown in Fig. 2. It can be seen nearly triangular shape which is additional indicator of capacitive behavior.

Table 1 shows specific capacitance values of the sample at calculated from CV and GCD experiments.

TABLE I
SPECIFIC CAPACITANCES AT 5, 10, 50 AND 100 MVs^{-1}

ν (mVs^{-1})	C_s (Fg^{-1})
5	190.67
10	136.66
50	61.26
100	44.96

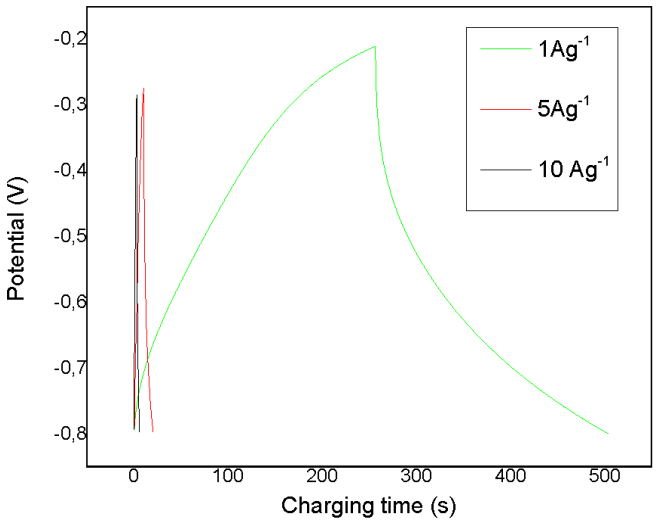
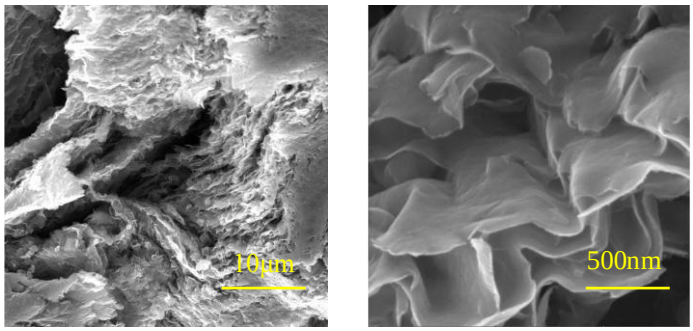


Fig. 2. Galvanostatic charge/dischagre diagrams at three different current densities. 2a - 1Ag^{-1}, 2b - 5Ag^{-1}, 10Ag^{-1}



1a 1b

Fig. 3 - SEM photographs of the sample: 1a 5000x magnification, 1b 100000x magnification

The Fig 3. shows SEM photographs of our sample. On photograph 1a the lamellar structure can be spotted. It can be seen that sheets of MXene are stacked but not 'melted'. Also, they are not completely flat, but that can be a good characteristic of the supercapacitve performance [10]. This phenomenon is more noticeable at photograph 1b (greater magnification). Interlayer space and sheets itself are more clearly visible.

On Fig 4, sharp peak at approximately 7º (corresponds to 002 diffraction) confirmes the interlayered distance between sheets of MXene [11-13]. Also, it can be noticed that 'main' peak (~7º) is significantly smaller after 1 day of aging. It loses intensity in time but the main drop is after the first day [14]. This loss of intensity can be interpretated as degradation of structure probably due to formation of TiO_2. The peak at 14 corresponds to 004 plane and it is almost completely lost after 2 days. This reveals one big disadvantage of MXenes and that is relatively fast degradation of the structure.

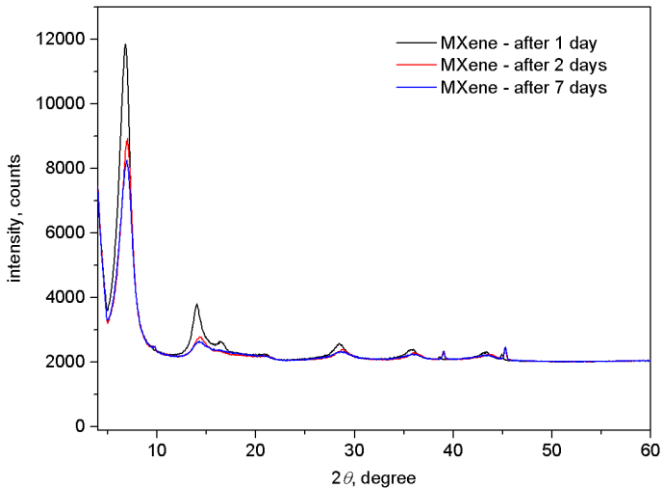


Fig. 4 - Three XRD patterns of our sample collected at different aging period: Black- after 1 day, Red after 2 days and Blue after 7 days

IV. Conclusion

Herein we report a successful synthesis of Ti_3C_2 MXene. XRD results clearly indicate MXene structure, SEM photographs show delaminated structure which is crucial for supercapacitive performance. Rectangular like shape also shows that our sample have capacitive properties. This specific combination of the synthesis and characterization of this material contributes to a better understanding of the process of obtaining MXenes, which are increasingly used to produce nanocomposites with very specific properties. Composites can be with a polymer matrix (PVA, PMMA, PE, PVDA ...) and can also be combined with ceramic and/or metal nanotubes (Active Carbon, Si, S, Pt ...)

References

[1] Maria R. Lukatskaya, Sankalp Kota, Zifeng Lin, Meng-Qiang Zhao, Netanel Shpigel, Mikhael D. Levi, Joseph Halim Pierre-Louis Taberna, MichelW. Barsoum, Patrice Simon, Yury Gogotsi, 'Ultra-high-rate pseudocapacitive energy storage in two-dimensional transition metal carbides', 'Nature Energy', 10.1038/nenergy.2017.105

[2] Ohta, T., Bostwick, A., Seyller, T., Horn, K. & Rotenberg, E. Controlling the electronic structure of bilayer graphene. Science 313, 951–954 (2006)

[3] Lang, X.-Y. et al. Ultrahigh-power pseudocapacitors based on ordered porous heterostructures of electron-correlated oxides. Adv. Sci. 3, 1500319 (2016).

[4] S.S.K. Mallineni, Y. Dong, H. Behlow, A.M. Rao, R. Podila, A wireless triboelectric nanogenerator, Adv. Energy Mater. 1702736 (2017)

[5] T.L. Tan, H.M. Jin, M.B. Sullivan, B. Anasori, Y. Gogotsi, High-throughput survey of ordering configurations in MXene alloys across compositions and temperatures ACS Nano 11, 4407–4418 (2017)

[6] C.J. Zhang, B. Anasori, A. Seral-Ascaso, S.-H. Park, N. McEvoy, A. Shmeliov, G.S. Duesberg, J.N. Coleman, Y. Gogotsi, V. Nicolosi, Transparent, flexible, and conductive 2D titanium carbide (MXene) films with high volumetric capacitance, Adv. Mater. 1702678 (2017)

[7] B. Anasori, M.R. Lukatskaya, Y. Gogotsi, 2D metal carbides and nitrides (MXenes) for energy storage, Nat. Rev. Mater. 2 16098 (2017)

[8] C.J. Zhang, B. Anasori, A. Seral-Ascaso, S.-H. Park, N. McEvoy, A. Shmeliov, G.S. Duesberg, J.N. Coleman, Y. Gogotsi, V. Nicolosi, Transparent, flexible, and conductive 2D titanium carbide (MXene) films with high volumetric capacitance, Adv. Mater. 1702678 (2017)

[9] M. Alhabeb, K. Maleski, B. Anasori, P. Lelyukh, L. Clark, S. Sin, Y. Gogotsi, Guidelines for synthesis and processing of two-dimensional titanium carbide (Ti3C2Tx MXene), Chem. Mater. (2017)

[10] M.A. Hope, A.C. Forse, K.J. Griffith, M.R. Lukatskaya, M. Ghidiu, Y. Gogotsi, C.P. Grey, NMR reveals the surface functionalisation of Ti3C2 MXene, Chem. Phys. 18 5099–5102 (2016)

[11] Ji, X.; Xu, K.; Chen, C.; Zhang, B.; Ruan, Y. J.; Liu, J.; Miao, L.; Jiang, J. J. Probing the Electrochemical Capacitance of MXene Nanosheets for High-Performance Pseudocapacitors. *Phys. Chem. Chem. Phys. 18*, 4460-4467 (2016).

[12] Bai, Y. L.; Zhou, K.; Srikanth, N.; Pang, J. H. L.; He, X. D.; Wang, R. G. Dependence of Elastic and Optical Properties on Surface Terminated Groups in Two-dimensional MXene Monolayers: A Firstprinciples Study. *RSC Adv. 35731-35739* (2016).

[13] Lanyong Yu, Longfeng Hu, Babak Anasori, Yi-Tao Liu, Qizhen Zhu, Peng Zhang, Yury Gogotsi, and Bin Xu,ACS Energy Lett. 3, 1597−1603, (2018)

[14] Lin, Zifeng and Rozier, Patrick and Duployer, Benjamin and Taberna, Pierre-Louis and Anasori, Babak and Gogotsi, Yury and Simon, Patrice Electrochemical and in-situ X-ray diffraction studies of Ti3C2Tx MXene in ionic liquid electrolyte. Electrochemistry Communications, vol. 72. pp. 50-53.1388- 2481, (2016)

Soft polymeric networks based on poly(methacrylic acid),itaconic acid, casein and liposomes for targeted delivery and controlled release of poorly water-soluble active substance

Maja Marković, Vesna Panić, Sanja Šešlija, Pavle Spasojević, Vukašin Ugrinović, Nevenka Bošković-Vragolović and Rada Pjanović

***Abstract*—Soft polymeric networks based on poly(methacrylic acid) (PMAA) are attractive candidates for targeted and controlled drug release due to their non-toxicity, biocompatibility and pH-sensitivity. The highly hydrophilic nature of PMAA networks enables transport of hydrophilic drugs only. This limitation has been overcome in present work by a PMAA modification with casein and liposomes. Casein is a natural amphiphilic protein which enabled the encapsulation and targeted and controlled release of the model drug- caffeine. The FTIR spectra showed that the hydrophobic interactions and hydrogen bonds were established between the casein and caffeine. The caffeine *in vitro* release was monitored in two media at 37°C: phosphate buffer pH=6.8, which simulated the pH environment in the human intestines and 0.1M HCl pH=1.2, which simulated the pH environment in the human stomach. The presence of liposomes with the encapsulated caffeine in the carriers caused the decrease in the speed of caffeine release. Introduction of itaconic acid (IA) as a hydrophilic and pH-sensitive substance with two carboxylic groups resulted in a non-regular structure of the carriers with large voids which caused an increase in swelling rate of the carriers and an increase in speed of caffeine release. All obtained results showed that the targeted and controlled release of a poorly water-soluble substance was achieved.**

***Index Terms*—Poly(methacrylic acid); itaconic acid; casein; liposomes; controlled release; poorly water-soluble drug; release kinetic**

Maja Marković is with the Innovation Center of Faculty of Technology and Metallurgy, University of Belgrade, 4 Karnegijeva Street, 11000 Belgrade, Serbia (e-mail: mmarkovic@tmf.bg.ac.rs).

Vesna Panić is with the Innovation Center of Faculty of Technology and Metallurgy, University of Belgrade, 4 Karnegijeva Street, 11000 Belgrade, Serbia (e-mail: vpanic@tmf.bg.ac.rs).

Sanja Šešlija is with the Institute of Chemistry, Technology and Metallurgy, University of Belgrade, 12 Njegoseva Street, 11000 Belgrade, Serbia (e-mail: sseslija@tmf.bg.ac.rs).

Pavle Spasojević is with the Faculty of Technical Sciences, University of Kragujevac, 65 Svetog Save Street, 32000 Čačak, Serbia (e-mail: pspasojevic@tmf.bg.ac.rs).

Vukašin Ugrinović is with the Innovation Center of Faculty of Technology and Metallurgy, University of Belgrade, 4 Karnegijeva Street, 11000 Belgrade, Serbia (e-mail: vugrinovic@tmf.bg.ac.rs).

Nevenka Bošković-Vragolović is with the Faculty of Technology and Metallurgy, University of Belgrade, 4 Karnegijeva Street, 11000 Belgrade, Serbia (e-mail: nevenka@tmf.bg.ac.rs).

Rada Pjanović is with the Faculty of Technology and Metallurgy, University of Belgrade, 4 Karnegijeva Street, 11000 Belgrade, Serbia (e-mail: rada@tmf.bg.ac.rs).

I. Introduction

DRUG delivery carriers for the targeted delivery of the active substances and their controlled release are intensively developed and used in treatment of some serious diseases [1]. pH sensitive soft polymeric networks- hydrogels have a great potential in this field of application because they respond to the pH changes in the external medium by swelling or contracting due to which they release their loadings. pH sensitive hydrogels of much interest for targeted drug delivery are based on poly(methacrylic acid) (PMAA). These hydrogels are biocompatible, non-toxic and contain a large number of ionisable –COOH groups. Ionization of carboxylic groups and generation of the negative charges on them occurs if the pH of the external medium is above the pKa of PMAA (4.6) causing the repulsion of the PMAA polymeric chains and swelling of the PMAA [2]. Although PMAA hydrogels have been shown to be good carriers of a hydrophilic active substance, the limitation for the use of these carriers in controlled release of the poorly water-soluble active substances is imposed by the hydrophilic nature of PMAA and relates to the poor interactions with a poorly water-soluble active substance [3]. In order to overcome this limitation PMAA hydrogels must be modified with amphiphilic substances such as some proteins and phospholipidic nanoparticles. Casein (the major milk protein) is a great candidate for targeted delivery of poorly water-soluble active substances due to its amphiphilic nature, pH sensitivity, non-toxicity and biocompatibility. It is also recognized as a GRAS protein and approved by American Food and Drug Administration. Spherical phospholipidic nanoparticles, such as liposomes which consist of one or more lipidic layers and an aqueous core, could be used for delivery and controlled release of both, hydrophilic and poorly water-soluble substances.

The goal of this research was to develop a hydrophilic polymer carrier for targeted delivery and controlled release of a poorly water-soluble substance. Hence, carriers for poorly water-soluble model drug-caffeine based on poly(methacrylic acid) and casein (PMAC), PMAC with itaconic acid (PMAC/IA) and PMAC with incorporated liposomes(PMAC-L) were synthesized. The successful entrapment of a drug, its

delivery to a specific place in the human body and the controlled release depend on a successful design of a carrier. Therefore, an investigation of the structure of synthesized carriers and of interactions established between the carrier and the drug was conducted. In addition, the influences of the presence of the itaconic acid and the presence of the liposomes on the caffeine release and on the swelling rate of the carriers were also analyzed. In order to control the drug release and to release a desirable concentration of a drug in the specific place in the human body it is important to get an insight into the drug release kinetic, hence two mathematical models were used for the investigation of caffeine release kinetic.

II. Materials and Methods

A. Materials

Methacrylic acid (MAA) (99.5%) was purchased from Merck, Germany. Sodium caseinate (C) powder, containing 88.9 wt. % of protein (the rest was the proteins, lipids, attached moisture and ashes) was supplied from Lactoprot, Deutchland GmbH (Germany). Itaconic acid (IA) (≥99%) was supplied from Aldrich Chemical Co. (USA). NATIPIDE® II containing phospholipids from soybean >20% (with 3-sn-phosphatidylcholine 76+ 3%) was purchased from Lipoid (Germany). Caffeine (Cf) was supplied from Merck (Germany). N, N'-Methylenebisacrylamide (MBA) (p.a.) and sodium hydroxide (p.a.) (NaOH) were supplied from Aldrich Chemical Co. (USA). The initiator, 2,2'-azobis-[2-(2-imidazolin-2-yl)propane] dihydrochloride (VA-044) (99.8%) was purchased from Wako Pure Chemical Industries (Japan). Monobasic sodium phosphate (anhydrous) (NaH_2PO4) and dibasic sodium phosphate (anhydrous) (Na_2HPO4) were purchased from Centrohem (Serbia). Hydrochloric acid (37%) was supplied from Zorka Pharma (Serbia). All chemicals were used as received.

B. Preparation of the samples

The PMAC and PMAC/L carriers were obtained via the free-radical polymerization mechanism using the procedure previously described by M. Markovic et al [4]. The PMAC samples were obtained as follow. Firstly, 4 cm^3 of MAA and caffeine were dissolved into an adequate amount of distilled water (Table 1.). For the PMAC samples with itaconic acid had different only the first step of the synthesis was different: 3 cm^3 of MAA, 1 cm^3 of itaconic acid and caffeine were dissolved into an adequate amount of distilled water (Table 1.). Then, after the total neutralization of MAA (or MAA and IA) with NaOH, the mixture was heated to 60°C and 4 g of casein was added and dissolved. Thereafter, crosslinker MBA (Table 1.) was added and after 10 minutes the initiator VA-044 (0.9cm^3 of 1wt% aqueous solution) was added and the polymerization process began immediately. The mixture was poured quickly into the glass moulds (plates 12 cm × 12 cm, separated by a 2 mm tick PVC hose) and left in the air oven at 60°C for 5h. After the polymerization process ended, the disc-shaped samples (7mm in diameter) were cut and dried at room temperature. All obtained samples were stored in an exicator until they were used for further investigation.

The synthesis path of the PMAC/L samples was similar to the synthesis path of the PMAC samples, but the synthesis was carried out at 40°C in order to prevent liposomes degradation. The first step was the addition of the liposomes with the encapsulated caffeine in a drop wise manner to 4 cm^3 of MAA or to 3 cm^3 of MAA and 1 cm^3 of IA in synthesis of the sample with itaconic acid. After total neutralization of MAA (or MAA and IA) with NaOH, the mixture was heated to 40°C and 4 g of casein was added and dissolved. The crosslinker and initiator were added in the same manner as previously. The liposomes with encapsulated caffeine were obtained by pro-liposomic method [5]. The caffeine solution in distilled water (20 mg/ml) was added to NATIPIDE® II (10wt% with respect to the final liposomic dispersion) in a drop wise manner under the constant stirring.

The synthesized samples were denoted as PMAC-xN-y, PMAC-xN-L, PMAC/IA-xN-y and PMAC/IA-xN-L,where N represented the symbol for neutralization, L was the symbol for liposomes, x denoted the neutralization degree of MAA and y denoted the caffeine amount in the carriers (g).

TABLE 1.
FEED COMPOSITION

Samples	H_20 (cm^3)	Caffeine (g)	Liposomes with caffeine (cm^3)
PMAC-100N-0.2	6.20	0.2	-
PMAC/IA-100N-0.2	6.00	0.2	-
PMAC-100N-L	3.10	-	3.10
PMAC/IA-100N-L	3.00	-	3.00

C. Fourier Transform Infrared Spectroscopy (FTIR) and Scanning Electron Microscopy (SEM)

The FT-IR spectra of xerogel disks were recorded in transmittance mode for the wavelength range of 600–4000 cm^{-1} with a resolution of 4 cm^{-1}, using Nicolet™ iS10 FTIR Spectrometer. The SEM analyses were performed using a Tescan MIRA 3 XMU field-emission gun scanning electron microscope (FEG-SEM) with an acceleration voltage of 20 kV.

D. Monitoring of the carriers swelling and caffeine release

Swelling and caffeine release measurements were carried out at 37°C in two media during a 24h period: 0.2 M phosphate buffer (pH=6.8) (PB) (as a simulation of duodenum pH environment) and 0.1 M solution of HCl (as a simulation

of stomach pH environment) [6]. Dry hydrogel disks with known initial weight (m_0) were entirely immersed in the specified solution and left to swell. At predetermined time intervals the disks were removed from solutions and weighed (m_t). This was repeated until the equilibrium was reached (m_{eq}). Swelling degree (SD) at the equilibrium state (SD_{eq}) was calculated as:

$$SD_{eq} = \frac{(m_{eq} - m_0)}{m_0}. \quad (1)$$

The absorbances of the caffeine solutions in release experiments were measured at predefined time intervals at 273 nm using the UV-Vis Shimadzu UV-1800 spectrophotometer. The caffeine release kinetics was investigated by the Kopcha model (2), which involves both diffusion and polymer chains relaxation effects on the drug release [7]:

$$\alpha = \frac{M_t}{M_\infty} = k_D t^{0.5} + k_R t^1, \quad (2)$$

Where α represents the fractional drug release, M_t is the released concentration of drug at time t, M_∞ is the released drug concentration at equilibrium, k_D is the constant of the speed of drug diffusion and k_R is the constant of the speed of drug release by the process of the polymer chains relaxation.

III. Main Results

The FTIR spectra of the synthesized PMAC carriers are presented in Fig. 1. and they have all characteristic peaks of poly(methacrylic acid) and casein.

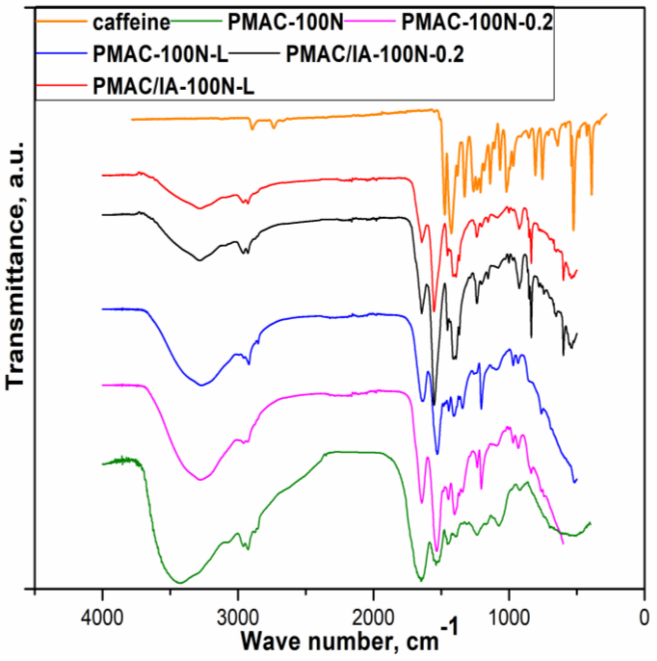


Fig. 1. The FTIR spectra of caffeine and of the PMAC carriers of different formulation

The FTIR spectrum of the PMAC samples without the caffeine was described in our previous work [4]. Compared to the FTIR spectrum of the PMAC carriers without the itaconic acid the FTIR spectrum of PMAC/IA carriers has more intensive peaks at 1540 cm^{-1} (symmetric stretching vibration of C(=O)–O^-) and at 1645 cm^{-1} (C(=O)–OH symmetric stretching vibrations) as a consequence of the presence of the higher number of the –COO^- and –COOH groups due to the presence of two carboxylic groups in the itaconic acid structure [8]. The presence of the caffeine was confirmed by the presence of the caffeine characteristic peaks: the peak at 973cm^{-1} (C-C stretching), 1357 cm^{-1} (C-H stretching) and at 1700 cm^{-1}(stretching of C=O) [9]. The shifts of the characteristic peaks of casein at 1235 cm^{-1} (saturated C–C stretching) and at 1452 cm^{-1} (C-C stretching of aromatic ring) to 1245 cm^{-1} and to 1442 cm^{-1}, respectively, suggested that hydrophobic interactions between the casein and caffeine were established [10]. Also, the shift of the characteristic peak of casein at 1398 cm^{-1} (C=O stretching of aspartic acid and glutamic acid residues) to 1408 cm^{-1} could be caused by the hydrogen bonds established between the casein and caffeine [10]. The FTIR spectra of the PMAC-100N-L and PMAC/IA-100N-L showed the shift of the characteristic peak of casein observed at 1398 cm^{-1} to 1388 cm^{-1} which could be caused by the hydrogen bonds established between the casein and the liposomes (carbonyl group or N-H group of amide II of protein and oxygen group of phospholipid nanoparticles). The hydrophobic bonds established between the casein and liposomes may cause the shifts of the casein characteristic peaks at 1235 cm^{-1} and at 1452 cm^{-1} to 1245cm^{-1} and to 1444cm^{-1}, respectively.

The SEM micrographs of synthesized PMAC carriers are presented in Fig. 2. The micrograph of PMAC-100N-0.2 (Fig. 2. a)) showed the regular porous structure. The PMAC-100N-L sample has the same structure as the PMAC-100N-0.2 and the micrograph confirmed the presence of the liposomes which were marked with the white circles (Fig. 2. b)). The micrograph of the PMAC/IA-100N-0.2 (Fig. 2. c)) showed a non-regular highly porous structure. This was expected due to the presence of the itaconic acid which caused higher value of swelling degree and higher swelling rate of the carrier compared to the samples without itaconic acid. The micrograph of the PMAC/IA-100N-L (Fig. 2. d)) showed the same structure as the analog sample without the liposomes. The liposomes are marked with white circles in Fig. 2. d).

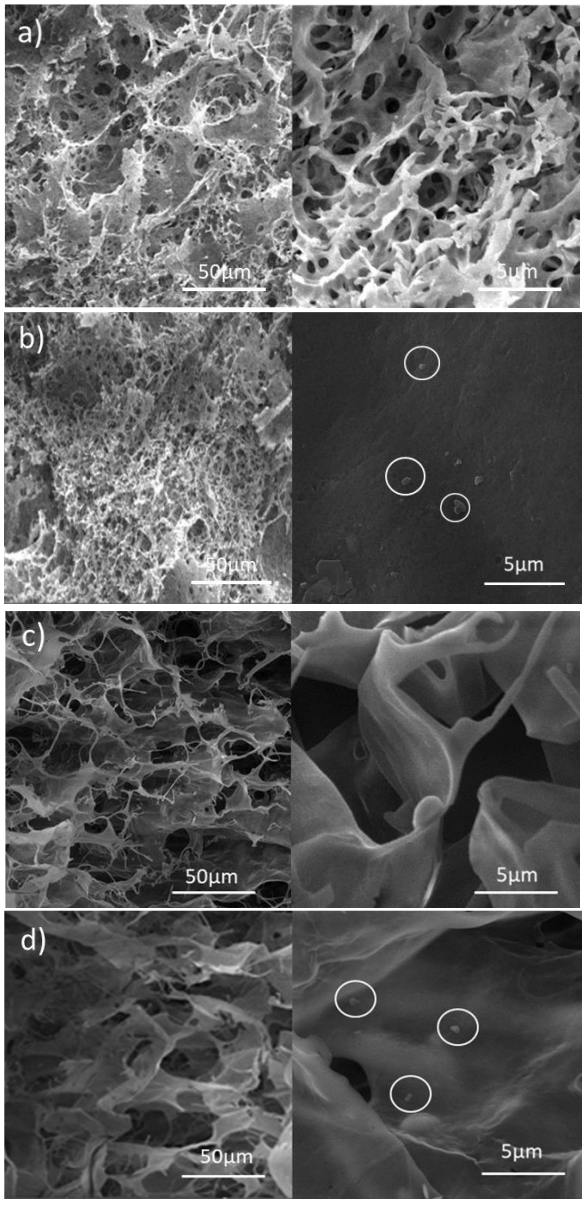


Fig. 2. The SEM micrographs of the carriers of the different formulation

The swelling curves of the synthesized carriers and caffeine release profiles in two media with different pH values are presented in Fig. 3. and in Fig. 4., respectively. From Fig. 3, it can be seen that the PMAC carriers with itaconic acid have higher equilibrium swelling degree and higher swelling rate than the analog samples without the itaconic acid (Table 2.) due to the presence of the higher number of carboxilyc groups. The values of the equilibrium swelling degree were higher in PB than in 0.1 M HCl for all samples (Table 2.) because the negative charges were generated on the carboxylic groups in PB medium which caused the repulsion between the polymer chains and the higher swelling rate of all carriers than the swelling rate of the carriers in 0.1M HCl. The presence of the liposomes in the carriers caused the decrease in the value of the swelling degree. The diffusion of the medium into the carriers could be slower due to the presence of the liposomes in the pores of the matrix of the carriers.

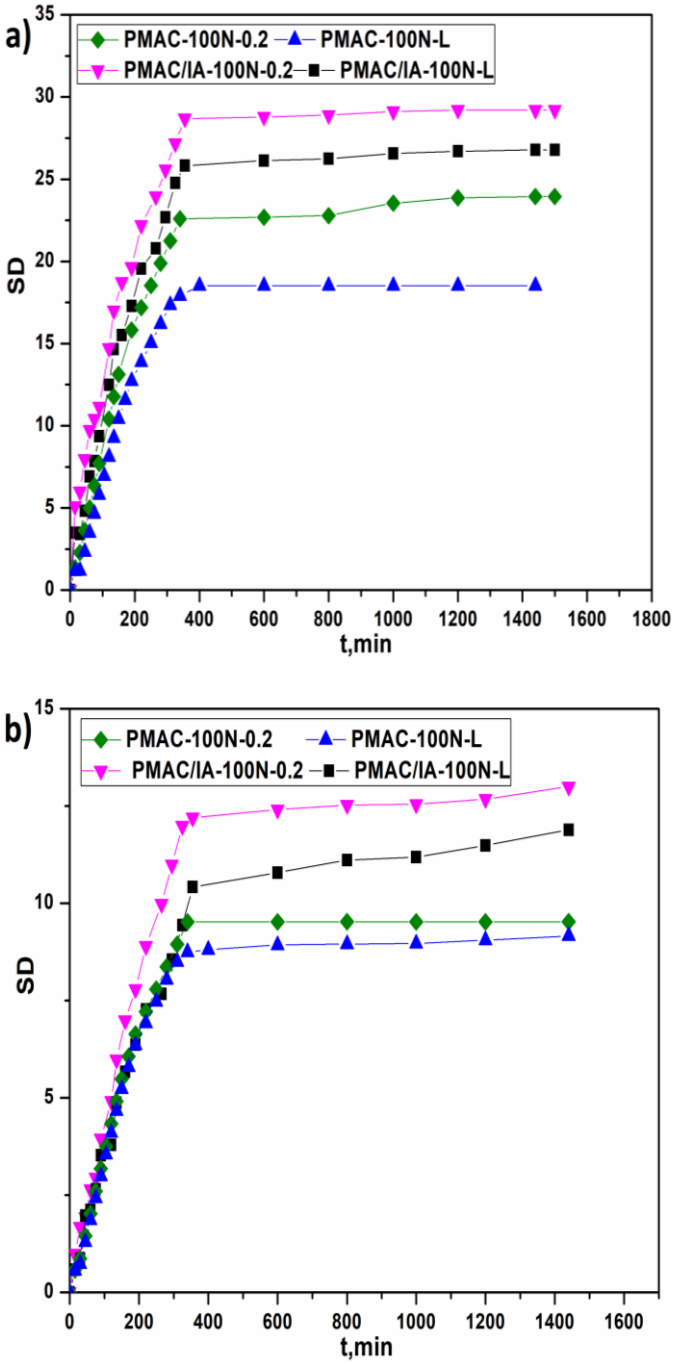


Fig. 3. The swelling curves of the carriers for: a) PB and b) 0.1 M HCl

The released caffeine concentration- c (mg/ml) during time- t (Fig. 4.) showed that the carriers with itaconic acid released caffeine faster than the carriers without it due to the presence of the higher number of carboxylic groups and higher swelling rate. All samples released higher concentration of caffeine in PB than in 0.1M HCl due to the aforementioned behavior of the samples in the PB medium. The caffeine was released more slowly from the samples with liposomes than from the samples without liposomes.

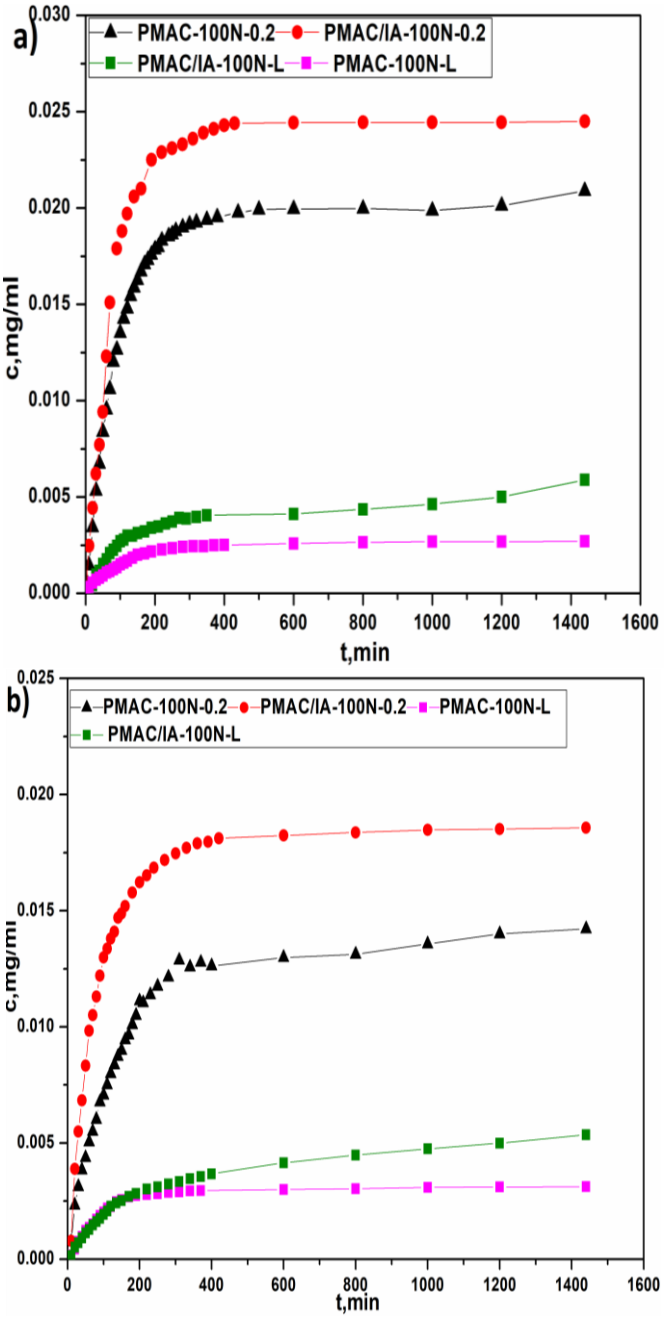


Fig. 4. The caffeine release profiles from the carriers in: a) PB and b) 0.1 M HCl

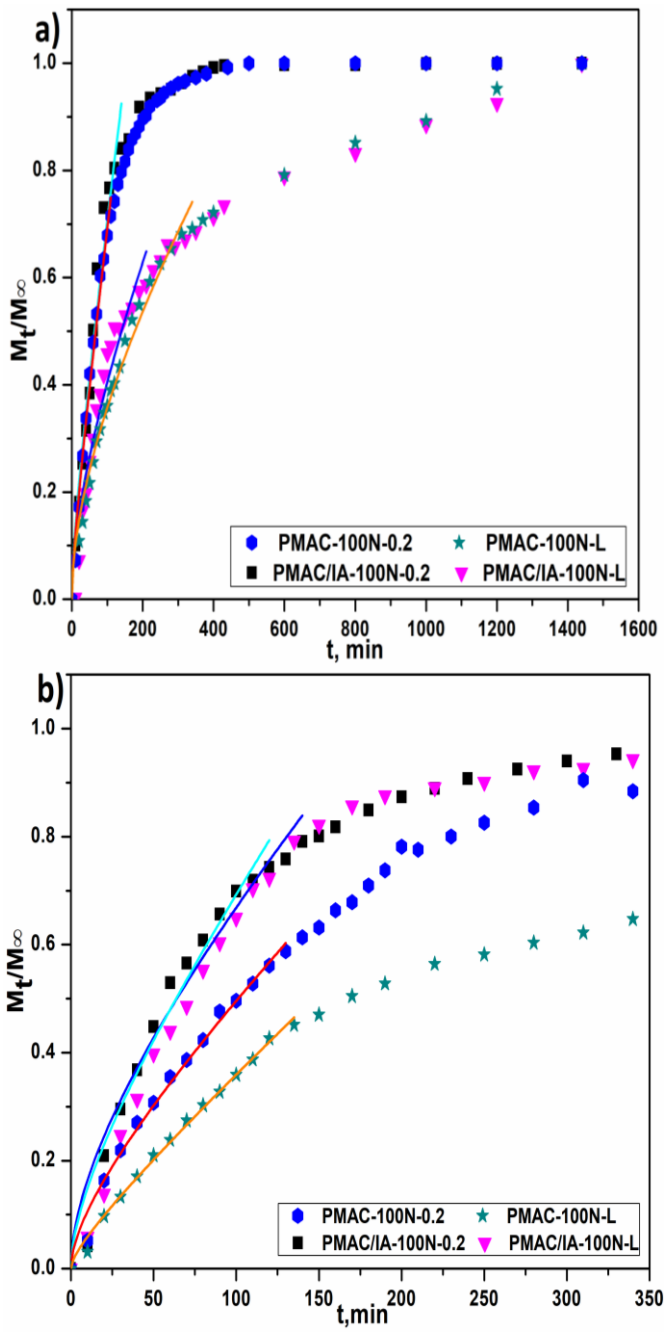


Fig. 5. The fractional caffeine release from the carriers in both media: the symbols represent the experimental data and the solid lines represent the Kopcha model

The fractional release data for PB and 0.1 M HCl to which the Kopcha model was applied is presented in Fig. 5. a) and b), respectively. The estimated values of the parameters of the Kopcha model, the field of applicability $\Delta\alpha$ and the values of R^2 are shown in Table 2. The first 60%-80% of release data fitted well to this model (R^2 ~0.980). The values of the Kopcha model parameter k_D were higher than the values of k_R for all samples, which suggested that the diffusion was the main mechanism of caffeine transport into the media.

TABLE 2.
THE VALUES OF THE EQUILIBRIUM SWELLING DEGREE AND THE KINETIC PARAMETERS OF THE KOPCHA MODEL

Sample	Media	The Kopcha model				SDeq
		k_D $*10^2$	k_R $*10^3$	Δα %	R^2	
PMAC-100N-0.2	HCl	2.62	2.34	58.78	0.991	9.5
	PB	2.40	4.52	71.53	0.989	23.95
PMAC/IA-100N-0.2	HCl	3.67	3.26	74.28	0.972	13.0
	PB	2.96	4.1	84.08	0.968	29.2
PMAC-100N-L	HCl	3.22	0.380	70.74	0.983	9.2
	PB	3.65	2.46	74.96	0.998	22.6
PMAC/IA-100N-L	HCl	2.65	3.65	79.24	0.986	11.9
	PB	3.09	0.96	57.53	0.939	26.8

IV. CONCLUSION

The PMAA based carriers of different formulations were synthesized for controlled and targeted delivery of a poorly water-soluble model drug-caffeine. The significance of these carriers is that they represent the fusion of hydrophilic polymers-PMAA and IA and one amphiphilic polymer-casein, which enabled bonding to poorly water-soluble substance. The FTIR spectra of these carriers showed that established interactions between the casein and caffeine were hydrophobic interactions and hydrogen bonds. The SEM micrographs showed the regular porous structure of the PMAC carriers without itaconic acid, whereas the PMAC carriers with itaconic acid had a non-regular structure with large voids. The presence of the liposomes in the carriers was confirmed by the SEM micrographs of the PMAC-L carriers, indicating that the degradation of the liposomes did not occur during the synthesis of the carriers.

The carriers with itaconic acid had higher equilibrium swelling degree and higher swelling rate than the samples without it. The presence of the liposomes caused a minor decrease in the values of the SDeq, which could be a consequence of the physical presence of the liposomes in the pores of the carriers which may cause the decrease in the diffusion speed of the media into the carriers. All carriers had higher swelling rate in the PB than in 0.1M HCl.

These pH sensitive drug delivery carriers were able to protect the model drug in 0.1M HCl at 37°C (as simulation of the pH condition in human stomach) and release higher caffeine concentration in a medium which simulated the conditions in human intestines- phosphate buffer pH=6.8 at 37°C. The carriers with itaconic acid released higher caffeine amount than the analog carriers without it due to the higher swelling rates. The mathematical model used for investigation of the release kinetics- the Kopcha model, fitted well to the experimental caffeine release data. The Kopcha model showed that the diffusion governed caffeine release from all samples and that the polymer chains relaxation was also present, but its influence on caffeine release was minor. The best control of caffeine release was achieved from the PMAC samples with incorporated liposomes. Incorporated liposomes in PMAC carriers decreased the speed of the caffeine release compared to the samples without them. Synthesized pH-sensitive PMAC carriers are promising candidates for controlled and targeted delivery of poorly water-soluble drugs.

ACKNOWLEDGMENT

The authors would like to acknowledge funding from the Ministry of Education, Science and Technological Development of the Republic of Serbia, through Projects No. 172062 and III 46010.

REFERENCES

[1] M. Murata, K. Tahara, H. Takeuchi, "Real-time in vivo imaging of surface-modified liposomes to evaluate their behavior after pulmonary administration", *European Journal of Pharmaceutics and Biopharmaceutics*, vol. 86, no. 1, pp. 115-119, 2014.

[2] A. Nesic, V. Panic, S. Ostojic, D. Micic, I. Pajic-Lijakovic, A. Onjia, S. Velickovic,"Physical-chemical behavior of novel copolymers composed of methacrylic acid and 2-acrylamido-2-methylpropane sulfonic acid", *Materials Chemistry and Physics*, vol. 174 pp. 156-163, 2016.

[3] E. Larrañeta, L. Barturen, M. Ervine, R.F. Donnelly, "Hydrogels based on poly(methyl vinyl ether-co-maleic acid) and Tween 85 for sustained delivery of hydrophobic drugs", International Journal of Pharmaceutics vol. 538, no. 1 pp. 147-158, 2018.

[4] M. D. Markovic, P. M. Spasojevic, S. I. Seslija, I. G. Popovic, Dj. N. Veljovic, R. V. Pjanovic, V. V. Panic, "Casein-poly(methacrylic acid) hybrid soft networks with easy tunable properties", *European Polymer Journal*, vol. 113, pp. 276-288, 2019.

[5] R. Pjanović, R. Stojanović, M. Šajber, J. Veljkovic, N. Bošković-Vragolović, S. Pejanović, "Diffusion of lidocaine hydrochloride from lipid microparticles", *Chemical Industry and Chemical Engineering Quarterly*, vol. 15, no. 1, pp. 33-35, 2009.

[6] G.P. Asnani, J. Bahekar, C.R. Kokare, "Development of novel pH–responsive dual crosslinked hydrogel beads based on Portulaca oleracea polysaccharide-alginate-borax for colon specific delivery of 5-fluorouracil", *Journal of Drug Delivery Science and Technology*, vol. 48, pp. 200-208, 2018.

[7] M. Kopcha, K.J. Tojo, N.G. Lordi, "Evaluation of Methodology for Assessing Release Characteristics of Thermosoftening Vehicles", *Journal of Pharmacy and Pharmacology*, vol. 42, no. 11, pp. 745-751, 1990.

[8] M. Sakthivel, D.S. Franklin, S. Sudarsan, G. Chitra, T.B. Sridharan, S. Guhanathan, "Investigation on pH/salt-responsive multifunctional itaconic acid based polymeric biocompatible, antimicrobial and biodegradable hydrogels", *Reactive and Functional Polymers*, vol. 122, pp. 9-21, 2018.

[9] N. Noor, A. Shah, A. Gani, A. Gani, F.A. Masoodi, "Microencapsulation of caffeine loaded in polysaccharide based delivery systems", *Food Hydrocolloids*, vol. 82, pp. 312-321, 2018.

[10] N. Chen, P. Di, S. Ning, W. Jiang, Q. Jing, G. Ren, Y. Liu, Y. Tang, Z. Xu, G. Liu, F. Ren, "Modified rivaroxaban microparticles for solid state properties improvement based on drug-protein/polymer supramolecular interactions", *Powder Technology*, vol. 344, pp. 819-829, 2019.

Swelling and bioactivity of poly (methacrylic acid)/ hydroxyapatite / bioactive glass composite hydrogels

Vukasin Ugrinovic, Vesna Panic, Sanja Seslija, Pavle Spasojevic, Ivanka Popovic, Djordje Janackovic, Djordje Veljovic

***Abstract*—The goal of the study was to examine the influence of bioactive glass (BG) particles incorporation on properties of poly(methacrylic acid)/hydroxyapatite (HA) composite hydrogels. Composite hydrogels were synthesized by free-radical polymerization. Theoretical amount of incorporated inorganic fillers was 60 wt%, while BG/HA ratio was varied. Composites were characterized by Scanning Electron Microscopy, and swelling behavior was determined in distilled water. During 28 days of in vitro bioactivity test, pH changes were constantly monitored. Equilibrium swelling increased by 110.9 % as the content of BG in composites increased. pH values of SBF were significantly higher in the case of the sample with higher amount of BG. Morphological investigations revealed localized bioactivity of the sample with 40 wt% of BG, while the one with 10 wt% exhibited no significant bioactivity.**

***Index Terms*—hydroxyapatite; scaffold; 45S5 bioactive glass; poly (methacrylic acid).**

I. Introduction

Bone tissue scaffolds are three-dimensional porous structures designed to, temporally or permanently, replace a critical bone tissue volume loss, and act as a structural support for adhesion, proliferation and differentiation of osteoblasts, which is the main precondition for healing and regeneration of damaged tissue. For the successful design of bone tissue scaffolds the following requirements have to be satisfied: biocompatibility, satisfactory mechanical properties, porosity and bioactivity [1], [2].

Composite hydrogels, the composites constituted of hydrogels - the three-dimensional, hydrophilic networks, and inorganic nano-filler particles such as hydroxyapatite (HA) and bioactive glass (BG), have gained increasing attention over the past few years due to their unique properties such us, high degree of swelling and improved mechanical properties compared to pure hydrogels [3]–[11].

For biomedical purposes, hydrophilic polymeric networks based on poly (methacrylic acid) (PMAA) have been mainly investigated as drug delivery vehicles due to pH-sensibility [12]–[14]. However, the possibility to easily change the volume in response to protonization/ deprotonization of carboxyl groups, allows simple tunability of PMAA hydrogels porosity [8].

HA is the main inorganic component of human bone tissue and accounts for ≈ 60 % of total bone mass. There have been numerous investigations of HA as the main component of bone tissue scaffolds. HA imparts stiffness and mechanical stability to scaffolds, as well as biocompatibility and osteoconductivity [15].

In our previous research we demonstrated significant improvement of PMAA hydrogels modulus by incorporation of nano-HA particles [8]. However, the lack of bioactivity could be the limiting factor in potential clinical application. Therefore, the aim of this study was further optimization of PMAA/HA scaffolds bioactivity by incorporation of 45S5 BG particles, which is well known, highly bioactive material [16].

II. The Method

Nano-sized HA powder was synthesized as in [17], [18]. Molar ratio of Ca/P was 1.67 which corresponds to stoichiometric HA.

Bioactive glass 45S5 was synthesized by melt-quenching method. Raw materials were mixed (Table 1), homogenized and subsequently melted by heating from room temperature to 1400 °C at 10 °C/min. After 1 h of maintaining BG at 1400° C, it was room-temperature water quenched and dried in an oven. To obtain fine particles, BG was ground in Planetary Ball Mill Retsch PM 100 (33 ±1 g of glass was ground for 1 h in 125 ml grinding jar, with 500 zirconium oxide balls 5 mm in diameter, under the wheel speed of 500 min^{-1}).

Vukasin Ugrinovic is with Innovation Center of Faculty of Technology and Metallurgy, University of Belgrade, Karnegijeva 4, 11000 Belgrade, Serbia (e-mail: vugrinovic@tmf.bg.ac.rs).

Vesna Panic is with Innovation Center of Faculty of Technology and Metallurgy, University of Belgrade, Karnegijeva 4, 11000 Belgrade, Serbia (e-mail: vpanic@tmf.bg.ac.rs).

Sanja Seslija is with Centre of Excellence in Environmental Chemistry and Engineering, Institute of Chemistry, Technology and Metallurgy, University of Belgrade, Njegoševa 12, 11000 Belgrade, Serbia (sseslija@tmf.bg.ac.rs).

Pavle Spasojevic is with Innovation Center of Faculty of Technology and Metallurgy, University of Belgrade, Karnegijeva 4, 11000 Belgrade, Serbia, and with Faculty of Technical Sciences, University of Kragujevac, Svetog Save 65, 32000 Cacak, Serbia (pspasojevic@tmf.bg.ac.rs).

Ivanka Popovic is with Faculty of Technology and Metallurgy, University of Belgrade, Karnegijeva 4, 11000 Belgrade, Serbia (ivanka@tmf.bg.ac.rs).

Djordje Janackovic is with Faculty of Technology and Metallurgy, University of Belgrade, Karnegijeva 4, 11000 Belgrade, Serbia (nht@tmf.bg.ac.rs).

Djordje Veljovic is with Faculty of Technology and Metallurgy, University of Belgrade, Karnegijeva 4, 11000 Belgrade, Serbia (djveljovic@tmf.bg.ac.rs).

TABLE I
THE WEIGHTS OF COMPONENTS IN THE PRECURSOR MIXTURE FOR BIOGLASS SYNTHESIS

SiO_2 [g]	$CaCO_3$ [g]	$NaH_2PO_4 \cdot 2H_2O$ [g]	Na_2CO_3 [g]
17.03	15.28	4.61	13.08

PMAA/HA/BG composite hydrogels were synthesized by free-radical polymerization by the modification of already described procedure [8]. Briefly, 0.6972 g of sodium hydroxide was dissolved in 1.50 ml of distilled water after which 1.50 ml of methacrylic acid was added. In the next step, 0.0107 g of crosslinker (N,N′-methylenebisacrylamide) was added, followed by the addition of previously synthesized nano-HA and 45S5 bioglass powders (Table 2), to obtain the composites with the 60 wt% of inorganic phase. Finally, 100 µl of 6.8 % w/v initiator (2,2′-Azobis[2-(2-imidazolin-2-yl)propane] dihydrochloride) solution was added to the reaction mixture. After 20 minutes of vigorously stirring, and 15 minutes of sonication, the mixture was poured into glass molds, and placed in an oven at 70 °C, for 4 h to complete reaction.

Theoretical content of inorganic phase (with regard to polymeric network) in composites was fixed at 60 wt%, while the amounts of HA and 45S5 bioglass were varied to prepare the samples containing 10 and 40 wt% of bioglass.

The obtained samples are denoted as CHX, where X= 10 or 40, denoting the weight fraction of bioglass in the composites.

TABLE II
THEORETICAL CONTENT OF INORGANIC PHASE IN THE SYNTHESIZED COMPOSITE HYDROGELS

Sample	*HA [g]*	*45S5 bioglass [g]*
CH10	2.3542	0.4709
CH40	0.9417	1.8834

Dynamical swelling behavior of composites was determined as a function of time, in distilled water at room temperature (20°C), and calculated using the equation:

$$SD = \frac{m_t - m_0}{m_0} \quad (1)$$

Where *SD* is the swelling degree of the hydrogel at time *t*, m_t is the weight of the swollen hydrogel sample at time *t*, and m_0 is the weight of the xerogel.

The equilibrium swelling degree (SD_{eq}) is the swelling degree of the hydrogel at equilibrium, i.e. when the hydrogel sample attained constant mass after swelling (m_{eq}):

$$SD_{eq} = \frac{m_{eq} - m_0}{m_0} \quad (2)$$

For each sample at least three swelling measurements were performed and the mean values were used.

Morphological examinations of the obtained samples were carried out using Field emission scanning electron microscope (SEM) *Tescan MIRA 3 XMU* operating at 20 kV. Prior to examination, all samples were swollen to equilibrium in distilled water, lyophilized, and sputter-coated with gold using a *POLARON SC502 sputter coater* in order to avoid electrostatic charge. Prior to lyophilization samples were frozen for two days at -20° C, after which lyophilization was carried out using *Martin Christ Freeze-dryer Alpha 1-2 LDplus*, under the vacuum of 0,310 mbar and -32° C temperature, during 48 h.

In vitro bioactivity behavior was investigated by soaking composite hydrogels in simulated body fluid (SBF), prepared as in [19]. The experiment was carried by immersing the samples of ≈150 mg in 15 ml of SBF (concentration ≈ 10 g/l) for 28 days at the temperature of 37° C, while renewing the solution of SBF each three days. At the same time (before the each SBF solution renewing), the pH values of the solutions were measured by pH meter (*WTW inoLab 7110, Weilheim, Germany*).

III. RESULTS AND DISCUSSION

The influence of BG incorporation on the swelling behavior of composite hydrogels was examined in distilled water. Fig. 1 presents obtained swelling curves.

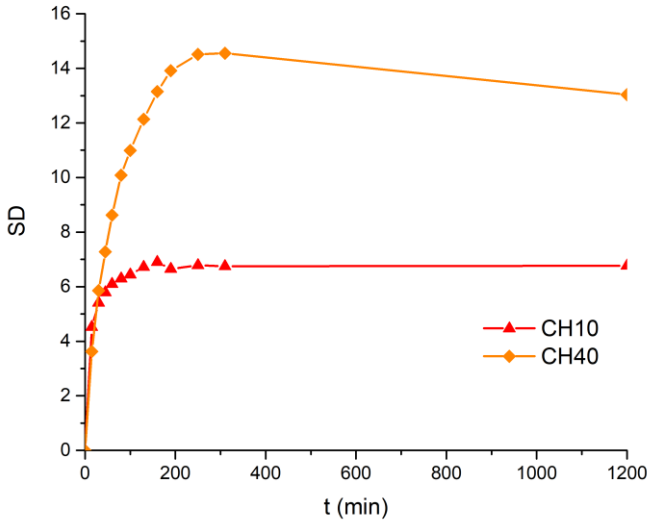


Fig. 1. Isothermal swelling curves of obtained composite hydrogels.

It can be noticed that increase in BG content increased SD_{eq} values by 110.9 %. Nano-sized particles of inorganic filers are characterized by large specific surface area and can act as a crosslinker through the interactions with polymer matrix. For example, Li has reported the formation of chelating and hydrogen bonds in polyacrylamide/HA composites between Ca^{2+} ions and C=O bonds and the PO_4^{3-} (in HA) and NH_2 (in polyacrylamide), respectively [10]. In our system, we can presume similar interactions: 1) chelating bonds between Ca^{2+} (in HA), and $–COO^-$ (in PMAA), and 2) hydrogen bonds between PO_4^{3-} (in HA) and –COOH (in PMAA). However,

since the precursor for polymer synthesis, methacrylic acid, is 100% neutralized by sodium hydroxide (-COO^- form prevails in the structure), probably, the majority of interactions will be chelating bonds. Therefore, the addition of BG instead of HA would decrease the content of crosslinker which in turn will cause the network to swell more. Also, HA nano-particles are smaller and have significantly higher specific surface than BG particles, which means that substitution of HA with BG would decrease the total surface area of filler, which will also lead to decreased crosslinking and increased swelling.

Noticeable deswelling of CH40 after 1200 min is also indicative of decreased crosslinking of CH40, as it is probably the consequence of partial dissolution of hydrogel network. On the other hand, CH10 demonstrated higher structural stability.

Fig. 2 plots the difference between the measured pH and the pH of the basic solution (ΔpH), versus time.

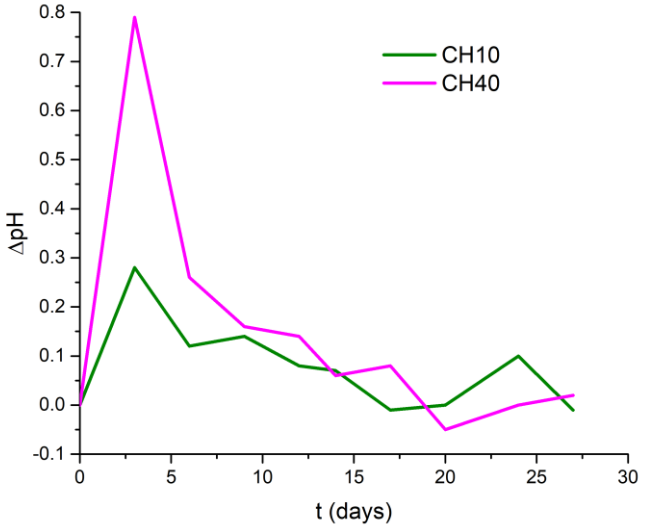


Fig. 2. pH difference between the measured pH and of the basic solution as a function of time.

The increase in pH value after first 3 days is significantly higher for CH40. The increase in the pH of SBF occurs as a result of the exchange of Ca^{2+} and Na^+ ions from glass, with H^+ ions from water, and is an indication of the glass dissolution. The similar trend is noticed after 6 days but with the significantly lower pH values, indicating the slowdown of the ion exchange reaction. Further experiments have demonstrated continuous decline of pH values suggesting the reducing of ion exchange activity. The pH values measurements have indicated the peak of the exchange after the first 3 days of immersion, while the majority of reaction is completed by 12 days. These results suggest higher bioactivity of CH40.

Fig. 3 shows SEM micrographs of lyophilized CH10 composites, after 28 days of immersion in SBF. Spherically agglomerated rod-like HA nano-particles, 50-100 nm in size, are uniformly distributed through PMAA network of composite. BG particles of different sizes are observable in the structure of composite hydrogel. However, there is no indication of significant surface coverage by newly-formed HA crystals, which is sign of bioactivity. SEM micrographs from our previous work evidenced good wetting of HA particles by PMAA matrix, which had a major influence on the inclusion of high amount of HA and uniform distribution of particles, hence allowing improvement of mechanical properties [8]. However, at the same time it limited the exposure of HA to SBF, thus preventing the bioactivity. PMAA is non-biodegradable polymer and thin layer of polymer over the bioactive particles could negatively affect the bioactivity of composites. From the Fig. 3B it is clearly visible that HA particles are coated by a thin layer of PMAA. Thus, substitution of 10 wt% of HA with BG particles didn't significantly affect the microstructure and thus no significant improvement of bioactivity has been achieved compared to PMAA/HA from [8].

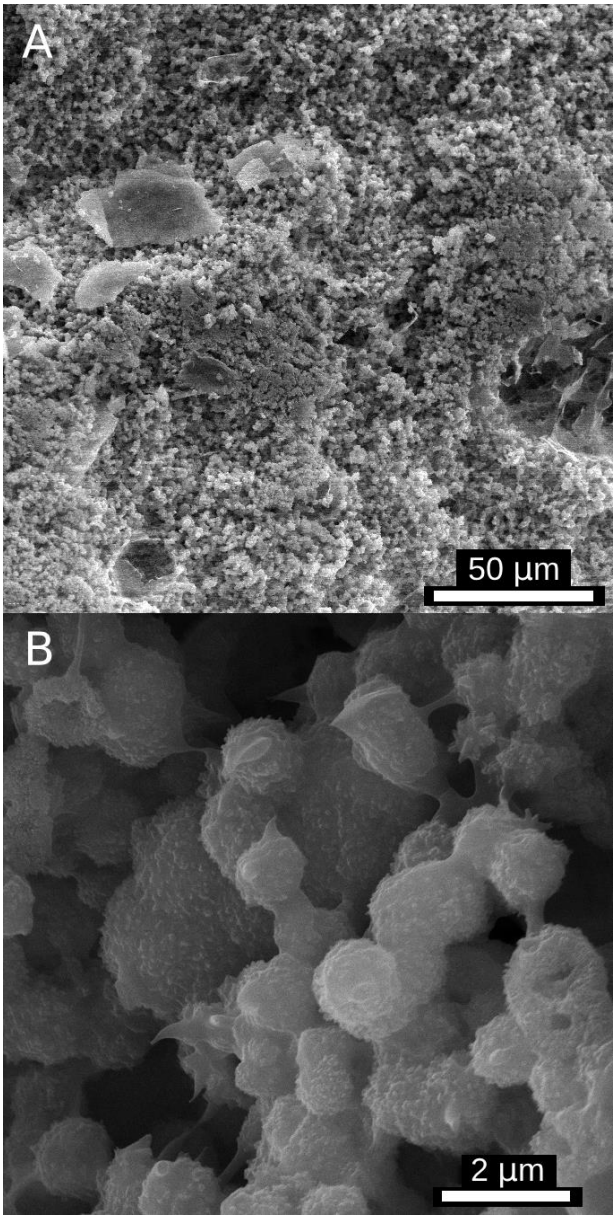


Fig. 3. SEM micrographs of lyophilized CH10 composite hydrogels after 28 days of in vitro bioactivity test. (A) 1000x magnification and (B) 20000x magnification.

On the other hand, Fig. 4 reveals a certain degree of bioactivity of CH40 composite hydrogels. There is no significant surface coverage by newly-formed HA, but localized bioactivity could be noticed, especially in the vicinity of BG particles. In the Fig. 4B is demonstrated a BG particle covered by a layer of HA. In the background of the particle, a thin coat of PMAA is noticeable, however it doesn't cover the particle itself, enabling unhindered contact between the particle and SBF. These results are in accordance with swelling results proving a lower degree of interactions between filler particles and polymer matrix.

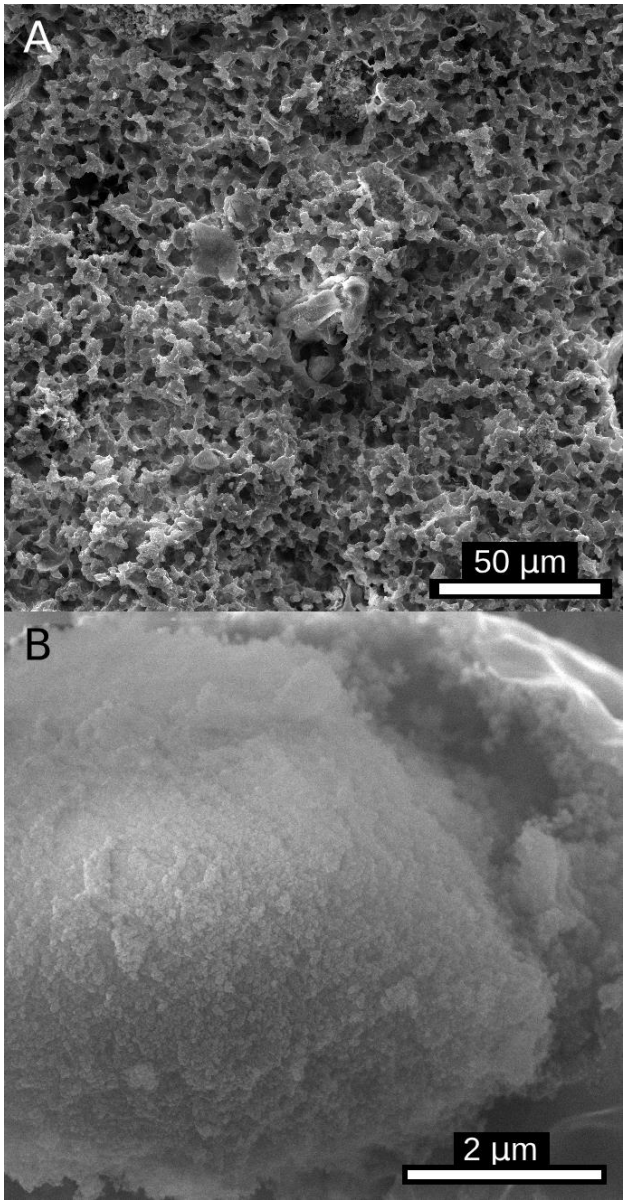


Fig. 4. SEM micrographs of lyophilized CH40 composite hydrogels after 28 days of in vitro bioactivity test. (A) 1000x magnification and (B) 30000x magnification.

IV. Conclusion

Composite hydrogels were synthesized by a simple method with high amount of filler particles incorporated. Swelling examinations gave an insight of composite structure, demonstrating higher degree of crosslinking of CH10 composites, which was ascribed to higher degree of mutual interactions between filler and matrix through chelating and hydrogen bonds. Changes in pH of SBF after immersion of composites showed higher ion exchange activity of CH40, indicating higher bioactivity, which was also confirmed by morphological investigations.

In general, this research gave us better understanding of how size and type of filler particles could influence the scaffold properties.

Acknowledgment

The authors would like to acknowledge funding from the Ministry of Education, Science and Technological Development of the Republic of Serbia, through Project No. 172062 "Synthesis and characterization of novel functional polymers and polymeric nanocomposites" and Project No. III45019 "Synthesis, processing and application of nanostructured multifunctional materials with defined properties".

References

[1] T. Ghassemi, A. Shahroodi, M. H. Ebrahimzadeh, A. Mousavian, J. Movaffagh, and A. Moradi, "Current Concepts in Scaffolding for Bone Tissue Engineering," *Arch. bone Jt. Surg.*, vol. 6, no. 2, pp. 90–99, Mar. 2018.

[2] D. W. Hutmacher, "Scaffolds in tissue engineering bone and cartilage," *Biomater. Silver Jubil. Compend.*, pp. 175–189, Jan. 2000.

[3] R. Arun Kumar A. Sivashanmugam, S. Deepthi, Sachiko Iseki, K. P. Chennazhi, Shantikumar V. Nair, and R. Jayakumar, "Injectable Chitin-Poly(ε-caprolactone)/Nanohydroxyapatite Composite Microgels Prepared by Simple Regeneration Technique for Bone Tissue Engineering," *ACS Appl. Mater. Interfaces*, vol. 7, no. 18, pp. 9399–9409, 2015.

[4] C. Chang, N. Peng, M. He, Y. Teramoto, Y. Nishio, and L. Zhang, "Fabrication and properties of chitin/hydroxyapatite hybrid hydrogels as scaffold nano-materials," *Carbohydr. Polym.*, vol. 91, no. 1, pp. 7–13, Jan. 2013.

[5] A. Gantar, L. da Silva, J. Oliveira, A. Marques, V. Correlo, S. Novak, and R. Reis, "Nanoparticulate bioactive-glass-reinforced gellan-gum hydrogels for bone-tissue engineering," *Mater. Sci. Eng. C*, vol. 43, pp. 27–36, Oct. 2014.

[6] J. Hu, Y. Zhu, H. Tong, X. Shen, L. Chen, and J. Ran, "A detailed study of homogeneous agarose/hydroxyapatite nanocomposites for load-bearing bone tissue," *Int. J. Biol. Macromol.*, vol. 82, pp. 134–143, Jan. 2016.

[7] J. A. Killion, S. Kehoe, L. M. Geever, D. M. Devine, E. Sheehan, D. Boyd, and C. L. Higginbotham, "Hydrogel/bioactive glass composites for bone regeneration applications: Synthesis and characterisation," *Mater. Sci. Eng. C*, vol. 33, no. 7, pp. 4203–4212, Oct. 2013.

[8] V. Đ. Ugrinović, V. V Panić, Đ. N. Veljović, P. M. Spasojević, S. I. Šešlija, and Đ. T. Janaćković, "Synthesis and properties of nano-hydroxyapatite/poly (methacrylic acid) composite hydrogels," *Tehnika*, vol. 73, no. 5, pp. 613–620, 2018.

[9] J. Lacroix, E. Jallot, and J. Lao, "Gelatin-bioactive glass composites scaffolds with controlled macroporosity," *Chem. Eng. J.*, vol. 256, pp. 9–13, Nov. 2014.

[10] Z. Li, W. Mi, H. Wang, Y. Su, and C. He, "Nano-hydroxyapatite/polyacrylamide composite hydrogels with high mechanical strengths and cell adhesion properties," *Colloids Surfaces B Biointerfaces*, vol. 123, pp. 959–964, Nov. 2014.

[11] M. Ribeiro, M. A. de Moraes, M. M. Beppu, M. P. Garcia, M. H. Fernandes, F. J. Monteiro, and M. P. Ferraz, "Development of silk fibroin/nanohydroxyapatite composite hydrogels for bone tissue engineering," *Eur. Polym. J.*, vol. 67, pp. 66–77, Jun. 2015.

[12] S. Sajeesh, K. Bouchemal, C. P. Sharma, and C. Vauthier, "Surface-functionalized polymethacrylic acid based hydrogel microparticles for oral drug delivery," *Eur. J. Pharm. Biopharm.*, vol. 74, no. 2, pp. 209–218, Feb. 2010.

[13] H. Ichikawa and N. A. Peppas, "Novel complexation hydrogels for oral peptide delivery: In vitro evaluation of their cytocompatibility and insulin-transport enhancing effects using Caco-2 cell monolayers," *J. Biomed. Mater. Res. Part A*, vol. 67A, no. 2, pp. 609–617, 2003.

[14] H. Pawar, D. Douroumis, and J. S. Boateng, "Preparation and optimization of PMAA–chitosan–PEG nanoparticles for oral drug delivery," *Colloids Surfaces B Biointerfaces*, vol. 90, pp. 102–108, Feb. 2012.

[15] K. Rezwan, Q. Z. Chen, J. J. Blaker, and A. R. Boccaccini, "Biodegradable and bioactive porous polymer/inorganic composite scaffolds for bone tissue engineering," *Biomaterials*, vol. 27, no. 18, pp. 3413–3431, Jun. 2006.

[16] L. L. Hench, "Bioceramics: From Concept to Clinic," *J. Am. Ceram. Soc.*, vol. 74, no. 7, pp. 1487–1510, 1991.

[17] D. Janaćković, I. Petrovic-Prelevic, L. Kostic-Gvozdenovic, R. Petrović, V. Jokanović, and D. P. Uskokovic, "Influence of Synthesis Parameters on the Particle Sizes of Nanostructured Calcium-Hydroxyapatite," in *Bioceramics 13*, 2000, vol. 192, pp. 203–206.

[18] D. Janaćković, I. Jankovic-Castvan, R. Petrović, L. Kostic-Gvozdenovic, S. K. Milonjić, and D. P. Uskokovic, "Surface Properties of HAp Particles Obtained by Hydrothermal Decomposition of Urea and Calcium-EDTA Chelates," in *Bioceramics 15*, 2002, vol. 240, pp. 437–440.

[19] T. Kokubo, H. Kushitani, C. Ohtsuki, S. Sakka, and T. Yamamuro, "Chemical reaction of bioactive glass and glass-ceramics with a simulated body fluid," *J. Mater. Sci. Mater. Med.*, vol. 3, no. 2, pp. 79–83, Mar. 1992.

Sinthesis and Characterization of Hydroxyapatite and Fluorapatite Powders

Željko Radovanović, Abdulmoneim Mohamed Kazuz, Predrag Vulić, Lidija Radovanović, Đorđe Veljović, Rada Petrović, Đorđe Janaćković

Abstract— **The biomaterial powders of hydroxyapatite (HAp) and fluorapatite (FAp) were synthesized by a hydrothermal method. Powders were analyzed by energy-dispersive X-ray spectroscopy (EDS), field emission scanning electron microscopy (FESEM), and X-ray powder diffraction analysis (XRPD). EDS analysis shows the presence of non-stoichiometries FAp and HAp with molar ratio CA/P < 1.67. FESEM analysis of both powders indicates the presence of agglomerates of micrometric dimensions, while primary nanoparticles are rod-like. The Rietveld refinement of XRPD data showed that the single phase powders of FAp and HAp were synthesized. The results showed that obtained nanomaterials can be potentially applied in dentistry.**

Index Terms—**Biomaterial; Hydroxyapatite; Fluoroapatite; Nanoparticles; Rietveld refinement.**

I. Introduction

A new trend in the treatment of teeth is the application of material that would fill the cavity after removal of damaged dental tissue and also remineralize the surrounding tissue. The appropriate material could be a hydroxyapatite (HAp) as well as fluorapatite (FAp). Synthetic HAp, $Ca_5(PO_4)_3(OH)$, is similar to the inorganic part of bones and the dentine of teeth. It is biocompatible, bioactive, nontoxic, and osteoconductive [1, 2]. Replacement of the OH^- groups in HAp with F^- gives FAp, an implant material with better hardness, greater stability, less solubility and better antimicrobial effect than HAp [1].

Taheri *et al.* [3] synthesized the FAp by hydrothermal method at different pH values and temperatures. They found that the pH value of hydrothermal solution is more significant factor than temperature in terms of shape and dimension of the synthesized FAp. Ge *et al*. [4] obtained FAp after heat treatment at 600 °C in a water vapor environment for 3 h started from the as-deposited amorphous fluoridated calcium phosphate. They revealed that FAp had significantly better antibacterial activity than HAp. Stanić *et al.* [5] synthesized FAp powders by neutralization method. They found that the antimicrobial activity of the samples increases with the increase of concentration of $F^-$and decrease of pH value of saline solution. Zhao *et al*. [6] have shown that by the solution combustion method is possible to obtain single phase FAp and HAp starting from inexpensive raw materials and applying relatively simple preparation process and low-cost experimental installation.

The aim of this study was to synthesize the pure nanosized powders of FAp and HAp and investigate the properties of these materials. For this purpose, modified hydrothermal synthesis was applied.

II. Materials and Method

FAp and HAp powders were synthesized by a previously described modified hydrothermal method [7–10]. Pure HAp was synthesized from stoichiometric quantities of the chemicals: $NaH_2PO_4 \cdot 2H_2O$ (VWR Prolabo, 99.8%), NH_4OH (Zorka Pharma, p. a.) and $Ca(NO_3)_2 \cdot 2H_2O$ (Roth, ≥ 98%). In the synthesis of FAp, beside aforementioned chemicals, NaF (Riedel de Haen, 99%), were also used. Synthesis was performed with a constant molar ratio Ca/P = 1.67. After dissolution of the chemicals in 2 L of distilled water, the dish with the solution was inserted into an autoclave, previously filled with the 1.5 L of distilled water. The solution was heated at 160 °C for 3 h. After slow cooling, the obtained suspension was filtered, and the residue was washed with distilled water and dried at 105 °C for 4 h.

Energy-dispersive X-ray spectroscopy (EDS) of the powders was performed on a Jeol JSM 5800 SEM with a SiLi X-Ray detector (Oxford Link Isis series 300, UK).

The morphologies of the powders were observed by Tescan Mira 3 XMU field emission scanning electron microscopy (FESEM). Before analysis, the powders were coated with Au using a Polaron SC502 sputter coater. The particle size distribution was determined and presented using Mira software and Microsoft Excel programme, respectively.

The X-ray powder diffraction (XRPD) measurements were performed on a Rigaku SmartLab diffractometer using Cu *Kα* radiation, at 40 KV and 30 MA, in Bragg–Brentano geometry.

Željko Radovanović is with the Innovation Center of the Faculty of Technology and Metallurgy, University of Belgrade, Karnegijeva 4, 11120 Belgrade, Serbia (e-mail: zradovanovic@tmf.bg.ac.rs).

Abdulmoneim Mohamed Kazuz is with the Faculty of Technology and Metallurgy, University of Belgrade, Karnegijeva 4, 11120 Belgrade, Serbia (e-mail: abdulmohamedabdul520@gmail.com).

Predrag Vulić is with the Faculty of Mining And Geology, University of Belgrade, Đušina 4, 11000 Belgrade, Serbia (e-mail: predrag.vulic@rgf.bg.ac.rs).

Lidija Radovanović is with the Innovation Center of the Faculty of Technology and Metallurgy, University of Belgrade, Karnegijeva 4, 11120 Belgrade, Serbia (e-mail: lradovanovic@tmf.bg.ac.rs).

Đorđe Veljović is with the Faculty of Technology and Metallurgy, University of Belgrade, Karnegijeva 4, 11120 Belgrade, Serbia (e-mail: djveljovic@tmf.bg.ac.rs).

Rada Petrović is with the Faculty of Technology and Metallurgy, University of Belgrade, Karnegijeva 4, 11120 Belgrade, Serbia (e-mail: radaab@tmf.bg.ac.rs).

Đorđe Janaćković is with the Faculty of Technology and Metallurgy, University of Belgrade, Karnegijeva 4, 11120 Belgrade, Serbia (e-mail: nht@tmf.bg.ac.rs).

Diffraction data were collected in the range $5° < 2\theta < 120°$ (scan speed: 1° min^{-1}, step width: 0.01° 2θ) at room temperature.

III. Main Results

The atomic % of O, Ca, P and F were determined from the results of EDS analysis (Table I). This analysis shows the presence of non-stoichiometric HAp and FAp with molar ratio Ca/P < 1.67, which means that in both cases Ca-deficient apatite is synthesized.

TABLE I
EDS results for FAp and HAp powders.

Atomic %	Powder	
	HAp	**FAp**
O	72.71±0.23	66.16±1.06
F	–	5.00±0.58
P	11.53±0.05	11.58±0.27
Ca	15.75±0.20	17.26±0.69
Ca/P	1.37	1.49

The FESEM micrographs of the hydrothermally obtained HAp and FAp powders are shown in Fig. 1. Both powders consist of rod-like nanosized particles with average particle size of 87±17 nm for HAp and 87±20 nm for FAp. Also, the particle size distribution (Fig. 2.) are very similar and more than 60 % of particles are in the range of 70–90 nm for both powders. The particles form the agglomerates of micrometric dimensions. In the case of HAp, the agglomeration is more pronounced.

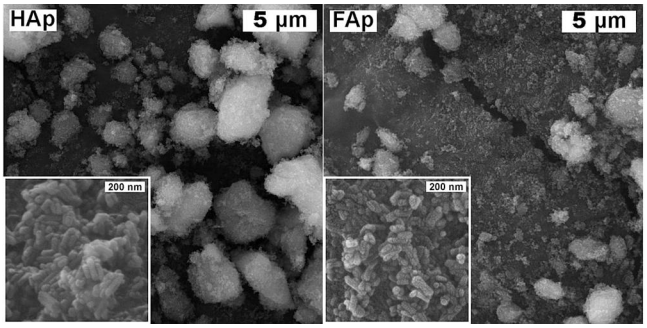


Fig. 1. FESEM micrographs of the HAp and FAp powders.

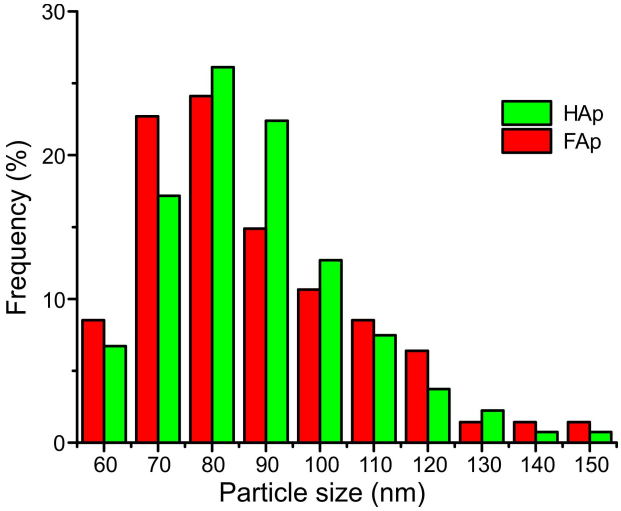


Fig. 2. Particle size distribution of the HAp and FAp powders.

TABLE II
Crystallographic and Rietveld refinement parameters of HAp and FAp.

Phase	**HAp**	**FAp**
Crystal system	hexagonal	hexagonal
Space group	$P6_3/c$	$P6_3/c$
a [Å]	9.4205(1)	9.3760(1)
c [Å]	6.88151(9)	6.88276(9)
V [Å^3]	528.9(1)	524.0(1)
Crystallite size [Å]	346(1) [–0.356, 0.935, 0]	330(1) [0.356, –0.935, 0]
Crystallite size [Å]	346(1) [0.935, 0.356, 0]	330(1) [0.935, –0.356, 0]
Crystallite size [Å]	886(9) [0, 0, 1]	983(1) [0, 0, 1]
Strain [%]	0.096(2)	0.105(1)
R_{wp} [%]	4.87	4.92
R_p [%]	3.76	3.82
R_e [%]	3.92	3.92
χ^2	1.5414	1.5744
S	1.2451	1.2548
Maximum shift/e.s.d.	0.081	0.022

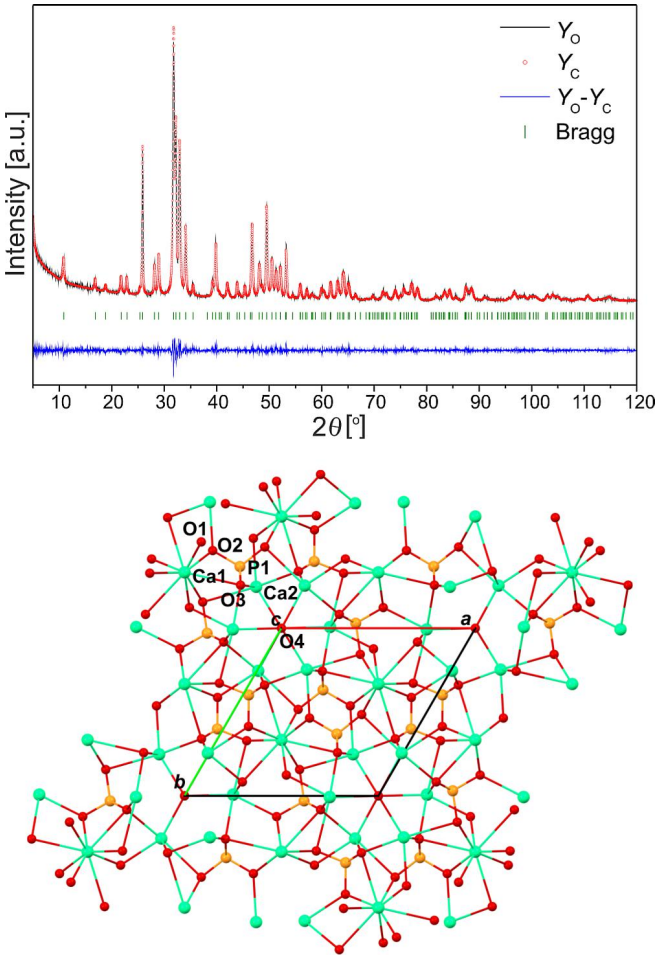


Fig. 3. Rietveld refinement (up) and crystal packing diagram in *ab* plane (down) of HAp.

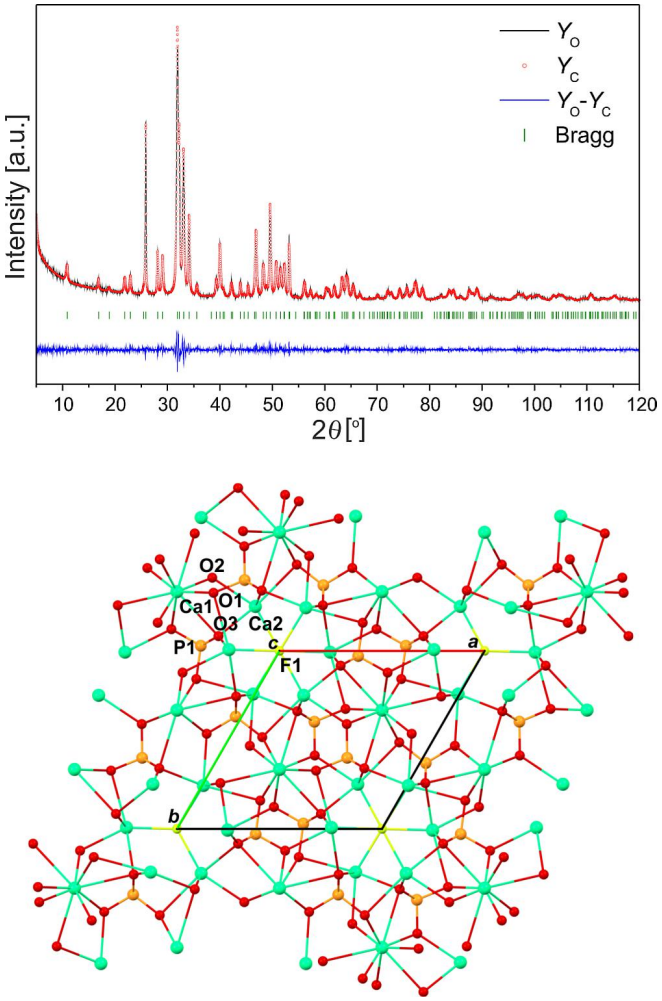


Fig. 4. Rietveld refinement (up) and crystal packing diagram in *ab* plane (down) of FAp.

XRPD patterns of HAp and FAp (Figs. 3 and 4, respectively) are very similar, but the peaks in the XRPD pattern of FAp are mildly shifting to higher values of 2θ angles indicating that unit cell of FAp is smaller.

The structures of HAP and FAp are presented in Figs. 3 and 4, respectively. The both structures crystallize in the hexagonal space group $P6_3/c$, with two formula units $Ca_5(PO_4)_3OH$ per unit cell, for HAp and with two formula units $Ca_5(PO_4)_3F$ for FAp. The lattice parameters for both structures are presented in Table II, from which it can be seen that unit cell of FAp is slightly smaller because of the substitution of OH^- group with F atom. Also, the crystallites of HAp and FAp are more elongated along the *c* axis (Table II) which is in accordance with rod–like morphology of particles observed by FESEM.

Rietveld refinement showed that there is no deficiency of Ca atoms in the structures of HAp and FAp, so the ratio Ca/P < 1.67 obtained by EDS analysis can be attributed to the errors of this method.

TABLE III
SELECTED BOND LENGTHS (Å)FOR HAP AND FAP

HAp	FAp
Ca1–O1 2.401	Ca1–O1 2.392
Ca1–O2 2.452	Ca1–O2 2.449
Ca1–O3 2.836	Ca1–O3 2.820
Ca2–O1 2.703	Ca2–O1 2.689
Ca2–O2 2.365	Ca2–O2 2.367
Ca2–O3 2.329	Ca2–O3 2.338
Ca2–**O4** 2.396	Ca2–**F1** 2.312
P1–O2 1.536	P1–O1 1.537
P1–O1 1.535	P1–O2 1.542
P1–O3 1.537	P1–O3 1.542

The selected bond lengths for HAp and FAp are listed in Table III. The Ca2–F1 bond length in FAp is shorter than Ca2–O4 (O4 is from the OH^- group) bond length in HAp, which can possibly be the reason for smaller unit cell of FAp.

IV. CONCLUSION

The pure, rod-like nanoparticles of HAp and FAp, suitable for application in teeth treatment, have been prepared by simple hydrothermal synthesis. Due to the presence of very small particles, using these materials in filling the tooth's canal is more preferable in comparison with similar materials but with microsized particles .

The future investigations will be oriented towards the synthesis of nanosized Ca-deficient HAp and FAp powders that could be doped further with different metal ions (Mg^{2+}, Si^{4+}, Na^{+}, Cu^{2+}, etc.). Also, the composite materials will be prepared by mixing of the appropriate ratio of HAp and FAp and their potential use as dental material will be examined.

ACKNOWLEDGMENT

The authors wish to acknowledge the financial support for this research from the Ministry of Education, Science and Technological Development of the Republic of Serbia through project III 45019.

REFERENCES

[1] S. V. Dorozhkin, "Calcium orthophosphates Occurrence, properties, biomineralization, pathological calcification and biomimetic applications," Biomatter, vol. 1, no. 2, pp. 121–164, 2011.

[2] L. L Hench, "Bioceramics," J. Am. Ceram. Soc., vol. 81, pp. 1705–1728, 1998.

[3] M. M. Taheri, M. R. Shirdar, A. Keyvanfar, A. Shafaghat, "Evaluating hydrothermal synthesis of fl uorapatite nanorods: pH and temperature," J Exp Nanosci., vol. 12, no. 1, pp. 83–93, 2017.

[4] X. Ge, Y. Leng, C. Bao, S. L. Xu, R. Wang, F. Ren, "Antibacterial coatings of fluoridated hydroxyapatite for percutaneous implants," J Biomed Mater Res A, vol. 95, pp. 588–599, 2010.

[5] V. Stanić, S. Dimitrijević, D. G. Antonović, B. M. Jokić, S. P. Zec, S. T. Tanasković, S. Raičević, "Synthesis of fluorine substituted hydroxyapatite nanopowders and application of the central composite design for determination of its antimicrobial effects," Appl Surf Sci., vol. 290, pp. 346–352, 2014.

[6] J. Zhao, X. Dong, M. Bian, J. Zhao, Y. Zhang, Y. Sun, J.H. Chen, X.H., Wang, "Solution combustion method for synthesis of nanostructured hydroxyapatite, fluorapatite and chlorapatite," Appl Surf Sci., vol. 314, pp. 1026–1033, 2014.

[7] Đ. Janaćković, I. Petrović-Prelević, Lj. Kostić-Gvozdenović, R. Petrović, V. Jokanović, D. Uskoković, "Influence of synthesis parameters on the particle sizes of nanostructured calciumhydroxyapatite," Key Eng. Mater, vol. 203, pp. 192–195, 2001.

[8] Đ. Veljović, E. Palcevskis, A. Dindune, S. Putić, I. Balać, R. Petrović, Đ. Janaćković, "Microwave sintering improves the mechanical properties of biphasic calcium phosphates from hydroxyapatite microspheres produced from hydrothermal processing," J Mater Sci., vol. 45, pp. 3175–3183, 2010.

[9] B. Jokić, D. Radmilović, D. Drmanić, S. Drmanić, R. Petrović, Đ. Janaćković, "Synthesis and characterization of monetite and hydroxyapatite whiskers obtained by a hydrothermal method," Ceram Int., vol. 37, pp. 167–173, 2011.

[10] Ž. Radovanović, B. Jokić, Đ. Veljović, S. Dimitrijević, V. Kojić, R. Petrović, Đ. Janaćković, "Antimicrobial Activity and Biocompatibility of Ag^{+} and Cu^{2+} doped biphasic Hydroxyapatite/α-Tricalcium phosphate Obtained from Hydrothermally Synthesized Ag^{+} and Cu^{2+} doped Hydroxyapatite, Appl Surf Sci., vol. 307, pp. 513–519, 2014.

The fabrication of dental insert based on magnesium doped hydroxyapatite and its shear bond strength with Maxcem dental cement

Tamara Matić, Maja Ležaja Zebić, Vesna Miletić, Sanja Jevtić, Rada Petrović, Djordje Janaćković, Djordje Veljović

***Abstract*—The polymerization shrinkage (PS) presents the cause of the secondary caries, one of the most common reasons for high failure rate of dental restorations. In order to lower PS and improve the lifespan of dental restorations, inorganic dental inserts based on hydroxyapatite (HAP) have been proposed. The aim of this study was to fabricate dental inserts based on hydroxyapatite doped with 5 mol. % of magnesium ions (Mg-HAP) and investigate its bonding ability with commercially available dental cement for possible application in restorative dentistry. The Mg-HAP inserts were characterized using X-ray diffraction (XRD) analysis, energy dispersive X-ray (EDX) analysis and scanning electron microscopy (FE-SEM). Bonding ability of the untreated inserts with Maxcem cement was measured by shear bond strength (SBS) test, and the type of fracture was analysed. The obtained average SBS value of the untreated Mg-HAP inserts with Maxcem cement was 3.1 MPa, with "adhesive" fracture type.**

***Index Terms*— hydroxyapatite; magnesium; shear bond strength; dental insert**

I. Introduction

DENTAL resin-based composites (RBCs) have been widely used in restorative dentistry for repairing decayed or damaged tooth structure [1]. They have gained popularity owing to their extraordinary aesthetic appearance, especially their ability to vary restorations nuance in accordance with patients natural teeth colour. On the other hand, majority of performed restorations are in fact replacements of the failed restorations, due to high failure rate of the RBCs (up to 15 %) [2], and their relatively short lifespan (around 6 years) [3]. The most common reason for restoration failure is connected to the formation of secondary caries at the tooth-restoration margins [4-6].

During the curing of the RBCs, the distance between monomers decreases, which manifests as contraction of the polymer, known as a polymerization shrinkage (PS). The PS causes a micro-gap formation at the tooth-restoration margins which enables penetration of oral fluids and microorganisms toward the dental pulp. Bacteria (e.g. *Streptococcus mutans*) metabolize sugars to lactic acid and lowers pH, which demineralizes hydroxyapatite in the tooth structure and causes the caries formation. Moreover, large restorations in the posterior region have high failure rate due to RBCs inability to bear a high cyclic occlusal loading during mastication [7]. In addition, polymerization shrinkage imposes stresses on the tooth-restoration interface, which increases the chance of an interfacial failure of the restoration.

In order to lower the polymerization shrinkage of RBCs, inorganic dental inserts based on hydroxyapatite (HAP) were introduced [8,9]. Application of HAP based inserts, not only may improve biological response to the foreign material in the body, owing to HAPs bioactivity, biocompatibility and structural similarity to the inorganic part of the natural human bones and teeth, but also could improve restoration lifespan by decreasing the polymerization shrinkage. It was previously reported that utilization of HAP inserts has lowered the polymerization shrinkage by nearly three times [8]. In addition, HAP has intrinsic radio-opaque property and similar hardness to the natural tooth [10].

Magnesium ions were reported to stabilize β– tricalcium phosphate (β–TCP) phase during sintering process of HAP, by increasing its transition temperature into α–polymorph [11,12], which is undesirable phase, as it lowers the mechanical properties of the end product. Moreover, doping HAP with magnesium ions leads to a higher cell proliferation as well as to increased bioactivity, biocompatibility and improved mechanical properties [13-15].

The aim of this study was to fabricate dental inserts based on hydroxyapatite doped with 5 mol. % of magnesium ions and test the shear bond strength with commercially available dental cement for possible application in restorative dentistry.

Tamara Matić is with the Faculty of Technology and Metallurgy, University of Belgrade, Karnegijeva 4, 11120 Belgrade, Serbia (e-mail: tmatic@tmf.bg.ac.rs).

Maja Ležaja Zebić is with the School of Dental Medicine, University of Belgrade, Rankeova 4, 11000 Belgrade, Serbia (e-mail: dr.maja.zebic@gmail.com).

Vesna Miletić is with the School of Dental Medicine, University of Belgrade, Rankeova 4, 11000 Belgrade, Serbia (vesna.miletic@stomf.bg.ac.rs).

Sanja Jevtić is with the Faculty of Technology and Metallurgy, University of Belgrade, Karnegijeva 4, 11120 Belgrade, Serbia (e-mail: sanja@tmf.bg.ac.rs).

Rada Petrović is with the Faculty of Technology and Metallurgy, University of Belgrade, Karnegijeva 4, 11120 Belgrade, Serbia (radaab@tmf.bg.ac.rs).

Djordje Janaćković is with the Faculty of Technology and Metallurgy, University of Belgrade, Karnegijeva 4, 11120 Belgrade, Serbia (nht@tmf.bg.ac.rs).

Djordje Veljović is with the Faculty of Technology and Metallurgy, University of Belgrade, Karnegijeva 4, 11120 Belgrade, Serbia (djveljovic@tmf.bg.ac.rs).

II. Materials and Methods

Nanosized calcium-hydroxyapatite powder doped with 5 mol. % magnesium (Mg-HAP) was synthesized by a previously described modified hydrothermal method [16-18], starting from a solution of $Ca(NO_3)_2 \cdot 4H_2O$ (11.80 g), $Na_2H_2EDTA \cdot 2H_2O$ (11.18 g), $NaH_2PO_4 \cdot 2H_2O$ (4.68 g), urea (12.00 g) and $Mg(NO_3)_2 \cdot 6H_2O$ (0.641 g). The precursor solution was thermally treated in an autoclave at 160 °C for 3 hours, under pressure of 8 bar. The obtained precipitates were collected by vacuum filtration, flushed with deionized water and dried at 105 °C.

Synthesized Mg-HAP powder was pressed into cylindrical green compacts (d = 6 mm, h = 1.6 mm) by two-step pressing process using a high-quality cylindrical steel mould. The powder was firstly uniaxially pressed at 100 MPa and then isostatically pressed at 400 MPa (CIP-15, MTI Corporation, Richmond, USA), in duration of 1 min per step. Green compacts were then sintered at 1200 °C in high-temperature furnace (Elektron, Banja Koviljaca, Serbia) for 2 h at the heating rate of 20 °C/min. After sintering, the Mg-HAP inserts were naturally cooled down to the room temperature.

In order to determine elemental composition of the Mg-HAP powder, an energy dispersive X-ray (EDX) analysis was performed on Oxford Inca 3.2 coupled with the SEM Jeol JSM 5800, operated at 20 keV. The results of EDX analysis were presented as average arithmetical values of three measurements taken at different areas of the sample, at the 1000 times magnification.

Phase composition of Mg-HAP powder and sintered insert was determined by X-ray diffraction analysis (XRD) conducted on Ital Structure APD 2000 X-ray diffractometer with CuKα radiation (1.54 Å) in the 2θ angle ranging from 20° to 60° with a scan step 0.02° s^{-1}. Identification of the phases was accomplished by comparing the experimental XRD patterns with standards cards: JCPDS 09-0169 and JCPDS 09-0432 for β–TCP and HAP respectively.

A scanning electron microscope (Tescan FE-SEM Mira 3 XMU, Tescan a.s., Brno, Czech Republic) operated at 20 keV was used to characterize the surface microstructure and cross section of the Mg-HAP inserts. The samples were coated with gold using a sputter coater (Polaron SC503, Fisons Instruments, Ipswich, UK).

Shear bond strength between Mg-HAP inserts commercially available restorative material – Maxcem cement (Maxcem Elite, Kerr, Italy) was measured by SBS test. As Maxcem cement belongs to group of "self-etch" and "self-adhesive" restorative materials, it was applied on untreated inserts surface, solely copiously rinsed with water and mildly air-dried. A custom-made mould used in this study was of the following dimensions: 6 mm in lower diameter and 2 mm high for the insert placement; 3 mm in upper diameter and 2 mm high for application of the restorative material. Each sample was prepared by placing an insert into the mould and applying Maxcem cement on the inserts top surface. In order to initiate polymerization LED light-curing unit (LEDition, Ivoclar Vivadent, Schaan, Liechtenstein) was used during 10 s at an intensity of 800 mW/cm2.

SBS test was performed on universal testing machine (Force Gauge PCE-FM200, Southampton, United Kingdom), using a knife-edge shearing blade. Force was applied at 1 mm distance of the restorative material-insert interface at 1 mm/min speed, until fracture. SBS (τ) was calculated using the maximum force reached (*F* [N]) and the bonded surface area (*A* [mm^2]) following the equation:

$$\tau = F/A \text{ [MPa]} \quad (1)$$

The type of fracture was analysed and classified as: "adhesive" (fracture at the material-insert interface), "cohesive" (fracture in the insert or material) or "mixed".

III. Results

The results of EDX analysis of the Mg-HAP powder is shown in the Table I. The presence of magnesium as dopant in Mg-HAP insert was confirmed. The Ca/P ratio was less than stoichiometric (1.67), implying that obtained hydroxyapatite was deficient in calcium and expected to undergo HAP-β–TCP phase transition during sintering. Illustration must be inserted in the text.

TABLE I
THE COMPOSITION OF THE MG-HAP POWDER

Ca [at.%]	P [at.%]	Mg [at.%]	Ca/P ratio
11.59	8.00	0.27	1.45

The XRD pattern of the as-synthesized Mg-HAP powder sample show characteristic peaks for calcium-hydroxyapatite, suggesting that monophasic powder was obtained (Fig. 1.a). The XRD pattern of the Mg-HAP sample upon sintering at 1200 °C is shown in Fig. 1.b. It is well documented that calcium-hydroxyapatite undergoes phase transition into β–TCP at 800 °C, and further transforms into α–TCP at 1125 °C [11,19]. The β–α phase transition is non desirable as it lowers mechanical properties of the end product [12]. According to the XRD pattern of the Mg-HAP sintered insert it may be concluded that β–α phase transition did not take place. Magnesium is reported to be able to stabilize β–TCP structure by postponing the β–α transition to the higher temperatures [11,13,15], which explains the absence of the α–TCP phase upon sintering at 1200 °C. The phase composition of the Mg-HAP inserts indicated that biphasic structure with equally dominant HAP and β–TCP phases was obtained.

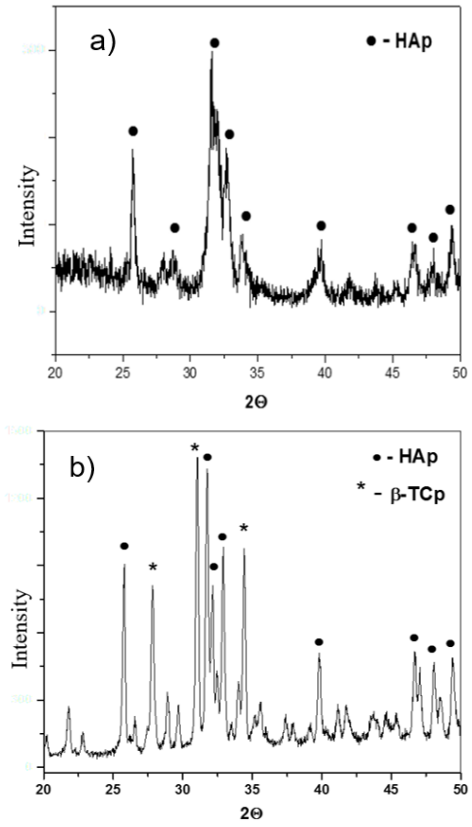


Fig. 1. XRD patterns of: a) Mg-HAP powder b) Mg-HAP insert sintered at 1200 °C

The results of the SEM analysis of the surface microstructure and cross section of the Mg-HAP inserts are shown in Fig. 2. As seen in the micrograph of the insert cross section (Fig. 2.a), the remained porosity upon sintering was relatively low, with spherical pores around 500 nm in size. In Fig. 2.b it is noticeable that the Mg-HAP inserts consisted of polyhedral grains varying in size from few hundred nanometres to about 1.5 micrometres.

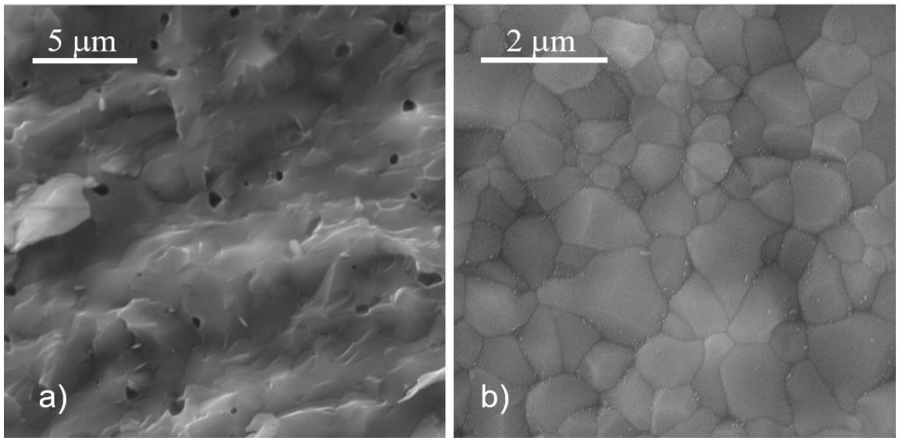


Fig. 2. SEM micrographs of the Mg-HAP inserts sintered at 1200 °C: a) cross section; b) surface microstructure

The results of shear bond strength (SBS) test shown in Table II indicate a relatively low bonding ability of the Maxcem cement and the untreated Mg-HAP inserts, with an average SBS value of 3.1 MPa. Although Maxcem cement is a "self-etch" and "self-adhesive" cement, when applied to the untreated surface of the Mg-HAP inserts, the bonding ability was relatively poor. This is further confirmed with the type of fracture being mostly "adhesive". The obtained results are in good agreement with previously reported study [8] where different restorative materials have manifested low bonding ability (avg. SBS values ranging from 0.7-2.2 MPa) with the HAP inserts and "adhesive" fracture type, when applied without an adhesive pre-treatment of the inserts surface. Possible improvement of the SBS should involve pre-treatment of the inserts surface, which will be an object of the future studies.

TABLE II
THE RESULTS OF THE SBS TEST

Sample	SBS [MPa]	Fracture type
1	4.36	adhesive
2	4.25	adhesive
3	2.47	mixed
4	1.96	adhesive
5	2.29	adhesive
avg.	**3.07**	

IV. CONCLUSION

Dental inserts based on hydroxyapatite doped with 5 mol. % of magnesium ions were successfully fabricated. EDX analysis confirmed presence of magnesium ions as dopant species. The results of XRD analysis of the Mg-HAP inserts sintered at 1200 °C indicated that biphasic HAP-β–TCP structure was obtained. According to SEM analysis, the obtained inserts consisted of polyhedral grains varying in size from few hundred nanometres to about 1.5 micrometres. The remained porosity was relatively low, with spherical pores around 500 nm in size. Shear bond strength (SBS) value of the untreated Mg-HAP inserts with Maxcem cement was 3.1 MPa, with "adhesive" type of the fracture. Although Maxcem cement is a "self-etch" and "self-adhesive" cement, the obtained results indicate that there was low adhesion of the cement with untreated surface of the insert. Possible improvement of the SBS should involve pre-treatment of the inserts surface, which will be an object of the future studies.

ACKNOWLEDGMENT

The authors wish to acknowledge the financial support for this research from the Ministry of Education, Science and

Technological Development, Republic of Serbia through the project III45019 and ON172007.

REFERENCES

[1] R.L. Bowen, "Properties of a silica-reinforced polymer for dental restorations," J Am Dental Assoc, vol. 66, pp. 57–64, Jan, 1963.

[2] R. Hickel, C. Kaaden, E. Paschos, V. Buerkle, F. Garcia-Godoy, J. Manhart, "Longevity of occlusally-stressed restorations in posterior primary teeth," Am J Dent, vol. 18, no. 3, pp. 198–211, Jun, 2005.

[3] M. Downer, N. Azli, R. Bedi, D. Moles, D. Setchell, "Dental restorations: how long do routine dental restorations last? A systematic review," Br Dent J, vol. 187, no. 8, pp. 432–439, Oct, 1999.

[4] L. Marks, K. Weerheijm, W. van Amerongen, H. Groen, L. Martens, "Dyract versus Tytin class II restorations in primary molars: 36 months evaluation," Caries Res, vol. 33, no. 5, pp. 387–392, Sep, 1999.

[5] I. Mjör, "Glass-ionomer cement restorations and secondary caries: a preliminary report," Quintessence Int, vol. 27, no. 3, pp. 171–174, Mar, 1996.

[6] I. Mjör, O. Toffenetti, "Secondary caries: a literature review with case reports," Quintessence Int, vol. 31, no. 3, pp. 165–179, Mar, 2000.

[7] P. E. G. A. Campos, M. O. Barceleiro, H. R. Sampaio-Filho, and L. R. M. Martins, "Evaluation of the Cervical Integrity During Occlusal Loading of Class II Restorations," Oper Dent, vol. 33, no. 1, pp. 59-64, Jan, 2008.

[8] M. Lezaja, Dj. Veljovic, Dj. Manojlovic, M. Milosevic, N. Mitrovic, Dj. Janackovic, V. Miletic, "Bond strength of restorative materials to hydroxyapatite inserts and dimensional changes of insert-containing restorations during polymerization," Dent Mater, vol. 31, no. 2, pp. 171–181, Feb, 2015.

[9] G. Ayoub, Dj. Veljovic, M.L. Zebic, V. Miletic, E. Palcevskis, R. Petrovic, Dj. Janackovic, "Composite nanostructured hydroxyapatite/yttrium stabilized zirconia dental inserts – The processing and application as dentin substitutes," Ceram Int, vol. 44, no. 15, pp. 18200–18208, 2018.

[10] C. Santos, Z.B. Luklinska, R.L. Clarke, K.W. Davy, "Hydroxyapatite as a filler for dental composite materials: mechanical properties and in vitro bioactivity of composites," J Mater Sci Mater Med, vol. 12, no. 7 pp. 565–573, Jul, 2001.

[11] M. Frasnelli, V.M. Sglavo, "Effect of Mg doping on beta-alpha phase transition in TCP bioceramics," Acta Biomater, vol. 33, pp. 283-289, Mar, 2016.

[12] R. Enderle, F. Götz-Neunhoeffer, M. Göbbels, F. Müller, P. Greil, "Influence of magnesium doping on the phase transformation temperature of β–TCP ceramics examined by Rietveld refinement," Biomaterials, vol. 26, no. 17, pp. 3379–3384, Jun, 2005.

[13] H.S. Ryu, K.S. Hong, J.K. Lee, D.J. Kim, J.H. Lee, B.S. Chang, D.H. Lee, C.K. Lee, S.S. Chung, "Magnesia-doped HA/beta-TCP ceramics and evaluation of their biocompatibility", Biomaterials, vol. 25, no. 3, pp. 393–401, Feb, 2004.

[14] W. Hue, K. Dahlquist, A. Banarjee, A. Bandyopadhyay, S. Bose, "Synthesis and characterization of tricalcium phosphate with Zn and Mg based dopants," J Mater Sci Mater Med, vol. 19, no. 7, pp. 2669–2677, Jul, 2008.

[15] I. Cacciotti, A. Bianco, M. Lombardi, L. Montanaro, "Mg-substituted hydroxyapatie nanopowders: Synthesis, thermal stability and sintering behavior," J Eur Cer Soc, vol. 29, no. 14, pp. 2969-2978, Nov, 2009.

[16] Dj. Janackovic, I. Petrovic-Prelevic, Lj. Kostic-Gvozdenovic, R. Petrovic, V. Jokanovic, D. Uskokovic, "Influence of synthesis parameters on the particle sizes of nanostructured calcium-hydroxyapatite," Key Eng Mater, vols. 192–195, no. 203, pp. 203–206, Sept, 2000.

[17] Dj. Veljovic, E. Palcevskis, A. Dindune, S. Putic, I. Balac, R. Petrovic, Dj. Janackovic, "Microwave sintering improves the mechanical properties of biphasic calcium phosphates from hydroxyapatite microspheres produced from hydrothermal processing," J Mater Sci, vol. 45, no. 12, pp. 3175–3183, Aug, 2010.

[18] B. Jokic, D. Radmilovic, D. Drmanic, S. Drmanic, R. Petrovic, Dj. Janackovic, "Synthesis and characterization of monetite and hydroxyapatite whiskers obtained by a hydrothermal method," Ceram Int, vol. 37, no. 1, pp. 167–173, Jan, 2011.

[19] S.V. Dorozhkin, "Calcium orthophosphate bioceramics," Ceram Int, vol. 41, no. 10, pp. 13913–13966, Dec, 2015.

Nova metoda za odgrevanje uzoraka amorfnih legura povorkom pravouganih strujnih impulsa modulisanog trajanja

Jelena Orelj, Nebojša Mitrović

Apstrakt— Nalaženje optimalnih svojstava magnetno mekih amorfnih/nanokristalnih legura je neophodno radi dostizanja najboljih funkcionalnih karakteristika savremenih električnih naprava u kojima se one koriste. Za postizanje ovog cilja neophodna je optimizacija termičkih tretmana koje je potrebno prilagoditi primenama ispitivanih legura. Magnetni senzori na bazi magnetoimpedansnog MI-efekta načinjeni od amorfnih / nanokristalnih žica - mikrožica zahtevaju potpuno specifične termičke tretmane, najvećim delom zasnovane na odgrevanju strujnim impulsima. U tu svrhu je razvijena nova metoda odgrevanja pomoću povorke pravougaonih strujnih impulsa modulisanog trajanja.

Ključne reči— Amorfne legure, nanokristalne legure, odgrevanje, povorka strujnih impulsa, modulisano trajanje

I. UVOD

OPTIMIZACIJA parametara termičkih tretmana sa stanovišta postizanja najboljih funkcionalnih svojstava (magnetnih i mehaničkih karakteristika) amorfnih i nanokristalnih legura se može izvoditi ili klasičnom metodom (odgrevanje u peći [1-3]) ili brojnim alternativnim metodama (odgrevanje strujnim impulsima [3-10], laserom [11]), mikrotalasnim zračenjem [12]). Sve navedene metode imaju razne varijante kod kojih se varira: prisustvo spoljašnjeg magnetnog polja [13]/mehaničkog naprezanja [14], maksimalno dostignuta temperatura odgrevanja, brzina grejanja, dužina tretmana, itd.

Cilj termičkih tretmana je ili dostizanje optimalno relaksirane mikrostrukture (klasične legure na bazi gvoždja Fe i kobalta Co) ili formiranje nanostrukture (sistemi nanokristalnih legura na bazi gvoždja Fe). Amorfne legure na bazi kobalta poseduju najbolja magnetna svojstva u stanju optimalno relaksirane amorfne strukture gde su otklonjene slobodne zapremine i zaostala naprezanja nastala tokom procesa brzog hladjenja prilikom dobijanja samih žica. Prilikom ispitivanja optimalnih svojstava nanokristalnih legura na bazi gvoždja evidentirano je očuvanje dobrih mehaničkih svojstava pri odgrevanju jednosmernim strujnim impulsima [5-7] nasuprot uzoraka odgrevanih u peći na temperaturama nešto nižim od temperature kristalizacije.

Magnetni senzori na bazi magnetoimpedansnog efekta kod amorfnih/nanokristalnih legura (u obliku traka ili žica) su posebno interesantni za ispitivanje optimizacije njihovih svojstava primenom metoda odgrevanja strujnim impulsima [15-17].

Jelena Orelj – Univerzitet u Kragujevcu, Fakultet tehničkih nauka Čačak, Svetog Save 65, Srbija (e-mail: jelena.orelj@ ftn.kg.ac.rs.)

Nebojša Mitrović – Univerzitet u Kragujevcu, Fakultet tehničkih nauka Čačak, Svetog Save 65, Srbija (e-mail: nebojsa.mitrovic@ ftn.kg.ac.rs).

Metode odgrevanja na osnovu Džulovog efekta baziraju se na korišćenju ili jednosmerne (DC) ili naizmenične (AC) struje.

II. EKSPERIMENTALNI DEO

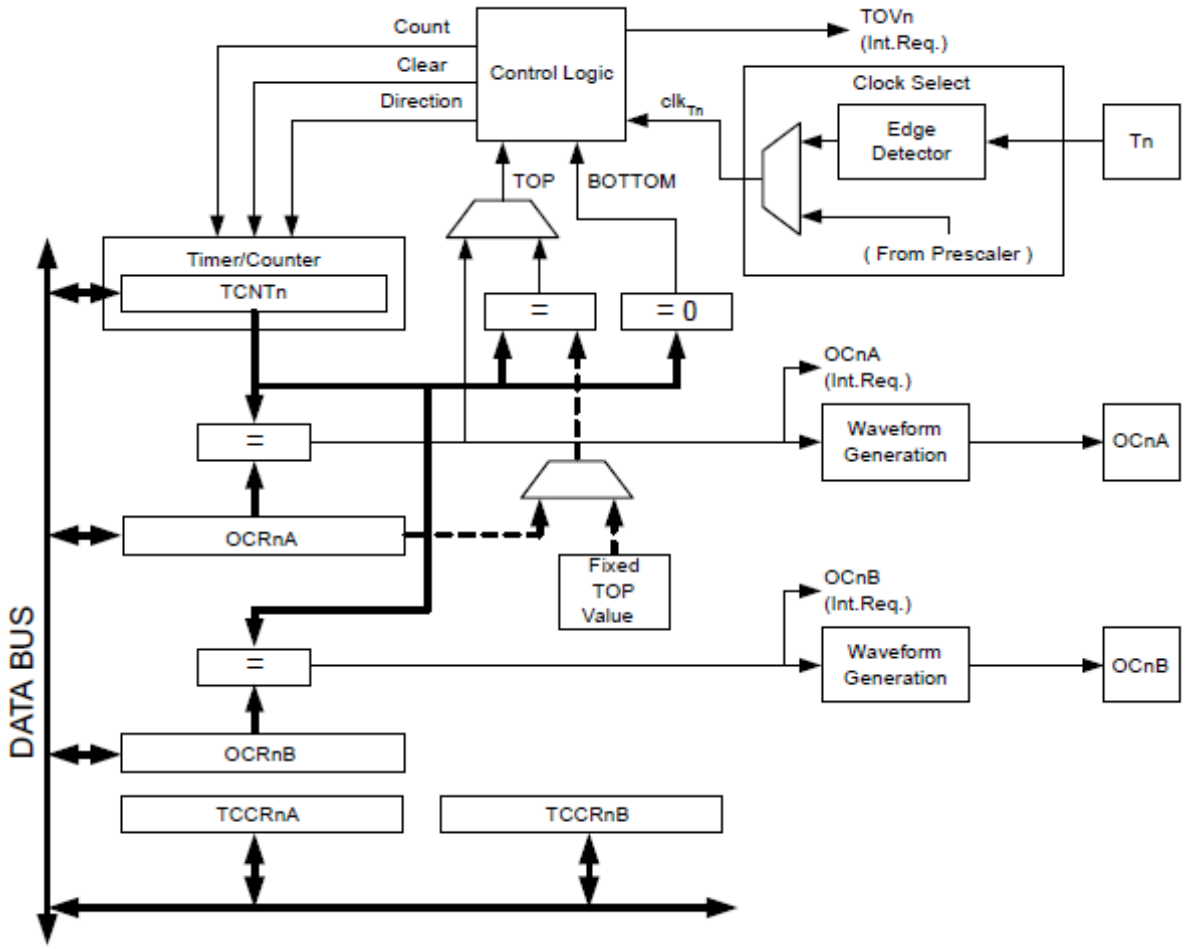
Pripremljena je potpuno nova aparatura za odgrevanje uzoraka amorfnih legura (mikrožica ili traka) tzv. metodom odgrevanja sa povorkom pravouganih strujnih impulsa modulisanog trajanja, odnosno modulisanog faktora popune. Radi se o cikličnom odgrevanju uzoraka širinski modulisanim strujnim impulsima, pri čemu je modulacija trajanja impulsa izvršena po kosinusnom zakonu trajanja četvrtine peroioda. Dakle, kroz uzorak se u ciklusima propušta PWM (eng. Pulse Width Modulation) strujni signal određenog trajanja, koje se utvrđuje na osnovu kritične vrednosti temperature do koje sme da se odgreva uzorak, te se na taj način vrši jednostavno, efikasno i kontrolisano odgrevanje uzorka. S obzirom na opisani postupak, metoda je nazvana Cyclic Current Annealing by Square Pulse Width Modulation – CCASPWM.

Suština algoritma za odgrevanje datog uzorka strujnim PWM signalom se sastoji u tome da se na osnovu odredjene temperature kristalizacije T_x (npr. korišćenjem DSC/DTA analize), odredi broj impulsa različitog faktora ispune (eng. *Duty Cycle*) i to od najšireg do najužeg impulsa, kojima će se uzorku predati odredjena količina toplote, odnosno povećati temperatura približno do temperature T_x. Pri tome treba voditi računa da se omogući postepeno približavanje temperaturi kristalizacije T_x, kako ne bi došlo do kristalizacije uzorka, koja je nepovoljan efekat za magnetna svojstva amorfnih žica na bazi kobalta Co. Dakle, u ovom slučaju cilj metode odgrevanja je dostizanje optimalno relaksirane amorfne mikrostrukture.

Uspešnost ove metode će se ogledati upravo u što boljoj kontroli procesa odgrevanja i efikasnom dobijanju optimalnih magnetnih svojstava u odnosu na do sada korišćene metode strujnog odgrevanja. Promenljivošću faktora ispune povorke impulsa upravo se obezbeđuje efikasna kontrola nad procesom odgrevanja.

S obzirom da realizacija PWM signalnog generatora u diskretnoj analognoj tehnici može da da loše i nepouzdane rezultate usled promena parametara ovih kola, generisanje PWM strujnog signala, mnogo preciznije i otpornije na smetnje, postignuto je u digitalnoj tehnici.

U tu svrhu korišćen je 8-bitni AVR ATmega328p mikrokontroler, koji kao i mnogi drugi mikrokontroleri ima već implementiran generator PWM signala u vidu tajmera/brojača i odgovarajućih upravljačkih registara, što je prikazano na Sl. 1.



Sl. 1. Pojednostavljena blok šema PWM modula unutar ATmega328p mikrokontrolera [18]

Na ovaj način su izbegnuti svi nedostaci analognih PWM generatora, kao što je nedefinisano stanje signala – glich, šumovi i slično. ATmega328p mikrokontroler ima tri 8 – bitna tajmera/brojača (Timer 0, Timer 1, Timer 2), koji mogu da rade u četiri različita režima rada (Fast mode, Phase Correct mode, Phase and Frequency Correct mode i CTC mode). Pri tome generisani PWM signal može da ima promenljiv faktor ispune i/ili frekvenciju signala.

U konkretnom slučaju, a s obzirom na potrebe tehnike odgrevanja, iskorišćen je ***Phase Correct* (PC)** režim rada koji je realizovan na "nultom" tajmeru (**Timer 0**). Metoda odgrevanja je realizovana sa konstantnom frekvencijom i promenljivim faktorom ispune. Primenjeni režim rada obezbedjuje tačno definisanje trenutaka pojavljivanja i trajanje impulsa kojima se odgreva uzorak, a s obzirom da je zahtevana tačnost reda ms, ovaj režim rada se pokazuje kao odgovarajući izbor.

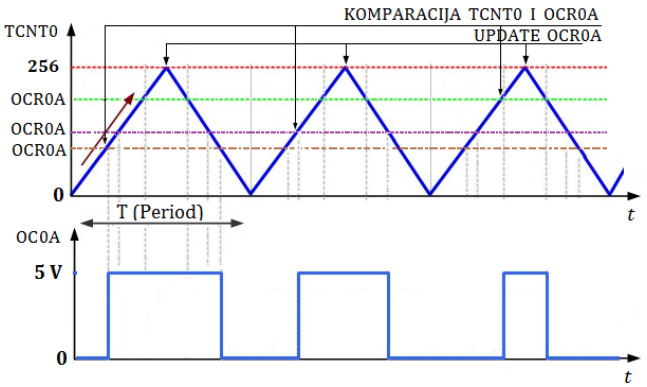
Korišćenjem brojača sa visokom rezolucijom, odnos impuls/pauza se moduliše da odgovara specificiranom nivou analognog signala, pa za dati programski kôd kojim se menja stanje odgovarajućih registara (OCRnA/OCRnB) sa Sl. 1. dobija se željeni PWM signal, Sl. 2 (OCnA/OCnB).

Kako je prilikom odgrevanja uzoraka korišćen niz impulsa PWM signala od najšireg do najužeg impulsa, signal je generisan na OCnA, odnosno na OC0A izlazu (OCnB predstavlja invertovan OCnA).

Frekvencija (a samim tim i perioda) željenog PWM signala za *PC* režim rada može se izračunati na osnovu sledećeg obrasca:

$$f_{OCnxPCPWM} = \frac{f_{clk}}{N \times 510}, \qquad (1)$$

gde je N faktor skaliranja: 1, 8, 64, 256 ili 1024.



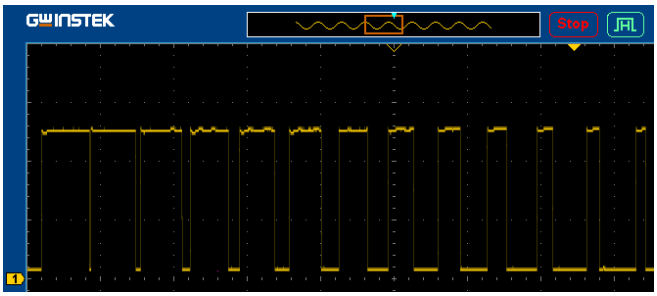
Sl. 2. Vremenski dijagram PWM signala za Phase Correct režim rada na OC0A *izlazu [18]*

Generisanom PWM signalu se pristupa preko digitalnih pinova 5 (OC0B) i 6 (OC0A), mikrokontrolera ATMega328P (Sl. 5). To je zapravo izlaz "nultog" tajmera/brojača. U konkretnom slučaju signalu se pristupa sa pina 6. Menjanjem stanja u komparacionim registrima OCR0A i

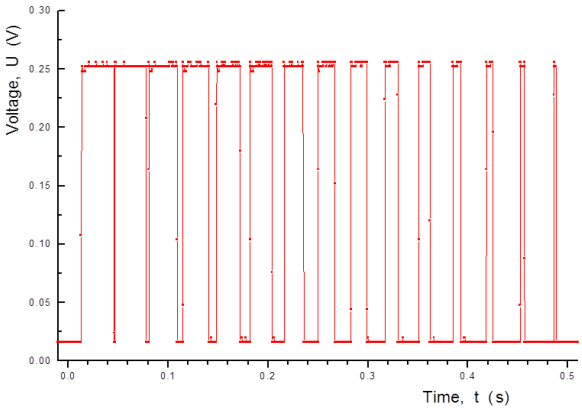
OCR0B (koja se porede sa sadržajem tajmera/brojača TCNT0) dobija se odgovarajuća sekvenca impulsa različitog faktora popune. Programski kôd kojim je generisan željeni PWM strujni signal na pinu 6:

```
#include <avr/io.h>
#include <util/delay.h>

int main(void){

 // definisanje izlaza
  pinMode(5, OUTPUT);
  pinMode(6, OUTPUT);

  // inicijalizacija tajmera  TIMER0
  // fphase_correctPWM=16 MHz/(1024*510)=30.64 Hz
  //Period Tphase_correctPWM=32.64 ms

  TCCR0A = 0b10110001; //Phase Correct rezima rada
  TCCR0B = 0b0000101; //faktor skaliranja 1024
  TCNT0 = 0;    // Reset TCNT0 (Timer0/brojac)
  OCR0A = 0;    // Inicijalizacija registara A i B
  OCR0B = 0;

unsigned char duty_cyc_a, duty_cyc_b;
duty_cyc_a=255;      //Inicijalizacija registra OC0A      duty_cyc_b=0;
//Inicijalizacija registra OC0B
int n=16;
  for(int br_T=0;br_T<n;br_T++){
    if(duty_cyc_a>=0) {
      OCR0A=duty_cyc_a;
        duty_cyc_a-=n+1;
    }
 }
return 0;
}
```

Za potrebe strujnog odgrevanja amorfne žice generisan je PWM signal periode ~ 32 ms (faktor skaliranja N=1024), promenljivog faktora ispune u toku petnaest perioda (koliko traje jedna PWM sekvenca), kao što se može videti na ekranu digitalnog osciloskopa na Sl. 3. ili grafički na Sl. 4.



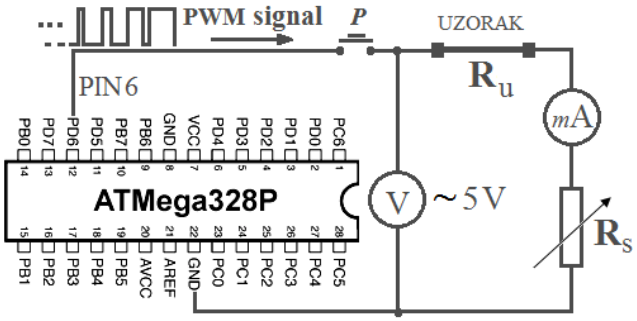
Sl. 3. Povorka pravouganih impulsa modulisanog trajanja na ekranu digitalnog osciloskopa.



Sl. 4. Grafički prikaz povorke pravouganih impulsa modulisanog trajanja.

Tu promenu je moguće kontrolisati po sinusnom ili nekom drugom zakonu, što ovoj metodi daje prednost u pogledu kontrole procesa odgrevanja uzorka.

Formirana povorka naponskih impulsa se dovodi na uzorak sa električnom otpornošću R_u (Sl 5.) i na taj način formira povorka strujnih impulsa na ispitivanom uzorku amorfne legure.



Sl. 5. Blok šema formiranja povorke pravouganih strujnih impulsa modulisanog trajanja na uzorku amorfne legure - električna otpornost R_u.

Za buduće zahtevnije, superiornije, pa i komplikovanije algoritme strujnog odgrevanja uzoraka amorfnih legura (žice), verovatno će biti interesantna preostala dva režima rada tajmera/brojača ili uopšte generisanje PWM signala sa promeljivom frekvencijom, tj. promenljivom periodom unutar PWM signala. Za sada su se autori odlučili da koriste što jednostavniju, a veoma preciznu tehniku.

III. Zaključak

Savremeni razvoj digitalne tehnike za generisanje PWM signala je omogućio značajno unapredjenje specifičnih termičkih tretmana amorfnih legura. U ovom radu je pokazano korišćenje 8-bitnog mikrokontrolera za razvoj nove metode odgrevanja pomoću povorke pravougaonih strujnih impulsa modulisanog trajanja. Za potrebe odgrevanja strujnim impulsima, amorfnih magnetno mekih žica koje će se koristiti za izradu magnetnog senzora na bazi magnetoimpedasnog efekta, uspešno je generisana povorka pravougaonih PWM signala periode ~ 32 ms.

Zahvalnica

Ovaj rad je delimično finansiran od strane Ministarstva prosvete, nauke i tehnološkog razvoja Republike Srbije (projekat br. OI 172057).

Literatura

[1] K. Suzuki, N. Ito, J. S. Garitaonandia, J. D. Cashion, G. Herzer, "Local random magnetocrystalline and macroscopic induced anisotropies in magnetic nanostructures", *Journal of Non-Crystalline Solids*, Vol. 354, pp. 5089- 5092. 2008.

[2] L. Zhu, H. Zheng, S.S. Jiang, Y.G. Wang, „Modulating the crystallization process of $Fe_{82}B_{12}C_6$ amorphous alloy via rapid annealing", *Journal of Alloys and Compounds*, Vol. 785, pp. 328-334, 2019.

[3] M.A.,Willard, M. Daniil, „Nanocrystalline Soft Magnetic Alloys Two Decades of Progress", Handbook of Magnetic Materials, 21, Elsevier, pp. 173-342, 2013.

[4] N. Mitrović, S. Djukić and S. Djurić, "Crystallization of the Fe-Cu-M-Si-B (M = Nb, V) amorphous alloys by direct current Joule heating", *IEEE Transaction on Magnetics*, Vol. MAG-36, pp. 3858-3862, 2000.

[5] P. Allia, P. Tiberto, M. Baricco, F. Vinai, „Improved ductility of nanocrystalline $Fe_{73.5}Nb_3Cu_1Si_{13.5}B_9$ obtained by direct-current Joule heating“, *Applied Physics Letters*, Vol. 63 (20), pp. 2759-2761, 1993.

[6] N. Mitrović, “Magnetoresistance of the $Fe_{72}Cu_1V_3Si_{16}B_8$ amorphous alloys annealed by direct current Joule heating”, *Journal of Magnetism and Magnetic Materials*, Vol. 262, pp. 302–307, 2003.

[7] N. Mitrović, S. Roth, S. Djukić and J. Eckert, “Magnetic softening of metallic glasses by current annealing technique”, ARW PROSIZE “*Properties and Applications of Nanocrystalline Alloys from Amorphous Precursors*”, Kluwer Academic Publishers, pp.331-344, 2005.

[8] J. A. Moya, “Structural and magnetic properties evolution study method using a single ribbon-shaped sample“, *Journal of Magnetism and Magnetic Materials*, Vol. 432, pp. 300-303, 2017.

[9] N. Mitrović, S. Roth and M. Stoica, “Magnetic softening of bulk amorphous FeCrMoGaPCB rods by current annealing technique”, *Journal of Alloys and Compounds*, Vol. 434-435, pp. 618-622, 2007.

[10] N. Mitrović, S. Kane, S. Roth, A. Kalezić-Glišović, C. Mickel and J. Eckert, “The precipitation of nanocrystalline structure in the Joule heated $Fe_{72}Al_5Ga_2P_{11}C_6B_4$ metallic glasses”, *Journal of Mining and Metallurgy, Section B,* Vol. 48. B pp. 319-324, 2012.

[11] S. Katakam, A. Devaraj, M. Bowden, S. Santhanakrishnan, C. Smith, R. V. Ramanujan, S. Thevuthasan, R. Banerjee, and N. B. Dahotre “Laser assisted crystallization of ferromagnetic amorphous ribbons: A multimodal characterization and thermal model study”, *Journal of Applied Physics* Vol. 114, article numb. 184901, 2013.

[12] G. Kotagiri, S.D. Ramarao, G. Markandeyulu, “Magnetoimpedance studies on laser and microwave annealed $Fe_{66}Ni_7Si_7B_{20}$ ribbons”, *Journal of Magnetism and Magnetic Materials*, Vol. 382, pp. 43-48, 2015.

[13] Hu Li, Aina He, Anding Wang, Lei Xie, Qiang Li, Chengliang Zhao, Guoyang Zhang, Pingbo Chen, “Improvement of soft magnetic properties for distinctly high Fe content amorphous alloys via longitudinal magnetic field annealing“, *Journal of Magnetism and Magnetic Materials* Vol. 471, pp. 110-115, 2019.

[14] V. Zhukova, J. M. Blanco, M. Ipatov, M. Churyukanova, S. Taskaev, A. Zhukov „Tailoring of magnetoimpedance effect and magnetic softness of Fe-rich glass-coated microwires by stress annealing“ *Scientific Reports* Vol. 8, article numb. 3202, 2018.

[15] N.S Mitrović, S.N. Kane, P.V. Tyagi, S. Roth, „Effect of dc-Joule-heating thermal processing on magnetoimpedance of $Fe_{72}Al_5Ga_2P_{11}C_6B_4$ amorphous alloy“, *Journal of Magnetism and Magnetic Materials* Vol. 320, e792-e796, 2008.

[16] J. Liu, H. Shen, D. Xing, and J. Sun**, “**Optimization of GMI properties by AC Joule annealing in melt extracted Co-rich amorphous wires for sensor applications”, *Phys. Status Solidi A* Vol. 211, pp. 1577–1582, 2014.

[17] J. Liu, Z. Du, S. Jiang, H. Shen, Z. Li, D. Xing, W. Ma, J. Sun, “Tailoring giant magnetoimpedance effect of Co-based microwires for optimum efficiency by self-designed square-wave pulse current annealing”, *Journal of Magnetism and Magnetic Materials* Vol. 385, pp. 145–150, 2015.

[18] ATmega328P, 8-bit AVR Microcontroller, Datasheet, available on: http://ww1.microchip.com/downloads/en/DeviceDoc/Atmel-7810-Automotive-Microcontrollers-ATmega328P_Datasheet.pdf

ABSTRACT

In order to attain the best functional characteristics of electric devices made by amorphous or nanocrystalline soft magnetic alloys it is necessary to perform detailed research of their properties. Optimization of applied thermal treatments is the crucial task devoted to development magnetic sensors based on magnetoimpedance (MI) effect in amorphous or nanocrystalline wires-microwires. Different current annealing techniques with a lot of varieties were widely used. Therefore, the new method for cyclic current annealing of amorphous alloy samples by applying square pulse with modulation (CCASPWM) and 32 ms duration was developed.

New Method for Cyclic Current Annealing of Amorphous Alloys by Square Pulse Width Modulation CCASPWM

Jelena Orelj, Nebojša Mitrović

METHODS OF COSMIC MUON IMAGING

(INVITED PAPER)

Istvan Bikit, Department of Physics, Faculty of Sciences, University of Novi Sad, Serbia
Dusan Mrdja, Department of Physics, Faculty of Sciences, University of Novi Sad, Serbia
Kristina Bikit-Schroeder, Department of Physics, Faculty of Sciences, University of Novi Sad, Serbia

ABSTRACT

Cosmic-ray muons can be used for imaging of large structures, or high-density objects with high atomic number. The first task can be performed by measurement of muon absorption within very thick material layers, while the second approach is based on muon multiple scattering. However, the muon imaging of small structures with low atomic number and density was not yet solved appropriately. Our research group has demonstrated recently completely new imaging method by cosmic-ray muons, based on the detection of secondary particles produced by muons in object material (I. Bikit et al, Novel approach to imaging by cosmic-ray muons, EPL 113 (2016) 58001). Novel imaging technique by cosmic-ray muons is based on the detection of secondary produced particles generated within materials and objects by passage of cosmic-ray muons. This method opens up possibility to obtain 2D and 3D images of small objects made of materials with low atomic number. The advances of "conventional" muon imaging systems based on muon absorption and scattering and detection of incoming and scattered muons will be presented, and compared with the new imaging technique. The possible applications of the new imaging technique will be discussed.

Thoron ^{220}Rn Exhalation Rate Measurement: Dependence of the Grain Size

Dunja Antonijević, Luka Rubinjoni, Andrija Janković, Igor Čeliković, Aleksandar Kandić, Boris Lončar

***Abstract*— Radon, ^{222}Rn with its progeny is considered as the second cause of lung cancer after smoking. On the other hand, thoron ^{220}Rn was often neglected since its concentrations in the indoor environment were considered to be smaller compared to radon concentrations and due to its short lifetime compared to radon. Nevertheless, there are regions that have thoron concentrations higher or comparable to radon concentrations. In this contribution, measurements of thoron exhalation rate of different materials as a function of grain size is presented. Measurements were performed using RTM1688-2 of Sarad Company.**

***Index Terms*—Thoron; Exhalation rate measurements; building material; grain size.**

I. Introduction

Radon is a noble gas with all its isotopes being radioactive. It is colourless, tasteless and odourless and therefore not detectable by human senses. Radon itself is not hazardous for health since majority of inhaled radon is exhaled before decaying in lungs. The main health hazard are its progeny which tend to attach to aerosols and when inhaled they tend depose in the lungs and irradiate surrounding tissue as they decay.

In 1988, International Agency for Research on Cancer has identified radon as a human carcinogen [1]. Based on the more recent epidemiological studies performed in Europe, Asia and America, World Health Organisation (WHO) has identified radon as the second cause of lung cancer after smoking. Based on pooled studies, it is estimated that between 3 and 14% of all lung cancers are due to exposure to radon and its progeny [2].

Influence of ^{220}Rn (thoron) to population exposure has often been neglected. It is estimated that thoron and its progenies contribute only 10% of the population exposure to radon [3]. Firstly, thoron half-life ($T_{1/2}$ = 55.6 s) is significantly shorter than half-life of radon ($T_{1/2}$ = 3.82 days) having therefore substantially shorter diffusion length. While the main source of indoor radon is soil beneath dwelling and the second one is building material, the main source of thoron is building material. Soil can be significant source of thoron in dwellings that does not have foundation or it is poor condition. Furthermore, due to short diffusion length of thoron, its concentration rapidly decreases several centimeters from its source, i.e. building material while radon concentration in the room is considered to be uniform. In spite of the abovementioned, the regions with higher effective doses due to thoron compared to radon, could be found [4-7].

In the era of increasing of energy-efficiency of homes, in old houses, doors and windows are being replaced with the new ones that are more tight, while new buildings are being built with new materials with better thermal isolation. This leads to decrease of air exchange rate and consequently to increase of indoor radon concentration [8, 9].

Therefore, there is an interest to perform exhalation rate measurements of thoron from building materials. Some of the typical methods are based on the use of continuous active devices such as semiconductors [10], or passive ones based on solid state nuclear track detectors [11].

It is known that exhalation rate from material does not depend only on the concentration of its predecessor, (^{226}Ra in the case of radon and ^{224}Ra in the case of thoron), but on the material properties such as: porosity, particle size, moisture content etc.

In this paper we are presenting the results of the thoron exhalation rate as a function of the grain size of the selected material. Measurements were performed using active device RTM1688-2 of Sarad Company.

II. Material and method

Measurement of thoron exhalation rate was performed on three different materials: soil sample, natural stone and brick block. Each sample was prepared in the same way: mechanically crashed and then sorted using sieves with different sizes. Thus, several particle size has been obtained: less than 0.5 mm, from 0.5-0.7 mm, from 0.7 to 1.2 mm, from

Dunja Antonijević is with the Faculty of Technology and Metallurgy, University of Belgrade, Karnegijeva 4, 11000 Belgrade, Serbia (email: dunjaantonijevic88@gmail.com).

Luka Rubinjoni is with the Faculty of Technology and Metallurgy, University of Belgrade, Karnegijeva 4, 11000 Belgrade, Serbia (e-mail: rubinjoni@gmail.com).

Andrija Janković is with the Faculty of Tehnology and Metallurgy, University of Belgrade, Karnegijeva 4, 11000 Belgrade, Serbia (e-mail: andrija366@gmail.com).

Igor Čeliković is with the "Vinča" Institute of Nuclear Sciences, University of Belgrade, Mike Petrovića Alasa 12-14, 11001 Belgrade, Serbia (e-mail: icelikovic@vin.bg.ac.rs).

Aleksandar Kandić is with the "Vinča" Institute of Nuclear Sciences, University of Belgrade, Mike Petrovića Alasa 12-14, 11001 Belgrade, Serbia (e-mail: akandic@vin.bg.ac.rs).

Boris Lončar is with the Faculty of Technology and Metallurgy, University of Belgrade, Karnegijeva 4, 11000 Belgrade, Serbia (email: bloncar@tmf.bg.ac.rs)

1.2-1.6 mm and from 1.6-2.1 mm.

Thoron mass exhalation rate measurements has been performed using the close-chamber method [12, 13]. Measuring sample was put in the 1.5 dm^3 volume chamber, which is in the close loop, using tubes, connected with the active device. Consequently thoron concentration in the chamber was measured. In Fig. 1, schematic representation of the measuring system is shown.

Thoron measurements were performed using active RTM1688-2 device. Thoron concentration was determined from the detection of the alpha particle emitted by ^{216}Po. Since half-life of the ^{216}Po is of the order of 0.1 s, equilibrium between thoron and ^{216}Po is established immediately. Duration of measurement was 20 minutes with sampling time of 5 minutes.

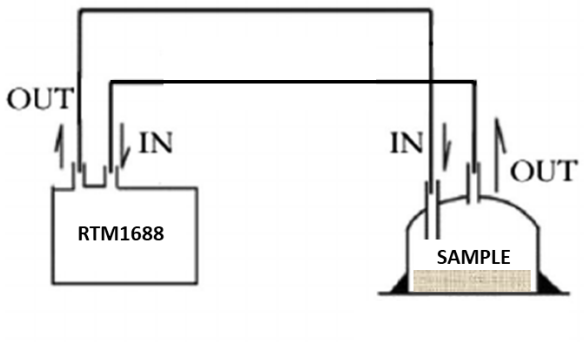


Fig. 1. Schematic representation of thoron exhalation rate measuring system.

Build-up of thoron in accumulation chamber can be expressed by following formula [11]:

$$C(t) = \frac{E_m m}{V\lambda}(1 - e^{-\lambda t}) + C_0 e^{-\lambda t} \quad (1)$$

Where $C(t)$ is the thoron concentration (Bq m^{-3}) in time t, E_m is the thoron mass exhalation rate (Bq kg^{-1} s^{-1}), m is the mass of the sample (kg), λ is the thoron decay constant, V (m^3) is the total volume of the measurement system including: volume of the chamber, active volume of the measuring device, tubes and C_0 is the initial thoron concentration (Bq m^{-3}) in the chamber.

Since measurement time is much longer than half-life of thoron, exponential term in equation (1) can be neglected and thoron concentration in the chamber can be expressed as:

$$C = \frac{E_m m}{V\lambda} \quad (2)$$

Showing that steady state thoron concentration in the chamber is achieved.

III. Main Results

Measurement of each sample lasted 20 minutes, with a sampling time of 5 minutes. Thoron concentration is therefore extracted from the dose received during 20 minutes measurement period. In Figure 2, one series of 4 thoron measurements for each sample type is presented.

Although, results within each thoron series agree within the uncertainty, there is large variation of measured thoron concentration, which leads to the large uncertainty of the deduced averaged thoron concentration.

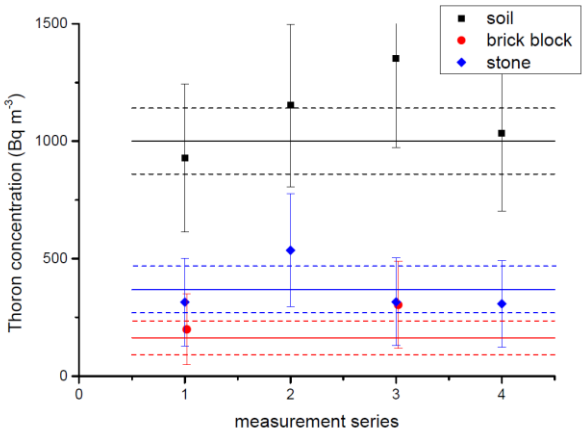


Fig. 2. Thoron measurement of each sample type. Deduced thoron concentration and its uncertainty are indicated with full and dashed lines, respectively.

In Table 1, an averaged thoron concentration for each sample type and particle size is presented together with the mass of each sample.

TABLE I
RESULTS OF MEASURED THORON CONCENTRATIONS FOR THREE DIFFERENT SAMPLES: SOIL, BRICK BLOCK AND NATURAL STONE AND FIVE DIFFERENT PARTICLE SIZES: EACH SAMPLE TYPE AND PARTICLE SIZES: 1) <0.5 MM, 2) 0.5-0.7 MM, 3) 0.7-1.2 MM, 4) 1.2-1.6 MM, AND 5) 1.6-2.1

Sample	Particle size [mm]	Mass [g]	Thoron concentration [Bq m^{-3}]
Soil	<0.7	244.77	1000±140
	0.7-1.2	198.86	990±170
	1.2-1.6	176.36	400±110
	1.6-2.1	154.02	500±120
Brick Block	<0.5	159.62	68±48
	0.5-0.7	157.05	93±55
	0.7-1.2	160.70	163±44
	1.2-1.6	151.09	39±38
	1.6-2.1	157.06	55±36
Natural Stone	<0.5	160.46	410±100
	0.5-0.7	163.72	370±100
	0.7-1.2	160.97	580±130
	1.2-1.6	162.26	770±150
	1.6-2.1	162.10	420±110

Thoron mass exhalation rate was estimated from the measured thoron concentration given in Table 1, using the

equation (1). Results are given in Figure 3, showing the dependence of mass exhalation rate as a function of particle size, for three different samples: soil, brick block and natural stone.

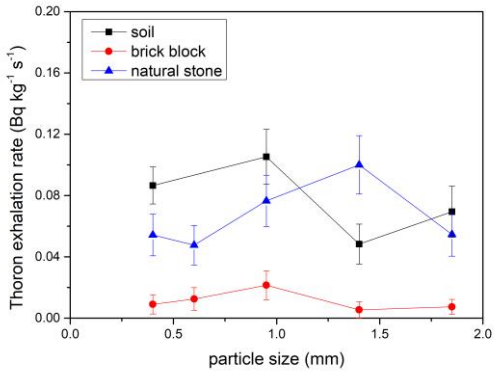


Fig. 3. Thoron mass exhalation rate measurements as a function of 5 different particle sizes: : 1) <0.5 mm, 2) 0.5 – 0.7 mm, 3) 0.7 – 1.2 mm, 4) 1.2-1.6 mm, and 5) 1.6-2.1 mm for three different samples: soil, brick block and natural stone.

Systematics of thoron exhalation rate as a function of particle size is given with a large measurement uncertainty, as shown in Figure 3. Although data indicate slight variation of thoron mass exhalation with particle size, these variation are within measurement uncertainty in the case of brick block and natural stone. Similar results were reported, elsewhere [14]. In the case of soil sample, there is statistical difference between exhalation rate of the first two samples and the last two samples. First two soil samples with the smaller particle diameter, has a higher exhalation rate, compared to the larger particles. This behaviour is envisaged, since for grains of smaller size there is higher probability that decay of ^{224}Ra will lead to thoron emanation from that grain.

.

IV. Conclusion

In this contribution, dependence of thoron mass exhalation rate on the grain size was investigated. For that purpose, close chamber method was used. Thoron concentration in the chamber was measured using active RTM1688-2 device. For two samples: brick block and natural stone it was not observed an increase of thoron exhalation rate with a decrease of grain/particle size as expected, possibly due to large measurement uncertainty and the size of particles. An increase of exhalation rate for two soil samples with the smallest grain size, compared to samples with larger grains was observed. Since all samples were measured at ambient humidity, some effect due to humidity could be observed. Therefore, it would be interesting to further investigate whether exhalation rate would be observed or more pronounced for smaller grain sizes and at different humidities at the same time.

Acknowledgment

The authors acknowledge the support of the Ministry of Education, Science and Technological Development of the Republic of Serbia [P171018, P171007].

References

[1] IARC, Man-made Mineral Fibres and Radon. Geneva, Switzerland, WHO, 1988.
[2] WHO. *Handbook on Indoor Radon*, Geneva, Switzerland WHO, 2009.
[3] UNSCEAR. *Sources-to-effects assessment for radon in homes and workplaces.* Vol. 1, Vienna, UNSCEAR, 2006.
[4] A. Kumar, S. Sharma, R. Mehra, S. Narang, R. Mishra, "Assessment of indoor radon, thoron concentrations, and their relationship with seasonal variation and geology of Udhampur district, Jammu & Kashmir, India", *Int J Occup Environ Health*. vol. 23, no. 3, pp 202-214, Jul, 2017.
[5] R. C. Ramola, G. S. Gusain, B. S. Rautela, D. V. Sagar, G. Prasad, S. K. Shahoo, T. Ishikawa, Y. Omori, M. Janik, A. Sorimachi, S. Tokonami, Levels of thoron and progeny in high background radiation area of southeastern coast of Odisha, India. *Radiat. Prot. Dosim.* vol. 152, no. 1-3, pp. 62–65, Nov, 2012.
[6] Z. S. Žunić, I. Čeliković, S. Tokonami, T. Ishikawa, P. Ujić, A. Onischenko, M. Zhukovsky, G. Milić, B. Jakupi, O. Čuknić, N. Veselinović, K. Fujimoto, S. K. Sahoo, I. Yarmoshenko, "Collaborative investigations on thoron and radon in some rural communities of Balkans", *Radiat. Prot. Dosim*., vol 141, no. 4, pp. 346-350, Oct, 2010.
[7] M. Doi, K. Fujimoto, S. Kobayashi, H. Yonehara, "Spatial distribution of thoron and radon concentrations in the indoor air of a traditional Japanese wooden house." *Health Phys*. vol. 66, no. 1, pp: 43-9, Jan, 1994.
[8] I. Yarmoshenko, A. Vasilyev, A. Onishchenko, S. Kiselev, M. Zhukovsky, Indoor radon problem in energy efficient multi-storey buildings. *Radiat. Prot. Dosim*. vol. 160, no. 1-3, pp. 53-56, 2014.
[9] J. Milner, C. Shrubsole, P. Das, B. Jones, I. Ridley, Z. Chalabi, I. Hamilton, B. Armstrong, M. Davies, P. Wilkinson, Home energy efficiency and radon related risk of lung cancer: modelling study. *Brit. Med. J.* vol. 348, f7493, Jan, 2014.
[10] P. Tuccimei, M. Moroni, D. Norcia, Simultaneous determination of 222Rn and 220Rn exhalation rates from building materials used in Central Italy with accumulation chambers and a continuous solid state alpha detector: influence of particle size, humidity and precursors concentration. Appl. Radiat. Isotopes vol. 64, no. 2, pp. 254–263, Feb, 2006.
[11] P. Ujić, I. Čeliković, A., Kandić, Z. Žunić, Standardization and difficulties of the thoron exhalation rate measurements using an accumulation chamber. *Radiat. Meas.* vol. 43, no. 8, pp. 1396–1401, Sept. 2008.
[12] F.A. Abu-Jarad Application of nuclear track detectors for radon related measurements, Nucl. Tracks Radiat. Meas. vol. 15, no. 1-4, pp. 525-534, 1988.
[13] N.P. Petropoulos, M.J. Anagnostakis, S.E. Simopoulos, Building materials radon exhalation rate: ERRICCA intercomparison exercise results. Sci. Total Environ. vol. 272, pp. 109–118 , 2001.
[14] D. Avramović, I. Čeliković, P. Ujić, I. Vukanac, A. Kandić, A. Jevremović, D. Antonijević, B. Lončar, Radon exhalation rate of some building materials common in Serbia, RAD2018 Sixth International Conference Proceedings, Ohrid, Macedonia, vol. 3, pp. 119-122, 18.06-22.06., 2018.

Determination of Surface Contamination with Handheld Equipment

Marija M. Janković, Jelena D. Krneta Nikolić
Predrag M. Božović, Nataša B. Sarap, Milica M. Rajačić

***Abstract*—Surface contamination meters are used to detect the presence of radioactive substances on some surfaces. The level of contamination can be measured using portable handheld detectors. Factors that can affect the results are source to detector distance, source geometry. This paper presents the results of measuring alpha, beta and mixed alpha-beta surface contamination presented on printed spiked papers.**

***Index Terms*—surface contamination, alpha activity, beta activity, handheld equipment.**

I. Introduction

Radioactive surface contamination is unwanted radioactive material deposited in an uncontrolled manner in or on animate or inanimate objects so that their concentrations present either operational inconvenience or radiological hazard [1]. There are several ways in which radioactive contamination can be classified: (1) by the type of radiation emitted (alpha, beta and low energy beta emitters), (2) by the availability for or ease of transfer to other objects (fixed or loose), (3) or by the physical form of the contaminant.

Monitoring techniques for determination of surface contamination can be direct or indirect. Direct method implies direct measurement on the surface and indirect means using smears and measuring them on gamma spectrometric detector or gas proportional detector.

Due to the short range of alpha particles in air, direct monitoring of alpha activity presents several problems. To ensure the detection of alpha contamination by direct method, distance from surface to detector should be less than 5 mm [1]. The same situation is with low energy beta radiation. Beta particles have longer range in the air than alpha particles and the detection is easier, but the response on detector depends on the energy of beta radiation.

Indirect methods for analysis of surface contamination can be used when the geometry of the surface is not appropriate for the measurement on other types of spectrometers. For the contamination to be measured on the spectrometer, the radioactive material has to be removed from the surface using smears and monitored on a detector. This is not always possible or practical, especially if the measurement is made in the field. That is when using handheld measuring device is applied.

Handheld equipment has shield, which when closed prevents alpha and beta radiation from reaching the detector, enabling a gamma-only measurement to be made.

This paper presents the results of surface radioactivity contamination measurement done within the inter laboratory proficiency test under the Project RER/7/008 - Strengthening Capabilities for Radionuclide Measurement in the Environment and Enhancing Quality Assurance/Quality Control System for Environmental Monitoring.

II. The Method

Response of detectors on contamination depends on the type and energy of the radiation, the counting efficiency, the detection geometry, instrument electrical noise etc.

Handheld detector used for this investigation was Automess 6150AD-k (Fig.1) [2]. The probe 6150AD-k uses a sealed proportional counter which does not require refilling or flushing from external gas reservoirs [3]. It is sensitive to alpha, beta, and gamma radiation, however due to large sensitive area it makes surveying areas much easier. Additionally, it provides an electronic switch to the operating mode "alpha" where only alpha radiation is recognized and detected very sensitively because the background is much lower in this mode. A removable discriminator plate (1 mm thick stainless steel) allows distinguishing between beta and gamma radiation.

Adequate calibration of the field instruments is necessary for accurate measurements of total surface activity [4,5]. Alpha radiation emitters with energies similar to those expected of the contaminant in the field should be used as calibration sources. Sources used for calibration were ^{241}Am and ^{90}Sr (Table I).

Instruments have decay mode: alpha, beta and gamma. Detectors generally have efficiencies from 0 to 30 %.

Marija M. Janković, University of Belgrade, Radiation and Environmental Protection Department, Vinča Institute of Nuclear Sciences, Mike Petrovića Alasa 12-14, 11001 Belgrade, Serbia (e-mail: marijam@vinca.rs).

Jelena D. Krneta Nikolić, University of Belgrade, Radiation and Environmental Protection Department, Vinča Institute of Nuclear Sciences, Mike Petrovića Alasa 12-14, 11001 Belgrade, Serbia (e-mail: jnikolic@vinca.rs).

Predrag M. Božović, University of Belgrade, Radiation and Environmental Protection Department, Vinča Institute of Nuclear Sciences, Mike Petrovića Alasa 12-14, 11001 Belgrade, Serbia (e-mail:bozovic@vinca.rs).

Nataša B. Sarap, University of Belgrade, Radiation and Environmental Protection Department, Vinča Institute of Nuclear Sciences, Mike Petrovića Alasa 12-14, 11001 Belgrade, Serbia (e-mail: natasas@vinca.rs).

Milica M. Rajačić, University of Belgrade, Radiation and Environmental Protection Department, Vinča Institute of Nuclear Sciences, Mike Petrovića Alasa 12-14, 11001 Belgrade, Serbia (e-mail: milica100@vinca.rs).

TABLE I
AUTOMESS 6150 AD-K TEHCNICAL SPECIFICATIONS [P1]

Measuring instrument	6150AD-k surface contamination probe
Producer	Automess
Detector type	Proportional counter
Sensitive area of detector	17×10 = 170 cm^2
Calibration factor for Alpha ^{241}Am	0.074 (Bq/cm^2)/s^{-1}
Beta ^{90}Sr	0.011 (Bq/cm^2)/s^{-1}

Factor that may affect the detector efficiency is distance between source and detector because changes in the distance produce changes in the detector response especially for the alpha radiation. To minimize the effect of surface to detector distance it is recommended that the distance should be as small as possible. For alpha contamination it is recommended that the distance should not exceed 5 mm [1]

Source geometry can also affect the results. Calibration source should be similar to point source geometry. In our case printed sources have the same geometry as handheld detector.

Background count rate must be measured on the place where there is no potential high activity from local area (for example granite, or ceramic tiles…).

Surface activity can be calculated using the Equation (1) [6]:

$$A_s = \frac{N - N_0}{E_i E_s W} = \frac{N - N_0}{Es} \cdot C_s \qquad (1)$$

where:

As is surface activity in Bq cm^{-2}

N is count rate of the measurement in cps

N_0 is background count rate in cps

E_i is instrument efficiency in count per emission

E_s source efficiency in emission per Bq

W is the detector area in cm^2

C_S is the calibration factor equal to $1/E_iW$ [Bq cm^{-2}/s^{-1}]

In this measurement it is assumed that the source efficiency E_S is equal to 1.

Fig. 1. Handheld detector Automess 6150AD-k

III. THE RESULTS

Samples obtained by IAEA were spiked with ^{90}Sr (pure beta emitter) (sample 1), ^{244}Cm (almost pure alpha emitter) (sample 2) and mixture ^{90}Sr + ^{244}Cm (sample 3). Sample area for printed paper was 14 cm × 20 cm (Fig.2). Measurements were performed in the Radiation and Environmental Protection Department, Vinča Institute of Nuclear Sciences, within the Project: RER/7/008 - Strengthening Capabilities for Radionuclide Measurement in the Environment and Enhancing Quality Assurance/Quality Control System for Environmental Monitoring. On Figure 2 we can see the surface contamination samples delivered to the Laboratory. All samples were covered with plastic foil in order to prevent the damage or cross contamination of the samples during the transport. Also, samples were protected with foam frame that did not cover the spiked area, so the measurement could be conducted without removing it.

Fig. 2. Samples for determination of surface contamination

Firstly, for the sample 1 handheld detector was placed above the surface at 1 cm distance. Net count rate for sample was taken as average value of 10-15 measurements. For the sample 2 handheld detector was placed as close to the surface as possible without contact with surface. Net count rate for this sample was also taken as average value of 10-15 measurements. In case of sample 3, which contained mixture of alpha and beta emitters, firstly net count rate for alpha was measured without shield on detector in alpha mode and then for beta emitters measurement, the shield was placed back on the detectors, enabling us to measure in beta mode.

Background was measured 5 times in one point.

Since all the measurements were conducted in the same room, held on the room temperature of approximately 20°C and without changing the environmental parameters, the influence of the temperature on the measurement quality is minimized.

Positioning of the instrument above the sample could present a problem, so any potential support for the instrument should be utilized. Also, the distance for the measurement in alpha mode should be as small as possible without touching the surface itself in order to avoid the attenuation of the alpha particles in air. In this situation, that was accomplished by placing the detector on the protective frame of the sample, enabling us to achieve the distance of 1mm.

Since not all of the area of the sample was covered by the

surface of the instrument, the dimensions of the sample area that was covered by the detector had to be taken into account. This was achieved by multiplying the calibration factor with the ratio of the detector area and the area of the sample.

Measurement uncertainty was calculated as the combined measurement uncertainty with the coverage factor 1.

$$u(A) = \sqrt{u^2(N_0) + u^2(N) + u^2(C_s) + u^2(\mathrm{d})} \qquad (2)$$

where $u(N_0)$ is the relative uncertainty of the background count, $u(N)$ is the relative uncertainty of the measured sample count, $u(C_s)$ is the relative uncertainty of the calibration factor and $u(d)$ is the relative uncertainty of the detector-sample distance.

The relative uncertainty of the background count and measured sample count was determined as the standard deviation of the repeated measurements. For beta emission background, the standard deviation was around 5% and for the alpha emission it was around 2%. The relative uncertainty of the sample count ranged from 2.2% for alpha emission count in sample 2 to 7.9% for beta emission count in sample 3.

Since the calibration factor was obtained from the instrument manual, the uncertainty of this parameter was considered as negligible. The relative uncertainty of the detector-sample distance was estimated to be 1%.

The final results were obtained by taking the mean value of the repeated count measurements and applying the Equation (1).

Results with the appropriate measurement uncertainties are given in Table II.

TABLE II
RESULT FOR SURFACE ACTIVITY

Sample	Emitter	Activity (Bq cm^{-2})
1.	beta	1,14 ± 0,08
2.	alpha	0,64 ± 0,04
3.	alpha	0,27 ± 0,02
	beta	0,46 ± 0,05

Target values were: 1,21 Bq cm^{-2} for sample 1, 0,71 Bq cm^{-2} for sample 2 and for sample 3 for beta, target value was 0,6 Bq cm^{-2} and for alpha 0,37 Bq cm^{-2}. Results were evaluated as acceptable according to the z-score.

IV. CONCLUSION

This paper presents the results of surface radioactivity contamination measurement done within the inter laboratory proficiency test. This measurement enabled us to see firsthand the advantages and problems that are encountered when conducting surface contamination measurements with handheld instruments. Special attention was committed to the measurement conditions leading to the situation that the values obtained in this proficiency test had good agreement with the target values.

ACKNOWLEDGMENT

The authors would like to thank to the Ministry of Education, Science and Technological Development of the Republic of Serbia (Grant no. III 43009).

REFERENCES

[1] INTERNATIONAL ATOMIC ENERGY AGENCY, "Monitoring of Radioactive Contamination on Surfaces," Technical Reports Series 120, Vienna, 1970.

[2] https://www.automess.de/Download/Prospekt_AD17k_E.pdf (Last accessed on 29.04.2019.)

[3] G. F. Knoll, "Radiation detection and measurement," 3rd edition, John Wiley & Sons, New York, 2010.

[4] International Organization for Standardization ISO-7503-1, Evaluation of surface contamination-part 1: Beta emitters (maximum beta energy greater than 0,15 MeV) and alpha emitters, 1988.

[5] E. W. Abelquist, W. S. Brown, G. E. Powers, A. M. Huffert, "Minimum Detectable Concentrations with Typical Radiation Survey Instruments for Various Contaminants and Field Conditions," NUREG-1507, Nuclear Regulatory Commision, Office of Nuclear Regulatory Research, Division of Regulatory Applications, USA, 1997.

[6] Y. Kodama, F. F. Suzuki, M. P. Sanches, D. L. Rodrigues, "Discussion on surface contamination monitoring using portable zinc sulfide scintillation detectors," Nuclear future: thinking for building Proceedings of the 5 Brazilian national meeting on nuclear applications; 8 General congress on nuclear energy; 12 Brazilian national meeting on reactor physics and thermal hydraulics, (p. 1450). Brazil

The Effects of X-Radiation in a Quasi-Low-Dropout Voltage Regulator

Vladimir Dj. Vukić

***Abstract*—The aim of this paper was to test the possibility of implementation of the commercial-off-the-shelf (COTS) quasi-low-dropout voltage regulator in a harsh bremsstrahlung environment (in the case of a lifelong exposure to moderate total ionising doses). Results of examination of the LT1086CT5 voltage regulator, performed in the field of X-rays, are presented in this paper. Biased and loaded circuits demonstrated acceptable characteristics in the bremsstrahlung environment for total doses up to 433 Gy, but some samples of unbiased circuits demonstrated unacceptable decline of the output voltage even after absorption of ionising dose of 260 Gy. One-week room temperature annealing led to a further degradation of all of the irradiated voltage regulators, both unbiased and biased, during irradiation. All the examined samples remained functional, but pointed to numerous limitations for implementation of these COTS voltage regulators in an ionising radiation environment.**

***Index Terms*—Voltage regulator; X-rays; bipolar transistor; radiation effects.**

I. Introduction

IMPLEMENTATION of the commercial-off-the-shelf (COTS) integrated circuits, instead of the specially designed "rad-hard" components, is now a long-lasting trend [1]. The main reason is related to the much lower price of the COTS components, as well as the considerably wider portfolio of the available integrated circuit made by the use of modern technological processes [1]. Among the serious candidates for implementation of COTS solutions are power integrated circuits. Low-dropout voltage regulators belong to this class of electronic devices [2]-[5], being particularly important to the point-of-load, low-current power supply of small load (being in the order of several hundred milliamps), as well as the battery powered electronic circuits [6].

In previous years, many efforts were made to examine the possibility of implementation of the COTS low-dropout voltage regulators with various power transistors and control circuit topologies [7]-[9]. Nevertheless, among three different topologies of low-dropout voltage regulators, also one quasi-low-dropout circuit, LT1086CT5, was examined, utilising the NPN power transistor as a pass element [10]. This circuit was a potential candidate for implementation in a moderate radiation environment.

In this paper, examination of characteristics of the LT1086CT5 voltage regulator was presented. Radiation effects caused by total absorbed doses of the braking ionising radiation (bremsstrahlung), up to 433 Gy, were described, followed by a short term 168-hour room temperature annealing.

Vladimir Dj. Vukić is with University of Belgrade, Electrical Engineering Institute "Nikola Tesla", Koste Glavinića 8a, PO Box 139, 11000 Belgrade, Serbia (e-mail: vvukic@ieent.org).

II. Theory

A. Configuration of Low-Dropout Voltage Regulators

One of the basic characteristics for evaluation of voltage regulators is a dropout voltage that is a minimum voltage across a serial (pass) transistor that enables this integrated circuit to maintain a stable voltage on its output terminal [11]. The main difference between configuration of a low-dropout (a) and a quasi-low-dropout (b) voltage regulator is presented in Fig. 1. In a low-dropout voltage regulator, a PNP power transistor may be independent pass element (Qp), so a dropout voltage is equal only to the collector – emitter voltage of a serial transistor [11]. In a quasi-low-dropout voltage regulator, an NPN transistor (Qp1) should be driven by a collector current of accompanying PNP transistor [11] (Qp2). Thus, dropout voltage is now the sum of a PNP driver transistor emitter-base voltage and a serial NPN transistor collector – emitter voltage [11].

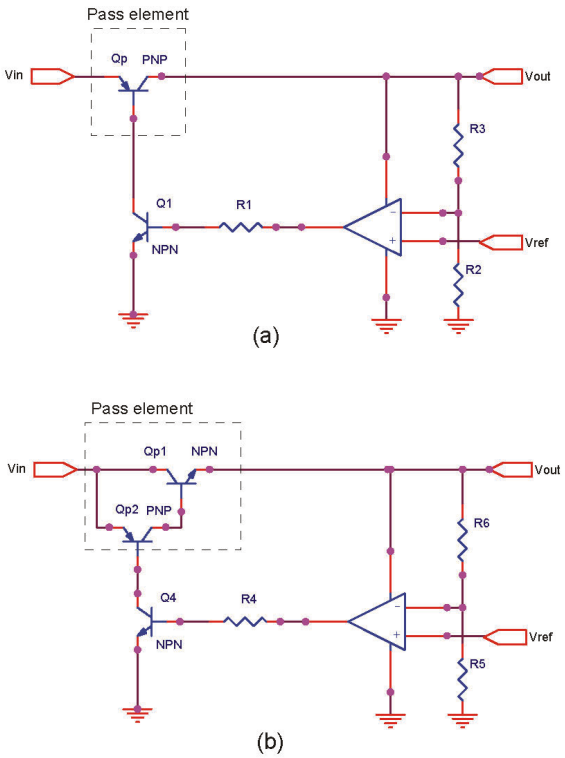


Fig. 1. Topologies of voltage regulators: a) low-dropout; b) quasi-low-dropout.

A negative consequence of implementing a combination of an ordinary PNP and an NPN power transistor as a pass element is primarily related to the increase of a dropout voltage (approximately 1.2-1.5 V for a nominal load, in comparison with approximately 0.6-0.8 V in a case when a PNP power transistor was utilised [11]). On the other hand, the advantage of using an NPN power transistor is related to its operation with much lower base current and, consequently, much lower quiescent current of the entire voltage regulator.

B. Radiation and Post-Irradiation Effects in Bipolar Transistors and Integrated Circuits

The primary effect that ionising radiation has on a bipolar junction transistor is related to the increase of its base current, consequently leading to reduction of the forward emitter current gain [1]. This excess base current is primarily a result of the charge becoming trapped in the isolation oxide and at the semiconductor-oxide interface above the base area [1]. In general form, the relation for the excess base current (ΔI_B) in an irradiated bipolar transistor may be presented as [12]:

$$\Delta I_B = I_B - I_{B0} = \frac{1}{2} q n_i P_E s(N_{ot}, N_{it}, V_{BE}) \gamma(N_{ot}, V_{BE}) e^{\frac{V_{BE}}{2V_T}} \quad (1)$$

where: I_B – base current after irradiation, I_{B0} – base current before irradiation, q – elementary electron charge ($1.6 \cdot 10^{-19}$ C), n_i – intrinsic carrier concentration in silicon, P_E – length of the emitter perimeter, N_{ot} – concentration of the oxide traps, N_{it} – concentration of the interface traps, V_{BE} – voltage on the junction base – emitter, $s(N_{ot}, N_{it}, V_{BE})$ - surface recombination velocity, a function of the specifed electrical values, $\gamma(N_{ot}, V_{BE})$ – an integral describing the depletion region spreading over the intrinsic base [13], and V_T – thermal voltage (26 mV at 20°C).

The other mechanism that primarily affects a PNP bipolar transistor forward emitter current gain is related to similar effects above the emitter area, causing the spread of the base-emitter depletion region deep into the P-type emitter area [14]. Thus, oxide-trapped charge and interface traps above the base and emitter areas are the main cause of the characteristic degradation of a bipolar transistor operating in an ionising radiation environment [14]. Above the P-type area, the effects of the interface and oxide traps are complementary, both leading to degradation of a semiconductor device [14]. Usually, the excess base current of the PNP transistor is considered to be primarily affected by the influence of interface traps [14], while NPN transistors are considered to be mostly affected by the oxide-trapped charge [12]. The following empirical relations are a good illustration of the mentioned phenomena [15]:

$$\Delta I_B \sim N_{it} e^{N_{ot}^2}, \text{ for an NPN transistor} \quad (2)$$

$$\Delta I_B \sim N_{it} \frac{1}{N_{ot}^2}, \text{ for an PNP transistor} \quad (3)$$

Bias conditions also have great impact on the radiation response of a bipolar transistor or integrated circuit based on such an element. While the positive electrical field in the oxide may suppress build-up of the oxide-trapped charge, the negative bias voltage may suppress or even prevent build-up of interface traps [16]. Usually, the most damaging effect may be observed in bipolar transistors and integrated circuits exposed to the influence of ionising radiation without any bias voltage [1]. In these cases, very high excess base currents may be recorded, and both interface traps and the oxide-trapped charge may have great influence on the degradation of a bipolar transistor current gain. On the other hand, operation of the bipolar transistor with high dissipation may significantly reduce its radiation damage and even prevent the integrated circuit failure [3, 9], both due to the high increase of a chip temperature (followed by an annealing of the trapped charge) or operation with high-current-density at the base-emitter area [3].

After the end of ionising radiation exposure, bipolar transistors may demonstrate recovery, or further degradation of their electrical characteristics [1]. Owing to a type of its post-irradiation response, often a trustworthy conclusion may be drawn regarding the mechanism of the bipolar transistor radiation response [1]. The post-irradiation response of an irradiated bipolar transistor is highly temperature dependant. Thus, high temperature annealing may lead to recovery of both oxide-trapped charge (this process becomes more expressed as temperature increases up to 100°C [16, 17]) and interface traps (dominates when temperatures are in the range 150-250°C [17]). Nevertheless, even at a room temperature of nearly 20°C, trapped charge may be recovered, primarily due to the tunnelling mechanism [16]. Thus, room temperature recovery of bipolar transistors is also described by the word "annealing". Recovery of the oxide-trapped charge may be accompanied by further interface trap build-up [18]. In such a case, forward emitter current gain of a bipolar transistor may be additionally reduced. On the other hand, if the oxide-trapped charge had significant influence on the excess base current then the annealing of this charge may lead to an expressed recovery of the forward emitter current gain of an irradiated bipolar transistor.

III. MATERIALS AND METHODS

A. Radiation Source and Dosimetry

The samples of LT1086CT5 voltage regulator, an integrated circuit made by *Linear Technology*®, were tested at the Vinča Institute of Nuclear Sciences, in the Metrology-Dosimetric laboratory. As a source of X-rays, the *Philips*® MG320 dosimetric generator was used. The continuous spectrum of braking radiation was obtained using a voltage of 300 kV and a current of 10 mA, with implementation of a tungsten target. Thus, photon spectrum, with energies up to 300 keV, was obtained [9]. Primary filtration of a photon spectrum was obtained with 4-mm-thick embedded aluminium foil [9]. In order to obtain additional filtration, another 0.47-mm-thick aluminium foil was added [9]. As a result, all the photons with energies lower than 37 keV were removed from the spectrum [19]. Consequently, a bremsstrahlung field was obtained, having a mean energy of 170 keV [9].

Exposure measurement was performed with a cavity ionising chamber *Dosimentor*® PTW M23361 [7]. The DI4 reader was used during the measurement [7]. With these instruments, exposure in Roentgen (R) was procured, and,

recalculated further for the total absorbed dose in silicon-dioxide (Gy(SiO_2)) [9] using X-ray mass attenuation coefficients for air and silicon-dioxide [20]. Using the dosimetric chamber, the exposure rate was measured to be F = 340 R/min = 20,400 R/h. In standard temperature and pressure (STP) dry air, the following relation is valid [17]: 1 R = 0.84 rad(air) = 0.84 cGy(air).

Since the integrated circuit is comprised of the multiple, micrometer-thick layers, only an approximate calculation was made. The kinetic energy of the primary charged particles released by photons, enhanced by the contribution of charged particles travelling through the silicon-dioxide [20], had the crucial influence to the total dose response of bipolar transistors in an integrated circuit. For the nearest available value of the mean photon energy in the table of mass-energy absorption coefficients, that is 150 keV, coefficients were extracted for the air and the silicon-dioxide (borosilicate glass) [20]: $\frac{\mu_{en}}{\rho}(SiO_2) = 2.727 \cdot 10^{-2} \frac{cm^2}{g}$, $\frac{\mu_{en}}{\rho}(air) = 2.496 \cdot 10^{-2} \frac{cm^2}{g}$. According to the available data, the dose rate in silicon-dioxide (reduced to cGy) was calculated as [17, 20]:

$$\frac{dD}{dt}(SiO_2) = F \cdot 0.84 \cdot \frac{\frac{\mu_{en}}{\rho}(SiO_2)}{\frac{\mu_{en}}{\rho}(air)} \quad (4)$$

Finally, total dose rate (in Gy(SiO_2)) was:

$$\frac{dD}{dt}(SiO_2) = 187.2 \frac{Gy}{h} \quad (5)$$

Therefore, in the 170-keV effective X-rays field (V_{TUBE} = 300 kV, I_{TUBE} = 10 mA), dose rate for silicon-dioxide was 187.2 Gy(SiO_2)/h. Measurements were performed after deposition of the following total ionising doses (in Gy(SiO_2)): 0; 43.3; 86.7; 130; 173.3; 260; 346.7; 433. One additional measurement was performed after 168-hour room temperature annealing (without implementation of any bias and load) of the irradiated samples.

B. Electrical Characteristics

Samples of LT1086CT5 voltage regulators were tested in a field of X-rays with two various operating conditions: without bias and load (V_{IN} = 0 V, I_{OUT} = 0 A), as well as with input bias voltage and load current (V_{IN} = 7 V, I_{OUT} = 100 mA) during operation in the ionising radiation environment. In order to enable the voltage regulator operation in the X-rays field, exposed samples were supplied with 10 m long cables [7]. In order to enable remote measurement of the output voltage at the terminals of voltage regulators, the sensing cables of the same length were laid alongside the power supply cables [7]. The power source of DC voltage enabled simultaneous supply up to four electrically isolated integrated circuits [7]. Irradiation of components and measurement were performed at a room temperature of 20°C [7].

Electrical characteristics used for evaluation of the voltage regulator degradation in the X-rays environment were the maximum output current and the minimum dropout voltage. The maximum output current was procured in the measurement point where the voltage regulator input voltage was set to 8 V DC, while the output voltage declined down to 4.7 V (this is a value close to the point when voltage regulator commence the shutdown procedure) [8]. The minimum dropout voltage was obtained for a constant load current of 400 mA and a constant output voltage of 4.9 V [8].

IV. Results

In Fig. 2 were presented variations of the maximum output current in the field of 170-keV X-rays, obtained for unbiased as well as biased and loaded voltage regulators. As initially expected, unbiased circuits demonstrated greater degradation of the maximum output current than the biased ones. So, while biased circuits demonstrated degradation of the maximum output current up to 7 % of the initial value, exposure to X-rays reduced the same parameter in unbiased voltage regulators up to 20 %. However, for a total dose of 433 Gy, such reduction may be acceptable if the integrated circuit kept its functionality as a power source with a stable five-volt output.

Fig. 3. shows the results on variations of the minimum dropout voltage obtained on the same samples of the LT1086CT5 voltage regulator in a bremsstrahlung environment. This test gave much more questioning results regarding this voltage regulator radiation hardness.

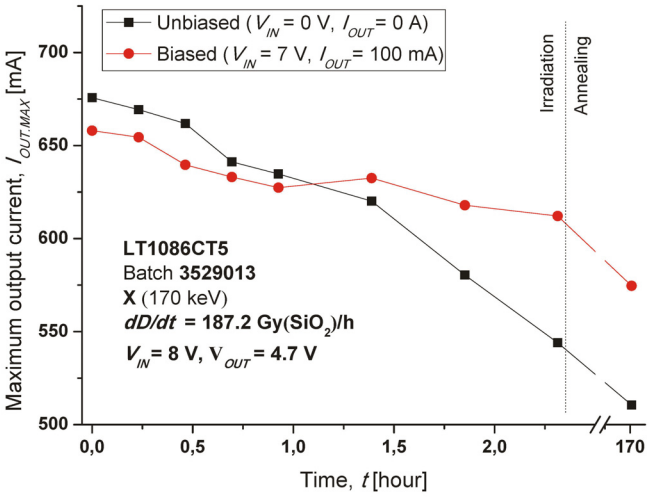


Fig. 2. Change in the maximum output current of LT1086CT5 voltage regulators in the X-rays field.

At first, a comment has to be made regarding the high initial values of the minimum dropout voltage, being approximately 3 V. This relatively high value is a consequence of the low filter capacitance (nominally 330 µF [8]) at the output of the Gretz diode bridge, being a converter of the (transformed) mains AC voltage to a DC voltage at the input terminal of a voltage regulator. Therefore, input voltage on a voltage regulator had a very high AC component affecting the voltage regulator operation, particularly with

high load currents. Nonetheless, such a measurement configuration did not negatively affect procured results on the integrated circuit radiation tolerance.

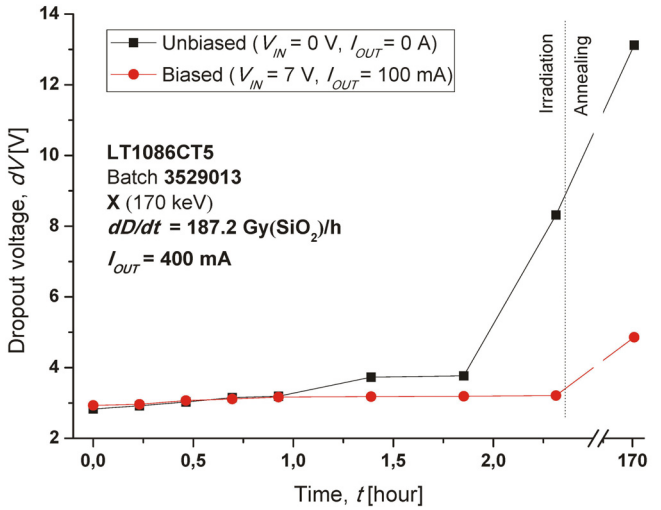


Fig. 3. Change in the minimum dropout voltage of LT1086CT5 voltage regulators in the X-rays field.

The data obtained for unbiased voltage regulators presented in Fig. 3 point to the significant increase of the minimum dropout voltage after deposition of the total dose of 260 Gy. Dropout voltage reached unacceptably high values after the final dose of 433 Gy was achieved. The more careful examination of individual irradiated samples points to the inability of one sample (out of a total of five) to obtain an output voltage of 4.9 V (following absorption of the total dose of 260 Gy). Then, after deposition of 433 Gy, another unbiased sample decreased its output voltage below the mentioned threshold of 4.9 V, bringing the number of unacceptable circuits up to two. Therefore, even to obtain such a reduced voltage, the much higher input voltage had to be applied. Therefore, it led to an increase of the mean value of the minimum dropout voltage of unbiased samples up to 8 V (Fig. 3). It is important to emphasise that none of the biased and loaded samples expressed a similar phenomenon, with all five samples being completely functional, simultaneously having a dropout voltage across a serial transistor being marginally higher from its pre-irradiation value.

V. Discussion

As may be seen from Figs. 2 and 3, in both cases one-week, room temperature isothermal annealing led to a further degradation of irradiated voltage regulators, regardless of their bias conditions during bremsstrahlung exposure. Such a post-irradiation response suggests interface traps as the primary influence on degradation of the critical bipolar transistors. As previously mentioned, physical mechanism that affected the voltage regulator response is related with a partial recovery of the oxide-trapped charge, accompanied by a simultaneous build-up of interface traps. At a first glance, this is a little surprising, since it would be expected that the oxide-trapped charge would have a dominant influence on the serial power NPN transistor. Taking into account the perceived effects, a rise of the power NPN transistor excess base current may be expected, consequently leading to a decline of its forward emitter current gain. Additionally, radiation response of a voltage regulator is not only dictated by a serial power transistor and, furthermore, some small-signal transistors may have a decisive effect on its operation characteristics. It should be brought to mind that, in a quasi-low-dropout voltage regulator, a driver PNP transistor is also a part of the pass element.

Since the photons in bremsstrahlung spectrum contain energies from 37 keV up to 300 keV, there certainly exists possibility for violation of the charged-particle equilibrium (CPE) in the examined integrated circuit. Nonetheless, it was assumed that, looking chip-wide, CPE wouldn't be violated in the active areas of bipolar transistors, situated at the interface Si – SiO_2. Potentially critical areas for the dose enhancement are connections between metallizations and the chip of an integrated circuit [17]. Most of these connections are situated at the bottom of integrated circuit, thus being far away from the active area, right beneath the insulation oxide [17]. Nonetheless, metallizations from the top side of the case, connecting the output leads of voltage regulator with semiconductor, may cause dose enhancement in silicon, if the connection material has a much higher atomic number [16] (in this case, relative to silicon (Si): Z = 14). Usually, metals used for such connections are copper (Cu; Z = 29) or gold (Au; Z = 79). Since the copper doesn't have much higher atomic number than silicon, significant dose enhancement should not be expected. Nonetheless, completely different situation may appear if gold is used. In that case, charged-particle equilibrium may be locally violated. The author did not have insight into the exact material composition of the implemented TO-220 case, used for packaging of the LT1086CT5 voltage regulator. Thus, it was assumed that, chip-wide, for the calculation of the total ionising dose in the insulation layer of SiO_2, CPE was achieved. Nonetheless, it is not impossible that locally, at the connections of some of three output leads with chip, dose enhancement exists. However, it was assumed that the mentioned effect could be distinctively expressed only in some small-signal transistors, and that the serial NPN power transistor is too large to be notably affected by the violation of the CPE.

In general, quasi-low-dropout voltage regulator demonstrated significant sensitivity to the influence of X-rays. A total dose of 260 Gy(SiO_2), being a threshold value when unacceptable degradation was observed in unbiased samples, is lower than expected before the examination. Nonetheless, this was certainly not an example of the extremely low radiation hardness of an analogue integrated circuit. As already presented in the literature [17], threshold of radiation tolerance may vary for several orders of magnitude, depending on the batch of integrated circuit, manufacturer's technological process, bias conditions during the irradiation, as well as the characteristics of the ionising radiation environment (types of spectra, mean photon energy, dose rate,

etc.). The author of this article detected failure of some samples of LM2990T-5 LDO voltage regulators even at the bremsstrahlung total dose as low as 37 Gy(SiO_2) [9], regardless the fact that samples from the same lot in γ-radiation field remained completely functional even after absorption of the total dose of 500 Gy (SiO_2) [9].

In manuals and handbooks may be found approximate data on radiation tolerance of various types of semiconductor devices and integrated circuits. For instance, in the document [21], radiation tolerance of CMOS operational amplifiers (Op Amp) was estimated to be up to 100 Gy [21], while the same parameter for bipolar Op Amps was estimated to be 1 – 2.5 kGy [21]. In the same document, low-dropout voltage regulators were estimated to be radiation tolerant for total ionising doses up to 300 – 500 Gy [21]. Nonetheless, in the same document, there were specified several *National Semiconductor*® regulators, having the stated radiation hardness threshold exceeding 1 kGy [21]. Tolerance to very high total doses of ionising radiation is usually characteristic of the specially designed, rad-hard semiconductor devices and integrated circuits.

The main goal of the research presented in this paper was to examine the possibility of implementation of cheap COTS integrated circuits instead of the very expensive, specially designed rad-hard components. The additional motive for the presented research was the existence of a practically the same component, of the same manufacturer, LT1086MH/883, yet declared for implementation in the harsh radiation environments. According to the performed tests, a conclusion can be made that unbiased LT1086CT5 circuits are not suitable for bremsstrahlung environments where deposition of the total dose greater than 200 Gy(SiO_2) may be expected. Nevertheless, none of the examined circuit demonstrated a functional failure for total ionising doses up to 433 Gy(SiO_2).

Thus, a complex research has to be made, in various radiation fields, prior to qualification of some semiconductor component for exploitation in a radiation environment. In general, implemented technological processes for the synthesis of semiconductor chips (and particular quality of the insulation oxides) dominantly affect the radiation response of an analogue integrated circuit.

VI. Conclusion

Quasi-low-dropout voltage regulators, LT1086CT5, were tested for possibility of implementation in the ionising radiation environment. Regarding these quasi-LDO regulators, for the first time was presented data on the maximum output current and the minimum dropout voltage (recorded with a high load current). Also, results on the one-week, room-temperature annealing are novel data for the irradiated *Linear Technology*® LT1086CT5 COTS voltage regulators.

In the effective 170-keV X-rays field, for a medium dose rate (187.2 Gy(SiO_2)/h), samples of this COTS circuit were tested for two various modes of operation during the exposure: unbiased, as well as with bias and load (V_{IN} = 7 V, I_{OUT} = 100 mA). For the total ionising dose of 433 Gy(SiO_2), biased samples demonstrated higher radiation tolerance than the unbiased. Nevertheless, the maximum output current recorded in unbiased samples of LT1086CT5 voltage regulators did not fall below 80 % of their pre-irradiation values. Less acceptable were the values of the minimum dropout voltage (for I_{OUT} = 400 mA), particularly due to a decline of the output voltage below the minimum acceptable value of 4.9 V, recorded on some of the tested samples. Also, cause for concern was further degradation of electrical parameters used in all the tested circuits (regardless of their bias conditions during irradiation), noticed during the following room temperature, one-week isothermal annealing sequence. Experimental results support the conclusion that unbiased LT1086CT5 voltage regulators are not suitable for bremsstrahlung environments where lifelong absorption of the total dose greater than 200 Gy(SiO_2) may be expected. Nonetheless, recorded radiation response leaves a possibility that these voltage regulators may be satisfactory exploited in the radiation environment as continuously biased circuits, operating with load currents being, for an order of magnitude, lower than the voltage regulator nominal output current.

Procured results could not lead to an unambiguous conclusion on the radiation tolerance of these COTS voltage regulators. Presented results were related only to a medium–photon-energy bremsstrahlung field, and for two types of operating conditions. Thus, it was not sufficient to make a final conclusion on the possibility to implement this analogue integrated circuit in all kinds of ionising radiation environments. Therefore, in order to get a complete insight into this circuit's ionising radiation hardness, it would be necessary to perform also its detailed examination in the field of γ-radiation, with more different bias and load conditions, as well as for total ionising doses exceeding 433 Gy(SiO_2). Finally, only separation of radiation effects in small-signal transistors, on the one hand, and the power transistor, on the other, could enable the precise identification of the failure mechanisms of a quasi-low-dropout voltage regulator LT1086CT5, affected by the influence of an ionising radiation environment.

Acknowledgment

This work was supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia under the project 171007, "Physical and functional effects of the interaction of radiation with electrical and biological systems".

References

[1] A. Holmes-Siedle and L. Adams, *Handbook of radiation effects*. Oxford, U.K.: Oxford Univ. Press, 2004.

[2] J. Beaucour, T. Carribre, A. Gach, and P. Poirot, "Total dose effects on negative voltage regulator," *IEEE Trans. Nucl. Sci.*, vol. 41, no. 6, pp. 2420–2426, Dec. 1994.

[3] R. L. Pease, S. McClure, J. Gorelick, and S. C. Witczak, "Enhanced low-dose-rate sensitivity of a low-dropout voltage regulator," *IEEE Trans. Nucl. Sci.*, vol. 45, no. 6, pp. 2571–2576, Dec. 1998.

[4] P. C. Adell, R. D. Schrimpf, W. T. Holman, J. L. Todd, S. Caveriviere, R. R. Cizmarik, and K. F. Galloway, "Total dose effects in a linear voltage regulator," *IEEE Trans. Nucl. Sci.*, vol. 51, no. 6, pp. 3816–3821, Dec. 2004.

[5] V. Ramachandran, B. Narasimham, D. M. Fleetwood, R. D. Schrimpf, W. T. Holman, A. F. Witulski, R. L. Pease, G. W. Dunham, J. E. Seiler, and D. G. Platteter, "Modeling total-dose effects for a low-dropout voltage regulator," *IEEE Trans. Nucl. Sci.*, vol. 53, no. 6, pp. 3223–3231, Dec. 2006.

[6] G. A. Rincon-Mora, *Analog IC design with low-dropout regulators*. McGrawHill Educat., 2014.

[7] V. Dj. Vukić, "Minimum dropout voltage on a serial PNP transistor of a moderately loaded voltage regulator in a gamma radiation field," *Nucl. Technol. Radiat.*, vol. 27, no. 4, pp. 333-340, Dec. 2012.

[8] V. Dj. Vukić, "Computer simulation model for evaluation of radiation and post-irradiation effects in voltage regulator with vertical PNP power transistor," *Inform. MIDEM*, vol. 48, no. 4, pp. 181-193, Sep. 2018.

[9] V. Dj. Vukić, P. V. Osmokrović, "Failure of the negative voltage regulator in medium-photon-energy X radiation field," *Microelectron. Reliab.*, vol. 54, no. 1, pp. 79-89, Jan. 2014.

[10] V. Dj. Vukić, "Rapid and long-term gamma-radiation annealing in low-dropout voltage regulators," *Nucl. Technol. Radiat.*, vol. 32, no. 2, pp. 155-165, Jun. 2017.

[11] C. Simpson, "Linear and switching voltage regulator fundamentals," *National Semicond. Corp.,* USA, 2003.

[12] A. Wei, S. L. Kosier, R. D. Schrimpf, D. M. Fleetwood, W. E. Combs, "Dose-rate effects on radiation-induced bipolar junction gain degradation," *Appl. Phys. Lett.,* vol. 65, no. 15, pp. 1918-1920, Oct. 1994.

[13] S. L. Kosier, R. D. Schrimpf, R. N. Nowlin, D. M. Fleetwood, M. DeLaus, R. L. Pease, W. E. Combs, A. Wei, E Chai, "Charge separation for bipolar transistors," *IEEE Trans. Nucl. Sci.*, vol. 40, pp. 1276-1285, Dec. 1993.

[14] D. M. Schmidt, D. M. Fleetwood, R. D. Schrimpf, R. L. Pease, R. J. Graves, G. H. Johnson, K. E. Galloway, W. E. Combs, "Comparison of ionizing-radiation-induced gain degradation in lateral, substrate and vertical PNP BJTs," *IEEE. Trans. Nucl. Sci.*, vol. 42, pp. 1541-1549, Dec. 1995.

[15] V. S. Pershenkov, V. B. Maslov, S. V. Cherepko, I. N. Shevtzov-Shilovsky, V. V. Belyakov, A. V. Sogoyan, V. I. Rusanovsky, V. N. Ulimov, V. V. Emelianov, V. S. Nasibullin, "The effect of emitter junction bias on the low dose-rate radiation response of bipolar devices*, IEEE Trans. Nucl. Sci.*, vol. 44, no. 6, pp. 1840–1848, Dec. 1997.

[16] T. R. Oldham, F. B. McLean, "Total ionizing dose effects in MOS oxides and devices," *IEEE. Trans. Nucl. Sci.*, vol. 50, pp. 483-499, Jun. 2003.

[17] G. C. Messenger and M. S. Ash, *The Effects of Radiation on Electronic Systems*. Van Nostrand Reinhold, 1992.

[18] C. E. Barnes, D. M. Fleetwood, D. C. Shaw, and P. S. Winokur, "Post irradiation effects (PIE) in integrated circuits," *IEEE. Trans. Nucl. Sci.*, vol. 39, pp. 328-341, Jun. 1992.

[19] G. L. Rhinehart and N. F. Modine, "Half-value layers at photon from 10 keV to 10 MeV," *U. S. National Center for Radiological Health*, 1966.

[20] "X-ray mass attenuation coefficients", *National Institute for Standards and Technology*, 2007.

[21] M. Maher, "Radiation owners manual", *National Semiconductor Corp.,* 1999.

Start-up Approach and Proposal for Nuclear Safety Knowledge Management Strategy in the Republic of Serbia

Koviljka Stankovic, *Member, IEEE*

***Abstract*— The aim of this paper is to present the start-up approach and proposal for nuclear safety knowledge management strategy in the Republic of Serbia. The main role in maintaining and protecting existing nuclear knowledge has been taken by academic staff. In order to share nuclear safety knowledge, university teachers started with implementation of bilateral agreements on professional and technical cooperation between respective institutions. The main goals of such cooperation, but not limited to, are effective solving of problems that are met in practice at professional level as well as involving students in real nuclear safety practice. It should be pointed out that the initiators of such coordination are highly educated young professionals (in science and engineering) who have been well trained through international training courses. That way of cooperation could be suitable starting point for straightening approach and developing long-term nuclear knowledge safety strategy at national level.**

***Index Terms*— nuclear safety, knowledge management, strategy, public opinion.**

I. Introduction

KNOWLEDGE management (KM) consists of three fundamental components: people, processes and technology (PPT sheme). KM focuses on people and organizational culture to stimulate and nurture the sharing and use of knowledge; on processes or methods to find, create, capture and share knowledge; and on technology to store and make knowledge accessible and to allow people to work together without being together. People are the most important component, because managing knowledge depends upon people's willingness to share and reuse knowledge [1-5].

Nuclear knowledge management (NKM) is knowledge management as applied in the nuclear technology field. It supports the gathering and sharing of new knowledge and the updating of the existing knowledge base. Knowledge management is of particular importance in the nuclear sector, owing to the rapid development and complexity of nuclear technologies and their hazards and security implications. The International Atomic Energy Agency (IAEA) launched a nuclear knowledge management programme in 2002 [6].

Concerns about global climate change and the availability of economically exploitable fossil fuels are driving many countries to reconsider the use of nuclear energy. Yet, the innovations required to design, construct, operate and maintain nuclear power plants consistent with international needs and constraints must derive from a strong foundation of well-sustained nuclear knowledge [7].

It is probable that nuclear knowledge will continue to expand and change. Without diligence in managing nuclear knowledge, substantial portions of it could be lost due to personnel retirements and the likelihood that much of it could be disused or discarded as a result of either negligence or changing priorities. It will be as important to identify and properly treat obsolete, superseded knowledge as it will be to gather and share new knowledge. It is therefore necessary to maintain effective and efficient KM systems. NKM has become an increasingly important element of the nuclear sector in recent years, resulting from a number of challenges and trends [8]:

- Countries with expanding nuclear programmes require skilled and trained human resources to design and operate future nuclear installations. Capacity building through training and education and transferring knowledge from centres of knowledge to centres of growth are key issues.
- In countries with stagnating nuclear programmes, the challenge is to secure the human resources needed to sustain the safe operation of existing installations, including their decommissioning and related programmes for spent fuel and waste. Replacing retiring staff and attracting the young generation to a career in the nuclear field are key challenges.
- Non-power applications of nuclear technologies require a stable or even growing base of nuclear knowledge and trained human resources, be it for cancer treatment or for food and agriculture. This need is present in all States using nuclear technologies, independent of the use of nuclear power.

Nuclear safety knowledge management (NSKM) includes the safety of nuclear instalations, radiation safety, the safety of radioactive waste management and safety in the transport of radioactive materials for the protection of people and the environment against radiation risk and for safety of faciliteies and activities that give rise to radiation risks, under normal circumstances or as a concequence of incidents [9].

This paper is focuses on current situation at national level in the field of nuclear knowledge. The majority of data presented, author said at Technical Meeting on Managing

Koviljka Stanković is with the Faculty of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail: kstankovic@etf.bg.ac.rs).

Nuclear Safety Knowledge - National Approaches and Experience held in IAEA Vienna from 17 July to 21 July, 2017.

II. BACKGROUND FACTS

The Republic of Serbia is a landlocked country situated at the crossroads of Central and Southeast Europe and the central Balkans. Serbia numbers around 7 million residents. Its capital, Belgrade, ranks among the oldest and largest cities in Southeast Europe. During the breakup of Yugoslavia in 1990s, Serbia formed a union with Montenegro which dissolved peacefully in 2006, when Serbia reestablished its independence.

In golden age of Nuclear Sciences, in early1950s, the Faculty of Electrical Engineering at University of Belgrade introduced the Nuclear engineering graduate program. Up to early 1990s Faculty of Electrical Engineering produced majority of personnel and scientist who has been working at Vinca Institute of Nuclear Sciences. The Vinca Institute was established in 1948, in the same time as Faculty of Electrical Engineering did. Vinca Institute of nuclear sciences was, at that time, the most significant research institution at national level and still is. Vinca have had two research reactors, RA and RB. The first one was shouted down before former Yugoslavia adopted moratorium on Nuclear Power uses in 1989. The other one, RB is zero power research reactor and operated until 2012 when the license expired.

The Vinca Institute had been splited in two parts in 2009 when the Government established the Public Company Nuclear Facillities of Serbia. Both reactors instalations together with radiactive waste storages became properties of Nuclear Facillities of Serbia. Currently, Reactor RA is in the process of decommisioning. In the same year, 2009, the Government established the national regulatory body – Serbian Radiation Protection and Nuclear Safety Agency. The first low on ionizing radiation was adopted in 2011. That low has been replaced with newly introduced Low on radiation and nuclear safety and security in 2018 [10]. In complianse with this Low the national regulatory body became independant body, so Serbia has been approaching condition met in developed countries.

From 1990s, applications of ionizing radiation can be studied at almost all state universities, mainly through undergraduate and graduate programs on Biomedical and environmental engineering and Medical physics at: University of Belgrade (Faculty of Electrical Engineering, Faculty of Physics, and Physical Chemistry), University of Novi Sad (Faculty of Technical Sciencies and Faculty of Physics), University of Kragujevac (Faculty of Physics), University of Nis (Faculty of Electronic Engineering, Faculty of Physics), University of Kosovska Mitrovica (Faculty of Physics).

Nowadays, the Faculty of Electrical Engineering (Univ. of Belgrade) retained the core of the subjects from the period before the 1990s, such as Nuclear Physics, Nuclear Technique and Dosimetry and Radiation Protection, and thus remained unique in the field of education in nuclear science and radiation protection at national level. There is a clear tendency in changing orientation in education from the period of moratorium adoption, so all faculties direct education in the field of radiation physics and technology that does not include nuclear power. At present, university programs that has retained the study of nuclear technology are master program for Biomedical and Environmental Engineering and doctoral program for Nuclear, Medical and Environmental Eng. both at the Faculty of Electrical Engineering, University of Belgrade.

It is interesting to point out that the Faculty of Electrical Engineering reintroduced the following courses in 2011: Neutron Physics, Nuclear Power Engineering and Application of Radioisotopes in Industry through the master program on Biomedical and Environmental eng. For the period of last seven years, the course on Nuclear Power Engineering has been followed by 230 students.

III. EMIGRATION AND AGEING

Migraton (or mobility) of competent personel has, in general, a positive impact in knowledge share. In our specific case, starting from 1990s and up to this date The Republic of Serbia suffers the emigration of well-educated and competent personel, since the period of the state trasition is long lasting and causes the brain drain effect. From the other hand, older generation of experts undergone the proces of retirement, so the number of people involved in the field of nuclear scieces rapidly drops down. Problem is more expressed by taking in account senior experts, regarded as “singletons”, being retired with having no immediate successors [1]. It is easily to conclude that Serbia already lost all generations older than 50 in the field of nuclear sciences.

IV. CURRENT SITUATION

A. People and Buliding the Competence

People involved in field of nuclear and radiation safety are mostly of age between 30 and 45. All young professionals employed by operators, regulators and R&D organizations acquired fundamental and applicative education at respective state universities. Majority of them permanently participate in international training courses, mainly in those organized by IAEA. In that sence, our young professionals are well - educated to understand and apply safety culture and the systemic approach to nuclear safety as well as to cope with technology changes.

For junior employees the training course that is exceptionally important and useful for building the competence is IAEA PGEC - Postgraduate Educational Course in Radiation Protection and the Safety of Radiation Sources [11]. Beside having a formal education, the IAEA PGEC gives the technical, operational and legal knowledge, self-confidence, capability to share knowledge, to promote safety culture and to build informal regional network of young professionals.

B. Communities of Practices

Professional associations that deals with nuclear/radiological safety at national and/or regional level

are:

- Society for Radiation Protection of Serbia and Montenegro is sucssesor of The Yugoslav Society for Radiation Protection (founded in 1963). The society acts as a professional non-profit organization that is oriented to improve the safety of people and the environment from harmful effects of radiation. The main activity of the society is to support the organization of research in all fields of application and use of radiation sources, and to encourage and monitor the highest professional and ethical principles in the course of this research. The society actively influence the writing of the proposal for the regulation of radiation protection at the national level, working with other professional societies and associations in the field of radiation protection in other countries and with international organizations and associations. Every second year the society organises symposium. Symposia proceedings are implemented in IAEA database (http://dzz.org.rs).
- The oldest and the most prestigious Serbian professional society in the field of technical sciences is ETRAN (E - Electronics, T - Telecommunications, R - Computing, A – Automatics, N - Nuclear Technologies). This professional association of engineers and nuclear scientists was established in 1953 as non-governmental and non-profit organization. Conferences are held annually and supported by IEEE (https://www.etran.rs). The Nuclear Section of the ETRAN has been strongly binded to the Nuclear Society of Serbia (http://nss.vinca.rs). Nuclear Society of Serbia has been in the phase of inactivity for last few years.

There are few societies related to Health Care such as: Society of radiology and nuclear medicine technicians (https://radteh.org.rs), Society of Nuclear medicine (www.unms.rs), Society of Medical Physics, Association of Clinical Engineers the Republic of Serbia (http://www.akis.rs), Society of medical oncologists of Serbia (umos.org.rs) etc.

C. Coordination mechanisms and Level of integration between institutions

International documents that deals with the nuclear safety knowledge management suggest that the government should take the main role in forming the strategy for nuclear safety knowledge management. That is an ideal situation, easily applicable in developed counties.

As in case of Serbia, the main role in maintaining and protecting existing nuclear knowledge has been taken by academic staff. In order to share nuclear safety knowledge, university teachers started with implementation of bilateral agreements on professional and technical cooperation between respective institutions. It should be pointed out that the initiators of such coordination are highly educated young professionals (in science and engineering) who have been well trained through IAEA PGEC and other Agency's training courses. The example of good practice is agreement between Faculty of Electrical Engineering (Univ. of Belgrade) with Public Company Nuclear Facilities of Serbia, as well as with Radiation and Environmental Protection Department of the Vinca Institute of Nuclear Sciences. Those agreements established the solid bases for BSc, MSc and PhD students education, professional and scientific collaboration among staff within these institutions in the field of nuclear sciences and radiation protection. The main goals of such cooperation, but not limited to, are effective solving of problems that are met in practice at professional level as well as involving students in real nuclear safety practice. Some results of such cooperation are traditionally published through conferences organised by the Society for Radiation Protection of Serbia and Montenegro and ETRAN society. This should be used as suitable starting point for integration between institutions in Serbia as well as for straightening approach and developing long-term nuclear knowledge safety strategy at national level, supervised by national regulatory body. Benefit of such approach can be realized in terms of efficiency (cost) savings.

The suggested approach by the author was presented at Technical Meeting on Managing Nuclear Safety Knowledge - National Approaches and Experience held in IAEA Vienna from 17 July to 21 July, 2017. The approach has been evaluated by IAEA, in still unpublished TECDOC report, as following: In Serbia, which faces the challenges of ageing of human resources and emigration, the University of Belgrade proposed a unique approach for national nuclear safety knowledge management, via coordination mechanisms among relevant institutions, Communities of Practice (CoP) and building competences. A new strategy for knowledge management is envisaged, for which the established Communities of Practice of young professionals could be a starting point [9].

D. Proposal for Creating a Strategy for NSKM

The proposal for creating the national strategy for NSKM should consist following successive steps:

- Orientation – where the knowledge is? Starting points: university collaboration with industry and/or regulatory body as well as with communities of practices helps to define the orientation of NSKM with cost savings.
- Strategy formulations – to consolidate initial ideas on current and further needs at national level.
- Design and launch – to define clear methodology for specific plans, timed plan of the tasks as well as IT platform for creating a formal network. At this point, national regulatory body should take the key role.
- Expand and support - KM is a long lasting process and it should be harmonized, by time to time, with national specific strategies.

E. Coping with Societal Occasions and Public Acceptance of Radiation

Accidents occurred during history are the main reason for anti-nuclear organizations in preventing global use of nuclear energy. Two accidents in history have dominantly influenced the change of national policies regarding the application of nuclear energy, as well as the change in the opinion of the public about the application of this energy source. These are the Accident in Chernobyl, which took place on April 26,

1986, and the Fukushima Accident on March 11, 2011. Following the Fukushima Accident, the acceptance of nuclear energy by public opinion has diminished to a historical minimum. Shortly after this event, the European Union in 2012 published a Communication "Energy Roadmap by 2050" [12]. With this document, the Commission proposed a transformation of the energy sector with a reduction in greenhouse gas emissions by 2050, to 80 to 95% below the emission level in 1990. Such a scenario can not be achieved without the use of nuclear energy. Serbia also envisioned the possibility of using nuclear energy to reduce the greenhouse effect in a document published in 2015 [13]. After all, nuclear fission is the second-largest world's low-carbon electricity source after hydroelectric power. Given the growing need for electricity, it can be concluded that civilization is entering the era of "nuclear renaissance". However, without public acceptance, the future of nuclear energy is uncertain.

Serbia is also specific in this respect, since public opinion forms an attitude of ionizing radiation on a completely different basis. Namely, the decisive influence on the shaping the opinion on the nuclear energy application has depleted uranium. After the NATO “intervention” in the territory of Serbia in 1999, the influence of media and the few medical doctors contributed to the fact that almost entire population in Serbia, regardless of the level of formal education, believes that the predominant causative agens of the “cancer epidemia” is basically depleted uranium. So, we are facing the consequences of false results promotion performed by few individuals and their followers [14].

Scientific facts, the knowledge that technical and medical staff wants to transfer to the population, as well as publications intended for the general auditorium, such as Ref. [14], can not change the attitude of the population about the complete mystification of depleted uranium.

In such an environment, it is much easier to educate young generations than adults. For the education of the youth and for raising awareness about nuclear energy, we have a good national example in the region, which is Slovenia [15]. The students of the Faculty of Electrical Engineering in 2012 visited the Slovenian Educational Center.

In terms of adult education, there are huge problems in front of us, so we have to face the first challenge - education of journalists, and then all other population groups in Serbia. So, shaping the public opinion on nuclear energy is still questionable point.

V. CONCLUSION

The paper presented the start-up approach and proposal for nuclear safety knowledge management in the Republic of Serbia. The main role in NSKM at national level should be taken by the university and research staff since they have obligations to follow new trends in the field and to work permanently on implementation in university studies programs. University collaboration with operators, R&D organisations and communities of practice should be used as suitable starting point for integration between institutions in Serbia as well as for straightening approach and developing long-term nuclear knowledge safety strategy at national level, all supervised by national regulatory body. Benefit of such approach can be realized in terms of efficiency (cost) savings. Each institution has its own knowledge tradition, so relevant institutions should be kept together through the nuclear safety knowledge management in order to create and develop both the national memories and national investment for the future.

ACKNOWLEDGMENT

This work is supported by The Ministry of Education, Science and Technological development under the project 171007.

REFERENCES

[1] Knowledge Management for Nuclear Research and Development Organizations, IAEA TECDOC 1675
[2] INTERNATIONAL ATOMIC ENERGY AGENCY, Knowledge Management for Nuclear Industry Operating Organizations, IAEA-TECDOC-1510, IAEA, Vienna (2006).
[3] INTERNATIONAL ATOMIC ENERGY AGENCY, Risk Management of Knowledge Loss in Nuclear Industry Organizations; STI/PUB/1248, IAEA, Vienna (2006).
[4] INTERNATIONAL ATOMIC ENERGY AGENCY, Managing Nuclear Knowledge, IAEA Proceedings including CD-ROM, STI/PUB/1266, ISSN: 0074-1884, IAEA,Vienna (2006).
[5] INTERNATIONAL ATOMIC ENERGY AGENCY, Status and Trends in Nuclear Education, IAEA Nuclear Energy Series, No. NG-T-6.1, IAEA, Vienna (2011).
[6] https://www.iaea.org/topics/nuclear-knowledge-management (accesed May, 2019)
[7] https://www.iaea.org/sites/default/files/16/11/np-parisagreement.pdf (accesed May, 2019)
[8] https://www.iaea.org/publications/13453/nuclear-knowledge-management-challenges-and-approaches (accesed May, 2019)
[9] Managing Nuclear Safety Knowledge - National Approaches and Experience, IAEA TECDOC (Unpublished).
[10] LAW ON RADIATION AND NUCLEAR SAFETY AND SECURITY (Official Gazette of the Republic of Serbia, No. 95/18 and 10/19)
[11] https://www.iaea.org/services/education-and-training/training-courses/training-radiation-transport-waste-safety (accesed May, 2019)
[12] Communication "Energy Roadmap 2050", https://ec.europa.eu/energy/sites/ener/files/documents/2012_energy_roadmap_2050_en_0.pdf (accesed May, 2019)
[13] Strategija razvoja energetike Republike Srbije do 2015. godine sa projekcijama do 2030. godine http://www.pravno-informacioni-sistem.rs/SlGlasnikPortal/eli/rep/sgrs/skupstina/ostalo/2015/101/1/reg
[14] Z. Radovanovic, Istina o raku, Heliks, Beograd, 2108.
[15] https://www.djs.si/proc/nene2017/html/pdf/NENE2017_1108.pdf

Karakterizacija moderatora detektora brzih neutrona primenom Monte Carlo simulacije

Jovana Knežević, Miloš Vujisić

***Apstrakt*—U radu je, pomoću Monte Carlo metode, izvršena simulacija odziva detektora brzih neutrona sa Bonner-ovim sferama, odnosno proračun efikasnosti detekcije u zavisnosti od energije upadnih neutrona u rasponu od 10^{-8} MeV do 10^{2} MeV, uzimajući u obzir različite prečnike moderatorske sfere (od 2 do 12 inča), kao i različite materijale moderatora: polietilen i parafin. Simulacija je izvršena pomoću kôda razvijenog u softverskom paketu MATLAB. U fizičke modele neutronskih interakcija uvedene su aproksimacije koje omogućavaju efikasnije izvršavanje razvijenog kôda. Rezultati dobijeni na ovaj način poređeni su sa rezultatima datim u literaturi, dobijenim detaljnijim modelom.**

***Ključne reči*—detekcija brzih neutrona; elastično rasejanje; Bonner-ove sfere; moderator; Monte Carlo; MATLAB.**

I. Uvod

Neutronsko zračenje predstavlja čestično, nenaelektrisano jonizujuće zračenje koje potiče iz jezgra, iz kog se emituje pri raspadu spontanom fisijom ili u nuklearnim reakcijama. U prirodne izvore neutronskog zračenja spada kosmičko zračenje, gde neutroni nastaju u nuklearnim reakcijama primarnog kosmičkog zračenja sa jezgrima gasova u gornjim slojevima atmosfere, pri čemu se neutroni oslobađaju procesom kaskadnih nuklearnih reakcija i raspada i kao sekundarno kosmičko zračenje dospevaju na Zemlju. U veštačke izvore neutrona, za primene u nauci, tehnologiji materijala i industriji, spadaju istraživački nuklearni reaktori, nuklearni generatori i prenosivi izvori. U raznim neutronskim poljima sreću se neutroni sa energijama u širokom opsegu, od nekoliko eV do nekoliko desetina MeV. U Tabeli I je data klasifikacija neutrona u zavisnosti od energije [1].

TABELA I
PODELA NEUTRONA PO ENERGIJAMA

Hladni neutroni	< 0,001 eV
Termički neutroni	0,001 – 0,5 eV
Epitermički neutroni	0,5 – 500 eV
Intermedijarni neutroni	0,5 – 100 keV
Brzi neutroni	0,1 – 10 MeV
Ultra brzi neutroni	> 10 MeV

Jovana Knežević – Elektrotehnički fakultet, Univerzitet u Beogradu, Bulevar Kralja Aleksandra 73, 11020 Beograd, Srbija (e-mail: jovanadknezevic@yahoo.com); istraživanje podržalo Javno preduzeće „Nuklearni objekti Srbije“, Mike Petrovića Alasa 12-14, 11351 Vinča, Beograd, Srbija

Miloš Vujisić – Elektrotehnički fakultet, Univerzitet u Beogradu, Bulevar Kralja Aleksandra 73, 11020 Beograd, Srbija (e-mail: vujsa@etf.rs).

Zbog elektroneutralnosti, neutron pri interakciji sa materijalnom sredinom gotovo ne reaguje sa elektronskim omotačem atoma, već direktno sa jezgrom. Neutroni ne osećaju odbojnu kulonovsku silu jezgra, pa su u stanju da izazovu nuklearne reakcije i pri veoma niskim energijama.

Neutronske interakcije, na kojim se zasniva i njihova detekcija, mogu se podeliti na rasejanja i apsorpcije, kao što je predstavljeno u Tabeli II.

TABELA II
TIPOVI INTERAKCIJA NEUTRONA I JEZGRA

Rasejanje	Elastično rasejanje	Potencijalno
		Rezonantno
	Neelastično rasejanje	Na nivoima
		Na kontinuumu
Apsorpcija	Reakcije sa emisijom više neutrona	
	Reakcije sa emisijom drugih čestica	
	Fisija	

Materijali koji se koriste za detekciju neutrona najčešće predstavljaju kombinaciju onih koji pretvaraju neutrone u sekundarne naelektrisane čestice i poznatih detektora ovih naelektrisanih čestica. U zavisnosti od energije (brzine) upadnih neutrona, detektori se mogu klasifikovati u detektore sporih neutrona i detektore brzih neutrona. Detekcija sporih neutrona se bazira na nuklearnim reakcijama pri kojima se neutroni konvertuju u naelektrisane čestice koje se direktno identifikuju jonizacionim komorama, scintilacionim brojačima i slično.

Kod detekcije brzih neutrona, prvo se vrši usporavanje neutrona do energija koje odgovaraju sporim neutronima, a zatim se mehanizam detekcije odvija po postupku za detekciju sporih neutrona. Materijal koji usporava neutrone se naziva moderator. Na njemu se neutroni usporavaju reakcijom elastičnog rasejanja, pa se za moderator bira onaj materijal koji ima veliku verovatnoću za elastično rasejanje. Najčešći materijali koji se koriste za moderaciju su stoga laka i teška voda, grafit, polietilen, parafin i drugi.

Nuklearne reakcije pogodne za detekciju sporih neutrona su: $^{10}B(n,\alpha)^{7}Li$, $^{6}Li(n,\alpha)^{3}H$, $^{3}He(n,p)^{3}H$, kao i fisija nuklida ^{233}U, ^{235}U i ^{239}Pu sporim neutronima. U materijale koji se koriste za detektore sporih neutrona ubrajaju se gasni proporcionalni brojač punjen bor tri fluoridom BF_3, scintilacioni detektor od litijum jodida aktiviranog

europijumom LiI(Eu), proporcionalni brojač punjen gasom helijuma ^{3}He i drugi.

Najčešće korišćen detektor na bazi moderacije neutrona danas predstavlja detektor sa Bonner-ovom sferom. Konfiguracija detektora sa Bonner-ovom sferom podrazumeva moderatorsku sferu u čijem se centru nalazi mali sferični ili cilindrični detektor [3,4].

II. Opis Monte Carlo simulacije

Monte Carlo metoda predstavlja statističku tehniku uzorkovanja koja se uspešno primenjuje u raznovrsnim naučnim disciplinama. Interakcije koje određuju istoriju svakog pojedinačnog neutrona koji se prati modeluju se probabilističkim zakonima. Metoda je razvijena sredinom dvadesetog veka i prvenstveno je korišćena za istraživanje ponašanja neutronskih lančanih reakcija u fisionim uređajima.

Monte Carlo simulacija transporta neutrona za potrebe istraživanja u ovom radu sprovodi se u nekoliko koraka.

Najpre se uzorkuje pravac upadnog neutrona. Ukoliko se razmatra izotropno polje neutrona kome je izložen detektor na bazi Bonner-ovih sfera, na slučajan način se uzorkuju parovi tačaka na površini moderatorske sfere. Tetiva koja spaja te dve tačke predstavlja pravac upadnog neutrona.

Zatim se određuje mesto interakcije neutrona, pretpostavljajući da čestica ulazi u prostor sastavljen od dva regiona (moderatorske sfere i detektora) od homogenog materijala, od kojih je svaki okarakterisan makroskopskim efikasnim presekom za interakciju.

Posle mesta interakcije potrebno je odrediti na kom se atomu materijala dogodila interakcija. Nakon odabira jezgra, vrši se izbor interakcije na tom jezgru. Neutron će putem apsorpcije nestati sa verovatnoćom $\frac{\sigma_a}{\sigma_t}$, gde je σ_a mikroskopski efikasni presek za apsorpciju, a σ_t totalni mikroskopski efikasni presek. Ostale interakcije su reakcije rasejanja, u koje spadaju elastično i neelastično rasejanje [2].

Izbor elastičnog rasejanja se vrši u skladu sa verovatnoćom $\frac{\sigma_{el}}{\sigma_t-\sigma_a}$. Dalji koraci za modelovanje su:

1. Izabere se slučajan broj ξ_1
2. Odredi se kosinus ugla rasejanja u sistemu centra mase preko $cos\theta^{CM} = 1 - 2\xi_1$
3. Odredi se kosinus ugla rasejanja u laboratorijskom sistemu preko $cos\theta = \frac{1+Acos\theta^{CM}}{\sqrt{1+2Acos\theta^{CM}+A^2}}$
4. Odredi se kinetička energija neutrona nakon rasejanja $E_n' = E_n \frac{1+2Acos\theta^{CM}+A^2}{(A+1)^2}$
5. Izabere se drugi slučajni broj ξ_2
6. Odredi se azimutalni ugao u laboratorijskom sistemu preko $\varphi = 2\pi\xi_2$
7. Isporuče se φ, θ, E_n'.

Simulacija neelastičnog rasejanja na nivoima se vrši kroz sledeće korake:

1. Izabere se slučajan broj ξ_1
2. U slučaju izotropnog neelastičnog rasejanja u sistemu centra mase, određuje se $cos\theta^{CM} = 1 - 2\xi_1$
3. Određuje se $B = \sqrt{A^2 + QA\frac{A+1}{E_n}}$, gde vrednost $-Q$ određuje eksitacionu energiju jezgra
4. Određuje se ugao rasejanja u laboratorijskom sistemu preko $cos\theta = \frac{1+Bcos\theta^{CM}}{\sqrt{1+2Bcos\theta^{CM}+B^2}}$
5. Izračunava se kinetička energija neutrona nakon rasejanja kao $E_n' = E_n \frac{1+2Bcos\theta^{CM}+B^2}{(A+1)^2}$
6. Izabere se drugi slučajni broj ξ_2
7. Odredi se azimutalni ugao u laboratorijskom sistemu preko $\varphi = 2\pi\xi_2$
8. Isporuče se φ, θ, E_n'.

III. Rezultati simulacija i diskusija

U svrhu ovog numeričkog eksperimenta razvijen je MATLAB kôd pomoću kog je izvršena simulacija odziva detektora sa Bonner-ovom sferom Monte Carlo metodom. Geometrija problema uključuje sfere različitih prečnika (od 2 do 12 inča) koje se nalaze u izotropnom polju neutrona. Pored toga, izvršena je i varijacija materijala moderatorske sfere. Uzeti su u obzir polietilen i parafin. Detektor je scintilacioni kristal LiI(Eu), sferne geometrije, prečnika 4 mm. Centar moderatorske sfere se poklapa sa centrom detektora. U Tabeli III dati su osnovni podaci za materijale moderatorske sfere i detektora, koji su korišćeni prilikom Monte Carlo proračuna.

TABELA III
Osnovni podaci o materijalu moderatora i detektora

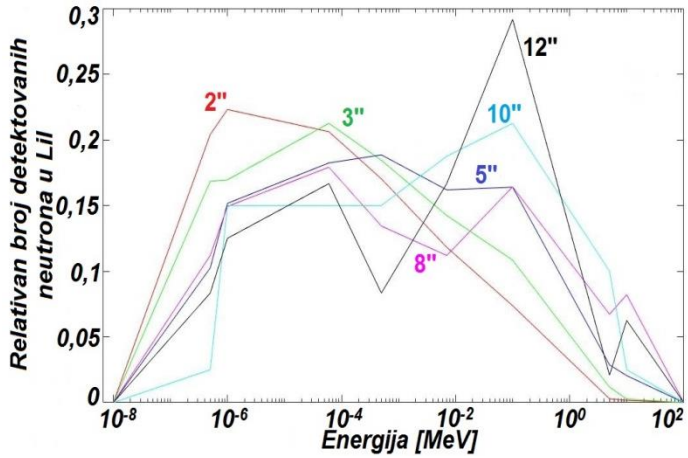
Materijal	Polietilen	Parafin	Litijum jodid (detektor)
Hemijska formula	CH_2	$C_{30}H_{62}$	LiI(Eu)
Molarna masa [g/mol]	14	422	133
Gustina [g/cm^3]	0,95	0,87	4,08

Izotropno neutronsko polje je simulirano slučajnim izborom parova tačaka na moderatorskoj sferi, gde tetiva koja spaja jedan par tačaka čini prvobitni pravac inicijalnog neutrona. Simulacija je izvršena za 500 000 istorija čestica, sa 11 diskretnih tačaka za energiju u rasponu od 10^{-8} MeV do 10^2 MeV za svaku česticu.

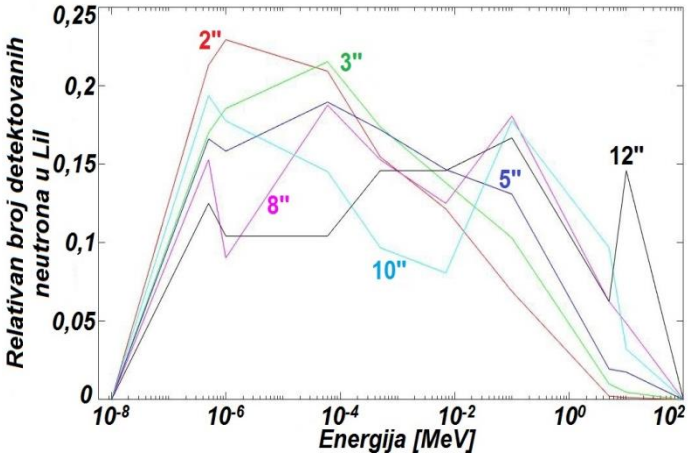
Podaci za mikroskopske efikasne preseke uzeti su iz biblioteke podataka ENDF/B-VII [5]. Razvijen MATLAB kôd se oslanja na podatke za eksperimentalno određene preseke koji opisuju interakciju neutrona sa jezgrom, za

energije neutrona niže od 20 MeV. Za neutrone energija preko 20 MeV ne postoji dovoljno podataka za preseke za interakciju, pa se stoga koriste nuklearni modeli (intranuklearna kaskada i evaporacioni modeli) u okviru Monte Carlo metode. Proračuni funkcije odziva Bonner-ove sfere pomoću Monte Carlo metode i nuklearnih modela mogu smanjiti pouzdanost, posebno pri energijama neutrona većim od 20 MeV.

Na slikama 1 i 2 dat je grafički prikaz rezultata simulacija odziva detektora, u vidu energetskih zavisnosti efikasnosti detekcije, koja je računata kao relativni broj detektovanih neutrona, za dva ispitivana materijala i za različite prečnike moderatorske sfere.



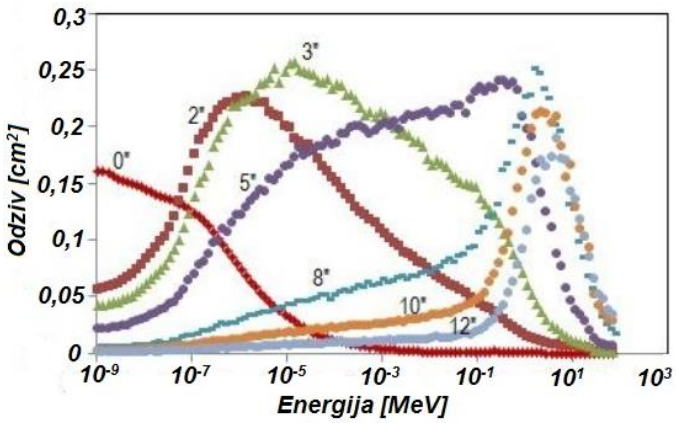
Sl.1. Poređenje efikasnosti detektora sa Bonner-ovim sferama različitih prečnika, materijal sfere je polietilen



Sl.2. Poređenje efikasnosti detektora sa Bonner-ovim sferama različitih prečnika, materijal sfere je parafin

Kriterijumi za okončanje istorije u simulacijama bili su: prag energije neutrona do kog se neutron prati (0,025 eV), apsorpcija unutar moderatorske sfere i izlazak neutrona iz moderatorske sfere. Smatra se da je neutron detektovan unutar scintilacionog kristala LiI ako je doživeo reakciju apsorpcije na jezgru ^{6}Li. Navedeni kriterijumi okončanja istorije i aproksimacije koje sadrži razvijeni MATLAB kôd daju relativno grube procene rezultata kada se dobijeni grafici porede sa graficima koji su nastali simulacijama nekih od detaljnijih modela iz literature [7,8,9].

Dobijeni rezultati mogu da se uporede sa rezultatima u radu [6], u kom je sprovedena simulacija izlaganja Bonner-ovih sfera (istih prečnika kao u ovom radu) neutronskom izvoru oblika diska, sa 10^{9} istorija i 111 diskretnih tačaka za energiju u rasponu od 10^{-9} MeV do 10^{2} MeV. Geometrija scintilacionog detektora LiI(Eu) je cilindrična. Navedena studija je za simulaciju transporta zračenja koristila softverski paket MCNP (Monte Carlo N-Particle). Odziv detektora sa Bonner-ovim sferama različitih prečnika, sa polietilenom kao materijalom moderatora, koji se može naći u [6], dat je na slici Sl.3. U navedenom radu je takođe izvršena simulacija odziva za sferu prečnika 0 inča, što odgovara detektoru bez moderatora, za koji oblik funkcije odziva direktno odražava mikroskopski presek za (n,α) reakciju na ^{6}Li.



Sl.3. Funkcija odziva Bonner-ovih sfera dobijena pomoću softverskog paketa MCNP [6]

Konfiguracija moderatora je najčešće sferna, kako bi se dobio izotropan odziv detektora. Za sfere malih dimenzija, stepen moderacije je mali, kao i zahvat termičkih neutrona unutar moderatora. Zbog toga, niskoenergetski neutroni imaju veliku verovatnoću da dođu do scintilacionog LiI kristala i da budu detektovani, a brzi neutroni uspevaju da umaknu iz moderatora. Kod sfera velikih dimenzija veći je i stepen moderacije, odnosno verovatnoća da će brzi neutroni biti usporeni pri elastičnim rasejanjima na jezgrima moderatorskog materijala. Takođe je veća i verovatnoća apsorpcije termičkih neutrona u moderatoru, pre nego što dospeju do detektora. Stoga su upravo neutroni visokih energija ti koji će prevashodno biti detektovani u LiI detektoru. Može se zaključiti da je za male dimenzije Bonner-ove sfere efikasnost detekcije veća za neutrone nižih energija, dok je za sfere većih prečnika efikasnost detekcije veća za visokoenergetske neutrone. Zbog toga se maksimumi efikasnosti detekcije pomeraju ka višim energijama, kako prečnik moderatorske sfere raste.

Sa grafika prikazanog na Sl.1, za sfere prečnika od 2 do 12 inča, glavne razlike su primećene u oblastima koje odgovaraju niskim energijama neutrona i u oblastima koje odgovaraju energijama preko 20 MeV. Funkcije odziva za sfere sa prečnicima od 2, 3 i 5 inča su značajne za oblast od 10^{-8} MeV do 10^{-3} MeV, odnosno za termičke i epitermičke neutrone. U opsegu od 10^{-2} MeV do 10^{2} MeV dominantne su funkcije

odziva sfera prečnika 8, 10 i 12 inča. Za sfere prečnika 2 i 3 inča oblici funkcije su međusobno slični i uticaj mikroskopskog efikasnog preseka za (n,α) reakciju na ^{6}Li polako postaje neznatan. U oblasti niskih energija, uočava se da dominira funkcija odziva sfere prečnika 2 inča u odnosu na prečnike od 3 i 5 inča, a vrednost maksimuma se nalazi u opsegu od 0,2 do 0,25. Dalje se može primetiti kako sa porastom dimenzija sfera efikasnost detekcije ima tendenciju opadanja za termičke i epitemičke neutrone, a maksimumi se pomeraju ka višim energijama. Maksimumi funkcije odziva za sfere od 5 i 8 inča imaju približne vrednosti, ali se oni ne nalaze u istom užem energetskom opsegu. Uočava se značajan porast relativnog broja detektovanih neutrona za sferu prečnika 12 inča koji dostiže vrednost nešto ispod 0,3 u oblasti energija oko 10^{-1} MeV, u odnosu na maksimume funkcija odziva za druge prečnike sfere moderatora.

Prema rezultatima prikazanim na Sl.2 takođe se može primetiti dominantnost funkcija odziva za sfere prečnika 2, 3 i 5 inča u oblasti niskih energija od 10^{-8} MeV do 10^{-3} MeV, i za sfere prečnika 8, 10 i 12 inča u visokoenergetskoj oblasti od 10^{-2} MeV do 10^{2} MeV. Funkcija ima sličan oblik za sfere prečnika od 2, 3 i 5 inča u oblasti niskih energija, gde se maksimum funkcije za sferu od 2 inča ističe u odnosu na funkcije odziva sfera od 3 i 5 inča i vrednost maksimuma se nalazi u opsegu od 0,2 do 0,25. Efikasnost detekcije zatim opada kako rastu energija neutrona, za ove tri sfere. U oblasti visokih energija neutrona, funkcije odziva sfera od 8, 10 i 12 inča dobijaju na značaju. Primećuje se da se maksimumi funkcija odziva ove tri sfere nalaze u energetskoj oblasti oko 10^{-1} MeV, gde je vrednost maksimuma najveća za sferu od 8 inča i nalazi se u opsegu od 0,15 do 0,2. Takođe, uzimajući u obzir maksimalne vrednosti relativnog broja detektovanih neutrona za sve sfere, uočava se da je najveća efikasnost detekcije za sferu prečnika 2 inča.

Ukoliko se izvrši poređenje dva grafika efikasnosti detekcije u zavisnosti od energije neutrona, dobijene za različite materijale moderatora, vidi se da u oblasti energija od 10^{-8} MeV do 10^{-3} MeV relativan broj detektovanih neutrona za sfere manjih dimenzija za parafin dostiže veće maksimalne vrednosti nego za polietilen. Za oblasti energija od 10^{-2} MeV do 10^{2} MeV maksimumi funkcije odziva imaju veće vrednosti za sfere većih dimenzija za polietilen u odnosu na parafin. Ukoliko se uzme u obzir kompletan grafički prikaz, primećuje se da se maksimalna vrednost efikasnosti detekcije dobija za polietilen, za sferu najvećih dimenzija u visokoenergetskom opsegu. Za parafin se maksimalna vrednost efikasnosti detekcije dobija za sferu najmanjih dimenzija u niskoenergetskom opsegu. Ako se posmatraju energije neutrona, može se reći da za visokoenergetske neutrone, bolju efikasnost detekcije ima moderatorska sfera većih dimenzija napravljena od polietilena. Za niskoenergetske neutrone, veću efikasnost detekcije ima moderatorska sfera manjih dimenzija napravljena od parafina.

Svi dobijeni rezultati pokazuju znatno slaganje sa rezultatima na Sl.3, dobijenim detaljnijim modelima neutronskih reakcija u paketu MCNP. S druge strane, za očekivati je da se kôd razvijen za aktuelno ispitivanje, koji je opisan u odeljku II, izvršava brže i na slabijim računarskim platformama od onih koje često zahteva MCNP, što je prednost kada je potrebno izvršiti niz ispitivanja odziva detektora sa raznim geometrijskim konfiguracijama i različitim materijalma moderatora i osetljive zapremine.

IV. ZAKLJUČAK

Cilj rada je razmatranje odziva detektora sa Bonner-ovom sferom u zavisnosti od energije upadnih neutrona, uzimajući u obzir različite prečnike i materijale moderatorske sfere. Analiza je sprovedena Monte Carlo simulacijom transporta neutrona kôdom razvijenim u softverskom paketu MATLAB.

Sa dobijenih grafika prikazanih na Sl.1 i Sl.2 može se uočiti da krive efikasnosti detekcije za sfere manjih dimenzija dostižu maksimume u energetskim oblastima od 10^{-8} MeV do 10^{-3} MeV. Kako prečnik sfere raste, tako se maksimumi krive efikasnosti detekcije pomeraju ka višim energijama neutrona, a maksimumi se nalaze u oblasti energija oko 10^{-1} MeV. Najveća vrednost relativnog broja detektovanih neutrona za polietilen ima sfera od 12 inča, dok za parafin ima sfera od 2 inča.

Moderatorska sfera od polietilena ima veću efikasnost detekcije za visokoenergetske neutrone u odnosu na parafinsku sferu, a parafin daje dobar odziv za niskoenergetske neutrone ukoliko se koristi sfera manjih dimenzija. Uzimajući u obzir hemijske formule polietilena (CH_2) i parafina ($C_{30}H_{62}$), ovo zapažanje se može pripisati manjem stepenu moderacije parafina zbog prisustva većeg broja atoma ugljenika u parafinu kao i manje gustine u odnosu na polietilen.

ZAHVALNICA

Ovaj rad je podržan od strane Javnog preduzeća „Nuklearni objekti Srbije“ i Ministarstva prosvete, nauke i tehnološkog razvoja u okviru projekta br. 171007.

LITERATURA

[1] D.Popović, “Osnovi nuklearne tehnike”, Naučna knjiga, 1970.
[2] P. Marinković, “Nuklearna medicinska tehnika”, skripta za predavanja, Univerzitet u Beogradu, Elektrotehnički fakultet, 2014.
[3] G.F.Knoll. “Radiation detection and measurement”, Wiley, 2010.
[4] M. Vujisić, P. Osmokrović, K. Stanković. “Dozimetrija i zaštita od zračenja”, skripta za predavanja, Univerzitet u Beogradu, Elektrotehnički fakultet, 2016.
[5] http://www.nndc.bnl.gov/sigma/index.jsp
[6] R. Tursinah, Bunawas, J. Kim. “Neutron response function of a Bonner sphere spectrometer with ^{6}LiI(Eu) detector”, Ganendra Journal of Nuclear Science and Technology Vol. 20, No. 2, Juli 2017 : 65-72
[7] S. Moratóa, B. Justea, R. Miróa, G. Verdúa, V. Guardiab, “Evaluation of the response of a Bonner Sphere Spectrometer with a ^{6}LiI detector using 3D meshed MCNP6.1.1 models”, Radiation Physics and Chemistry, Vol. 155, February 2019, pp. 221-224.
[8] Z.M. Hu, X.F. Xie, Z.J. Chen, X.Y. Peng, T.F. Du, Z.Q. Cui, L.J. Ge, T. Li, X. Yuan, X. Zhang, L.Q. Hu, G.Q. Zhong, S.Y. Lin, B.N. Wan, G. Gorini, X.Q. Li, G.H. Zhang, J.X. Chen, T.S. Fan, “Monte Carlo simulation of a Bonner sphere spectrometer for application to the determination of neutron field in the Experimental Advanced Superconducting Tokamak experimental hall”, Rev Sci Instrum. 2014, 85(11):11E417.
[9] R.M. Howell, E.A. Burgett, B. Wiegel, N.E. Hertelb, “Calibration of a Bonner sphere extension (BSE) for high-energy neutron spectrometry”, Radiat Meas. 2010, 45(10), pp. 1233-1237.

ABSTRACT

In this paper, using the Monte Carlo method, a simulation of the response of fast neutron detectors with Bonner spheres has been performed, in terms of detection efficiency dependence on neutron energy in the range from 10^{-8} MeV up to 10^{2} MeV, taking into account different diameters of the moderator sphere (from 2 to 12 inch), as well as various moderator materials: polyethylene and paraffin. The simulation was conducted using a code developed for this purpose in the MATLAB software package. Approximations have been introduced to the physical models of neutron interactions that enable more efficient code execution. Results obtained in this manner are compared to the results obtained by a more detailed model, found in literature.

Characterization of fast-neutron detector moderators based on Monte Carlo simulation

Jovana Knežević, Miloš Vujisić

Uloga Pavla Savića u otkriću fisije

Dragoslav Nikezić

***Apstrakt*—U ovom radu je opisano otkriće nuklearne fisije i uloga Pavla Savića u tom procesu. Dve istraživačke grupe, jedna u Berlinu i druga u Parizu, su ulagale ogromne napore ka dobijanju transuranskih elemenata. Ozračivan je uran u nadi da će niz beta minus raspada dovesti do stvaranja elemenata sa atomskim brojem većim od 92, ali je to ozračivanje dovodilo do fisije. Mnogobrojne poteškoće su dovodile do pogrešnih zaključaka što je otežalo razumevanje stvarnih procesa u ozračenom uranu.**

***Ključne reči* — Fisija, transuranski elementi, neutron, lantanoidi**

I. UVOD

Pošto je na Prirodno matematčkom fakultetu u Kragujevcu. 2019. godina proglašena godinom Pavla Savića, autor ovog rada je našao za shodno da se detaljnije osvetli njegova uloga u otkriću fisije. Otkriće fisije je posledica intenzivnih napora ka sintetizovanju transuranskih elemenata u periodu od 1934. godine do 1939. godine. Istorija ovog otkrića je detaljno opisana u većem broju publikacija [1],[2],[3].

Dve istraživačke grupe su se naročito istakle velikim radom i aktivnošću na ovom polju, jedna u Berlinu, koju su činili Otto Hahn, Fritz Strassmann i Lisa Meitner, i druga u Parizu, Irena Joliot Cuiri i Pavle Savić. Otkriće fisije je pratilo niz zabluda, pogrešnih tumačenja eksperimentalnih rezultata, čak i kad je objašnjenje bili skoro očigledno, kao i kontraverze oko dodele Nobelove Nagrade.

Sama reč "fisija" znači cepanje jezgra i u užem smislu podrazumeva se cepanje teških jezgara u dva nejednaka fragmenta posle zahvata neutrona. Cepanje jezgra se može odigrati na veći broj načina, tako da se u komadu urana ozračenog neutronima nalazi veliki broj različitih izotopa. Fisija je otkrivena na tri najteža elementa u prirodi, Th, Pa i U.

U periodu pre drugog svetskog rata nije postojala spektroskopija, te je identifikacija izotopa radjena na osnovu vremena poluraspada. Aktivnost izotopa je merena Gajger Milerovim brojačima. Poznato je da aktivnost svakog radioizotopa opada eksponencijalno sa vremenom; u logaritamskoj skali kriva raspada se pretvara u pravu čiji je koeficijenat pravca povezan sa konstantom radioaktivnog raspada, tj., sa vremenom poluraspada. Ovakav metod identifikacije izotopa je vrlo komplikovan i verovatno netačan u slučaju smeše većeg broja različitih izotopa, što se inače dobija pri fisiji. Problem identifikacije izotopa dobijenih u komadu urana izloženog neutronima, je jednim delom doprineo teškom otkrivanju fisije. Umesto gama spektroskopije, za identifikciju izotopa primenjivane su hemijske metode ekstrakcije, tzv. višestruka frakciona kristalizacija, posle čega je mereno njihovo vreme poluraspada.

Godine 1932. otkriven je neutron, mada je njegovo postojanje pretpostavio Raderford 12 godina pre eksperimentalne potvrde. Medjutim, Raderfordova pretpostavka je bila potpuno zaboravljena. Otkriće neutrona su pratile mnogobrojne teškoće i kontraverze, po pitanju nejasnoće dobijenih eksperimentalnih podataka.

Otkrićem neutrona eksperimentalni fizičari su dobili sredstvo kojim su mogli da bombarduju atome i na niskim energijama. Neutron je neutralan i lako može da udje u pozitivno naelektrisano jezgro, jer ne oseća električno polje naelektrisanih čestica kao u slučaju alfa čestice ili protona.

Enriko Fermi je prvi uvideo mogućnosti koje pružaju neutroni i krenuo je sa sistematskim ozračivanjem poznatih elemenata. Nije našao nikave nove nuklearne reakcije sve do fluora, F. Dobijena reakcija na fluoru ga je ohrabrila da nastavi rad sve do najtežeg elementa urana. Dobio je povećanje aktivnosti urana posle ozračivanja (uran je prirodna radioaktivan, ali je njegova aktivnost mala), i sasvim slučajno je otkrio da se efekat pojačava ako se umesto brzih, koriste spori neutroni. E. Fermi je u suštini proizveo fisiju u komadu urana, ali nije prepoznao značaj dobijenih rezultata. Rezultati koje je dobijao Fermi su bili kontradiktorni. Za svoj rad Fermi je dobio Nobelovu Nagradu 1938., a posle toga je napustio Italiju. U Italiji je u tom periodu narastao fašizam, a njegova supruga je bila jevrejskog porekla.

U to vreme pažnja istraživača je bila zaokupljena otkrićem transuranskih elemenata, tj., elemenata sa atomskim brojevima većim od 92. Fermi je dao sledeće smernice koje su bile potpuno pogrešne: *prva je bila da nuklearne reakcije izazivaju samo male promene jezgra*. Druga je da *svi elementi sa atomskim brojevima većim od urana (transurani) treba da budu slični elementima treće grupe prelaznih metala, Re, Os, Ir i Pt jer se nalaze u grupama periodnog sistema ovih elemenata*. U to vreme nije bilo poznato da postoje aktinidi i da svi imaju slične hemijske osobine, jer se popunjava unutrašnja 5f podljuska. Ove dve zablude su imale velikog uticaja na kasniji rad istraživača u ovoj oblasti, jer su uporno pokušavali da dobiju transurane, koji bi bili slični Re, Os, Ir i Pt. Čak su im dali i imena ekaRe, ekaOs, ekaIr i sl. Pored ovih zabluda, istraživače je zbunjivalo i to što su količine dobijenih elemenata vrlo male. Nisu znali da se fisija proizvodi na izotopu ^{235}U koji u ukupnom uranu učestvuje svega sa 0.72%.

Dve, ranije pomenute grupe, u Parizu i Berlinu su naročito intenzivno nastavile rad na dobijanju transuranskih elemenata.

Dragoslav Nikezić – Prirodno-matematički fakultet, Univerzitet u Kragujevcu, R. Domanovića 12, 34000 Kragujevac (e-mail nikezic@kg.ac.rs).

II. RAD GRUPE U BERLINU

Fermi je ranije dobio četiri beta emitujuće materije u uranu ozračenom neutronima sa vremenima poluraspada 10 s, 40 s, 13 min i 90 min. Grupa u Berlinu je ubrzo našla deset aktivnosti, dok su Cuiri i Savić našli osam aktivnosti u reakciji neutron-torijum. Hemijska analiza je obavljana rastvaranjem ozračenog urana u kiselinama, i hemijskom separacijom. Problem je bio mala količina stvorenih elemenata. Obe grupe su bile ubedjene da se ozračivanjem stvaraju transuranski elementi i da su analogni trećoj grupi prelaznih metala Re, Os, Ir, Ot, kako je to Fermi sugerisao. Zbog toga je korišćen renijum za taloženje radioaktivnih produkata u obliku sulfida. Ovaj način taloženja je bio problematičan jer je zajedno sa željenim, taložio i brojne druge elemente. Ovo je bila stranputica u istraživanjima izazvana pogrešnim zaključcima autoriteta Fermija. Drugi glavni eksperimentalni problem je bio nagomilavanje produkata raspada koji su preplavili male iznose novo formiranih produkata.

Takodje, obe grupe su smatrale da se neutron-uran reakcija odigrava na ^{238}U zbog njegove velike obilnosti u prirodnom uranu od 99.275 %. Tek 1939. godine, kada je shvaćeno da se radi o fisiji, prepoznato je da je fisibilan ^{235}U sa izotopskom obilnošću od svega 0.72 %. Ovim se objašnjava dobijanje vrlo malih količina novo proizvedenih izotopa. Značajan problem su i relativno kratka vremena poluraspada pojedinih izotopa (npr 10 s) što je onemogućavalo radiohemijsku analizu. Zbog toga je pažnja usmerena na radioaktivne elemente koji su dugoživeći i koje je moguće (bez obzira na minimalne dobijene količine) ipak analizirati radiohemijski.

Obe grupe su imale izuzetne eksperimentalne poteškoće.

Broj novodobijenih vremena poluraspada, tj., broj novodobijenih radioizotopa, se povećavao i to je dodatno komplikovalo analizu. Rasla je kompleksnost problema.

Hahn, Meitner i Strassmann (hemičar analitičar) su publikovali brojne radove u perodu 1935. do 1938. godine, izveštavajući o novo dobijenim aktivnostima i poboljšavajući podatke i eksperimentalne tehnike. U tim publikacijama često su menjali vremena poluraspada, identifikovali nova. Bili su čvrsto ubedjeni da se radi o transuranskim elementima i da su oni slični elementima treće grupe prelaznih metala. Razvili su niz šema raspada, i dali imena „novo otkrivenim“ elementima (pomenuta ranije). Grupa u Berlinu je bila sigurna u sledeće šeme raspada

$$n+{}_{92}U \rightarrow {}^{10s}_{92}U \xrightarrow{\beta} {}^{2.2\,\mathrm{min}}Eka_{93}\mathrm{Re} \xrightarrow{\beta} {}^{59\,\mathrm{min}}Eka_{94}Os \xrightarrow{\beta} {}^{66h}Eka_{95}Ir \xrightarrow{\beta} {}^{2.25h}Eka_{93}Pt \xrightarrow{\beta} Eka_{97}Au$$

$$n+{}_{92}U \rightarrow {}^{40s}_{92}U \xrightarrow{\beta} {}^{2.2\,\mathrm{min}}Eka_{93}\mathrm{Re} \xrightarrow{\beta} {}^{55.7h}Eka_{94}Os \xrightarrow{\beta} Eka_{95}Ir$$

$$n+{}_{92}U \rightarrow {}^{23\,\mathrm{min}}_{92}U \xrightarrow{\beta} Eka_{93}\mathrm{Re}$$

Postavljene su i druge hipoteze u vezi sa izomerizmom pojedinih elemenata, što je bilo poznato kao moguće. Izomerija je pojava da jedan izotop ima više vremena poluraspada. Prethodne šeme su bile u potpunosti pogrešne, kao i tumačenja oko izomerije.

U jednoj situaciji Hahn je obavio ozračivanje urana neutronima u toku od 80 dana. Ovakvo ozračivanje je moglo da proizvede dovoljnu količinu neptunijuma i plutonijuma, i da dovede do otkrića prva dva transurana. Nadjen je izotop sa vremenom poluraspada od 60 dana, ali nije nastavljen rad u ovom smeru.

U tom periodu došlo je do jakog porasta nacizma u Nemačkoj. Lisa Meitner je imala jevrejsko poreklo i uspela je da napusti Nemačku. Dospela je do Švedske tj., Štokholma gde nije imala skoro ničega ni za život ni za rad. Obrela se u laboratoriji Seighbahna.

U jednom periodu Hahn i Lisa su nastavili saradnju, eksperimenti su sprovodjeni prema Lisinim predlozima, i imali su više zajedničkih publikacija. Medjutim, koautorstvo je prekinuto, jer je sada Hahn bio u opasnosti zbog publikovanja sa jevrejskom kolegenicom.

III. RAD GRUPE U PARIZU

Grupa u Parizu. Irena i Pavle su radili na sličan način kao i u grupa u Berlinu. Oni su, 1935. u neutron-torijum reakciji našli četiri aktivnosti, koje su kasnije Hahn i Meitner povećali na šest. Od 1937. god. Irena i Pavle su počeli eksperimente ozračivanja urana neutronima. Ubrzo su našli aktivnost koja je opadala sa vremenom poluraspada od 3.5 h (koju Hahn i Meitner nisu nalazili). Za razliku od H&M, Pavle i Irena su koristili lantan kao nosilac. Lantan je izabran jer se u periodnom sistemu nalazi u istoj grupi kao i aktinijum, Ac.

Primenjujući frakcionu kristalizaciju Irena i Pavle su prvo sugerisali da je elemenat sa vremenom poluraspada od 3.5 h Aktinijum. Ponovo su obavili frakcionu kristalizaciju produkta očekujući da će se novodobijeni izotop odvojiti od lantana, ali se to nije desilo. Aktivnost je ostala sa lantanom, pri čemu Ac nije. Prvobitna pretpostavka da aktivnost od 3.5 h potiče od aktinijuma Ac, su ubrzo odbacili. Slika iz originalnog rada Cuiri i Savića je data na Slici 1.

Irena i Pavle su bili vrlo blizu otkriću fisije, jer su u eksperimentu zaista dobili lantan, zbog čega i nisu mogli da ga odvoje od lantana, koji su koristili kao nosač. Lantan je jedan od značajnih fisionih produkata i javlja se u više različitih izotopa u fisionoj smeši. I ova grupa je bila pod uticajem Fermijeve ideje da se dobijaju transurani ili elementi koji su blizu urana u periodnom sistemu. Ipak su ustvrdili da interpretacija rezulata stvara mnogobrojne teškoće. Publikovali su dva rada [4], [5]:

Ovo je prva jasna naznaka o dobijanju elemenata iz sredine periodnog sistema. I pored toga, nikom nije palo na pamet da dolazi do cepanja jezgra urana. Bilo je neshvatljivo da neutron energije reda 1 eV može da izazove raspad velikog jezgra urana koje ima 238 nukleona. Smatralo se da ako dolazi do cepanja jezgra, produkti cepanja moraju da tuneluju Coulumb-ovu potencijalnu barijeru, što je spor proces.

Grupa iz Berlina nije verovala ovim rezultatima, i Hahn je te rezultate komentarisao kao zbrku, neljubaznost i profesionalnu ljubomoru: produkt od 3.5 h je nazvao kuriozitetom. Hahn i Meitner su pisali Ireni Cuiri, sugerišući da je podatak o izotopu sa vremenom poluraspada 3.5 h pogrešan i da je potrebno da povuku publikovani rad. Rasla je napetost izmedju ovih dveju grupa i konačno je Hahn zamolio

Frederika da zaustavi Irenu i Savića u publikovanju rezultata koji pobijaju njegove i Meitnerove rezultate.

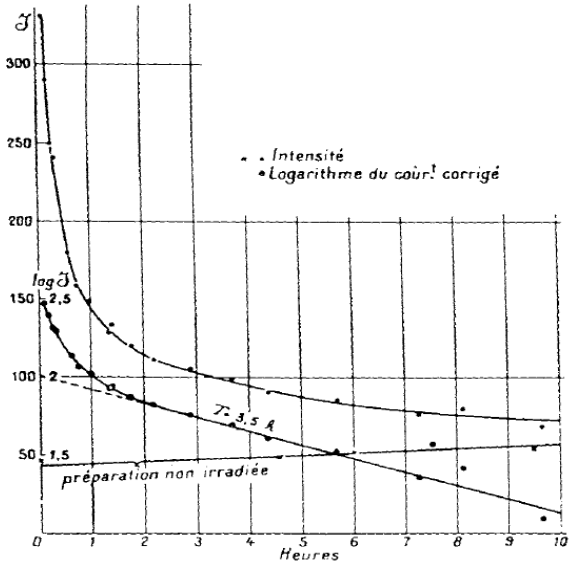


Fig. 8. The 3.5-h activity found in uranium irradiated by neutrons and measured directly as published by *I. Curie* and *P. Savitch* [86] in Paris, July 1938. Upper curve: activity measured with the sample covered by a copper foil of 0.48 g · cm^{-2} thickness, depicted linearly as a function of time; lower curve: build-up of natural decay products in uranium, also depicted linearly; middle curve: difference between the two measurements in semilogarithmic representation with 16-min "eka-rhenium" at the beginning followed by a pure decay with 3.5 h half-life.

Sl. 1. Slika iz originalnog rada Irene Cuiri i Pavla Savića. Rad je publikovan u Julu 1938. Gronja kriva: izmerena aktivnost uzorka pokrivenog sa bakarnom folijom debljine 0.48 g·cm^2 u linearnoj zavisnosti od vremena; donja kriva: nagomilavanje produkata prirodnog raspada urana, takodje prikazano linearno; srednja kriva: razlika izmedju dva merenja u semilogaritamskoj prezentaciji sa 16 min „eka-renijuma" na početku koji prati čist beta raspad sa 3.5 h poluraspada.

Produkt od 3.5 h se nikako nije mogao uklopiti u šeme transformacija koje su razvili Hahn i Strasman.

Cuiri i Savic su čak razmatrali mogućnost stvaranja lakših elemenata posle ozračivanja urana neutronima. Irena je jednom prilikom izjavila "*čini mi se da imamo ceo periodni sistem u ozračenom uranu*".

Hahn i Strassman su ubrzo publikovali rad tvrdeći da se posle ozračivanja urana neutronima dobija 16 različitih elemenata sa atomskim brojevima od 88 (Ra) do 90 (Th) i od 92 (U) do 96. Opet se vidi jak uticaj Fermijevih preporuka.

IV. FISIJA KONAČNO

Irena i Pavle su sa sigurnočću tvrdili da se formiraju transurani, kao i izotopi lakši od urana, ali bliski uranu po atomskom broju. Cuiri i Savić su hemijskom procedurom dobili produkt koji prati lantan (Z=57). Hahn i Strassmann (iako se nisu slagali sa rezultatima Cuiri i Savića) su ponovili postupak hemijske separacije, ali su umesto lantana koristili barijum. Barijum se u periodnom sistemu nalazi u istoj grupi kao i radijum, Ra. Dobili su produkt koji prati barijum (Z=56). Pretpostavili su (opet pogrešno) da su to izotopi radijuma Ra.

Ni jedni ni drugi nisu novodobijene rezultate prepoznali kao lantan ili barijum.

Do ove tačke, obe grupe su bile praktično na istom nivou u pogledu eksperimentalnog rada, ali i nerazumevanja stvarnih procesa, koji se dogadjaju ozračivanjem urana. Činjenica je da su Irena i Pavle prvi našli elemente iz sredine periodnog sistema, ali sami nisu verovali u svoj rezultat. Bolje rečeno, nisu bili u stanju da objasne dobijene rezultate, kao i grupa u Berlinu. Hahn i Strasman, kao i Irena i Pavle i dalje su sumnjali u rezultate dobijene dobrim eksperimentalnim radom. Svega mesec dana kasnije Hahn i Strassmann su bili sasvim sigurni da se barijumovi izotopi stvaraju nakon ozračivanja urana i torijuma. Pored toga nadjeni su i drugi fisioni produkti Sr, Y, Kr ili Xe, koji budući da su beta minus radioaktivni stvaraju Cs i Rb.

Grupa iz Pariza je na ovom mestu zaostala u "fotofinišu trke ka otkrićem fisije".

U Hahnovom pismu Lisi Meitner iz 19.12.1938. Hahn je pisao „*naš izotop radijuma se ne ponaša uopšte kao radijum već kao barijum. Možda ti (Lisa) možeš da daš neko fantastično objašnjenje. Mi shvatamo da uran ne može da sagori i pretvori se u barijum"*.

Tri dana kasnije, **22.12.1938**. Hahn i Strasman podnose rad koji je publikovan 1939. u Januaru. Hahn je insistirao na brzom publikovanju, a s obzirom da je bio vrlo uticajan, rad je izašao u roku od samo mesec dana [6].

U ovom radu Hahn i Strassmann su napisali

"*As chemists we really ought to revise the decay scheme given above and insert the symbols, Ba, La, Ce, in place of Ra, Ac, Th. However, as ''nuclear chemists'' working very close to the field of physics, we cannot bring ourselves yet to take such a drastic step which goes against all previous experience in nuclear physics. There could perhaps be a series of unusual coincidences which has given us false indications.*"

"Kao hemičari, zaista treba da preuredimo naše šeme raspada i da pišemo simbole Ba, La, Ce umesto Ra, Ac i Th. Medjutim, kao nuklearni hemičar, radeći vrlo blizu fizici, ne možemo da preduzmemo tako drastični korak, koji se protivi celokupnom prethodnom iskustvu u nuklearnoj fizici. Moguće je da je neki niz neobičnih slučajnosti doveo do pogrešne indikacije. "

Ovaj rad se smatra otkrićem fisije za koje su Hahn i Strassmann nagradjeni Nobelovom nagradom za hemiju 1944. (Drugi svetski rat je još uvek trajao u svoj svojoj žestini i pored imena Hahn i Strassmann, stoji Treći Rajh).

Lisa Meitner nije bila koautor na tom radu. Nije ni do sada sasvim jasno da li je ova sigurnost u stvaranju lakih elemenata njihova originalna ideja, ili je proistekla iz diskusije sa Lisom Meitner i poredjenjem sa rezultatima Irene Cuiri i Pavla Savića. Radovi Cuiri- Savić su citirani u tom radu, čime je priznat njihov odlučujući doprinos.

V. OBJAŠNJENJE EKSPERIMENTALNIH REZULTATA I SAME FISIJE

Hahn se više puta obraćao Lisi Meitner za pomoć u objašnjenju dobijenih eksperimentalnih rezultata. U tom periodu Lisa je bila u Štokholmu gde se našla sa svojim

rodjakom Otto Frishom (takodje fizičar, sestrić). Otto Friš je radio u Institutu za teorijsku fiziku Kopenhagenu u Danskoj.

U diskusiji Friš i Lisa su došli do zaključka da se radi o cepanju jezgra urana. Diskusija je vodjena katoličkog Božića 25.12.1938. u toku njihove šetnje po snegu u Švedskoj

Otto Friš se setio modela jezgra koji se popularno zove model tečne kapi- predložen od strane Gamova i Bohra. Jezgro je po nekim svojim osobinama slično tečnoj kapi i ovaj model može da objasni neke osobine jezgra. Ako kap tečnosti (npr, voda) perturbiramo, može doći do njenog raspadanja na više manjih kapljica. Ova analogija je poslužila Otto Frišu i Lisi Meitner za objašnjenje mnogobrojnih eksperimentalnih rezultata. Ispravno su pretpostavili da nakon apsorpcije neutrona od strane jezgra urana, jezgro postaje nestabilno, deformiše se i formiraju se dva entiteta. Konačno odbojna Kulonova sila nadjača privlačnu nuklearnu silu (koja drži nukleone na okupu u jezgru) i jezgro se podeli u dva lakša. Velika energija od oko 200 MeV, koja potiče iz različitih energija veze urana i fisionih produkata prema $E=mc^2$, se oslobadja kroz kinetičku energiju fisionih produkata. Za objašnjenje fisije nije potrebna kvantna mehanika, odnosno tunel efekat kojim se objašnjava alfa raspad. Takodje su shvatili da ako je jedan produkt Ba(Z=56) drugi mora biti Kr(Z=36) , 36+56=92 tj U(Z=92). Objašnjenje je kasnije poboljšano činjenicom da se pri zahvatu neutrona jezgrom ^{235}U oslobodi energija vezivanja koja je za oko 0.5 MeV veća od visine Coulumbove barijere, te zato nije potrebno tunelovanje.

Lisa Meitner u koautorstvu sa Otto Frischom je u radu podnetom u **16.01.1939**. upotrebila reč „**fisija**“. Rad je publikovan u časopisu Nature [7]

Abstrakt rada je sledeći

*It seems therefore possible that the uranium nucleus has only small stability of form, and may, after neutron capture, divide itself into two nuclei of roughly equal size (the precise ratio of sizes depending on finer structural features and perhaps partly on chance). These two nuclei will repel each other and should gain a total kinetic energy of c. 200 MeV, calculated from nuclear radius and charge. This amount of energy may actually be expected to be avaiable from the difference in packing fraction between uranium and the elements in the middle of the periodic system. The whole '**fission**' process can thus be described in an essentially classical way, without having to consider quantum-mechanical 'tunnel effects', which would actually be extremely small, on account of the large masses involved.*

Lise Meitner
Physical Institute,
Academy of Sciences,
Jan. 16 Stockholm.

O.R. Frisch
Institute of Theoretical Physics, Jan 16.
University, Copenhagen.

Slobodan prevod ovog, genijalno napisanog apstrakta je: *Čini se da jezgro urana ima neku formu male stabilnosti i može se posle zahvata neutrona podeliti u dva jezgra približno iste veličine (tačan odnos veličina zavisi od fine strukture i možda delimično od šanse). Ova dva jezgra se odbijaju i dobijaju kinetičku energiju oko 200 MeV, izračunato iz nuklearnog radijusa i naelektrisanja. Ovaj iznos energije se može očekivati iz razlike energije veze po nukleonu izmedju urana i elemenata iz sredine periodnog sistema. Ceo proces fisije se tako može opisati na klasičan način, bez razmatranja kvantno-mehaničkog tunel efekta, koji bi trebalo da je ekstremno mali s obzirom na velike mase koje učestvuju u procesu.*

U medjuvremenu Hahn i Strassman podnose još jedan rad u kome se u naslovu pominje reč fisija [8]. Ovaj rad je publikovan pre rada Maitner-Frish na zahtev Otta Hahna za brzim publikovanjem.

VI. ZAKLJUČAK

Put ka shvatanju da je cepanje jezgra urana moguće je trajao skoro pet godina. U tom periodu dve grupe, u Parizu i Berlinu su slepo verovale preporukama, koje je dao Fermi i koje su bile pogrešne. Irena Cuiri i Pavle Savić su prvi došli do saznanja da se u ozračenom uranu nalazi *lantan*, element iz sredine periodnog sistema. Hahn i Strassmann su ponovili njihov rad, ali sa *barijumom* kao nosiocem, i našli su veći broj elemenata iz sredine periodnog sistema. Krajnje objašnjenje procesa fisije dali su Lisa Meitner i Otto Frish.

Nobelovu Nagradu za hemiju dobili su Hahn i Strassman. Lisa Meitner nije dobila Nobelovu Nagradu, ali je transuranski element sa atomskim brojem 109 imenovan po njoj. Otto Frish nije dobio ni jedno priznanje sličnog značaja. Najveći gubitnici su Irena Cuiri i Pavle Savić, koji su bili na korak od otkrića fisije. Nedodeljivanje bar dela Nobelove Nagrade ovoj grupi se može tumačiti i time što je Irena Cuiri već dobila jednu Nobelovu Nagradu za otkriće veštačke radioaktivnosti.

Pored ovih dvaju grupa, postojale su i druge istraživačke grupe koje su ispustile ovo i još neka značajna otkrića u ovoj oblasti fizike.

Posle otkrića fisije, usledila je čitava lavina radova o karakteristikama fisije, od kojih su neki bili veoma značajni. Krajnji rezultati otkrića fisije su: konstrukcija nuklearnih reaktora, nuklearna energetika, proizvodnja nuklearnog oružja, otkriće velikog broja beta minus radioaktivnih izotopa koji se veoma mnogo koriste u medicini (Tc, Lu, Y, Mo i dr) i industriji.

ZAHVALNICA

Ministarstvu Prosvete Nauke i Tehnološkog razvoja koje je delimično finansirala ovaj rad kroz projekat 171021.

[1] Jack E. Fergusson, The history of the discovery of nuclear fission. Foundation of Chemistry, 13:145–166 , 2011.

[2] Gunter Herrmann, Five Decades Ago: From the "Transuranics" to Nuclear Fission, Angew. Chem. Int. Ed. Engl. 29, 481-508, 1990.

[3] O. R. Frisch: *What Little I Remember*, Cambridge University Press, Cambridge 1979, p. 118.

[4] Curie, I., Savitch, P. Concerning the nature of the radioactive element with 3.5-hour half-life, formed from uranium irradiated by neutrons. Comptes Rendus Acad. Sci. Paris 206, 1643–1644 (1938a)

[5] Curie, I., Savitch, P.: On the radioelements formed in uranium irradiated with neutrons—Part II. Journal de Physique et le Radium 9(7), 355–359 (1938b)

[6] O. Hahn, F. Strassmann, "Uher den Nachweis und das Verhalten der hei der Bestrahlung des Urans mittels Neutronen entstehenden Erdalkalimetalle," Naturwissenschaften 27 (1939a) 11 - 15 (December 22, 1938).

[7] Meitner, L., Frisch, O.R.: Disintegration of uranium by neutrons: a new type of nuclear reaction. Nature 143, 239–240 (1939).

[8] Hahn, O., Strassmann, F. Proof of the formation of active isotopes of barium from uranium and thorium irradiated with neutrons; proof of the existence of more active fragments produced by uranium fission. Naturwissenschaften 27(6), 89–95 (1939b).

ABSTRACT

Discovery of fission and role of Pavle Savić is described in this paper.

Discovery of fission and role of Pavle Savić

Dragoslav Nikezić

Preliminarni pregled početaka jugoslovenskog nuklearnog programa

Maja Korolija, *Istraživač-pripravnik, Institut za multidisciplinarna istraživanja, Beograd*

***Apstrakt*—Namera je da se u radu ukratko prikaže početak jugoslovenskog nuklearnog programa. U radu se ispituju kako specifičnosti konteksta Hladnog rata, tako i uloga i primena nauke, prvenstveno nuklearne fizike, u tom periodu. U radu se analizira geopolitički položaj FNRJ nakon prekida saradnje sa SSSR-om 1948. godine i fokus jugoslovenskog političkog vrha na razvoj nuklearnog programa, uz prikaz osnovnih podataka o institutima u kojima se on sprovodio. U ovom kontekstu prikazuje se dinamika odnosa između predstavnika vlasti i naučnika, i razmatra se značaj tih odnosa za jugoslovenski nuklearni program.**

Ključne reči — FNRJ; Hladni rat; Nuklearni program

Iz srca atoma rodiće se besklasno društvo.
Marko Ristić

NAUKA U KONTEKSTU HLADNOG RATA

Tek što se Drugi svetski rat završio, a konstantno odmeravanje snaga između dva nekadašnja saveznika u borbi protiv sila Osovine - SAD-a i SSSR-a - već je uzelo maha. Konfrontacija ove dve svetske sile dovela je do polarizacije država u svetu na dva sukobljena tabora. Ova situacija koja je potpuno dominirala svetskom političkom scenom druge polovine dvadesetog veka predstavljala je novu vrstu rata poznatog kao – Hladni rat (Hobsbaum 2002).

> „Kako je vreme prolazilo, bilo je sve više stvari koje su mogle da pođu pogrešno, i političkih i tehnoloških, u stalnoj nuklearnoj konfrontaciji koja se zasnivala na pretpostavci da jedino strah od „uzajamno zajamčenog uništenja“ (što je pravilno zgusnuto u akronim MAD (igra reči, na engleskom znači „lud“ prim. prev.) može sprečiti jednu ili drugu stranu da ne da stalno spreman signal za planirano samoubistvo civilizacije“ [Hobsbaum 2002: 174].

Prvi put u istoriji odvijao se sukob dva vojna bloka, naoružana nuklearnim raketama i taktičkim oružjem koje je u stanju da uništi čitavu ljudsku civilizaciju. Taj sukob, iako je dovodio situaciju do usijanja duže od četiri decenije, ipak se završio mirovnim sporazumom - onim potpisanim na samitu održanom u Parizu od 19. do 21. novembra 1990. godine[1]. Do tada, nikada nije postojala trka u naoružanju poput ove, a ipak to nije dovelo do izbijanja „klasičnog“ rata (Subrahmanyam 2008).

I SSSR i SAD su sve aspekte naučnog rada, nakon pobede u Drugom svetskom ratu, stavile u službu svoje strane u Hladnom rata (vidi Kojevnikov 2004). Proizvodnja atomske bombe je značajno doprinela uočavanju relevantnosti nauke u političkoj borbi. Takav društveni kontekst doveo je do toga da naučne aktivnosti širom sveta počnu drugačije da se posmatraju: u svetlu politike (Krementsov 1997). Percepcija uloge nauke - i u nemarksistčkim sistemima - približila se perspektivi Karla Marksa (Karl Marx), koji ju je posmatrao kao „istorijski dinamičnu i revolucionarnu silu u društvu“ (Marks prema Bernal 1952: 47). U oba bloka nauku su odlikovale sledeće zajedničke karakteristike: gigantomanija, državna podrška, kult nauke u društvu, (kon)fuzija između nauke i inženjerstva, multidisciplinarna istraživanja, kolektivni rad, složena birokratija i militarizacija (vidi Graham 1992 i vidi Kragh 1999). Tehnologije kakve su radar, nuklearna energija, mlazni motor, dalekometne balističke rakete, blizinski osigurač itd. zahtevaju koordiniranu aktivnost naučnika, inženjera, vojske, industrije kao i političku posvećenost stvaranju radnih sistema (Chertok 2010).

Sledstveno tome, neposredno nakon Drugog svetskog rata atomska bomba nalazila se u fokusu spoljene i unutrašnje politike SAD-a. „Atomska bomba je istovremeno simbolizovala i moć nauke i uništenje. I mada su neki naučnici, radeći unutar nacionalne državne bezbednosti, osećali da su sputani, institucionalna moć nauke je cvetala“ (Wolf 2013: 21).[23] Imajući u vidu izuzetan naučni, industrijski i tehnički uspeh na projektu

[1] https://www.osce.org/mc/39516

Menhetn, tamošnji kreatori politike su sumnjali da će bilo koja država uspeti da bude konkurencija „čudesima i užasima“ američke nauke. Američka vojska i političari, 1949. godine suočeni sa činjenicom da i SSSR ima atomsku bombu, ponovo su došli do zaključka da u cilju zaštite treba da se okrenu visokotehnološkom naoružanju, a ne diplomatiji. Nakon što je američki predsednik Truman objavio da SAD tragaju za termonuklearnim oružjem trka u naoružanju između ova dve države zvanično je počela (Wolf 2013). Nijedna evropska država, čak ni Francuska ni Engleska, nisu mogle da se porede sa SAD i SSSR kada je u pitanju uloga vojske u razvoju i primeni fizike (Kragh 1999).

FNRJ U KONTEKSTU HLADNOG RATA I POČECI JUGOSLOVENSKOG NUKLEARNOG PROGRAMA

Istoričar Hobsbaum (2002) posmatra period između 1947. i 1951. godine kao najburniji tokom Hladnog rata. U ovom razdoblju SAD je vodila agresivnu antikomunističku borbu. Sa druge strane SSSR je bio suočen sa pojavom prvih pukotina u svom bloku. Naime, 1948. godine, izbila je međunarodna kriza (Žanin Čalić 2013), tokom koje je Federativna narodna republika Jugoslavija (FNRJ) pod vođstvom Josipa Broza Tita napustila blok predvođen Sovjetskim savezom (Hobsbaum 2002). U Jugoistočnoj Evropi Jugoslavija se isticala kao samostalni centar komunističke moći što nije odgovaralo antikomunistički nastrojenom britanskom premijeru Čerčilu, ali ni komunističkom vođi Sovjetskog saveza Staljinu. Ovo drugo je za rezultat imalo da 28. juna. 1948. godine KPJ bude isključena iz Kominforma pod sumnjom da uvodi kapitalizam (Žanin Čalić 2013). U rezoluciji Informbiroa Jugoslavija se, između ostalog, optužuje i za nacionalističke tendencije[4] (Rezolucija Informacionog biroa komunističkih partija o stanju u Komunističkoj partiji Jugoslavije, 28. juna 1948).

U tom smislu zanimljivo je pomenuti razgovor između našeg istaknutog fizičara Pavla Savića, koji je radio u Institutu za fizičke probleme Sovjetske akademije nauka, čiji je direktor bio poznati naučnik Pjotr Kapica (Пётр Леони́дович Капи́ца), i Josipa Broza Tita, za vreme Titove diplomatske posete SSSR-u u proleće 1946. godine. Prilikom tog susreta Tito je rekao Saviću „Vrati se u zemlju, gradićemo i mi institut“ (Savić, 1978: 306). U jesen iste godine Savić se vratio u Jugoslaviju gde je po partijskoj dužnosti prihvatio da bude prorektor Beogradskog univerziteta. „Kada je došla 1948. godina nije bilo više mogućnosti za vraćanje u Sovjetski savez“ (Savić 1978: 306)[5].

Te 1948. godine uredbom Vlade FNRJ osnovana je Uprava za koordinaciju rada naučnih instituta sa ciljem (koji iz naziva nije lako dokučiv) da pokrene aktivnosti u vezi sa realizacijom najambicioznijeg programa u jugoslovenskoj nauci: razvoj nuklearne tehnologije. Za prvog direktora imenovan je Batrić Jovanović, dotadašnji pomoćnik saveznog ministra za obojenu metalurgiju (Spasić 2013). U oktobru 1952. godine Uprava se seli na novu adresu, a na osnovu nove uredbe menja svoj naziv u Uprava za rudarska istraživanja i rudarske studije. Odlukom Saveznog izvršnog veća (Službeni list FNRJ 18/1953) iz 1953. godine Uprava menja naziv u Zavod za geološko-rudarska i tehnološka istraživanja. Za direktora zavoda imenovan je dpl. hem. Miladin Radulović-Krcun, kasnije direktor Direkcije za nuklearne sirovine pri Saveznoj komisiji za nuklearnu energiju (Spasić 2013). Prolazeći kroz još nekoliko promena imena i formata, konačno januara 1966. godine Institut za tehnologiju nuklearnih i drugih mineralnih sirovina (ITNMS) dobija naziv koji nosi i danas.

1948. godine i Pavle Savić je po odluci Savezne vlade Jugoslavije pristupio izgradnji Instituta za fiziku u Vinči koji je 1950. godine dobio ime „Institut za ispitivanje

[2] Pre rata godišnje ulaganje u nauku od strane (američke) države je iznosilo oko pedeset miliona dolara (novac je pre svega bio namenjen za zdravstvo i poljoprivredu), no do 1950. godine država je za nauku na godišnjem nivou izdvajala više od jedne milijarde dolara (Wolf 2013).

[3] U SAD-u fizika je toliko bila povezana sa vojskom da su neki fizičari bili su strahu da fizika ne postane grana vojske. Filip Morison (Philip Morrison), fizičar angažovan na Menhetn projektu, već 1946. godine izražava zabrinutost zbog militarizacije fizike, ali i razumevanje za kompleksan položaj fizičara (Za više detalja videti Kragh 1999: 297).

[4] Uloga koja se daje državi, kao bitnom elementu marksističko-lenjinističke strategije, makar ona bila i „proleterska“ i nominalno internacionalistički orijentisana, predstavlja važnu metu mnogih kritika koje su sa leva upućene ovoj ideologiji. Segment tih kritika odnosi se i na to da je svaka državna forma, bez obzira na proklamovanu ideologiju, inherentno nacionalistička (Rocker 1997).

[5] Prema Saviću (1978) Kapica je verovao da će SSSR pomoći u poduhvatu pravljenja Instituta u Jugoslaviji. Savić je čak sastavio spisak materijala i aparature koje SSSR treba da da Jugoslaviji, no situacija nije ispalo kako su dvojica fizičara očekivala. Kapica je kasnije po Staljinovoj naredbi poslat u devetogodišnju internaciju (Savić 1978).

strukture materije“, a 1953. godine „Institut Boris Kidrič“[6] (Bondžić 2016). Danas je Institut za nuklearne nauke „Vinča“ institut od nacionalnog značaja za Republiku Srbiju. U izgradnji instituta Saviću je, pored državnih organa, pomagao njegov prijatelj, nuklearni fizičar iz Francuske, Robert Valen, a među prvim saradnicima Savić pominje i profesora Aleksandra Milojevića, koji je donosio materijal i aparaturu za prvu laboratoriju iz Nemačke (Savić 1978). U nabavci prvih instrumenata i opreme za laboratorije pomogao je i profesor Dragoljub Jovanović (Bondžić 2016).

> „Prva mašina za nuklearne reakcije bio je kaskadni akcelerator od milion i po elektronvolti, kupljen kod Erlikena u Švajcarskoj. Prvi kadar regrutovao sam među svojim đacima, kojih je bilo preko hiljadu: fizičara, fizikohemičara, hemičara. Svi su oni slušali atomistiku kod mene na Univerzitetu....Pokrenuli smo Bilten Instituta u Vinči koji sam uređivao....Preko toga Biltena su nas veoma brzo upoznali naučni krugovi izvan naše zemlje, tako da je naša ekipa na prvoj i drugoj međunarodnoj konferenciji za mirnodopsku primenu nuklearne energije u Ženevi nailazila na prijem sa velikim respektom“ [Savić 1978: 307].

Zahvaljujući Savićevom neumornom radu naučna zajednica je prvi put imala naučnu instituciju sa akceleratorima, reaktorima, anlizatorima jona, masenim spektometrima, plazmenim topovima kao i znanje o savremenoj atomistici, biologiji, elektronici, zaštiti od zračenja itd. (Senćanski 1986). „U „maloj“ Vinči, koja je iz dana u dan postajala sve veća, vrilo je kao u košnici, radilo se i danju i noću. Svetla se nisu gasila u biblioteci, u kojoj ni noću nije bilo slobodnih mesta za učenje. Nicale su nove zgrade, gradile se mašine...I nad svim tim poslovima, počev od nabavke hrane pa sve do pronalaženja neophodne, naučne opreme u ratom razorenoj Evropi, bdeo je profesor Savić“ (Senćanski 1986: 58).

Prema Bondžiću (2016), jugoslovenski politički vrh je u tri navrata od naučnika tražio da počnu da se bave razvojem nuklearnog oružja. Istraživanja pokazuju da je Saviću krajem 1950. godine obelodanjeno da partijski vrh insistira na pravljenju atomske bombe (do tada to nigde nije bilo eksplicirano). S obzirom na situaciju u kojoj se Jugoslavija našla posle raskida sa SSSR-om – u strahu od napada od strane SSSR-a (koji 1949. godine i sam dobija atomsku bombu) ali i ostalih država Informbiroa i nesigurna da li će pomoć od strane zapadnih kapitalističkih država uslediti - jasno je zašto je državno rukovodstvo Jugoslavije izlaz iz ove loše vojno-bezbednosne situacije videlo u projektu pravljenja nuklearnog oružja. Istaknuti partijski funkcioner Milovan Đilas tu logiku formuliše na sledeći način: „Ja sam za reč Lenjina: Među vukovima ja urličem. Dok smo okruženi vucima treba se braniti i imati najmoćnija oružja“ (Đilas prema Bondžić 2016: 105).

Pored toga jasno je da bi atomska bomba igrala veliku ulogu za održavanje moći trenutnog državnog rukovodstva Jugoslavije, uz nesumnjiv međunarodni prestiž koji posedovanje atomske bombe donosi. Pavle Savić je suptilno izbegavao otvoreno suprotstavljanje državnim rukovodiocima, nekad i hraneći njihove ambicije u vezi sa atomskom bombom, što je verovatno i dovelo do toga da njegovi stavovi ponekad budu protivrečni, ali sve ukazuje na to da on nikada nije ozbiljno prionuo na zadatak pravljenja oružja, iako je bio svestan da je i sam Institut „Boris Kidrič“ osnovan u te svrhe (Bondžić 2016).

Aleksandar Ranković, Edvard Kardelj i Milovan Đilas su poslali Slobodana Nakićenovića i Stevana Dedijera da kontrolišu Pavla Savića i pomognu (naročito Dedijer) oko pravljenja atomske bombe. Istraživački izvori ukazuju i na razilaženje u pogledima partijskog rukovodstva, koje je želelo bombu, i Savićevog viđenja uloge Instituta (i reaktora), prvenstveno u obuci kadrova, kako bismo u fundamentalnim, nuklearnim naukama bili ravnopravni sa ostatkom razvijenog sveta. Indikacije za ovakav Savićev stav možemo naći i u Izveštaju Uprave za koordinaciju rada naučnih instituta za 1948. godinu gde Savić naglašava „da je Institut u Vinči samo školska ustanova za izgradnju kadra i da se tu ne namerava izgrađivati nikakav krupniji objekat iz oblasti atomske energije izuzevši čisto školske objekte“ (AJ, 836, KMJ-II-6-a/4 prema Bondžić 2016: 106).

Ubrzo je i Dedijer počeo da zastupa shvatanja slična Savićevim. Nakon Staljinove smrti, krajem maja 1953. godine, u ime Instituta nuklearnih nauka Pavle Savić, Stevan Dedijer i Robert Valen sastavljaju *dopis* „O dva bitna uslova za razvitak atomske energije kod nas“ koji dostavljaju „Kabinetu Maršala i drugovima Kardelju,

[6] Institut je ime dobio po državnom i partijskom rukovodiocu Borisu Kidriču, koji je preminuo te iste godine, predsedniku privrednog saveta Jugoslavije koji je imao važnu ulogu i pružao punu podršku pri osnivanju i razvoju Instituta (Bondžić 2016).

Rankoviću i Vukmanoviću“. Na samom početku autori potvrđuju da su aktivnosti u vezi sa atomskom energijom kod nas započete sa ciljem proizvodnje atomskog oružja, kao i za korišćenje u privredne svrhe. U nastavku ukazuju na problem malih količina urana u zemlji, nedovoljnu snagu privrede, kritikuju slabu upućenost političkih i privrednih rukovodioca u pitanja razvoja nuklearne energije, kao i konspirativnost nametnutu stručnjacima, za koju ocenjuju da sprečava dobijanje neophodnih informacija o realnim mogućnostima zemlje.

> „Orijentišemo se na izgradnju jednog reaktora koji treba da bude instrument za naučna i tehnička istraživanja neophodna za razvoj atomske energije uopšte, ukoliko u zemlji postoje uslovi za ovo. Ukoliko ovih uslova nema onda građenje reaktora ne dolazi u obzir i mi se orijentišemo na pripremu minimalnog kadra sposobnog za eksploataciju uređaja za atomsku energiju, ukoliko u budućnosti međunarodna razmena omogući da dođemo do takvih uređaja. U slučaju realizacije reaktora naš Institut bi davao naučnu koncepciju i osnovne tehničke parametre, a naše rudarstvo, industrija i spoljna trgovina imali bi da reše svakako oko 95% svih materijalnih potreba“ [dopis prema Bondžić 2016: 108].

Prvi put se desilo da su se naučnici kritički osvrnuli na nuklearni program i upozorili rukovodstvo na ograničenja tog programa. Takođe, pored toga što su ukazali na štetnost „konspiracije“ naučnici su i zahtevali njeno ublažavanje tvrdeći da čak ni SAD ne skriva organizaciju metoda prospekcije kao ni nalazišta urana. Savić, Valen i Dedijer navode primer za ovo. Jaka konspiracija je bila jedan od razloga koji je sprečavao da se utvrdi koliko procenata urana se nalazi u nekoj rudi i time reši naučni problem koji je trajao tri godine. Na tom problemu se se radilo na više strana, ali zbog konspiracije to su činili ljudi izolovani jedni od drugih. Kada je konačno omogućena razmena iskustava problem je rešen u roku od mesec dana.

> „Na osnovu ovoga predlažemo da se za sada bar za ograničeni broj naučnika iz našeg Instituta, izvrši totalna dekonspiracija o organizaciji, metodama, rezultatima i investicijama na području prospekcije urana, u cilju poboljšanja brzine i kvaliteta ovog rada, koji je condito sine qua non za dalji rad na području atomske energije kod nas“ [*dopis* prema Bondžić 2016: 110]

Kasnije se ovakva praksa ponavljala. Prema Bondžiću (2016) ne postoji direktan odgovor rukovodilaca na *dopis* naučnika, ali može se reći da su dela sledećih godina sasvim dovoljna da bismo uvideli efekat. Nakon što je Đilas pao sa vlasti 1954. godine Stevan Dedijer, je prešao da radi u Institutu „Ruđer Bošković“ u Zagrebu. Robert Valen se vratio u Francusku. Na čelo Instituta, bez naučnih kvalifikacija, postavljen je diplomata i ekonomista Vojko Pavičić. Kardelj, Ranković i Vukmanović su bili zaduženi da oblikuju nuklearnu politiku zemlje, dok je na Institutu u Vinči od autora *dopisa* ostao samo Pavle Savić (Bondžić 2016). Naučnici iz Vinče su se usavršavali na istraživačkom reaktoru u Kjeleru u Norveškoj. Među prvim našim naučnicima sa zapaženijim rezultatima iz oblasti reaktorske fizike treba pomenuti profesora Elektrotehničkog fakulteta u Beogradu Dragoslava Popovića, čije je objavljivanje rada o preseku fisije uranovog izotopa bio senzacija, jer je u pitanju bila do tada nikada javno objavljena informacija važna za proizvodnju nuklearnog oružja (Hymans 2012).

Komisija za pomoć u naučnim istraživanjima održala je sastanak u junu 1954. godine kome su prisustvovali Pavle Savić, Svetozar Vukmanović Tempo, Anton Peterlin, Ivan Supek, Slobodan Nakićenović, Vojko Pavičić, Miladin Radulović, Marko Čančarević, i Milan Osredkar. Na sastanku je vođena i diskusija o izveštajima o radu instituta. Ukratko - navedeno je da je uočen napredak u radu i prvenstveno je stavljen fokus na značaj nuklearne energije i pripreme za njeno korišćenje. Takođe, uočena je i potreba za osamostaljivanjem instituta u budućnosti, kao i potreba za orijentacijom na domaće izvore u ekonomskom smislu. Na sastanku je potvrđeno i da će Komisija finansirati samo radove iz oblasti nuklearnih nauka. Drugim rečima „napori na nuklearnim istraživanjima pod državnim kontrolom i usmerenjem“ su se nastavila. Na sastanku je pomento i kako se Institut u Ljubljani usmerava ka nuklearnim naukama, kao i da Institut u Zagrebu treba da se u većoj meri koncentriše na problematiku iz oblasti nuklearnih nauka. U kasnijem periodu, od 1955. godine, ciljevi su bili usmereni na polje „privredne i mirnodopske primene nuklearne energije“, ali i na prikriveno polje „eventualne proizvodnje atomskog oružja“, „pod strogom kontrolom države i UDB-e“ (Bondžić 2016: 114). Savezno izvršno veće SFRJ formira 1955. godine Saveznu komisiju za nuklearnu energiju (SKNE) kojoj je u zadatak stavljeno

da kao savezni organ „programira, rukovodi i koordinira sve aktivnosti u oblasti nuklearnih nauka i nuklearnih tehnologija kao i u oblasti zaštite od jonizujućeg zračenja“ (Perović-Nešković 2000: 32). Prvi predsednik Komisije bio je državni sekretar za unutrašnju politiku Aleksandar Ranković (Bondžić 2016). Pre osnivanja SKNE država je direktno finansirala institute iz budžetskih sredstava, a sada je to rađeno posredstvom SKNE. Među prvim zadacima SKNE bio je i izbor i priprema za konstruisanje, izgradnju i puštanje u pogon nuklearnog reaktora. Smrću Staljina 1953. godine spor između Jugoslavije i SSSR-a kreće u pravcu razrešenja. 1955. godine Jugoslavija kupuje od SSSR teškovodni reaktor (RA), koji kreće sa radom 1959. godine. Nekoliko meseci pre toga u rad je pušten „nulti“ reaktor (RB), „i tako je ostvarena prva lančana reakcija fisije na Balkanu“ (Ristić 2000: 28). Može se reći da od 1955. godine počinje period procvata nuklearne energije kod nas.

Prema Nakićenoviću (1961) za Jugoslaviju, kao socijalističku zemlju, nauka, tehnologija i obrazovanje bili su veoma visoko pozicionirani na listi prioriteta. Samim tim je i očekivana silina kojom je državna politika usmerila čitavo društvo prema razvoju naučnih i obrazovnih institucija u ratom razorenoj, osiromašenoj, nerazvijenoj državi, do tada prevashodno orijentisanoj na poljoprivrednu proizvodnju. „Osnovna svrha socijalističkog društva ogleda se u nastojanju da se ljudski život učini što bolji i što srećniji“ (Nakićenović 1961: 4). Bez ulaganja u tehnologiju i nauku ovo nije moguće postići, jer nauka i tehnologija predstavljaju sredstva za ostvarenje ovog cilja. Na kraju karajeva ni naučna socijalistička misao ne bi postojala bez „epohalnih naučnih otkrića i progresa u nauci i tehnologiji“ (Nakićenović 1961: 4).

> „Treba uočiti da je Jugoslavija na vreme postala svesna potencijalne uloge i važnosti koja nova postignuća u nauci i tehnologiji mogu igrati u ekonomskom razvoju zemlje. Zato je, ubrzo nakon rata, Jugoslavija inicirala rad u polju nuklearne energije“ [Nakićenović 1961: 9].

ZAKLJUČAK

Ideološke karakteristike sistema, kao noseći uzrok insistiranja na razvoju nuklearne energije, treba posmatrati u kontekstu geopolitičke pozicije Jugoslavije. Ako se prepozna njen položaj u jeku Hladnog rata, nakon raskida sa SSSR-om, jasno je zašto je jugoslovenski državni vrh toliko insistirao na razvoju nuklearnog programa. On je to radio ne samo, kako se u kasnijoj zvaničnoj retorici navodi, u mirnodopske svrhe, već pretežno u vojne svrhe. Imajući u vidu period Hladnog rata, i činjenicu da su dva bloka raspolagala nuklearnim oružjem, i sve vreme radila na njegovom usavršavanju, jasno je da je Jugoslavija u nuklearnom naoružavanju, pored nacionalne bezbednosti, videla i perspektivu nezavisnosti i mogućnost da prema svom nahođenju kroji sopstvenu politiku.

Čini se stoga da sama državna forma, a ne ideologija, u real-političkim odnosima predstavlja uzrok fokusa sistema na određene segmente naučnog rada. U prilog tome govori i činjenica da su i liberalno-demokratski sistemi težili razvoju nuklearnog programa, ulažući svoje resurse prvenstveno u tom smeru, baš kao i Jugoslavija i SSSR.

Ipak, treba imati u vidu da je u osnovi ubrzanih modernizacijskih procesa, koji su nerazdvojni od razvoja nauke i tehnologije, u tadašnjom jugoslovenskom društvu bila neka od varijacija socijalističkih ideologija. Može se stoga zaključiti da, kada je u pitanju Jugoslavija, prvo marksističko-lenjinističku, a zatim samoupravnu socijalističku ideologiju, treba posmatrati kao idejne pretpostavke koje su omogućile postavljanje sistema na strukturne osnove koje su omogućile pokretanje nuklearnog programa.

ZAHVALNICA

Rad je nastao u okviru projekta Teorija i praksa nauke u društvu: multidisciplinarne, obrazovne i međugeneracijske perspektive (registarski broj 179048) koji finansira Ministarstvo prosvete, nauke i tehnološkog razvoja Republike Srbije.

LITERATURA

Bernal, John D. (1952). *Marx and Science.* New York: International Publishers.

Bondžić, Dragomir (2016). *Između ambicija i iluzija: nuklearna politika Jugoslavije 1945-1990.* Beograd: Institut za savremenu istoriju.

Charter of Paris for a New Europe < https://www.osce.org/mc/39516> Pristupljeno 15. 04. 2019.

Chertok, Boris E. (2010). *Rockets and People: Hot Days of the Cold War (Volume III).*Washington: NASA.

Čalić, Mari-Žanin (2013). *Istorija Jugoslavije u dvadesetom veku.* Beograd: Clio.

Graham, Loren R. „Big science in the last years of the big Soviet Union" in *Science after '40*, in Thackray, Arnold (ed.), Osiris: 1992. pp. 49–71.

Hobsbaum, Erik (2002). *Doba ekstrema: istorija kratkog dvadesetog veka 1914-1991*. Beograd: Dereta.

Hymans, Jacques (2012). *Achieving Nuclear Ambitions: Scientists, Politicians, and*

Proliferation. New York: Cambridge University Press.

Kojevnikov, Alexei B. (2004). *Stalin's Great Science: The Times and Adventures of Soviet Physicists.* London: Imperial College Press.

Krementsov, Nikolai (1997). *Stalinist Science.* : Princeton, New Jersey: Princeton University Press.

Kragh, Helge (1999). *Quantum Generations: A History of Physics in Twentieth Century.* Princeton, New Jersey: Princeton University Press.

Nakićenović, Slobodan (1961). *Nuclear Energy in Yugoslavia.* Beograd: Export Press.

Perović-Nešković, Branislava. „Sužavanje nuklearnog programa i orijentacija na rad za privredu i druge korisnike", u Perović-Nešković, Branislava (prir.). *Pola veka Instituta Vinča:* 2000. str. 32-37.

Rezolucija Informacionog biroa komunističkih partija o stanju u Komunističkoj partiji Jugoslavije: Informacioni biro komunističkih i radničkih partija, 28. juna 1948.

Ristić, Milorad. „Usmeravanje ka reaktorskim tehnologijama i nuklearnoj energetici", u Perović-Nešković, Branislava (prir.). *Pola veka Instituta Vinča*: 2000. str. 26-31.

Rocker, Rudolf (1997). *Nationalism and Culture*. Montreal : Black Rose Books

Savić, Pavle (1978). *Nauka i društvo.* Beograd: Srpska književna zadruga.

Senćanski,Tomislav (1986). *Iz kamena iskra: životni i naučni put Pavla Savića*. Beograd: Vuk Karadžić.

Službeni list Federativne Narodne Republike Jugoslavije. 18/1953. Beograd: Službeni list FNRJ.

Subrahmanyam, Krishnaswamy. „A historical overview of the Cold War", in Chari, Chandra (ed.). *Superpower Rivalry and Conflict: The long shadow of the Cold War on the twenty-first century:* 2010. pp. 15-33.

Spasić, Aleksandar M. (2013). *Šezdeset pet godina sa vama 1948-2013*. Beograd: ITNMS.

Wolf, Audra J. (2013). *Competing with the Soviets: Science, Technology, and the State in Cold War America.* Baltimore: Johns Hopkins University Press.

ABSTRACT

Our intention is to present basic outlines of the beginnings of Yugoslav nuclear program. In the paper we examine specifics of the Cold war context, as well as the role and application of science, notably nuclear physics, in that period. The paper is analyzing geopolitical position of FPRY after break with USSR in 1948. and the focus of the Yugoslav political leadership on the development of the nuclear program, while reviewing basic information on the institutes in which this program was developed. In this context dynamics of the relations between representatives of the government and scientists, and importance of those relations for the Yugoslav nuclear program is depicted.

Keywords - FPRY; Cold War; Nuclear Program

Preliminary review of the beginnings of the Yugoslav nuclear program

Maja Korolija

Uporedna analiza uticaja gama i X zračenja na karakteristike modela gasnog odvodnika prenapona u impulsnom režimu rada

Boris Lončar, Dušan Nikezić, Katarina Karadžić, Luka Rubinjoni i Andrija Janković

***Apstrakt*—Cilj ovog rada je da se ispita eksperimentalno uticaj gama i X zračenja na karakteristike fizičkog modela gasnog odvodnika prenapona za tri materijala elektroda pri tri brzine impulse. Dobijeni rezultati su pokazali da i gama i X zraci dovode do privremenog poboljšanja karakteristika gasnih odvodnika prenapona. Najbolji rezultati se postižu upotrebom elektroda od aluminijuma, pri najbržim impulsima.**

***Ključne reči*—Gasni odvodnik prenapona, histogram, gama zraci, X zraci.**

I. Uvod

Gasni odvodnici prenapona su nelinearni elementi, koji se isključivo koriste za zaštitu od prenapona. Sastoje se od dve ili tri elektrode, koje su zatopljene u keramičko ili stakleno kućište. Oni predstavljaju dvoelektrodnu ili troelektrodnu simetričnu konfiguraciju sa gasnom izolacijom. Kao izolacioni medijum se koristi plemeniti gas (najčešće argon, a može i neon, kripton ili ksenon) ili smeša plemenitih gasova na pritisku od 0,1 kPa do 70 kPa. Troelektrodni gasni odvodnici predstavljaju dva gasna odvodnika u istom balonu, u čijoj unutrašnjosti je plemeniti gas ili smeša plemenitih gasova. Rastojanje između elektroda je reda milimetra ili delova miilimetra [1,2]. Elektrode su tako postavljene da obezbeđuju postojanje pseudohomogenog električnog polja.

II. Pregled dosadašnjih rezultata

Rezultati dosadašnjih ispitivanja pokazuju da su gasni odvodnici prenapona najbolji prenaponski zaštitni elementi u polju radioaktivnog zračenja i stoga se oni mnogo češće koriste u tu svrhu od drugih elemenata za zaštitu od prenapona (prenaponske diode, varistori, kondenzatori) [3,4,5]. Ispitivan je i uticaj gama i X zračenja na karakteristike komercijalnih gasnih odvodnika prenapona [6]. Proučavan je i uticaj materijala elektroda, načina obrade elektrodnih površina, vrste gasa, pritiska i međuelektrodnog rastojanja na pretprobojnu struju komercijalnih odvodnika u polju gama i X zračenja [7,8]. Istraživan je i uticaj materijala elektroda i pritiska na karakteristike modela gasnog odvodnika prenapona u polju gama [9] i X zračenja [10] u jednosmernom režimu rada i uticaj gama i X zračenja na brzinu odziva modela gasnog odvodnika prenapona u impulsnom režimu [11]. Konačno, proučavana je i radijaciona otpornost modela gasnog odvodnika prenapona u polju neutronskog zračenja [12].

Boris Lončar – Tehnološko-metalurški fakultet, Univerzitet u Beogradu, Karnegijeva 4, 11000 Beograd, Srbija (e-mail: bloncar@ tmf.bg.ac.rs).

Dušan Nikezić – Institut za nuklearne nauke Vinča, Univerzitet u Beogradu, Mike Petrovića Alasa bb., 1100 Beograd, Srbija (e-mail: dusan@vinca.rs).

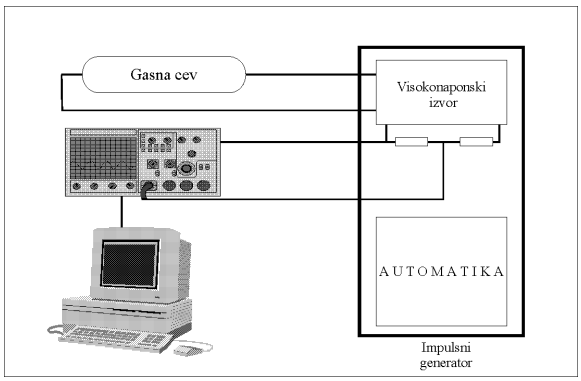
Katarina Karadžić – Tehnološko-metalurški fakultet, Univerzitet u Beogradu, Karnegijeva 4, 11000 Beograd, Srbija (e-mail: kkaradzic@tmf.bg.ac.rs).

Luka Rubinjoni – Inovacioni centar Tehnološko-metalurškog fakulteta, Karnegijeva 4, 11000 Beograd, Srbija (e-mail: rubinjoni@ tmf.bg.ac.rs).

Andrija Janković-Tehnološko-meatlurški fakultet, Univerzitet u Beogradu, Karnegijeva 4, 11000 Beograd, Srbija (e-mail: andrija366@gmail.com

III. Eksperiment

Instrumentacija korišćena u eksperimentalnim ispitivanjima sastojala se od sledećih osnovnih delova:
1) gasno-vakuumska komora; 2) merač pritiska SPEEDIVAC; 3) čelična boca sa argonom pod pritiskom; 4) vakuum pumpa EDWARDS 5; 5) impulsni test generator Haefely tip P6T VF-tel 202671 sa priborom; 6) osciloskop Tektronix TDS 220 SNB036675; 7) izolacioni transformator i 8) koaksijalni kablovi i priključci. Blok šema opreme za eksperimentalna ispitivanja prikazana je na slici 1.

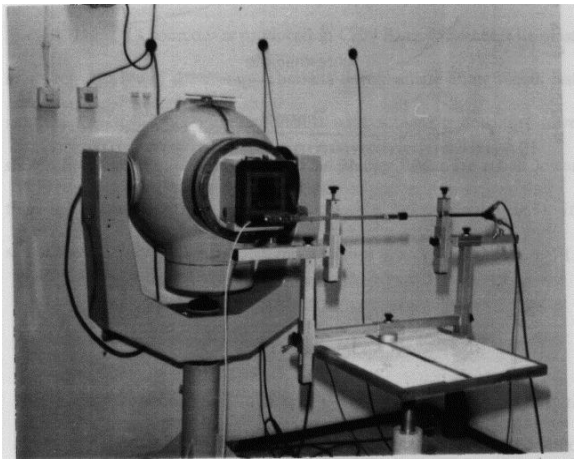


Sl. 1. Blok šema opreme za eksperimentalna ispitivanja

U cilju ispitivanja radijacione otpornosti modela gasnog odvodnika prenapona u radu su korišćeni sledeći izvori zračenja: 1) izvor gama zraka kobalt-60 u polju uređaja IRPIK-B; 2) izvor X zraka u polju dozimetrijskog generatora PHILIPS MG-320.

Izvor gama zraka kobalt-60 nalazi se u uređaju gama polja IRPIK-B koji je prikazan na slici 2 i nalazi se u Metrološko-dozimetrijskoj laboratoriji Instituta za nuklearne nauke Vinča. Ovaj uređaj se sastoji od kontejnera za smeštaj i čuvanje izvora i zračnika. Kontejner je tako postavljen da se uticaj

zračenja iz njega može zanemariti. Debljina zidova zračnika je takva da se intenzitet primarnog zračenja smanjuje hiljadu puta. Oblikovanje snopa se vrši kolimatorom, pri čemu je ugao kolimacije promenljiv. Ovaj uređaj služi za kalibraciju dozimetrijske instrumentacije. Maksimalna dimenzija polja na rastojanju od 1 m je 30 cm x 30 cm. Uzima se da je srednja energija gama fotona E = 1,225 MeV, jer se u spektru izdvajaju dve linije gama fotona, čije su energije 1,331 MeV i 1,173 MeV. Jačina apsorbovane doze gama zračenja u vazduhu iznosila je 96 cGy/h, 960 cGy/h i 1920 cGy/h, respektivno. Jačina ekspozicije iznosila je 7,17 x 10^{-6} C/kgs, 7,17 x 10^{-5} C/kgs i 1,43 x 10^{-4} C/kgs, respektivno.



Sl. 2. Uređaj za realizaciju polja gama zraka IRPIK B

Ispitivanja uticaja X zraka na karakteristike modela gasnog odvodnika prenapona vršena su u polju X zraka dozimetrijskog generatora PHILIPS MG-320, koji je prikazan na slici 3 i koji se nalazi u Metrološko-dozimetrijskoj laboratoriji Instituta za nuklearne nauke Vinča. Istraživanja su vršena u kalibrisanim poljima pri naponima X cevi od 60 kV, 150 kV i 300 kV, respektivno, kojima odgovaraju struje cevi od 15 mA, 10 mA i 10 mA, respektivno. Jačina ekspozicione doze X zraka iznosila je 2,83 x 10^{-6} C/kg s, 5,89 x 10^{-6} C/kg s i 3,46 x 10^{-6} C/kg s.

Sl. 3. Uređaj za realizaciju polja X zraka PHILIPS MG-32

Ispitivanja gasnih odvodnika prenapona vršena su prema sledećoj proceduri: 1) formiranje modela gasnog odvodnika prenapona. Pod tim se podrazumeva izbor odgovarajućeg materijala od koga su napravljene elektrode, smeštanje elektroda u gasno-vakuumsku komoru i podešavanje rastojanja između elektroda; 2) povezivanje formiranog modela (gasne cevi) u gasno - vakuumski sistem, pomoću odgovarajućih ventila, sa vakuumskom pumpom sa jedne strane i dovodom gasa iz čelične boce sa komprimovanim gasom sa druge strane, kao i povezivanje sa meračem pritiska; 3) vakuumiranje sistema, koje podrazumeva uspostavljenje stabilnog pritiska pomoću ventila ka vakuum pumpi i igličastih ventila, koji služe za doziranje pritiska. Pritisak mora biti stabilan, tj. njegova vrednost se ne sme menjati tokom eksperimenta; 4) pozicioniranje gasne komore na određenu jačinu doze; 5) povezivanje modela gasnog odvodnik u električno kolo prema šemi predstavljenoj na slici 1; 6) kondicioniranje elektrodnog sistema, tj. izazivanje serije proboja sa pauzama od 30 s između njih, što omogućava ponovljivost mernih rezultata, odnosno pouzdanost merenja; 7) merenje vrednosti impulsnog probojnog napona i njihovo očitavanje na osciloskopu (po pedeset vrednosti za svaku brzinu impulsa i konkretnu elektrodnu konfiguraciju); 8) promena položaja radne tačke modela gasnog odvodnika i ponavljanje postupka merenja. To podrazumeva promenu parametara gasno-vakuumske komore (materijala elektroda, jačine doze).

Uticaj γ i X zraka na karakteristike modela gasnog odvodnika prenapona ispitivan je za tri materijala elektroda i to: aluminijum, čelik i mesing i pri tri brzine impulsa i to: 1,2/50 μs, 10/700 μs i 100/700 μs.

IV. Rezultati merenja i diskusija rezultata

Dobijeni eksperimentalni rezultati u polju gama zraka za aluminijumske, čelične i mesingane elektrode prikazani su u tabelama 1 – 3, respektivno, a u polju X zraka u tabelama 4 – 6, respektivno.

TABELA I

STATISTIČKI PODACI ZA ELEKTRODE OD ALUMINIJUMA U POLJU GAMA ZRAKA

Promenljiva	Broj merenja	Srednja vrednost	-95%	+95%	Minimum	Maksimum	Standardna devijacija
VAR1	50	2076,200	878,644	3756,000	784,0000	3052,000	695,1388
VAR2	50	1148,420	1015,666	1281,174	448,0000	2436,000	467,1197
VAR3	50	607,600	547,259	667,941	350,0000	1120,000	212,3205
VAR4	50	1060,360	1003,787	1116,933	770,0000	1120,000	199,0640
VAR5	50	730,660	678,608	782,712	420,0000	1610,000	183,1551
VAR6	50	468,020	430,368	505,672	350,0000	840,000	132,4840
VAR7	50	767,900	751,687	784,113	665,0000	910,000	57,0482
VAR8	50	529,480	505,010	553,950	420,0000	700,000	86,1018
VAR9	50	381,220	372,617	389,823	357,0000	483,000	30,2724
VAR10	50	723,100	705,626	740,574	560,0000	840,000	61,4858
VAR11	50	492,240	476,141	508,339	434,0000	665,000	56,6470
VAR12	50	390,920	372,836	409,004	100,0000	560,000	63,6300

TABELA II

STATISTIČKI PODACI ZA ELEKTRODE OD ČELIKA U POLJU GAMA ZRAKA

Promenljiva	Broj merenja	Srednja vrednost	-95%	+95%	Minimum	Maksimum	Standardna devijacija
VAR1	50	1704,360	1599,822	1808,898	812,0000	2408,000	367,8376
VAR2	50	1384,600	1273,038	1496,162	448,0000	2072,000	392,5532
VAR3	50	1191,260	1073,033	1309,487	476,0000	1960,000	416,0041
VAR4	50	1126,160	1074,021	1178,299	742,0000	1554,000	183,4592
VAR5	50	897,120	845,902	948,338	476,0000	1260,000	180,2185
VAR6	50	709,800	656,182	763,418	392,0000	1092,000	188,6637
VAR7	50	801,080	761,500	840,660	630,0000	1330,000	139,2707
VAR8	50	578,340	543,015	613,665	413,0000	882,000	124,2971
VAR9	50	454,300	428,305	480,295	350,0000	672,000	91,4686
VAR10	50	718,620	701,321	735,919	602,0000	840,000	60,8705
VAR11	50	519,540	498,439	540,641	420,0000	742,000	74,2468
VAR12	50	403,760	389,084	418,436	329,0000	616,000	51,6418

TABELA III

STATISTIČKI PODACI ZA ELEKTRODE OD MESINGA U POLJU GAMA ZRAKA

Promenljiva	Broj merenja	Srednja vrednost	-95%	+95%	Minimum	Maksimum	Standardna devijacija
VAR1	50	1691,760	1607,575	1775,945	952,0000	2352,000	296,2210
VAR2	50	1151,920	1074,677	1229,163	588,0000	1820,000	271,7946
VAR3	50	1023,260	968,034	1078,486	448,0000	1526,000	194,3229
VAR4	50	1050,840	1010,976	1090,704	798,0000	1428,000	140,2686
VAR5	50	918,680	868,084	969,276	588,0000	1344,000	178,0312
VAR6	50	778,680	728,092	829,268	420,0000	1092,000	178,0031
VAR7	50	924,000	895,580	952,420	756,0000	1176,000	100,0000
VAR8	50	623,140	592,082	654,198	434,0000	868,000	109,2839
VAR9	50	485,520	458,084	512,956	378,0000	707,000	96,5377
VAR10	50	848,680	819,710	877,650	756,0000	1190,000	101,9368
VAR11	50	556,360	531,437	581,283	406,0000	784,000	87,6954
VAR12	50	439,040	424,756	453,324	378,0000	588,000	50,2601

TABELA IV

STATISTIČKI PODACI ZA ELEKTRODE OD ALUMINIJUMA U POLJU X ZRAKA

Promenljiva	Broj merenja	Srednja vrednost	-95%	+95%	Minimum	Maksimum	Standardna devijacija
VAR1	50	2076,200	1878,644	2273,756	784,0000	3052,000	695,1388
VAR2	50	1148,420	1015,666	1281,174	448,0000	2436,000	467,1197
VAR3	50	607,600	547,259	667,941	350,0000	1120,000	212,3205
VAR4	50	898,240	825,880	970,600	574,0000	1400,000	254,6112
VAR5	50	563,640	500,180	627,100	308,0000	1190,000	223,2946
VAR6	50	355,600	341,487	369,713	315,0000	693,000	49,6588
VAR7	50	729,540	675,545	783,535	490,0000	1260,000	189,9910
VAR8	50	510,160	457,236	563,084	378,0000	1106,000	186,2237
VAR9	50	370,860	346,966	394,754	315,0000	833,000	84,0772
VAR10	50	948,360	851,985	1044,735	504,0000	1736,000	339,1143
VAR11	50	561,120	499,566	622,674	364,0000	1344,000	216,5888
VAR12	50	419,860	389,603	450,117	343,0000	791,000	106,4659

TABELA V

STATISTIČKI PODACI ZA ELEKTRODE OD ČELIKA U POLJU X ZRAKA

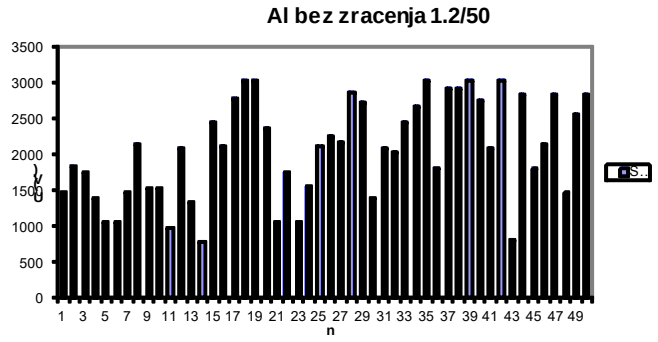
Promenljiva	Broj merenja	Srednja vrednost	-95%	+95%	Minimum	Maksimum	Standardna devijacija
VAR1	50	1704,360	1599,822	1808,898	812,0000	2408,000	367,8376
VAR2	50	1384,600	1273,038	1496,162	448,0000	2072,000	392,5532
VAR3	50	1191,260	1073,033	1309,487	476,0000	1960,000	416,0041
VAR4	50	920,920	863,750	978,090	686,0000	1540,000	201,1624
VAR5	50	755,300	698,711	811,889	434,0000	1092,000	199,1193
VAR6	50	524,440	467,415	581,465	336,0000	938,000	200,6531
VAR7	50	976,360	910,061	1042,659	700,0000	1498,000	233,2863
VAR8	50	731,360	672,699	790,021	448,0000	1218,000	206,4085
VAR9	50	458,500	414,455	502,545	336,0000	868,000	154,9790
VAR10	50	1234,940	1148,847	1321,033	700,0000	2100,000	302,9326
VAR11	50	1013,180	942,255	1084,105	462,0000	1680,000	249,5629
VAR12	50	734,580	668,830	800,330	350,0000	1190,000	231,3543

TABELA VI

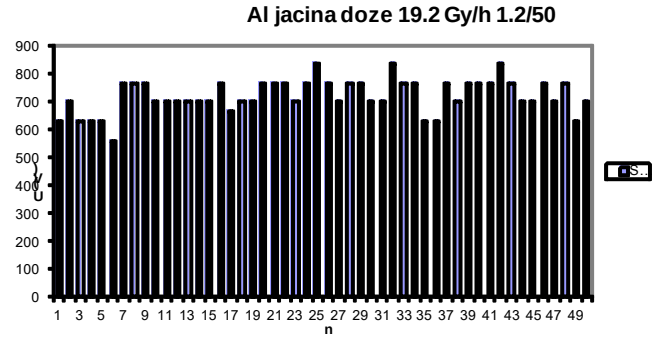
STATISTIČKI PODACI ZA ELEKTRODE OD MESINGA U POLJU X ZRAKA

Promenljiva	Broj merenja	Srednja vrednost	-95%	+95%	Minimum	Maksimum	Standardna devijacija
VAR1	50	1691,760	1607,575	1775,945	952,0000	2352,000	296,2210
VAR2	50	1151,920	1074,677	1229,163	588,0000	1820,000	271,7946
VAR3	50	1023,260	968,034	1078,486	448,0000	1526,000	194,3229
VAR4	50	718,060	698,894	737,226	595,0000	910,000	67,4402
VAR5	50	555,940	506,638	605,242	315,0000	910,000	173,4767
VAR6	50	368,480	345,281	391,679	315,0000	700,000	81,6304
VAR7	50	817,460	772,307	862,613	560,0000	1330,000	158,8791
VAR8	50	591,640	535,048	648,232	329,0000	980,000	199,1293
VAR9	50	362,460	351,480	373,440	329,0000	595,000	38,6339
VAR10	50	956,900	888,831	1024,969	595,0000	1540,000	239,5130
VAR11	50	851,060	768,717	933,403	392,0000	1540,000	289,7381
VAR12	50	521,780	463,319	580,241	350,0000	980,000	205,7047

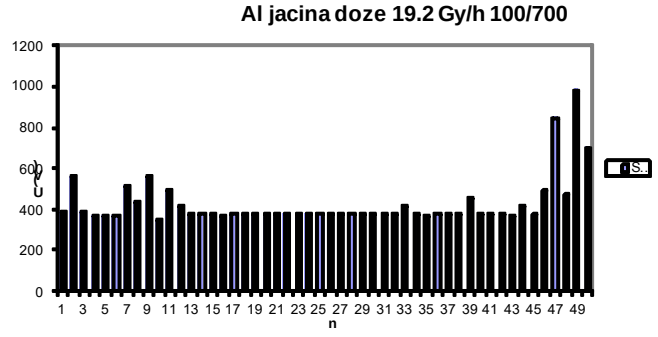
Hronološki nizovi vrednosti probojnog napona za aluminijumske elektrode bez zračenja i pri najvećoj primenjenoj jačini doze gama zračenja od 19,2 Gy/h, pri brzini impulsa 1,2/50 μs, prikazani su na slikama 4 i 5, respektivno, a u slučaju aluminijumskih elektroda pri najvećoj jačini gama doze i najsporijem impulsu 100/700 μs na slici 6.



Sl. 4. Hronološki niz vrednosti probojnog napona za Al elektrode bez zračenja pri impulsu 1,2/50

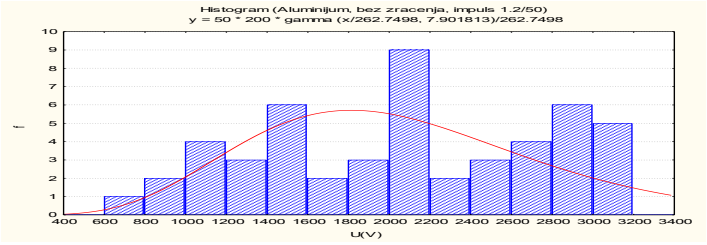


Sl. 5. Hronološki niz vrednosti probojnog napona za Al elektrode pri jačini doze gama zraka od 19,2 Gy/h pri impulsu 1,2/50

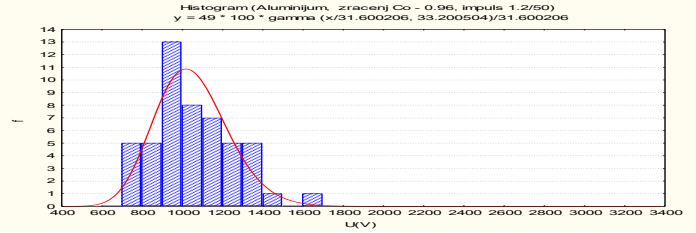


Sl. 6. Hronološki niz vrednosti probojnog napona za Al elektrode pri jačini doze gama zraka od 19, 2 Gy/h pri impulsu 100/700

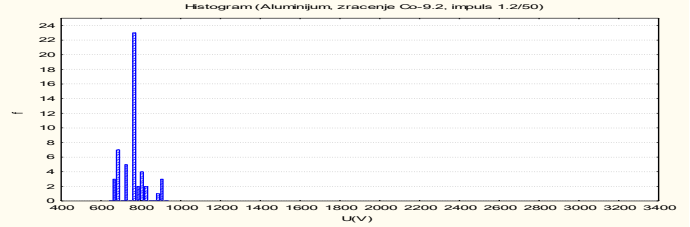
Histogrami vrednosti probojnog napona za aluminijumske elektrode bez zračenja i u polju gama zraka pri jačinama doza od 0,96 Gy/h, 9,6 Gy/h i 19,2 Gy/h, prikazani su na slikama 7 – 10, respektivno. Histogram koji odgovara hronološkom nizu prikazanom na slici 6 dat je na slici 11. Histogram vrednosti probojnog napona u polju gama zraka Co pri jačini doze od 9,6 Gy/h i brzini impulsa 1,2/50 μs u slučaju čeličnih, odnosno mesinganih elektroda prikazani su na slikama 12 i 13, respektivno.



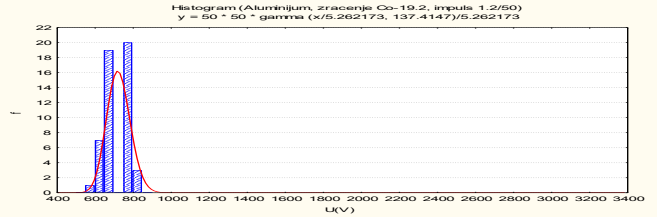
Sl. 7. Histogram vrednosti probojnog napona za Al elektrode bez zračenja pri brzini impulsa 1,2/50



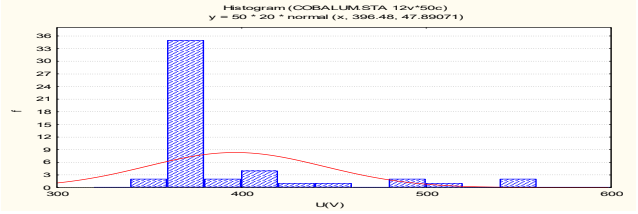
Sl. 8. Histogram vrednosti probojnog napona za Al elektrode u polju gama zraka pri jačini doze 0,96 Gy/h i brzini impulsa 1,2/50



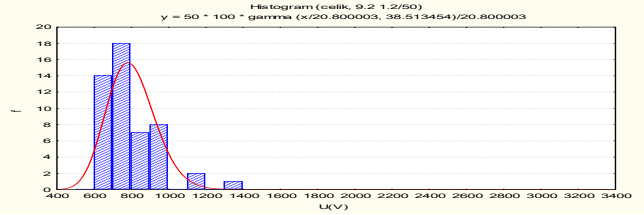
Sl. 9. Histogram vrednosti probojnog napona za Al elektrode u polju gama zraka pri jačini doze 9,6 Gy/h i brzini impulsa 1,2/50



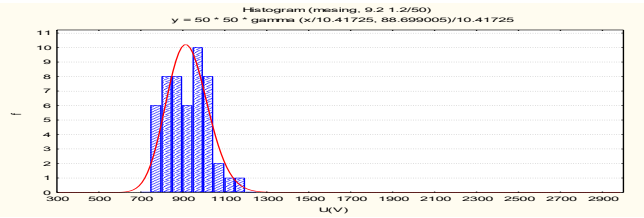
Sl. 10. Histogram vrednosti probojnog napona za Al elektrode u polju gama zraka pri jačini doze 19,2 Gy/h i brzini impulsa 1,2/50



Sl. 11. Histogram vrednosti probojnog napona za Al elektrode u polju gama zraka pri jačini doze 19,2 Gy/h i brzini impulsa 100/700



Sl. 12. Histogram vrednosti probojnog napona za čelične elektrode u polju gama zraka pri jačini doze 9,6 Gy/h i brzini impulsa 1,2/50



Sl. 13. Histogram vrednosti probojnog napona za mesingane elektrode u polju gama zraka pri jačini doze 9,6 Gy/h i brzini impulsa 1,2/50

Na osnovu dobijenih rezultata možemo zaključiti sledeće:

1) Gama zraci dovode do smanjivanja standardnog odstupanja. Dakle, ono dovodi do poboljšanja karakteristika odvodnika. Dobijeni efekti su najviše izraženi kod elektroda od aluminijuma, a najmanje kod čeličnih elektroda.

2) I u slučaju X zraka se najbolji rezultati postižu prilikom upotrebe aluminijumskih, a najlošiji pri upotrebi čeličnih elektroda.

3) Najmanje rasipanje vrednosti probojnog napona je pri najbržim impulsima 1,2/50 μs, a najveće pri najsporijim impulsima 100/700 μs.

Sve navedene promene su reverzibilne i odvodnici nakon prestanka dejstva zračenja ponovo imaju iste karakteristike, kao pre dejstva zračenja (dinamička radijaciona otpornost). Dakle, možemo zaključiti da je u gama polju veoma preporučljiva upotreba modela odvodnika sa elektrodama od aluminijuma i da ovo zračenje dovodi do privremenog poboljšanja karakteristika odvodnika. Pritom se najbolji rezultati postižu pri brzinama impulsa od 1,2/50 μs. I u polju X zraka najbolji rezultati se postižu primenom aluminijumskih elektroda pri impulsima 1,2/50 μs, samo što ono dovodi do neznatnog poboljšanja karakteristika modela gasnog odvodnika prenapona.

V. Zaključak

U ovom radu su prikazani eksperimentalni rezultati ispitivanja karakteristika modela gasnog odvodnika prenapona u polju gama i X zraka u impulsnom režimu rada za tri različita materijala elektroda. Pokazano je da se najveća brzina odziva odvodnika u polja gama i X zraka postiže primenom aluminijumskih elektroda pri najsporijim impulsima.

Zahvalnica

Ovaj rad napisan je u okviru projekata Ministarstva prosvete, nauke i tehnološkog razvoja ON171007 i III 43009.

Literatura

[1] Ž. Markov, Prenaponska zaštita u elektronici i telekomunikacijama, Beograd: Tehnička knjiga, 1983.

[2] Ž. Markov, "Upoređenje savremenih prenaponskih elemenata," *Elektrotehnika 10*, str. 961-963, 1987.

[3] P. Osmokrović, M. Stojanović, B. Lončar, N. Kartalović, I. Krivokapić, "Radioactive resistance of elements for over - voltage protection of low-voltage systems," *Nuclear Instruments and Methods in Physics Research B*, no. 140, pp. 143-151, 1998.

[4] P. Osmokrović, B. Lončar, S. Stanković, "Investigation the optimal method for improvement the protective characteristics of gas filled surge arresters-w/o the built in radioactive sources," *IEEE Trans. Plasma Science*, vol. 30, no. 5, pp. 1876-1880, 2002.

[5] B. Lončar, P. Osmokrović, S. Stanković, "Radioactive reliability of gas filled surge arresters," *IEEE Trans. Nuclear Science*, vol. 50, no. 5, pp. 1725-1731, 2003.

[6] B. Lončar, S. Stanković, A. Vasić, P. Osmokrović, "Uporedna analiza uticaja gama i X zračenja na karakteristike nekih komercijalnih gasnih odvodnika prenapona," u Zborniku radova *XLVIII Konferencije ETRAN*, 2004, tom IV, str. 68-71.

[7] B. Lončar, P. Osmokrović, A. Vasić, S. Stanković, "Influence of gamma and X radiation on gas-filled surge arrester characteristics," *IEEE Trans. Plasma Science*, vol. 34, no. 4, pp. 1561-1565, 2006.

[8] B. Lončar, P. Osmokrović, S. Stanković, R. Šašić, "Influence of electrode material on gas-filled surge arrester characteristics in gamma and X radiation field, " *Journal of Optoelectronics and Advanced Materials*, vol. 8, no. 2, pp. 863-866, 2006.

[9] B. Lončar, N. Kartalović, S. Stanković, A. Vasić, P. Osmokrović, "Uticaj materijala elektroda i pritiska na karakteristike modela gasnog odvodnika prenapona u polju gama zračenja," u Zborniku radova *XLIX Konferencije ETRAN*, 2005, tom IV, str. 68-71.

[10] B. Lončar, N. Kartalović, S. Stanković, M. Vukčević, M. Kovačević, "Uticaj materijala elektroda i pritiska na karakteristike modela gasnog odvodnika prenapona u polju X zračenja u jednosmernom režimu rada," u Zborniku radova *XXIII Simpozijuma društva za zaštitu od zračenja Srbije i Crne gore*, 2005, str.149-152.

[11] B. Lončar, N. Kartalović, S. Stanković, M. Vujisiić, P. Osmokrović, "Uticaj materijala elektroda na brzinu odziva gasnog odvodnika prenapona u polju gama i X zračenja ," u Zborniku radova *LI Konferencije ETRAN, 2007*,tom IV, str. NT3.3.1-3.3.4.

[12] B. Lončar, M. Vujisić, A.Vasić, P. Osmokrović, "Radijaciona otpornost modela gasnog odvodnika prenapona u polju neutronskog zračenja," u Zborniku radova *L Konferencije ETRAN*, 2006, tom IV, str. 61-64.

Abstract

The aim of this paper is to experimentally examine the influence of gamma and X radiation on the gas filled surge arresters physical model characteristics in pulse regime for three materials of electrode and three different pulse. The obtained results showed that gamma and X radiation temporally improved characteristic of gas filled surge arresters. The best results is obtained for aluminium electrode and the fastest pulse.

Comparative analysis of the influence of gamma and X radiation on the gas filled surge arresters model characteristics in pulse regime

Boris Lončar, Dušan Nikezić, Katarina Karadžić, Luka Rubinjoni, Andrija Janković

Robot Task Extraction and Replication from Raw Video Using Reinforcement Learning

Milivoje Majstorović, Zaviša Gordić, Kosta Jovanović

***Abstract*— This paper presents the underlying concept behind an industrial robot replicating a task demonstrated by a human. The task is extracted from the raw sample video in order to obtain the overall goal. The data is prepared for reward function extraction, after which state-action space is obtained in order to choose most suitable movement. To complete the cycle, the optimal policy for achieving the task is obtained and the task replication is evaluated. The initial tests are performed on a ball throwing motion task without an active gripping element with a set target for the ball – a basketball hoop.**

***Index Terms*—Industrial robotics; Reinforcement learning; Machine learning;**

I. Introduction

Facilitating robot programming and making it more intuitive for the human user is an ongoing and constantly evolving effort towards wider adoption of robotics. With regards to industrial robotics, it is targeted towards enabling faster reprogramming and increased adaptability required by the SMEs and Industry 4.0 in general.

However, analysis and breakdown of a human executed task is a complex and tedious task. Moreover, translating it to an industrial robot movement introduces additional multivariable challenges. For a vast majority of automation problems, the initial goal is to replicate an operator's articulated action, which is easily executed by a human, although complex to break down as a model for robot control. This is where the human demonstration examination and replication comes in handy. Today, some industrial robots already have kinesthetic teaching. In essence, it is a form of supervised learning where it is possible to manipulate the robot by hand, and then store the external path coordinates, which enable the robot to replicate the performed movement. Programming by demonstration [1] is focused on analyzing human actions and action sequences which can be obtained by interacting with a wide range of sensors, while on the other hand monitoring robot execution of the given task is represented as a programming sequence.

The vast majority of industrial robots which are used in an automation production are usually programmed in a static, artificial environment, where the robot has to follow a pre-planned route and to rely on the previous sequence, precise placement, without any deviations and environmental noises. By considering human demonstration footage as input for robot control, object path modeling and robot motion planning are not necessary for the given approach.

In order to demonstrate a concept of an agent observing its environment and adapting after some level of interaction with the environment an exemplary scenario of making a basketball shot was conducted, relying on programming by demonstration. [2]

The second section explains the environment as well as equipment which were used to perform the initial experiments.

The third section explains how the robot tasks are extracted from video material, and two different methods related to the identification of the object of interest.

The architecture of the system needed for reinforcement learning is explained in the fourth section, along with the description of its constituents.

The conclusion is provided in the fifth section, together with some thoughts for future work.

II. Environment Setup

The main component of the system is a 6-axis industrial robot, Denso VP-6242. Additional components are RGB and an RGB-D camera which is used for different aspects of the implementation test. The end-effector mounted on the robot manipulator is a passive 3D printed head cavity which loosely fits the ball in place. A basketball hoop is presented on the right as a target for the algorithm. The environment setup is depicted in figure 1, where the frame of the demonstration video sample was captured.

Figure 1. Demonstration of the task. The task is extracted from the motion performed by a human.

RGB and RGB-D cameras are taken into consideration. Video quality of the RGB camera is 1920 x 1080, 30fps, 16:9 aspect ratio, while the RGB-D camera is a Microsoft Kinect 1, with an RGB resolution of 640 x 480, 30fps and depth resolution of 320 x 240, with a minimum distance of

Milivoje Majstorović is with the School of Electrical Engineering, Signals and Systems Department, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail: majstormapp@gmail.com).

Zaviša Gordić is a research associate at the School of Electrical Engineering, Signals and Systems Department, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail: zavisa@etf.rs).

Kosta Jovanović is an assistant professor at the School of Electrical Engineering, Signals and Systems Department, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail: kostaj@etf.rs).

40cm and a maximum distance of 4.5m. The aforementioned cameras enable two different scenarios.

The first scenario involves using only the RGB camera. The RGB camera enables a 2D setup in which 3 mutually parallel robot axis, are controlled. In this scenario, the hoop position modification is limited by the throwing range of the robot and restricted to the plane orthogonal to the controlled axis of the robot.

The second scenario is based on RGB-D camera, which obtains a 3D point cloud and therefore enables moving the target wherever the robot had theoretical throwing reach. However, in order to achieve that, control of a greater number of robot's axis is needed, increasing the complexity.

For initial implementation tests and proof of concept, it was opted to use only the RGB camera due to its larger resolution and simplicity of the scenario implementation.

Communication with the robot is established by TCP/IP protocol between the robot controller (server side) and a PC running a Linux Virtual Machine (client side). On the client side, a Python script is running which feeds the output data to the robot control in terms of joint positions, velocity and acceleration. Both cameras are directly communicating with the PC where the script is obtaining each frame during the task execution.

III. TASK EXTRACTION

The first phase of an approach to the programming by demonstration task is to translate the demonstration into something meaningful to the machines. This phase is applicable for various tasks, may it be a door opening [3], autonomous driving [4] or pouring a liquid in a cup [5]. In essence, it means that the task should be extracted from all irrelevant background activities and clearly defined. For the particular example, the position of the center of the ball and the hoop on the image is important, since the trajectory of the ball, its distance and/or relative position to the hoop can be used to formulate the task. On the other side, the human arm throwing the ball and all the background is irrelevant.

In order to obtain any reasonable data from raw video footage, it is first needed to extract relevant object information. To this end, two possible approaches were chosen.

The first approach relies on defining an object of interest by analyzing the given environment in order to extract the relevant data.

The second approach uses an off-the-shelf pre-trained model for object detection [6][7] in order to extract the trajectory which the object of interest had.

For the first approach, in order to extract the object of interest, it is first needed to evaluate some of its features and establish thresholds for detecting the object against the rest of the environment. Initially, it was expected to focus on color difference as well as shape difference. Therefore, the emphasis was on defining the HSV range of lower and upper values of the object, thus excluding the complex surrounding and providing a mask for the object of interest. Additional image processing was performed for removing any residual blobs or noise, by using dilation and erosion on the defined mask. After post-processing, it was needed to extract contours from the frame and focus on values related to the largest contour. These values were further used to define a minimum threshold for the contour to be considered, which corresponds to the object of interest. For every frame, x and y values of the detected object in terms of centroid were appended. Appended values from multiple successive frames were interpolated and used for defining a path of interest for replication. The resulting frame is depicted in figure 2, where the interpolated trajectory is colored in red.

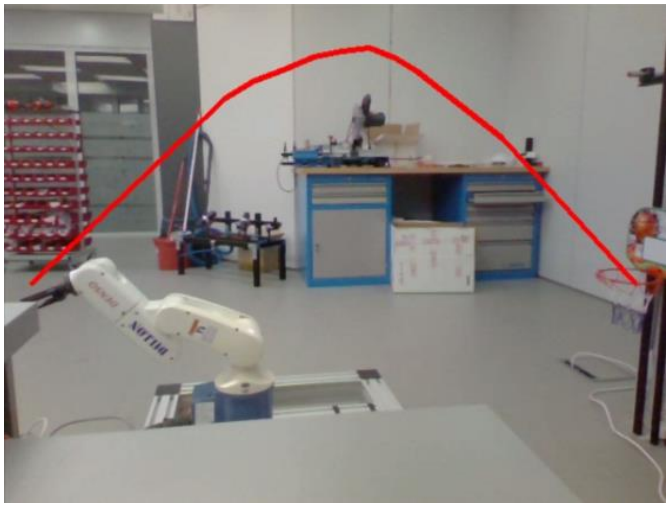


Figure 2. Task extraction by defining a feature to be monitored

The second approach would be to skip the desired object feature definition and rely on a pre-trained model of a sports ball classification. One of those models is YOLO – You Only Look Once. Convolutional Neural Networks are a common starting point in object detection, thus it is reasonable to expect the same to be used in YOLO. It consists of a three-step approach: resizing the image, running the image through a single convolutional neural network and outputting the classification result based on the confidence level threshold of the model. The initial model is trained on the ImageNet 1000-class dataset [8] which also has different kinds of sports balls in its training set. In order to generalize the data and lower the time needed for computation around 80 classes which can be detected were defined and presented with a confidence level. In order to extract the task, the same principle of tracking the ball classified object by centroid was used. The trajectory of the given object was used to form a reward function for providing the agent feedback from real-time object detection. Described method is depicted in figure 3 with corresponding confidence level of detection.

Figure 3. Task extraction by defining a feature to be monitored

Additional information which could be extracted using RGB-D camera is a point cloud, where it is possible to join the localized object in the RGB frame with the same localization of the depth matrix.

Taking everything into consideration, task extraction and reward function engineering are focused on determining object position and where it is relative to the target object. The target object is located based on the backboard square and established prior to the initial training. In future training sets, it will be tracked in real-time in order to adapt to the modifications of the hoop position.

IV. Task Replication

Obtaining the environmental feedback and establishing a robust task extraction sequence is a primary task for feeding the data to a reinforcement learning algorithm. Reinforcement learning presents a type of machine learning method in which the agent receives a reward with delay to evaluate its previous action. Typical reinforcement learning method consists of an Agent which makes a given action which has an impact on the Environment. , in turn, generates an updated Reward value and State upon which the Agent will react with a new Action.

In presented setup Action represents all of the possible movements, in other words, joint values. The state is presented as a current frame captured by the RGB or RGB-D camera. The reward function is constructed based on the difference of the current position of the ball relative to the target trajectory.

Reinforcement learning model can rely on the transition probability where the agent will likely know how to enter a specific state given the current state and action. Opposite to the aforementioned algorithm, model-free algorithm updates parameters based on trial-and-error, as it is the case with Q-learning [8] or deep Q networks (DQN). In order to achieve the continuous control of the Agent and update the robot joint positions, it is needed to refer to an actor-critic model-free algorithm, the algorithm is based on the deep deterministic policy gradient (DDPG) [9], whose architecture is shown in figure 4

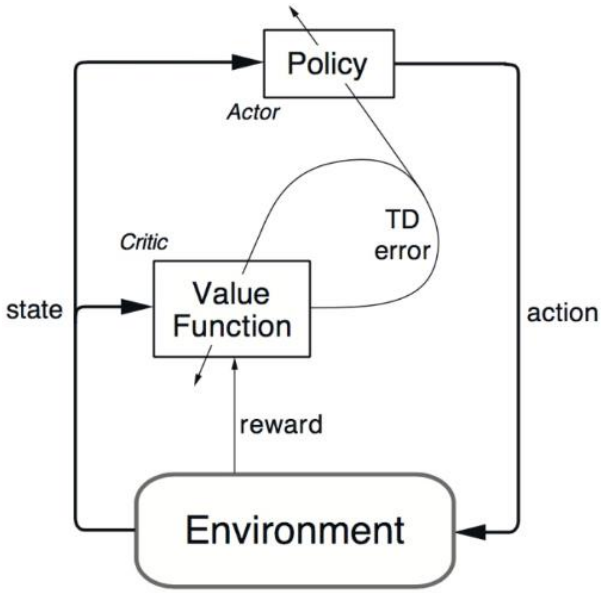


Figure 4. DDPG actor-critic architecture

Initial robot control consists of controlling end positions of 3 robot joints, J2, J3 and J5, which had an impact on the throwing movement. Velocity and acceleration were constant. While learning the policy initial problem was the deceleration of the robot while achieving the specified point. In order to achieve a continuous motion, robot joint actions have to be merged and interpolated in order to minimize the deceleration of the robot when achieving a desired position of the joints. Reward function was constructed to provide additional extreme value for achieving the centroid of the ball in close proximity of few pixels to the center of the basketball board. Figure 5 depicts trajectory obtained from a human demonstration video and it is presented in red, while the robot replication action is shown in blue.

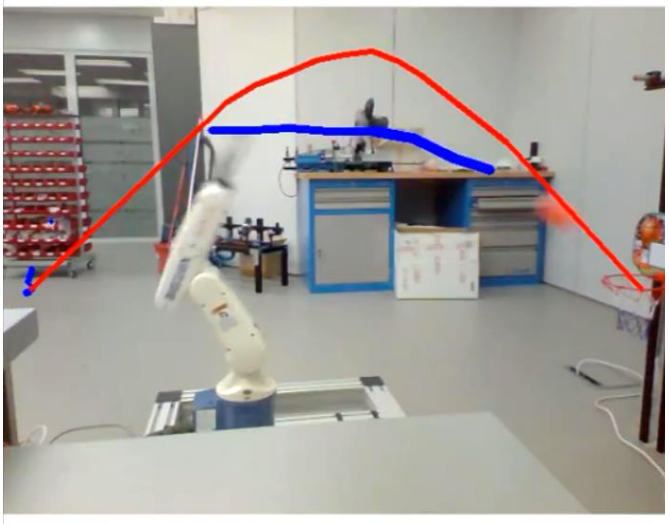


Figure 5. Robot action (blue) and target path (red)

The discrepancy between the given paths was impacted by the fact that the acceleration, deceleration, and velocity were constant. Although an active gripping element can introduce an additional variable and inconsistency, the catapult ejection of the ball is somewhat an unpredictable event. Therefore, it is not possible to maintain the control of the ejection, which was a problem in the region where the

ball had the position update without any connection to the Actor.

With the implementation of the RGB-D camera, we will introduce the control of all 6 joints and train the model on different positions of the hoop in order to establish a sequence-based target movement. While expanding to all joints it is of utmost importance to introduce the velocity, acceleration, and deceleration to the action vector, as the policy couldn't be fully optimized due to the fact that the robot wasn't able to replicate the exact same movement of the human demonstrator.

V. Conclusion

Complex tasks are increasingly being conducted by industrial robots and thus the programming has become more demanding as well as intended to be shorter. Programming by demonstration and imitation learning has been a potential solution for numerous applications. Additional hardware could accelerate the process and achieve better results. Additionally, real-time object detection with a larger class dataset could improve the environmental setup and better map the ball trajectory as well as the target position update. While training and obtaining an action-state space, there are potential problems of achieving an irregular policy although with the achieved goal, for example when gripping a ball model can be trained to exploit the camera angle and appear to be successful without touching the ball.

With the given approach of throwing an object, it is possible to achieve a larger robot reach and not to restrict the work area of the robot. Additional research of grasping different objects could establish a more robust and complete solution as well as more versatile and consistent results.

Further development of the models for programming by demonstration can simplify the overall system implementation of a complex task and provide a more human-like approach and movement for the robot.

Acknowledgment

The work on this project was supported by the Ministry of education, science, and technological development, Republic of Serbia, grant No. TR35003.

References

[1] R. Dillmann, "Teaching and learning of robot tasks via observation of human performance" in *Robotics and Autonomous Systems 47*, 2004.

[2] R. Meyes, H. Tercan, S. Roggendorf, T. Thiele, C. Buscher, M. Obdenbusch, C. Brecher, S. Jeschke, T. Meisen, "Motion Planning for Industrial Robots using Reinforcement Learning", 2017.

[3] P. Sermanet, K. Xu, S. Levine, Unsupervised Perceptual Rewards for Imitation Learning", 2017.

[4] M. Kuderer, S. Gulati, W. Burgard, "Functional Safety Standards for Machinery" in *IEEE International Conference on Robotics and Automation*, 2015.

[5] P. Sermanet, C. Lynch, J. Hsu, S. Levine, "Time-Contrastive Networks: Self-Supervised Learning from Multi-View Observation" in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[6] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection", 2016.

[7] J. Redmon, A. Farhadi, "YOLOv3: An Incremental Improvement", 2018.

[8] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, M. Riedmiller, "Playing Atari with Deep Reinforcement Learning", 2015.

[9] T. Lillicrap, J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, D. Wierstra, "Continuous Control with Deep Reinforcement Learning" in *International Conference on Learning Representations*, 2016.

Underactuated Finger Design for Flexible Grasping in Robotic Assembly

Lazar Matijašević, *PhD Student*, Petar B. Petrović, *Full Professor*

***Abstract*—At present-day manufacturing and assembly lines, fairly simple mechanism grippers are used. On the other hand, growing demand for customized products that are mass produced, require flexible, multipurpose grippers that are capable of grasping complex objects of different sizes and shapes. In order for robotic hand to be industry acceptable it needs to be robust, easy to control and most importantly it needs to be affordable price-wise. With that in mind concept of multifingered underactuated robotic hand appears as a good candidate to be optimal, general purpose solution. Underactuation as a concept allows robotic hands to grip arbitrary shaped objects without the need for complex control and sensory systems. Also, with less actuators than degrees of freedom multifingered underactuated robotic hand is more affordable and from robot arm carrying capacity standpoint, actuators with less weight allows robotic systems to move faster or to carry heavier loads. For this research linkage driven underactuated mechanisms are chosen because of their rigidity and that trait makes control system more reliable and easier to make thus making whole robotic system more robust and reliable. This paper presents some aspects of design and grasping force analysis of three degrees of freedom (3-DoF) underactuated robotic finger with linkage driven mechanism for CMSysLab Robotic Hand.**

***Index Terms*—Robotic Assembly, Robotic Hand, Design; Underactuation;**

I. Introduction

In industry setting, grasping of various objects in a well-structured or unstructured environment present complex tasks that are still performed by human operators even if environment is dangerous. The new production paradigm of mass customization and extensive needs for application of robotic technology in domain of Small and Medium Enterprises (SME), imposes demands for highly flexible, multipurpose gripers that are capable of precise and reliable grasping complex objects of different sizes and shapes, including in hand manipulation. In order to accommodate demands of mass customization and market push for ubiquitous use of robotic technology in SME operations, both robotic arms and robotic hands that approach performances of humans in terms of dexterity and adaptation capabilities must be developed, [1].

Robotic hands that can, to some degree, mimic human hand capabilities are using principle of complete actuation. One example of fully-actuated robotic hand is Shadow Dexterous Hand [2]. These fully-actuated robotic hands have some disadvantages like: space needed to put all actuators in, increased weight of such hand, complexity of control system needed to operate properly and biggest one, from application point of view, is high price of these hands.

On the other hand, prototypes are made that involve a smaller number of actuators than degrees of freedom. This approach, called underactuation is implemented through the use of passive elements like mechanical limits and springs leading to a mechanical adaptation of the finger to the shape of the object to be grasped.

In case of robotic grasping, underactuation allows the robotic hand to adjust itself to a wide range of shapes, including irregularly shaped objects, without the need for complex control strategies, sensory systems for feedback and replaceable mechanical parts. These underactuated robotic hand systems don't need complex control systems to operate, also have smaller mass and on top of that they are less expensive than fully actuated robotic hands and can even compare to regular grippers when it comes to price. These traits makes them a good choice when considering development of more flexible griping systems to use in industry.

Underactuation in robotic hands leads to some intriguing properties. Underactuated robotic hands cannot always ensure full whole-hand grasping. Distribution of the forces onto the different phalanges is predetermined by the mechanical design of the hand and in some configurations phalanges may not be able to actually exert any force. This uncontrollable force distribution can also lead to unstable grasps: a continuous closing motion of the actuator tending to eject the object [3] as shown on Fig.1.

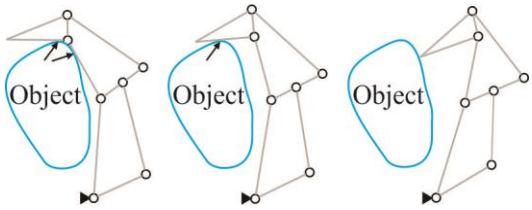


Fig. 1. Unstable grasp of object.

Designing and building underactuated robotic finger that can adapt to any object but can never actually grasp it is futile and to avoid that it is important to do a force analysis of that robotic finger.

In this paper, a new design of the robotic hand, i.e., the CMSysLab Hand, will be presented, focusing to its building block, an uderactuated finger, and within this scope it will be shown how to calculate force that each phalange exerts on object grasped. That information is important in understanding and prevention of above mentioned phenomenon of grasp degeneration and ejection of object.

PhD student Lazar Matijašević is with the Faculty of Mechanical Engineering, University of Belgrade, Kraljice Marije 16, 11120 Belgrade, Serbia (e-mail: lmatijasevic@mas.bg.ac.rs).

Full Professor Petar B. Petrović is with the Faculty of Mechanical Engineering, University of Belgrade, Kraljice Marije 16, 11120 Belgrade, Serbia (e-mail: pbpetrovic@mas.bg.ac.rs).

II. General Analytical Model

In this chapter, a method for obtaining information about force capabilities of robotic n-DoF fingers will be presented. This method is based on approach presented in [4] and [5]. This method will allow one to completely describe the relationship between the input torque of the finger actuator and the contact forces distribution on the phalanges.

For purpose of describing the proposed method, object of grasping will be considered fixed in space and friction will be ignored. Fixing object in space make proposed model predictable and allow for only one finger to be evaluated without the influence of another finger or moving object.

The neglecting friction by itself is very restrictive, but this assumption makes mathematical model relaxed and allows for primary aim of this paper which is to study the capability of the finger, and only finger, to exert contact force on the fixed object. In this stage of our research, our goal is to determine how is actuating torque transmitted through the kinematical chain of the finger and to determine its impact and impact of passive springs on contact force that is exerted on object. Obviously, above mentioned friction properties of contact between object and finger, as well as friction in joints of the finger have tremendous impact on grasping stability, and it will be part of second stage of our research. Friction properties of system are highly unpredictable and including those in this stage of research brings high level of uncertainty to the mathematical model. The significance of such model is questionable without proper experimental results and therefore it will be done with experimental setup that is in finishing stages of design.

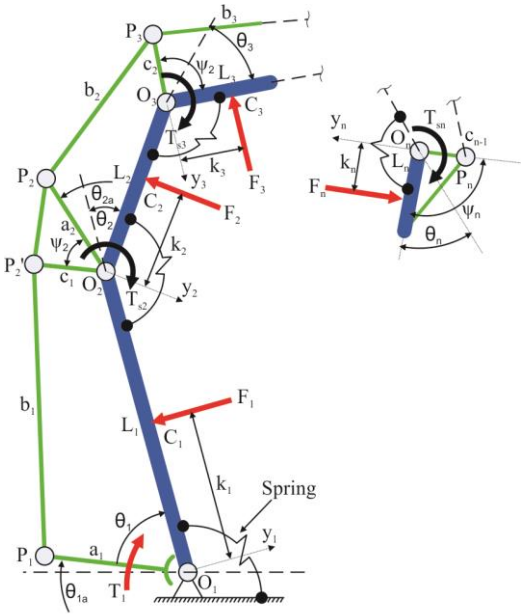


Fig. 2. Representation of n-DoF finger.

An underactuated robotic finger with *n* phalanges is illustrated in Fig. 2. The input torque from actuator is applied to the first joint of the finger, and it is transmitted to the phalanges through four-bar linkages (FBL).

Adding the springs to the joints results with fully adaptive finger with compliant joints. Passive elements are used to cinematically constrain the finger, and to ensure that finger will adapt to the shape of object being grasped.

The following parameters are presented on Fig. 2:

- L_i - the length of the i^{th} phalanx,
- a_i - the length of the first driving bar of the i^{th} FBL,
- b_i - the length of the i^{th} underactuated bar,
- c_i - the length of the second driving bar of the i^{th} FBL,
- θ_i - the rotating angle of the i^{th} phalanx,
- ψ_i - the angle between O_iP_i' and O_iP_i,
- T_1 - the torque of the actuator at the first joint,
- T_{si} - the spring torque of the i^{th} joint,
- F_i - the contact force of the i^{th} phalanx,
- k_i - position of contact point on the i^{th} phalanx.

In order to determine the distributions of the contact forces that depend on the contact point location and the joint torques inserted by springs, it is necessary to perform a quasi-static modeling of the finger. Equating the input and the output virtual powers of the finger, it yields:

$$T^T \omega_a = F^T v \tag{1}$$

where T represents the input torque vector from the actuator and springs, ω_a is the corresponding velocity vector, F is the contact force vector, and v is the projected velocity vector of the contact points:

$$T = \begin{bmatrix} T_1 \\ T_{s2} = -K_2\Delta\theta_2 \\ T_{s3} = -K_3\Delta\theta_3 \\ \dots \\ T_{sn} = -K_n\Delta\theta_n \end{bmatrix},\ \omega_n = \begin{bmatrix} \dot\theta_{1a} \\ \dot\theta_{2a} \\ \dot\theta_{3a} \\ \dots \\ \dot\theta_n \end{bmatrix},\ F = \begin{bmatrix} F_1 \\ F_2 \\ F_3 \\ \dots \\ F_n \end{bmatrix},\ v = \begin{bmatrix} v_{yc1} \\ v_{yc2} \\ v_{yc3} \\ \dots \\ v_{ycn} \end{bmatrix}, \tag{2}$$

where K_i is the stiffness of the torsional spring in joint O_i, and $\Delta\theta$ is the displacement between the current and initial angles of the joint O_i

Projected velocities can be expressed as a product of a Jacobian matrix J_T and the derivative vector of the phalanx joint coordinates, $\theta = [\theta_1, \theta_2, \theta_3, \dots, \theta_n]^T$:

$$v = J_T\dot\theta. \tag{3}$$

The Jacobian matrix J_T, of the projected velocities, can be obtained in a lower triangular form:

$$J_T = \begin{bmatrix} k_1 & 0 & 0 & \dots & 0 \\ \alpha_{12} & k_2 & 0 & \dots & 0 \\ \alpha_{13} & \alpha_{23} & k_3 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ \alpha_{1n} & \alpha_{2n} & \alpha_{3n} & \dots & k_n \end{bmatrix};\ \alpha_{ii} = k_i. \tag{4}$$

Matrix member α_{ij} can be calculated:

$$\alpha_{ij} = k_j + \sum_{k=i}^{j-1} L_k \cos\left(\sum_{m=k+1}^{j} \theta_m\right);\ i < j. \tag{5}$$

By using differential calculus, it is possible to relate the vector to the derivatives of the phalanx joint coordinates defined previously with an actuation Jacobian matrix J_a:

$$\dot\theta = J_a\,\omega_a. \tag{6}$$

In case of underactuated finger model, the four-bar linkage mechanism is used to transmit the actuator torque to each phalanx.

Principle of transmission provides the angular velocity ratio of four-bar linkage that is known as Kennedy's Theorem [6], which states that the three instantaneous centers of rotation shared by three rigid bodies in relative planar motion to another (whether or not connected) all lie on the same straight line.

Considering i^{th} four-bar linkage $O_iP_iP'_{i+1}O_{i+1}$:

$$\dot{\theta}_i = \dot{\theta}_{ia} + \dot{\theta}_{i+1a}\frac{c_i\left(L_i\sin\left(\theta_{i+1a}-\psi_{i+1}\right)-a_i\sin\left(\theta_i-\theta_{ia}+\theta_{i+1a}-\psi_{i+1}\right)\right)}{a_i\left(L_i\sin\left(\theta_i-\theta_{ia}\right)+c_i\sin\left(\theta_i-\theta_{ia}+\theta_{i+1a}-\psi_{i+1}\right)\right)}.$$

Considering last four-bar linkage $O_{n-1}P_{n-1}P_nO'_n$:

$$\dot{\theta}_{n-1} = \dot{\theta}_{n-1a} + \dot{\theta}_n\frac{c_{n-1}\left(L_{n-1}\sin\left(\theta_n-\psi_n\right)-a_{n-1}\sin\left(\theta_{n-1}-\theta_{n-1a}+\theta_n-\psi_n\right)\right)}{a_{n-1}\left(L_{n-1}\sin\left(\theta_{n-1}-\theta_{n-1a}\right)+c_{n-1}\sin\left(\theta_{n-1}-\theta_{n-1a}+\theta_n-\psi_n\right)\right)}.$$

Equation (7) is obtained by substituting equations for $\dot{\theta}_i$ and $\dot{\theta}_{n-1}$ into (6).

$$\begin{bmatrix}\dot{\theta}_1\\ \dot{\theta}_2\\ \dot{\theta}_3\\ \cdots\\ \dot{\theta}_n\end{bmatrix} = \begin{bmatrix}1 & X_1 & 0 & \cdots & 0\\ 0 & 1 & X_2 & \cdots & 0\\ \cdots & \cdots & \cdots & \cdots & \cdots\\ 0 & 0 & 0 & \cdots & X_{n-1}\\ 0 & 0 & 0 & \cdots & 1\end{bmatrix}\begin{bmatrix}\dot{\theta}_{1a}\\ \dot{\theta}_{2a}\\ \dot{\theta}_{3a}\\ \cdots\\ \dot{\theta}_{na}\end{bmatrix}, \quad (7)$$

where

$$X_i = \frac{c_i\left(L_i\sin\left(\theta_{i+1a}-\psi_{i+1}\right)-a_i\sin\left(\theta_i-\theta_{ia}+\theta_{i+1a}-\psi_{i+1}\right)\right)}{a_i\left(L_i\sin\left(\theta_i-\theta_{ia}\right)+c_i\sin\left(\theta_i-\theta_{ia}+\theta_{i+1a}-\psi_{i+1}\right)\right)}, \quad (8)$$

$$X_{n-1} = \frac{c_{n-1}\left(L_{n-1}\sin\left(\theta_n-\psi_n\right)-a_{n-1}\sin\left(\theta_{n-1}-\theta_{n-1a}+\theta_n-\psi_n\right)\right)}{a_{n-1}\left(L_{n-1}\sin\left(\theta_{n-1}-\theta_{n-1a}\right)+c_{n-1}\sin\left(\theta_{n-1}-\theta_{n-1a}+\theta_n-\psi_n\right)\right)}. \quad (9)$$

Function X_i is function that describes the transmition of actuator torque to the i^{th} phalanx.

Equation (10) that provides a practical relationship between the actuator torques and contact forces is derived from (1), (3) and (6):

$$F = J_T^{-T}J_a^{-T}T \quad (10)$$

This equation is only valid in case when $k_1k_2k_3\ldots k_n \neq 0$, which is the condition of singularity for matrix J_T. Matrix J_a cannot be singular, but, the finger may perform contact with the object in case that number of phalanges in contact with object is fewer than n. This assumption results in a singularity of the matrix J_T, so that (10) is not applicable.

III. KINETOSTATIC FORCE DISTRIBUTION MODEL

In order for a less-than-n phalanx grasp to be stable, every phalanx in contact with the object should have a strictly positive corresponding force. In grasping process, the contact appear not only with all phalanges, but also with fewer than n phalanges.

The corresponding generated forces for phalanges which are not in contact with the object should be zero, because that forces can also be seen as the external forces needed to counter the actuation torque. However, calculation of contact forces in case of fewer-than-n phalanges touching by object by using Equation (10) can be a problem because of the singularity of the matrix J_T.

This problem can be solved by proposing a general method to determine the distributions of contact forces in all cases of gripper behaviors in object grasping. In order to do that, it is assumed that the stability of the grasp must be satisfied in all cases.

From Equations (10) it is obtained:

$$J_T^TF = J_a^{-T}T, \quad (11)$$

where the component $J_a^{-T}T$ on the right side is the torque vector $\tau = [\tau_1, \tau_2, \ldots, \tau_n]^T$ at all joints of the finger relating to the actuator, spring torques and functions of torque transmission between actuator and phalanges, which is described by following equation:

$$\begin{bmatrix}\tau_1\\ \tau_2\\ \tau_3\\ \cdots\\ \tau_n\end{bmatrix} = J_a^{-T}T = \begin{bmatrix}1 & X_1 & 0 & \cdots & 0\\ 0 & 1 & X_2 & \cdots & 0\\ \cdots & \cdots & \cdots & \cdots & \cdots\\ 0 & 0 & 0 & \cdots & X_{n-1}\\ 0 & 0 & 0 & \cdots & 1\end{bmatrix}^{-T}\begin{bmatrix}T_1\\ T_{s2}\\ T_{s3}\\ \cdots\\ T_{sn}\end{bmatrix} \quad (12a)$$

which leads to:

$$\begin{bmatrix}\tau_1\\ \tau_2\\ \tau_3\\ \cdots\\ \tau_n\end{bmatrix} = \begin{bmatrix}T_1\\ T_{s2}-X_1T_1\\ T_{s3}-X_2T_{s2}+X_1X_2T_1\\ \cdots\\ T_{sn}+\sum_{j=1}^{n-1}\left[(-1)^{n-j}\prod_{i=j}^{n-1}X_iT_{si}\right]\end{bmatrix}, T_{s1}\equiv T_1. \quad (12)$$

The left part of (11) can be expressed as:

$$J_T^TF = \begin{bmatrix}k_1 & 0 & 0 & \cdots & 0\\ \alpha_{12} & k_2 & 0 & \cdots & 0\\ \alpha_{13} & \alpha_{23} & k_3 & \cdots & 0\\ \cdots & \cdots & \cdots & \cdots & \cdots\\ \alpha_{1n} & \alpha_{2n} & \alpha_{3n} & \cdots & k_n\end{bmatrix}^T\begin{bmatrix}F_1\\ F_2\\ F_3\\ \cdots\\ F_n\end{bmatrix}. \quad (13)$$

By substituting Equations (12) and (13) in (11), it follows:

$$\begin{bmatrix}\tau_1\\ \tau_2\\ \tau_3\\ \cdots\\ \tau_n\end{bmatrix} = \begin{bmatrix}k_1 & 0 & 0 & \cdots & 0\\ \alpha_{12} & k_2 & 0 & \cdots & 0\\ \alpha_{13} & \alpha_{23} & k_3 & \cdots & 0\\ \cdots & \cdots & \cdots & \cdots & \cdots\\ \alpha_{1n} & \alpha_{2n} & \alpha_{3n} & \cdots & k_n\end{bmatrix}^T\begin{bmatrix}F_1\\ F_2\\ F_3\\ \cdots\\ F_n\end{bmatrix}. \quad (14)$$

This equation shows that the torque τ_i at the i^{th} joint of the finger is calculated with respect to the contact forces vector F and parameters α_{ij}, as shown in following equation:

$$\sum_{j=i}^{n}\alpha_{ij}F_j = \tau_i, \quad \alpha_{ii} = k_i. \quad (15)$$

In case when number of phalanges in contact with object is fewer than n (e.g., when the i^{th} phalanx is not touching the object), the parameters α_{ij} in (15) are not relevant and F_i is zero. This means that (15) does not meet this condition, so it is not considered when computing the torque τ_i in case when the i^{th} phalanx is not touching the object. In order to calculate the contact forces vector F in (14), except for F_i, the following process must be used:

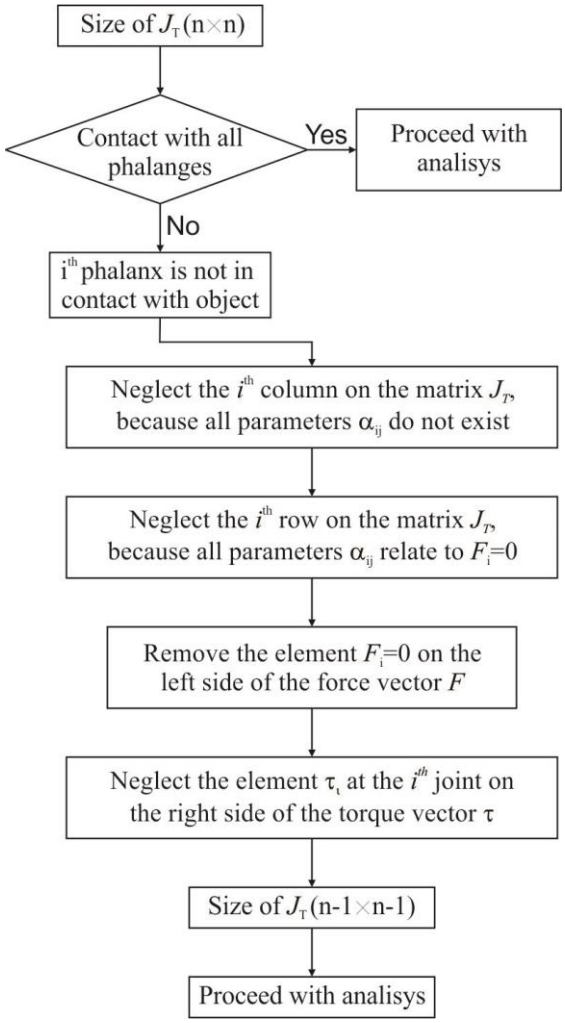


Fig. 3. Algorithm for calculating force parameters in case when not all phalanges are in contact with object.

After neglecting the i^{th} column and i^{th} row, dimension of the matrix J_T is reduced by $n-1\times n-1$, while it is guaranteed that matrix J_T will not be singular. Consequently, Equation (14) can be used in order to calculate the contact forces, except for force F_i on the i^{th} phalanx. The same process is also used in case when more than one phalanx is not in contact with the object.

IV. CMSysLab Hand finger design

Aim of this research is to develop multifingered underactuated hand for use in industry setting for tasks of extremely diversified robotic assembly and enable ubiquitous use in industry, in particular SME compatibility. For this purpose we have entered in development of the underactuated robotic finger that uses sets of linkages to transmit torque from actuator to phalanges. CMSysLab Hand finger is designed in such a manner that it allows easy installation of various sensors and building various configurations of the multifingered hands, including reconfigurable hands, making it optimal building block for various research purposes. Underactuated robotic finger design for CMSysLab Hand has three phalanges and therefore 3-DoF.

Parameters of finger for CMSysLab robotic hand are based on actual proportions of finger of human hand. The set of parameters presented in Table 1 is taking into account the mechanical joint limits, which are key elements in the design of underactuated fingers, when considering stability issues, because they limit the shape adaptation to reasonable configurations.

TABLE I
PARAMETERS OF CMSYSLAB UNDERACTUATED FINGER

a_1 [mm]	30	a_2 [mm]	23
b_1 [mm]	60.5	b_2 [mm]	37
c_1 [mm]	15	c_2 [mm]	14
L_1 [mm]	64.5	ψ_2 [°]	52
L_2 [mm]	37.5	ψ_3 [°]	90
L_3 [mm]	34.5		

Geometric and contact force parameters of underactuated 3-DoF finger are described in Fig. 4.

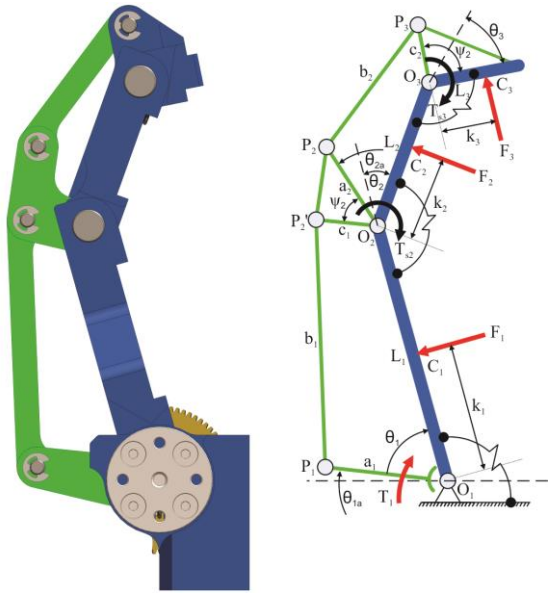


Fig. 4. Geometric and force parameters of underactuated 3-DoF finger.

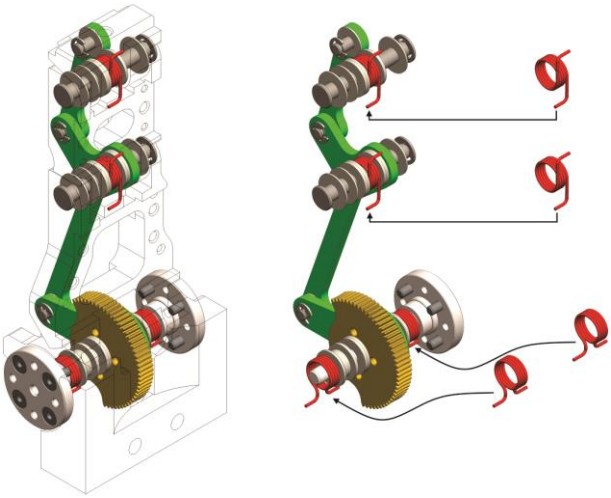


Fig. 5. Mechanical structure of underactuated 3-DoF finger.

The behavior of the finger is mostly determined by its geometry. Depending on the geometric parameters of the

mechanism, it is possible to obtain the final stability of the grasp. In structure of the finger, shown on Fig. 5, mechanical limit is used, which allows a pre-loading of the spring , shown with red color, to prevent any undesirable motion of the medial and distal phalanges, due to its own weight and/or inertial effects, as well to prevent hyperflexion of the finger. In Fig. 6 workspace as well as pre-forming stages of such finger are shown.

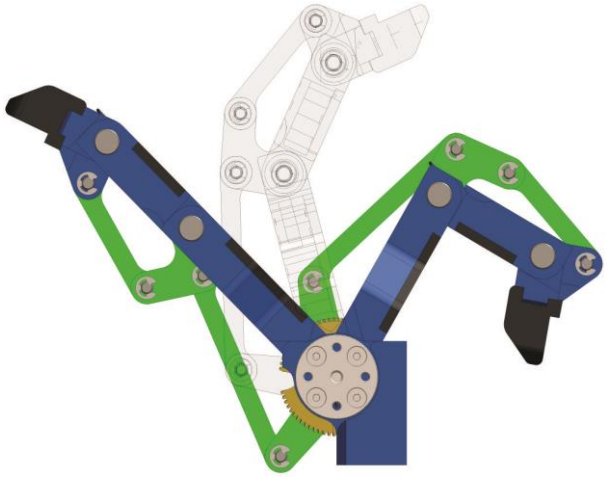


Fig. 6. Workspace of underactuated 3-DoF finger.

V. CONTACT FORCE ANALYSIS

In case of the underactuated finger with 3-DoF, (10) holds if and only if $k_1 k_2 k_3 \neq 0$, which represents the condition of singularity for the matrix J_T, as shown in Fig. 5. There are however other cases, where finger can contact the object when one or two phalanges of the finger are not touching the object, which is shown in Fig. 7, 8,9 and 10.

In order to calculate the contact forces F_1, F_2 and F_3 on the grasping object, it is necessary to separate four cases of possible behaviors between the finger and the object during grasping process.

Case 1: All three phalanges of the finger are in contact with the object, so $k_1 k_2 k_3 \neq 0$, as shown in Fig. 7.

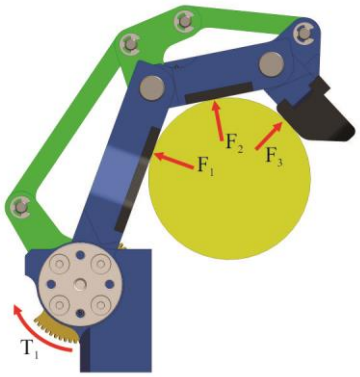


Fig. 7. Representation of 3-DoF robotic finger in case were all phalanges are in contact with grasped object.

The relationship between the actuator torques and contact forces can be derived from (14):

$$\begin{bmatrix} k_1 & k_2 + L_1C_2 & k_3 + L_1C_{23} + L_2C_3 \\ 0 & k_2 & k_3 + L_2C_3 \\ 0 & 0 & k_3 \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \\ F_3 \end{bmatrix} = \begin{bmatrix} T_1 \\ T_{s2} - X_1T_1 \\ T_{s3} - X_2T_{s2} + X_1X_2T_1 \end{bmatrix}. \quad (16)$$

From previous equation, the three contact forces F_1, F_2 and F_3 can be computed by using Equations (17), (18) and (19), respectively:

$$F_1 = \frac{T_1}{k_1} - \frac{(k_2 + L_1C_2)(T_{s2} - X_1T_1)}{k_1k_2} - \frac{(k_3 + L_1C_{23} + L_2C_3)(T_{s3} - X_2T_{s2} + X_1X_2T_1)}{k_1k_3} + \frac{(k_2 + L_1C_2)(k_3 + L_2C_3)(T_{s3} - X_2T_{s2} + X_1X_2T_1)}{k_1k_2k_3} \quad (17)$$

$$F_2 = \frac{T_{s2} - X_1T_1}{k_2} - \frac{(k_3 + L_2C_3)(T_{s3} - X_2T_{s2} + X_1X_2T_1)}{k_2k_3} \quad (18)$$

$$F_3 = \frac{T_{s3} - X_2T_{s2} + X_1X_2T_1}{k_3} \quad (19)$$

Case 2: The proximal and distal phalanges are in contact with the object, which means that parameter k_2 does not exist, while force F_2 is zero, as illustrated in Fig. 8.

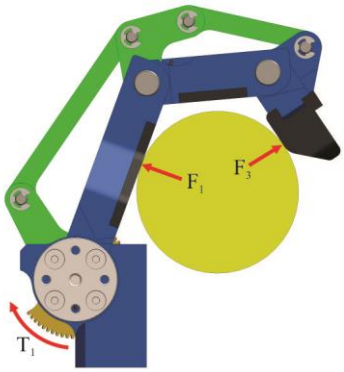


Fig. 8. Representation of 3-DoF robotic finger in case were proximal and distal phalanges are in contact with grasped object.

In Equation (16), the second column and second row in the matrix J_T relating to the medial phalanx are removed, as well as the force F_2 and the torque $\tau_2 = T_{s2} - X_1T_1$ in the vectors F and τ. After removal of aforementioned elements, (16) obtains the following form:

$$\begin{bmatrix} k_1 & k_3 + L_1C_{23} + L_2C_3 \\ 0 & k_3 \end{bmatrix} \begin{bmatrix} F_1 \\ F_3 \end{bmatrix} = \begin{bmatrix} T_1 \\ T_{s3} - X_2T_{s2} + X_1X_2T_1 \end{bmatrix}, \quad (20)$$

where forces F_1 and F_3 are calculated using (21) and (19), respectively:

$$F_1 = \frac{T_1}{k_1} - \frac{(k_3 + L_1C_{23} + L_2C_3)(T_{s3} - X_2T_{s2} + X_1X_2T_1)}{k_1k_3}. \quad (21)$$

Case 3: The medial and distal phalanges are in contact with the object, which means that parameter k_1 does not exist, while force F_1 is zero, as illustrated in Fig. 9.

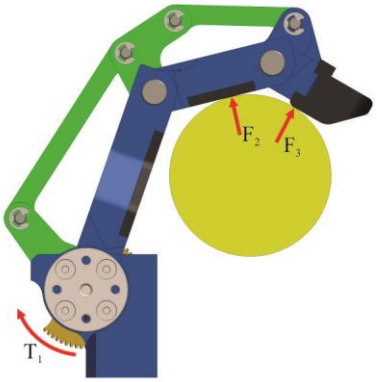


Fig. 9. Representation of 3-DoF robotic finger in case were medial and distal phalanges are in contact with grasped object.

In Equation (16), the first column and first row in the matrix J_T relating to the proximal phalanx are removed. Also, the elements F_1 and τ_1 in the force vector F and torque vector τ are removed. Then (16) becomes:

$$\begin{bmatrix} k_2 & k_3 + L_2C_3 \\ 0 & k_3 \end{bmatrix} \begin{bmatrix} F_2 \\ F_3 \end{bmatrix} = \begin{bmatrix} T_{s2} - X_1T_1 \\ T_{s3} - X_2T_{s2} + X_1X_2T_1 \end{bmatrix}, \quad (22)$$

where F_2 and F_3 are calculated by using (18) and (19), respectively.

Case 4: Only the distal phalanx is in contact with the object, which means that parameters k_1 and k_2 do not exist, while elements F_1 and F_2 of force vector F are zero, as illustrated in Fig. 10.

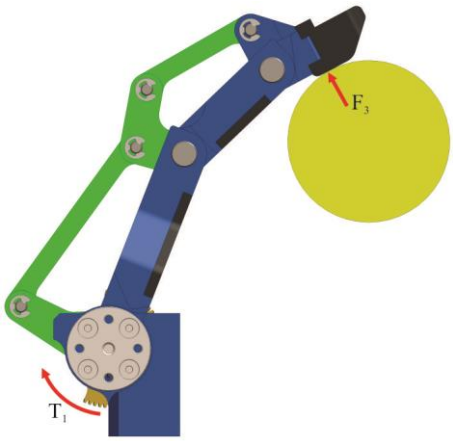


Fig. 10. Representation of 3-DoF robotic finger in case were only distal phalanx is in contact with grasped object.

In Equation (16), the first and second column and row in the matrix J_T relating to the proximal and medial phalanges are removed, as well as the elements F_1 and F_2 of the force vector F, and element τ_2 of the torque vector τ. Then, Equation (16) becomes:

$$k_3F_3 = T_{s3} - X_2T_{s2} + X_1X_2T_1, \quad (23)$$

where F_3 is calculated by using Equation (19).

VI. Conclusion

In this paper, a method for obtaining information of the force acting upon phalanges was presented. This kind of force analysis is important so that researchers and designers of underactuated robotic hands can design proper underactuated finger that has stable grasp of object. These hands have to be robust enough and to have stable grasp if they were to be used in industry setting.

As part of future research this proposed mathematical model for analysis of force distribution on phalanges is going to be supplemented with actual experimental results from experiments with first two finger underactuated CMSysLab Hand.

Future work on CMSysLab Hand will also include simulation of this hand in various softwares and authentication of results through experiments. Design of CMSysLab Hand allows the hand to be equipped with different sensory systems so that experiments can be done.

Future research also must include optimization of four-bar linkages and calculation and optimization of springs that are used to counter inertial effects due to phalanges own weight and friction.

Acknowledgment

This research work is supported by the Serbian Ministry for Education, Science and Technology Development through the projects titled 'Smart Robotics for Customized Manufacturing', grant No.: TR35007, and 'Development and Experiments of Mobile Collaborative Robot With Dual-Arm', grant No.: 401-00-00589/2018-09, which is jointly realized by Anhui University of Technology, Ma'anshan, Anhui, and Tsinghua University, Beijing, from one side, and Institute Mihajlo Pupin – Belgrade (IMP), Faculty of Mechanical Engineering, University of Belgrade (MFB), and Faculty of technical sciences of Novi Sad, University of Novi Sad (FTN), from the other side, supported by a cluster of industrial partners from both sides, as a bilateral scientific project of high national relevance with the Public Republic of China.

References

[1] Bullock, M., Ma, R., Dollar, A.:"A Hand-Centric Classification of Human and Robot Dexterous Manipulation", *IEEE Transactions on Haptics*, vol. 6, pp. 129-144, April-June 2013.

[2] Paul Tuffield, Hugo Elias "The Shadow robot mimics human actions", *Industrial Robot: An International Journal*, vol. 30 Issue: 1, pp.56-60, 2003.

[3] L. Birglen ; C.M. Gosselin "Kinetostatic Analysis of Underactuated Fingers", *IEEE Transactions on Robotics and Automation*, vol. 20 Issue: 1, pp.211 -221, april 2004

[4] Birglen, Lionel, Laliberté, Thierry, Gosselin, Clément M. "Kinetostatic Analysis of Underactuated Fingers" in Underactuated Robotic Hands, Springer-Verlag Berlin Heidelberg, 2008, ch. 3, pp. 33-60.

[5] Xuan Vinh Ha, Cheolkeun Ha and Dang Khoa Nguyen: "A General Contact Force Analysis of an Under-actuated Finger in Robot Hand Grasping" *International Journal of Advanced Robotic System*, vol. 13, Issue 1, December 2015.

[6] McCarthy, J. Michael, Soh, Gim Song "Analysis of Planar Linkages" in Geometric Design of Linkages, Springer-Verlag New York, 2011, ch. 2, pp. 15-53.

End-Effector Cartesian Stiffness Optimization: Sequential Quadratic Programming Approach

Nikola Knežević, Branko Lukić, Kosta Jovanović, Tadej Petrič and Leon Žlajpah

***Abstract*—Control of end-effector stiffness (or the whole mechanical impedance) is still a critical open issue in physical human-robot interaction. This paper presents an optimization approach for shaping the Cartesian stiffness of a robot end-effector. This research targets collaborative robots with intrinsic compliance - serial elastic actuators. Although robots with serial elastic actuators have constant joint stiffness, kinematic redundancy for a specific task (null-space) could be used for robot reconfiguration and shaping Cartesian stiffness matrix while still keeping the end-effector Cartesian position unchanged. The method proposed in this paper is Sequential Least SQuares Programming (SLSQP) algorithm, which presents an expansion of the quadratic programming algorithm for nonlinear functions with constraints. The method is tested in simulations for 4 DOF planar robot. Results are presented for control of the end-effector Cartesian stiffness initially along one axis, and then control of stiffness along both axis planarly - shaping the main diagonal of the end-effector Cartesian stiffness matrix.**

***Index Terms*—Cartesian Stiffness Control, Robot Redundancy, Physical Human-Robot Interaction, Sequential Least SQuares Programming.**

Nikola Knežević – School of Electrical Engineering, University of Belgrade, Bulevar kralja Aleksandra 73, 11020 Belgrade, Serbia (e-mail: knezevic@etf.rs).

Branko Lukić – School of Electrical Engineering, University of Belgrade, Bulevar kralja Aleksandra 73, 11020 Belgrade, Serbia (e-mail: branko@etf.rs).

Kosta Jovanović – School of Electrical Engineering, University of Belgrade, Bulevar kralja Aleksandra 73, 11020 Belgrade, Serbia (e-mail: kostaj@etf.rs).

Tadej Petrič – Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia (email: tadej.petric@ijs.si).

Leon Žlajpah – Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia (email: leon.zlajpah@ijs.si).

Parameterized Occupancy Grid as a Base for Perception Applications in ROS Environment

Stevan Stević, Marko Krnjetin, Nenad Četić and Nives Kaprocki

***Abstract*—Method to accomplish autonomous driving can be roughly described by three main capabilities of such vehicle, which are sensing, computing and actuation. Computing as a main component consists of number of different algorithms grouped by their functionalities such as: perception of environment, path planning and following, etc. These functions arose from well-studied approaches and architectures that proved best for self-driven vehicles. Larger community, mostly because of the commercial nature of the subject, does not know more in-depth details of computing platforms and libraries. *Autoware* is a *Robot Operating System (ROS*) based platform that aims to simplify and optimize the development and prototyping of the autonomous systems and features by providing modules and functionalities of autonomous driving to open-source community. In this paper, we extended the perception module of *Autoware* by implementing occupancy grid as generalized and parametrized *ROS* package that can be used as a base for preprocessing in variety of real-time autonomous applications. We also implemented *Proof of Concept* (*PoC*) applications applying *Test Driven Development (TDD)* and *Clean Code* programming principles in context of cutting-edge *Agile* Automotive software development processes. Performances have been evaluated on embedded NVIDIA Xavier SoC and high-end laptops.**

***Index Terms*—automotive; autonomous driving; ROS; Autoware; occupancy grids; Adaptive AUTOSAR; Agile;**

I. Introduction

Self-driven vehicles are having great impact on current society and will do so, on even larger scale, in the following decade. Original Equipment Manufacturers (OEMs) together with the hardware and software suppliers are creating alliances in order to bring fifth level of autonomous driving, where the vehicle is in complete control without the need for human interactions or even control interfaces. Collaboration between OEMs is very unusual in automotive industry because so far they have been competing against each other to profit as much as possible. This is conditioned by the fact, that vehicles at fifth level are becoming complex robotics systems that require multidisciplinary knowledge and skills from different fields to create such machines. One of the most important partnerships – AUTOSAR [1], is working on *Adaptive Platform* [2], which aims to provide standardized foundation for further development of autonomous driving. Pursuit of this requirement is also making changes in cities and roads infrastructure, industries, laws and lifestyle of society to provide support for deployment of these systems in their full potential. These major changes can be justified by the ultimate goal of the autonomous vehicles. The goal is to create safe driving environment, both for pedestrians and passengers, with reduced traffic accidents and death rate.

Looking at autonomous vehicle as a robot system we can dived it in three major segments: sensing, computing and actuation, as depicted in Figure 1. Challenges as the DARPA Grand Challenge [3] with papers that were published from competing teams [4], [5], [6] were they explained similar approaches, have confirmed that autonomous vehicles can achieve autonomous movement with this workflow.

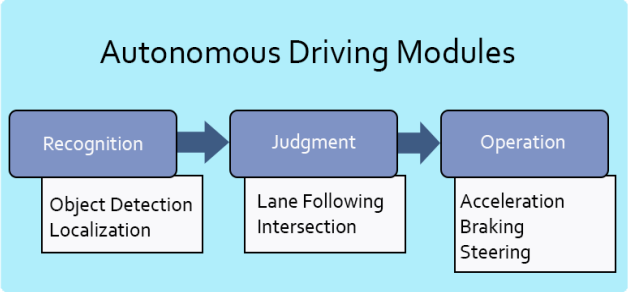


Figure 1. Workflow and examples of autonomous driving modules

Sensing implies usage of large number of sensors like cameras, LiDARs and RADARs for vehicles to be able to gather environmental data and act upon current situation and surroundings. This data is provided to computing modules, which are further divided to scene recognition, path planning, and vehicle control algorithms. Creating reliable image of surroundings in dynamic environment consists of precise simultaneous localization and mapping (SLAM) and detection and tracking of moving objects (DATMO). Many algorithms in this phase use occupancy grids as a preprocessing base for their functionalities. Occupancy grid is a multidimensional random field that maintains stochastic estimates of occupancy state of the cells in a spatial lattice as first introduced and explained by Elfes [7]. Advantages in their usage is reflected by similarities with humans in how environment is perceived,

Stevan Stević is with the RT-RK Institute for Computer Based Systems, Novi Sad and Faculty of Technical Sciences, University of Novi Sad, Serbia (e-mail: stevan.stevic@rt-rk.com).

Marko Krnjetin is with the RT-RK Institute for Computer Based Systems, Novi Sad, Serbia (e-mail: marko.krnjetin@rt-rk.com).

Nenad Četić is with the RT-RK Institute for Computer Based Systems, Novi Sad, Serbia (e-mail: nenad.cetic@rt-rk.com).

Nives Kaprocki is with the RT-RK Institute for Computer Based Systems, Novi Sad and Faculty of Technical Sciences, University of Novi Sad, Serbia (e-mail: nives.kaprocki@rt-rk.com).

better differentiation between moving and static objects, and accurate, real-time representation of dynamic surroundings, also suitable for noisy sensors.

This is followed by decision-making and path planning module where great engineering efforts are made to create more efficient algorithms like [8]. From rule-based and pattern recognition algorithms, to the latest which are relying extensively on artificial intelligence.

Achieving autonomous driving by connecting described components creates large code-base that grows exponentially in complexity. OEMs want to retain good practices that have secured a high level of quality and ultimate software products for years. Through the implementation of the ASPICE standard, and alongside practices like TDD and Clean Code, this requirement is fulfilled. Despite that, in order to meet the challenges posed by the market, they had to adapt software development processes by adopting cutting-edge trends such as scaled agile frameworks and their most influential implementations: LeSS and SAFe in contrast to classical V-model.

In this paper, we implemented simple yet effective library that creates occupancy grids. Library is implemented as a ROS package that can be utilized in different ways with creating nodes that will provide grid data or integrating it with custom nodes already applied in specific applications. In section II, we elaborated papers that have influenced the idea for this paper. Section III presents the solution and explains used software stack. Finally, in last section in order to evaluate usage of this library on embedded platforms in Autoware environment we implemented simple PoC applications and deployed them on embedded SoCs to perform measurements of performance in the real systems.

II. Related Work

Scene recognition manages data fusion from multiple sensors and tends to provide credible information in order for vehicle to be aware of its location and detailed surroundings.

Extensive academic work has been done in terms of vehicle localization and mapping - SLAM. One such solution [9] describes how the map-aided algorithms and classical methods like iterative closest point (IPC) and usage of global navigation satellite systems (GNSS) can be improved with occupancy grids. They solve localization problems imposed by highly dynamic, crowded urban environments and obstacles like trees or buildings that can that interfere with GPS signals.

In [10] and [11] experiments were conducted to explain how to track boundaries using occupancy grids and LiDAR data instead of classical Hough transform or Radon transformation for road segments. Tracking of boundaries plays a crucial role in computational modules to follow. Knowledge of obstacle position is needed when determining logical boundary that represents area through which vehicle can move for creation of local waypoints within path planning modules. Furthermore, this can be used for detection of other vehicles or unexpected obstacles like a running animals or humans crossing the road off the pedestrian crossing.

Free space detection is one of the features required in autonomous parking applications. Interesting research [12] presented how fusion of occupancy grid, created from camera and LiDAR sensors, together with static cameras on parking lot, can be utilized in order to detect parked vehicles and even track movement of pedestrians. Combined with information on free space, autonomous vehicle can park without incidents. Similar solution [13] also addressed detection of parked vehicles but using a different approach. This approach was based on increasingly used classifiers and feed from camera on moving vehicles Preparation work has been done with the test drives and extraction of parked vehicles from frames to create labeled data sets for training of deep convolutional networks.

Artificial intelligence was subject of another work [14] where they have gone a step further and combined occupancy grids and deep convolutional networks to detect and act on moving vehicles.

Discussed papers and many others -, have a wide range of applications and all of them are designed with occupancy grids as preprocessing solution for further implementation of their specific functionalities. We wanted to implement simple generic solution that can be reused in such applications, depending on the nature of approach, to avoid unnecessary waste of effort on re-implementation of same functionality. However, in order for this to be useful, platform that provides comfortable changes in sets of computational algorithms and their communication is imposed indirectly.

Platforms for autonomous driving similar to Autoware also exist, some of them being open-source oriented like Apollo [15] and proprietary ones like NVIDADriveWorks [16]. Adaptive AUTOSAR, however, will probably be the most implemented platform and found in most production-ready vehicles. Platform is oriented towards computational power and connectivity needed to deploy self-driven vehicles. Service oriented architecture makes it very similar to ROS and Autoware and one of the reasons we decided to use these platforms. These platforms are even utilized by OEMs and growing number of robotic projects.

III. Solution

In previous section, we stated the reasons for our decision to use ROS and Autoware. In order to understand their advantages properly, first part of this section covers brief introduction of these concepts. Second part explains the steps of occupancy grids package processing and connection with the rest of the system.

A. Robot Operating System - ROS

Robot Operating System [17] is flexible middleware designed to control robotic systems. Before idea of systems like this, people working on robotic projects often had to reinvent the wheel and implement all components of the system from scratch. ROS provided a way for these kinds of projects to have horizontal architecture. Benefits from

horizontal module organization are seen in code reusability, more efforts spent on optimization and extending of existing layers and focus is on development of new features. Layers are roughly divided in hardware abstraction, middleware and application layers. Furthermore, ROS relies on existing OS and provides structured communication layer for inter-process communication as a main feature. Acts like a *peer-to-peer* distributed system using publisher and subscriber model, as well as service-oriented communication. ROS Master node is main component of the system, responsible for registration and subscriptions of new nodes. Drawback introduced with this principle is the single point of failure, but it is handled in ROS 2 as well as support for real-time systems.

Node in ROS vocabulary represents a single process. Every node has a specific role in the system and communicates with other nodes by broadcasting data on advertised topics, by consuming from subscribed topics or both. ROS offers multilingual client libraries that provide easy integration of nodes with the system. This way, core functionality of most of the nodes is not bound to this middleware and can be reused in other environments, which created flexible and robust platform.

Inter-process communication is done through ROS communication tools - topics, services and actions. These tools have varying types depending on type of data they describe. Types are described using Interface Description Language (IDL), out of which actual code for different language bindings is generated to support multiformity of messages.

Ecosystem represented by ROS, which includes tools, capabilities and large community, created a basis for open-source autonomous platform like Autoware.

B. Autoware

Autoware [18] is an ROS-based open-source platform for autonomous driving developed by Nagoya Institute [19]. Platform architecture is designed to be able to support three main functionalities of self-driven vehicles discussed before. Set of implemented sensor drivers and algorithm nodes made this middleware well suited for urban driving. Necessary functions like 3-D map generation, localization, object recognition and vehicle control are all provided within this platform. Other driving areas can also be supported by easily expanding the functionalities of the platform.

Furthermore, platform is utilizing well-known open source libraries and frameworks like OpenCV or CUDA, which gives us the ability to harness their power with relative ease, and implement complex algorithms and functionalities.

C. Occupancy Grid Package

During the recent years, the occupancy grid framework that was out of the scope because of high resource demands, started to find its use in perception applications. Cause for this change is found in increased computing power and memory limits on embedded platforms introduced with the use of SoCs.

Objects in grids are logical representation of environment that imposes the space affected to be split and that every field carries information about probability of object in that field. This can be utilized in different applications, but they all need different parameters in order to harvest the full benefits.

We implemented ROS package that takes input parameters like dimensions and center of the grid and size of a single cell to create function specific occupancy lattice. In addition, source of the data for the processing can be configured by specifying the correct ROS topic. Library processes data in point cloud format. Point cloud represents a set of data points in space, scanned and produced by lasers, which are LiDARs in this case. Special Point Cloud Library (PCL) [20] is used by the package to optimize processing of these 2D/3D point clouds.

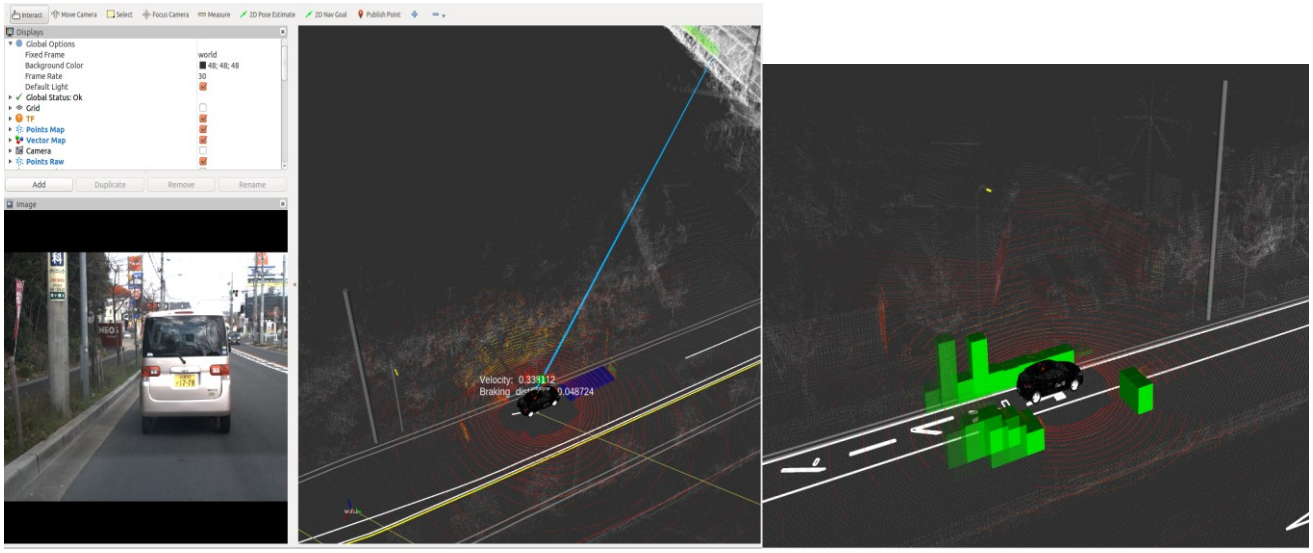
After obtaining data for processing, coordinate transformation of point cloud is required. Center of the recorded point cloud is represented by LiDARs placement, which can be on different parts of the vehicle depending on purpose and area aimed to cover. In order for fusion algorithms to provide realistic model of surroundings, the data processed, has to have uniform coordinate system for correct calculations. This base coordinate frame is also configurable and transforamtion is performed to put the points in given coordinate system.

Computing of probabilistic value for every cell can be resource consuming depending on LiDAR precision, so in order to prevent this from happening point cloud is filtered to region of interest (ROI) which is constructed based on given parameters. Creating ROI consist of two parts which are reduction of point cloud to fit the dimensions of the grid and extraction of ground points. Reducing size was effectively implemented with the usage of previously mentioned PCL that offers range conditions. Conditions were set on all three dimensions to create pass filters. Next step is extraction of ground points as they can affect the accuracy of detection. We used two-step filtering [21] with angle and distance-based filters. This way besides extracting the valid points, we also provided information of ground points, and this could be effectively used for free-space detection.

Preprocessed cloud is the input for the next phase where occupancy grid is created. Probability of objet found in a cell is expressed as a sum of points affected by the cell with the possibility of defining the threshold. We used vector to store the sum of the points and represented cells as vector indexes. Indices are simply calculated after diving current point positions with the cell size. To avoid negative indices whole matrix is displaced to positive range. At the end of processing library can provide occupancy grid structured as a vector carrying information about point occupation in every cell with displaced center and information about ground points. In next section with utilized this data through proof of concept applications.

IV. Evaluation

In order to properly evaluate library grid package in context of automotive development we utilized continuous integration tools in combination with TDD methodology to implement demonstrative ROS nodes working in Autoware environment. Applications also enabled us to execute performance measurements. As LiDARs work on 10MHz - 100 MHz we



a) b)

Figure 2. a) Stopping distance node that uses occupancy grid with center displaced from vehicle's rear axis to detect obstacles
b) Visualization of occupancy grid in context of SLAM

excluded this factor and defined measurements as end-to-end time required to process point cloud messages from rosbag to useful data like probabilistic value of detected object or free space, which is represented by extracted ground points.

Rosbag is a file used to log timestamped messages from all topics and it can be used to reply the scenarios for analysis and debugging. We used open-source visualization tool RViz while implementing perception applications as this tool provides offline visualization of the sensor data from rosbag file while other nodes can process the data as if they are part of the experimental system.

Two applications seen in Figure 2 are implemented to test the library. For the stopping distance application, we implemented the separate node that would publish the grid data. Application was subscribed to the topic advertised by created node and implemented calculations of distance required to stop. Simplified logic is based on vehicle velocity and provided data of object probability in front. Other application was set in context of SLAM and integrated library in the node itself. Both of them published different kind of markers also seen in Figure 2 in order to visualize and informally validate the process in RViz.

During the development as stressed before we wanted to create complete automotive environment, which requires usage of continuous integration tools, high code coverage, traceability and other parts of the automotive processes. Usage of Docker containers with Jenkins as a CI tool and Autoware environment enabled us to illustrate part of the real automotive development workflow on small-scale project.

The goal of these measurements was to gain sense of effectiveness of the library in real embedded automotive hardware as realistic as possible. Furthermore, to illustrate power of today's SoCs, defined measurements were conducted on NVIDIA Drive PX embedded platform with 2xXavier SoCs. In contrast to this, high-end laptop with i7-7500U processor, 16GB of RAM and GeForce 940MX dedicated GPU was used for comparison. Average measurement results are displayed in Figure 3. As expected, with faster execution on embedded platform, we can understand why, the resource consuming autonomous driving can be achieved with current technology.

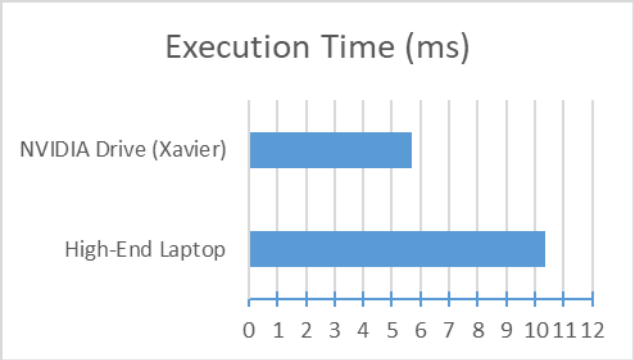


Figure 3. Average execution time of occupancy grid library on different platforms

V. Conclusion

Main goal of this paper and research has been to reflect current trends and problems of the automotive industry. Implementation of different sorts of algorithms is just one aspect of the autonomous driving, but many neglect importance of conducting all work and experiments under demanding automotive processes and principles. Ultimate reason for their existence, when correctly realized, is to facilitate development and minimize human error in the

process in order to deploy safe and secure products to the market. In this paper, we used good practices of automotive development, by using TDD methodology found suitable usage of CI tools, and covered part of the evaluation process. Furthermore, we briefly mention new agile processes that are coming to automotive industry and aim to merge best of classical and agile frameworks.

However, focus was to illustrate importance of modular autonomous platforms that is why we used ROS and Autoware platform to implement simple occupancy grid library wrapped in a ROS package. Occupancy grid carry probabilistic information of object presence in their scope, which makes them suitable both for mapping and detection. Implementing them as part of Autoware makes them usable in many perception applications, which are part of scene recognition module.

Finally, we evaluated this simple occupancy grid module by implementing pair of perception applications and measured times needed in order to obtain valid occupancy grid.

VI. Acknowledgment

This work was partially supported by the Ministry of Education, Science and Technological Development of Republic of Serbia under Grant III_044009_1.

References

[1] "AUTOSAR Partnership", https://www.autosar.org/ [Accessed April 2019].

[2] Adaptive AUTOSAR Release 19-03, AUTOSAR_EXP_ARAComAPI, March 2019.

[3] "Darpa Urban Challenge" https://www.darpa.mil/about-us/timeline/darpa-urban-challenge [Accessed April 2019].

[4] C. Urmson, J. Anhalt, H. Bae, J. A. D. Bagnell, C. R. Baker, R. E.Bittner, T. Brown, M. N. Clark, M. Forms, D. Demitrish, J. M. Dolan,D. Duggins, D. Ferguson , T. Galatali, C. M. Geyer, M. Gittleman,S. Harbaugh, M. Hebert , T. Howard, S. Kolski, M. Likhachev , B. Litkouhi,A. Kelly , M. McNaughton, N. Miller, J. Nickolaou, K. Peterson,B. Pilnick, R. Rajkumar, P. Rybski, V. Sadekar, B. Salesky, Y.-W. Seo,S. Singh, J. M. Snider, J. C. Struble, A. T. Stentz , M. Taylor , W. R. L.Whittaker, Z. Wolkowicki, W. Zhang, and J. Ziglar, "Autonomous driving in urban environments: Boss and the Urban Challenge," Journal Of Field Robotics Special Issue on the 2007 DARPA Urban Challenge, Part I, vol. 25, no. 8, pp. 425–466, June 2008.

[5] J. Leonard, D. Barrett, J. How, S. Teller, M. Antone, S. Campbell, A. Epstein, G. Fiore, L. Fletcher, E. Frazzoli, A. S. Huang, T. Jones, O. Koch, Y. Kuwata, K. Mahelona, D. Moore, K. Moyer, E. Olson, S. Peters, C. Sanders, J. Teo, and M. Walter, "Team MIT Urban Challenge Technical Report," Tech. Rep., 2007.

[6] M. Montemerlo et al., "Junior: The stanford entry in the urban challenge," in The DARPA Urban Challenge: Autonomous Vehicles in City Traffic, M. Buehler, K. Iagnemma, S. Singh, Eds. Berlin, Germany: Springer, 2009, pp. 91-123

[7] A. Elfes, "Using occupancy grids for mobile robot perception and navigation," in *Computer*, vol. 22, no. 6, pp. 46-57, June 1989.

[8] Hiroki Ohta, Naoki Akai, Eijiro Takeuchi, Shinpei Kato and Masato Edahiro, "Pure Pursuit Revisited: Field Testing of Autonomous Vehicles in Urban Areas," 4th International Conference on Cyber-Physical Systems, Networks, and Applications (CPSNA), IEEE, 2016.

[9] Trung-Dung Vu, Julien Burlet and Olivier Aycard, "Mapping of environment, Detection and Tracking of Moving Objects using Occupancy Grids," Journal Information Fusion, 2011, 12 (1), pp.58-69.

[10] K. Thormann, J. Honer and M. Baum, "Fast road boundary detection and tracking in occupancy grids from laser scans," 2017 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI), Daegu, 2017, pp. 348-353.

[11] W. S. Wijesoma, K. R. S. Kodagoda, A. P. Balasuriya and E. K. Teoh, "Road edge and lane boundary detection using laser and vision," Proceedings 2001 IEEE/RSJ International Conference on Intelligent Robots and Systems. Expanding the Societal Role of Robotics in the Next Millennium (Cat. No.01CH37180), Maui, HI, USA, 2001, pp. 1440-1445 vol.3.

[12] O. Aycard et al., "Grid Based Fusion & Tracking," 2006 IEEE Intelligent Transportation Systems Conference, Toronto, Ont., 2006, pp. 450-455.

[13] R. Dubé, M. Hahn, M. Schütz, J. Dickmann and D. Gingras, "Detection of parked vehicles from a radar based occupancy grid," 2014 IEEE Intelligent Vehicles Symposium Proceedings, Dearborn, MI, 2014, pp. 1415-1420.

[14] S. Wirges, T. Fischer, C. Stiller and J. B. Frias, "Object Detection and Classification in Occupancy Grid Maps Using Deep Convolutional Networks," 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, 2018, pp. 3530-3535.

[15] "Apollo", http://apollo.auto/ [Accessed April 2019].

[16] "NVIDIADriveWorks", https://developer.nvidia.com/drive [Accessed April 2019].

[17] "Robot Operating System – ROS", https://www.ros.org/ [Accessed April 2019].

[18] "Autoware", https://www.autoware.org/ [Accessed April 2019].

[19] "Nagoya Institute of Technology", https://www.nitech.ac.jp/eng/index.html [Accessed April 2019].

[20] "Point Cloud Library", http://pointclouds.org/ [Accessed April 2019].

[21] "Ground Filter", https://github.com/CPFL/Autoware-Manuals/blob/master/en/pdfs/ground_filter.pdf [Accessed April 2019].

Simulation of humanoid movements of the NAO robot using the Virtual Robot Experimentation Platform V-REP

Slađan Kantar, *School of Computing, University Union,* Miloš D Jovanović, *Institute Mihajlo Pupin, University of Belgrade*

***Abstract*— This paper is based on the use of a free virtual robot experimentation platform (V-REP) in order to design and test humanoid movements of the NAO robot. The simulation platform contains a complete NAO-robot model with integrated sensors. The document lists the necessary configurations of the environment to simulate the behavior defined in the tabular document. In addition, several scenes have been implemented, in various ways, to illustrate the platform's capabilities used for simulation. All examples are freely available to readers.**

***Key words*— Simulators, V-REP, Humanoid robot, NAO robot**

I. INTRODUCTION

Robotics, as a multidisciplinary field, has been introduced to all levels of education in the last few years. Automation, informatics, electronics, mechanics, physics and mathematics are just some of the sciences which learned in robotics. It is precisely this fact that makes robotics very versatile from a pedagogical perspective. Also, the scope of this field allows the experiments within it to be performed at very simple levels (simple motion, rotation, data input from sensors) as well as at much more complex ones (much more demanding experiments such as simultaneous localization and mapping (SLAM), navigation with image processing and the like). [1]

Until just a few years ago, the word robot was associated with a large and costly machine, which in many cases had to be connected to a computer with cables. This was one of the main reasons why robots stayed away from classrooms. Today, it is very easy to make a small mobile robot with cheap actuators, sensors, motors and built-in systems. Currently, mobile robots are used in many courses in the fields of IT, automation, mechanical engineering and mechatronics.

In most research areas, simulators are very important tools for testing the theorems, ideas, designs, before implementing a real experiment. In the field of robotics, simulators are even more important given that advanced robots are still expensive and there is no possibility that all students can perform their experiments on one device simultaneously. Currently, there is a great number of simulators for different robotics fields, and they all give very good results. Some of the most used ones are: ARGoS, Webots, Gazebo, RFCSIM, and V-REP. [2-8]

The V-REP simulator is one of the most used simulators. This simulator is a versatile, scalable tool for creating 3D simulations in a relatively short period.

The V-REP simulator contains an integrated development environment (IDE) based on a distributed and scripted architecture: each object within a unique simulation can include a script related to it; all scripts are executed simultaneously. The V-REP simulator comes with a large number of examples, models of robots, sensors, actuators for creating a virtual world and interacting with it. Also, the tool allows the creation of new, specific models.

This paper describes a simple simulation of the movement of the NAO robot using the V-REP simulator. The rest of the paper is organized as follows: Section 2 gives a better insight into the used simulator (V-REP); Section 3 describes a humanoid NAO robot; Section 4 shows the design, implementation, and results of the study; Section 5 gives the main conclusions and suggestions for further research.

II. THE V-REP SIMULATOR

The Virtual Robot Experimenting Platform (V-REP), which is the literal translation of the full name of this tool, is a versatile and scalable environment for the simple development of 3D simulations. The simulator was created in 2010 and has been experiencing tremendous growth in the last few years. Today, this tool is one of the most used simulators in education in the field of robotics (the manufacturer offers a free license for education purposes). [9]

The V-REP contains an integrated development environment (Figure 1) based on a distributive control of architecture: each object/model can be individually controlled using a script, an accessory (plugin), a robotic operating system (ROS) node, a remote API client, or some other utility tool. Controllers can be written in different languages such as Java, Lua, Matlab, Octave or Urbi.

The V-REP comes with a large number of examples, a robot models [10], a sensor, actuators for creating a virtual world and interacting with it in real time. Figure 1 shows one of the examples: a Nao robot with two cameras and images of the surroundings which the cameras read.

Slađan Kantar – School of computing, University Union, Nemanjina 12 11000 Belgrade, Serbia (e-mail: skantar12@gmail.com).
Milos D Jovanovic – Institute Mihajlo Pupin, Volgina 15, 11000 Belgrade, Serbia (e-mail: milos.jovanovic@pupin.rs).

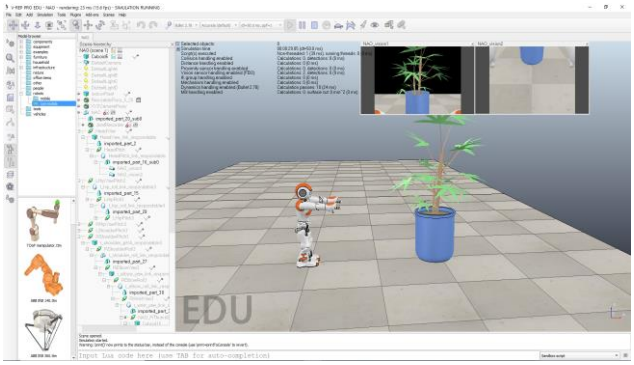


Fig. 1. The layout of the integrated development environment of the V-REP. The figure shows a scene including a NAO robot and an arbitrary object. In this figure, we can also see information that are registered by sensors (cameras) on the NAO robot.

As the NAO robot is one of the more famous humanoid robots, the V-REP platform, in its factory version, contains the model of this robot. Considering this advantage, the simulation for the NAO robot was designed and developed exactly by this simulator. The following section describes the model details.

III. THE NAO ROBOT

NAO robot is a programmable, humanoid robot 58 cm high, developed by the French company Aldebaran Robotics in 2006 (Figure 2). The NAO robot has the ability to move, to see with visual sensor, and to hear with audio sensors integrated into it. In addition to the above-mentioned features, it can talk. Until a few years ago, the NAO robot was highly known in the field of education. It has been officially used in more than 70 countries. [11]

Fig. 2. The look of the NAO robot. The figure shows two robots where several degrees of freedom can be detected

A. *Key components of the NAO robot*

Nao is a small humanoid robot with a simple combination of hardware and software [12-13]: sensors, motors and the NAOqi operating system which includes a software for controlling. The NAO robot has several key components:

- the NAO robot has a body with as many as 25 degrees of freedom. The key elements are electric motors and actuators (figure 3).
- the NAO robot contains a network of sensors consisting of two cameras, four directional microphones, a sonar remote sensor, two IR emitters and receivers, one inertial board, nine touch sensors and eight pressure sensors.
- the NAO robot has several communication devices, including a voice synthesis, LED lights, and two HiFi speakers. As such, the NAO robot can hear, talk and blink with the lights.
- the main processor unit is the INTEL Atom 1.6 GHz CPU, which is localized inside the head and run by the Linux kernel.
- the NAO robot also has a second processor in its torso.
- the NAO robot is run by a 48.6 Wh battery which allows the NAO robot up to 1.5 hours of fully independent autonomy.

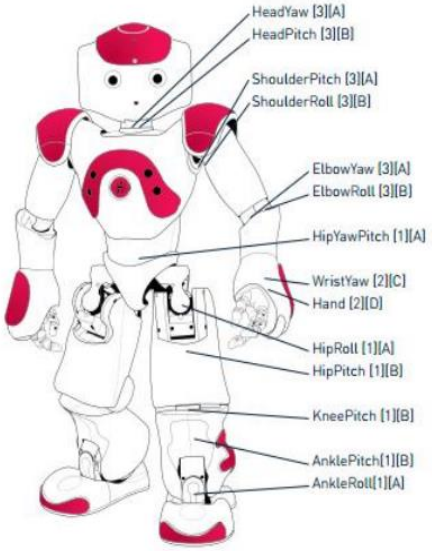


Fig. 3. Joints for NAO robot

B. *The NAO robot model in the V-REP simulator*

The NAO robot model, within the V-REP simulator, contains all the sensors and actuators of the real robot listed in the previous chapter. The model also contains an example of the use of the robot. The example shows a simple robot walk along a closed square path within a scene. From the example, one can clearly see how each of the degrees of freedom is manipulated.

IV. IMPLEMENTATION AND RESULTS OF THE STUDY

The basis of each model within the V-REP environment consists of Joints and Paths that can be assigned to each of the angles. Paths provide for defining of the complex movements of each object.

Also, the basic model of the NAO robot comes with a predefined script for scene recording, as well as the generated

sequence of the joints' positions that allow the robot to walk. By defining the paths, or by associating them with the joints, it is possible to define arbitrary movements, and the position of the joints can be saved in a text document.

Since most existing simulators allow the user to export the state of each of the degrees of freedom at certain time points, the aim of this paper was to implement a library for simulating the movement of the NAO robot defined in the text file. This is exactly what will enable us to run a previously defined simulation in another simulator or on some other way. [14-24]

A. *The appearance of the file with the positions of the joints*

Each line inside the tabular document contains information on each of the degrees of freedom. An example of a tabular document is shown in Figure 4. [32]

Time	RHipYawPitch	LHipYawPitch	RSho
0	0	0	
1.64	-0.366584	-0.366584	
3.12	-0.253067	-0.253067	
4.28	-0.340507	-0.340507	
5.92	-0.0551819	-0.0551819	
6.76	-0.0551819	-0.0551819	
7.52	-0.0551819	-0.0551819	

Fig. 4. Appearance of the file containing the sequence of behavior of the NAO robot. In the picture, we see some of the degrees of freedom defined in Figure 3.

As the NAO robot model within the V-REP environment comes with a predefined script for recording the condition of the joints, the file itself can be made using the simulator itself. The simulations performed in this work are exported from the scenes defined in the Choreography tool, made by the NAO robot manufacturer, which is unfortunately not free.

Note that, at the initial moment, each of the degrees of freedom has an initial value of 0. At that point, the robot has an upright position (Figure 5).

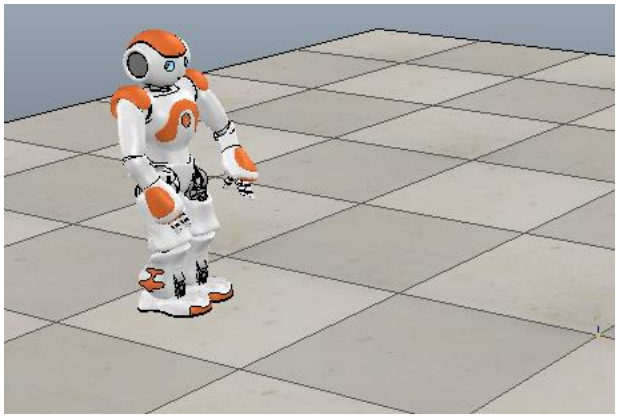


Fig. 5. The initial position of the NAO robot at the moment of t = 0

B. *Creating a file within the V-REP environment*

Within the V-REP simulator, the model of each robot consists of joints and sensors arranged in a tree-structure for easier manipulation. In addition to the existing elements, manually defined paths can be added within such a structure. Paths are added by inserting "Path" objects, or by connecting them to some of the joints.

There are two predefined paths, circular and rectilinear, but the tool allows the creation of complex paths by defining arbitrary points.

An example of a circular path assigned to the neck of the NAO robot is shown in Figure 6.

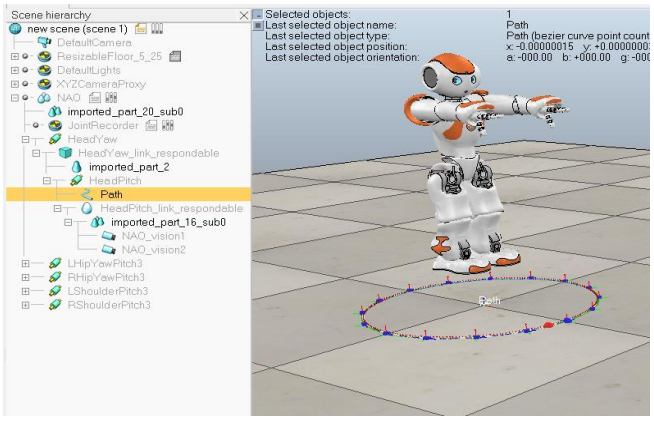


Fig.6 The circular pathway assigned to the joint of the neck of the NAO robot. Within the tree structure we see the Path object, in the hierarchy positioned as a Child to the joint that we want to manipulate.

C. *Process of sequence execution*

Controlling the model within the environment is performed by the scripts written in the LUA programming language. For the purposes of this paper, a script for loading the positions of the joints was written, and their execution. The process of executing the sequences defined in the file is in Figure 7.

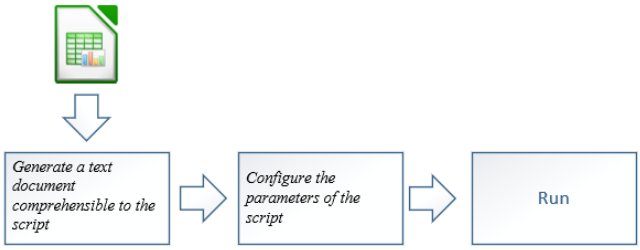


Fig. 7. The step diagram required to perform the simulation defined in the tabular document.

D. *Generating a text document comprehensible to the script*

The main reason for the existence of a tabular document is that it is easy for a man to manipulate it. The LUA programming language does not support simple management of tabular documents; a script is written in the Python programming language, which transfers the tabular document into a plain text document, where each cell is in a new row. With this simple script, available on the GitHub page [26], we

create a text file that is understandable to the NAO robot model within the V-REP simulator.

E. Configuring the parameters of the script within the simulation and starting the simulation

The script defined within the scene, as an argument, receives a path to the file that contains information about the position of the joints, and executes the movements defined therein. Right before starting the simulation, it is necessary to change the *filePath* parameter so that it has a value of the path to the generated file [27]. After setting the path, the simulation is ready to start. [28]

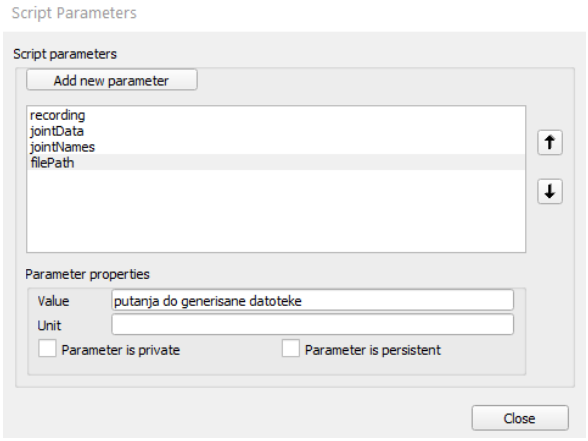


Fig. 8. An example of the window for changing the script parameters.

F. Case studies - waving and push-ups of the NAO robot

For the purposes of this paper, two humanoid movements were implemented: the waving of the NAO robot (Figure 9) [29-31] and push-ups of the NAO robot (Figure 10). [31-33]

Waving is one of the simplest movements of the NAO robot. The basis of this study is the movement of two joints, left shoulder and left elbow, with minimal movement of all other joints in order to gain effectiveness.

The NAO robot waving sequence is generated manually by changing the movement degree for low values. It was noted that this approach to generating the sequence was very slow, and it is almost impossible to generate sequences of more complex behavior by this method.

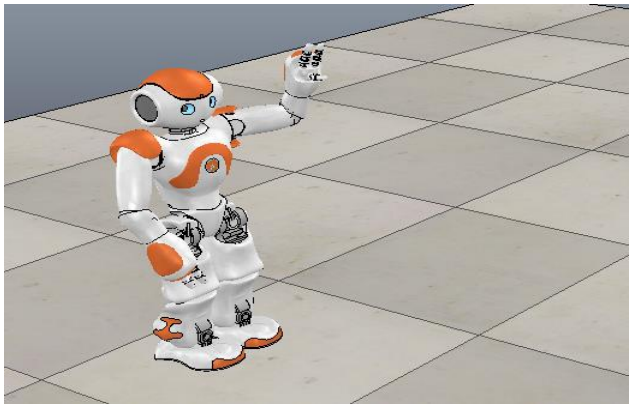


Fig. 9. An example of greeting (waving) of the NAO robot

Making the NAO robot do push-ups is a more complex study. The entire scene consists of several parts: getting into the push up position, doing push-ups, return to the starting position.

The sequence used for this study is extracted from the scene implemented within the Choreography tool. Choreography is one of the most popular simulation tools for NAO robots, so it is possible to find a large number of ready-made simulations on the Internet.

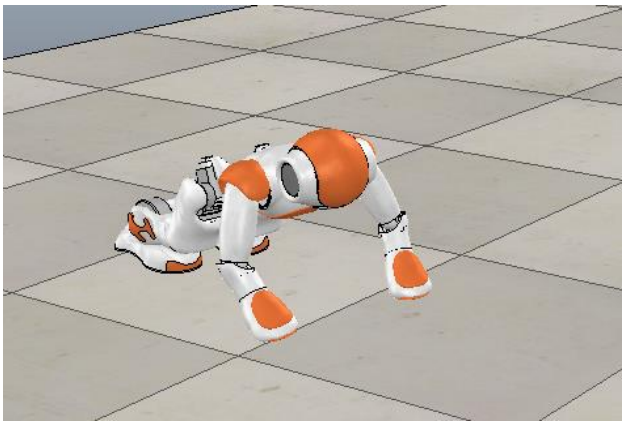


Fig. 10. An example of a NAO robot doing push-ups.

After starting the scene within the Choreography simulator, one can see the great advantages of the V-REP simulator. An example of how the identical simulation within the Choreography simulator looks is shown in Figure 10.

The greatest advantage is given to the influence of gravity, which within the V-REP environment gives a more realistic view of the real behavior of the NAO robot (the difference can be seen in Figures 10 and 11)

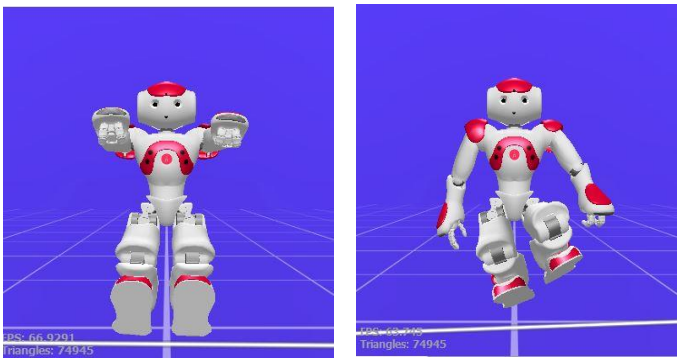


Fig. 11. An example of a NAO robot doing push-ups within the Choreography environment

V. CONCLUSION

In this paper, examples of NAO robot simulation using the V-REP simulator are presented. The examples show simple waving and simulation of push-ups on the original NAO robot model, which comes with the V-REP platform.

As a result of the development of this simulation, tools

were created that enable execution by defining behavior within a tabular document. By following these steps, it is possible to run many simulations without changing the code within the V-REP scene.

As a further research, the plan is to start one of the simulations created in the V-REP simulator on a real robot, and detect the observed differences.

REFERENCES

[1] Robotics [online]. Available: sh.wikipedia.org/wiki/Robotika
[2] V-REP simulator [online]. Available: coppeliarobotics.com/
[3] M. Freese, S. Singh, F. Ozaki, and N. Matsuhira. Virtual robot experimentation platform v-rep: a versatile 3d robot simulator. In Proceedings of the Second international conference on Simulation, modeling, and programming for autonomous robots, SIMPAR'10, pages 51–62, Berlin, Heidelberg, 2010. Springer-Verlag
[4] A. Staranowicz and G. L. Mariottini. A survey and comparison of commercial and open-source robotic simulator software. In Proceedings of the 4th International Conference on PErvasive Technologies Related to Assistive Environments, PETRA '11, pages 56:1–56:8, New York, NY, USA, 2011. ACM.New York, NY, USA, 2011. ACM
[5] ARGoS simulator [online]. Available: argos-sim.info
[6] Webots simulator [online]. Available: cyberbotics.com
[7] Gazebo simulator [online]. Available: gazebosim.org
[8] N. Koenig and A. Howard. Design and use paradigms for gazebo, an open-source multi-robot simulator. In Intelligent Robots and Systems, 2004. (IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on, volume 3, pages 2149–2154 vol.3, Sept.-2 Oct
[9] V-REP licence [online]. Available: coppeliarobotics.com/licensing.html
[10] V-REP models [online]. Available: coppeliarobotics.com/helpFiles/en/models.htm
[11] NAO robot [online]. Available: en.wikipedia.org/wiki/Nao_(robot)
[12] Jovanović, M. Vujović B., Rodić A., Potkonjak B., „Kinematic model of NAO humanoid robot", Int. Journal Robotica & Management, Ed. Robotics Society of Romania, Vol. 19, No. 1, pp. 21-26, ISSN: 1453-2069, June, 2014. [online]. Available: http://www.robotica-management.uem.ro/fileadmin/Robotica/2014_1/Pag_21_Jovanovic.pdf
[13] NAO robot (Characteristics) [online]. Available: wiki.ros.org/nao_description
[14] Gienger, M., Toussaint, M., and Goerick, C. Whole body motion planning–building blocks for intelligent systems. In Motion Planning for Humanoid Robots, pages 67–98. Springer. 2010
[15] Gouaillier, D., Collette, C., and Kilner, C. Omnidirectional closed-loop walk for nao. In Humanoid Robots (Humanoids), 2010 10th IEEE-RAS International Conference on, pages 448–454. IEEE
[16] Gouda, W. and Gomaa, W. Nao humanoid robot motion planning based on its own kinematics. In Press. 2014
[17] Kofinas, N. Forward and inverse kinematics for the NAO humanoid robot. PhD thesis, Diploma thesis, Technical University of Crete, Greece. 2012
[18] Shamsuddin, S., Ismail, L. I., Yussof, H., Ismarrubie Zahari, N., Bahari, S., Hashim, H., and Jaffar, A. (2011). Humanoid robot nao: Review of control and motion exploration. In Control System, Computing and Engineering (ICCSCE), 2011 IEEE International Conference on, pages 511–516. IEEE
[19] J. H. Strom, G. Slavov, and E. Chown.Omnidirectional walking using zmp and previewcontrol for the nao humanoid robot. In J. Baltes,M. G. Lagoudakis, T. Naruse, and S. S. Ghidary,editors,RoboCup, volume 5949 of Lecture Notes in Computer Science, pages 378–389. Springer, 2009.
[20] Ghaffari Jadidi, M., Hashemi, E., Zakeri Harandi, M.A., Sadjadian, H.: Kinematic Modeling Improvement and Trajectory Planning of the Nao Biped Robot. In proceedings of the Joint International Conference on Multibody System Dynamics, Finland, May 2010
[21] Kofinas N., Orfanoudakis E., Lagoudaki G M., "Complete Analytical Forward and Inverse Kinematics for the NAO Humanoid Robot", J Intell Robot Syst, DOI 10.1007/s10846-013-0015-4, 2014.
[22] Siciliano B., Khatib 0., "Springer Handbook of Robotics", Springer, 2008.
[23] Mebius J.E., "Derivation of the Euler-Rodrigues formula for three-dimensional rotations from the general formula for four-dimensional rotations", arXiv General Mathematics 2007. [online]. Available: http://arxiv.org/abs/math/0701759
[24] A. D. Ames. First steps toward automatically generating bipedal robotic walking from human data. In 8th International Workshop on Robotic Motion and Control, RoMoCo'11, Bukowy Dworek, Poland, 2011.
[25] Complete code [online]. Available: github.com/SKantar/NAOSimulation
[26] Script for text document generation [online]. Available: github.com/SKantar/NAOSimulation/blob/master/GenerateFile.py
[27] Configuring the parameters of the script [online]. Available: coppeliarobotics.com/helpFiles/en/scriptSimulationParameters.htm
[28] V-REP scene [online]. Available: github.com/SKantar/NAOSimulation/blob/master/NAOrobot.ttt
[29] Sequence (waving) [online]. Available: github.com/SKantar/NAOSimulation/blob/master/Hey_There.xlsx
[30] Generated text document (waving) [online]. Available: github.com/SKantar/NAOSimulation/blob/master/hey_there.txt
[31] Waving [online]. Available: https://youtu.be/2aKTajzN4po
[32] Sequence (push-ups) [online]. Available: github.com/SKantar/NAOSimulation/blob/master/PushUps.xlsx
[33] Generated text document (push-ups) [online]. Available: github.com/SKantar/NAOSimulation/blob/master/pushups.txt
[34] Push-ups – https://youtu.be/JMQsZ97pH-o

Benefits of residual networks in reinforcement learning using V-rep simulator

Aleksandar Pluškoski, Igor Ciganović and Miloš D. Jovanović

***Abstract*—There are two major problems in the reinforcement learning. One is the stability of the training algorithm. The other is the limited accessibility of the real world training. Reinforcement learning algorithms are notoriously unstable while training because of the very complex loss functions. Residual networks offer smoother gradient descent and thus more stability while training. The other problem is the practicality of the real world training. The reinforcement learning is used for planning and problem solving. While game-like environments are enough for the proof of concept algorithm benchmarks, practical implementation usually requires highly specialized and usually very expensive equipment. It is not practical to have the agent train on the equipment that can easily be damaged. Also the speed and the parallelizability of the physical robots are the additional limiting factors. All these limitations can be addressed by using the simulator. However, setting up the simulator to be used with the external reinforcement learning agent is not a simple task. The solution presented in this paper illustrates the process of setting up the V-rep simulator and presents the results of testing the residual network versus the simple convolutional network.**

***Index Terms*— Artificial intelligence; AI; neural network; reinforcement learning; simulation; v-rep; convolutional network; residual network**

I. Introduction

Reinforcement learning is the process of mapping states to actions. The problem that the agent is trying to solve is usually modelled as a Markov decision process. The goal of the reinforcement learning algorithm is to learn the transitions between the states, as described in [1]. Said states represent the input to the agent and are usually observed through some sensor device. The output is the action that the agent should take to achieve some predetermined goal.

The process of learning is modelled on how we learn by trial and error. The idea is that the agent should try many different solutions and observe the feedback, in a form of a predetermined reward, from the environment. Using that feedback the algorithm should try to find optimal solution by reinforcing the good behaviour and discouraging the bad [1].

Environments with finite amount of states and transitions could be solved using simple tabular methods. When the observation and action spaces are continual the only solution is to introduce the deep reinforcement learning. In deep reinforcement learning instead of a table, the neural network is used to approximate the function that maps points in the observation space to the points in the action space.

In contrast to supervised and unsupervised learning where the agent receives the feedback after every execution, the reinforcement learning agent can execute many steps before any, positive or negative, reward is observed. Reward engineering is the whole distinct field of study in itself and the details are beyond the scope of this paper. After observing the reward the algorithm propagates that reward backwards through time steps. The reward and the state – action pairs are then used as an input for the gradient descent algorithm used to train the neural network.

The new problem which exists in the reinforcement learning but not in the other fields of machine learning, defined in the [2], is the problem of balancing exploration and exploitation. The agent should try to maximize the reward but at the same time it should try to explore and find better solution.

Taking into account sparsity of the reward, complexity of the learning method, possible partial observability and stochasticity of the environment and other problems, it is easy to conclude that the reinforcement learning algorithms can be very unstable. This is described in [3] and a possible solution is presented. In practice that means that the results are often not repeatable. Running the same setup multiple times often yields different results. Often the training gets stuck in local optima or just fails to find any solution.

Many different approaches to partially solve this problem have been implemented, like in [3, 4, 5]. The solution presented in this paper tries to improve on the stability of the training process by making the neural network itself more stable during the training.

It had been shown in [6] that residual neural networks offer more stability during the training and improve capacity of the neural network by allowing much greater depth. The benchmark of using the residual neural network versus classical convolutional neural network in the reinforcement learning algorithm is presented in this paper.

Since the reinforcement learning is used to solve planning problems, it is logical to apply it to the field of robotics. Also it is easy to conclude that trial and error experimentation on the expensive and complex robotic systems is not acceptable. Because of that different tools in the form of simulators have been developed. Although these tools have been much improved in the recent years, it is still not trivial to set up the simulation for the particular problem. This is described in greater detail in [7].

Based on the surveys of different simulators [8, 9] it was decided that the Coppelia Robotics V-REP [10] simulator is to be used for the solution presented in this paper. The important factor that was taken into consideration was the existence of the Python bindings. Since the simulator will

Aleksandar Pluškoski is with the School of Computing, Union University, Knez Mihailova 6/VI, 11000 Belgrade, Serbia (e-mail: aleksandar.plu@gmail.com).

Igor Ciganović is with the School of Computing, Union University, Knez Mihailova 6/VI, 11000 Belgrade, Serbia (e-mail: igor.ciganovic@gmail.com).

Miloš D. Jovanović, Mihailo Pupin Institute, University of Belgrade, Volgina 15, 11000 Belgrade, Serbia, (e-mail: milos.jovanovic@pupin.rs).

have to be controlled from the external framework. An illustration of the aforementioned simulator can be seen in the Fig. 1.

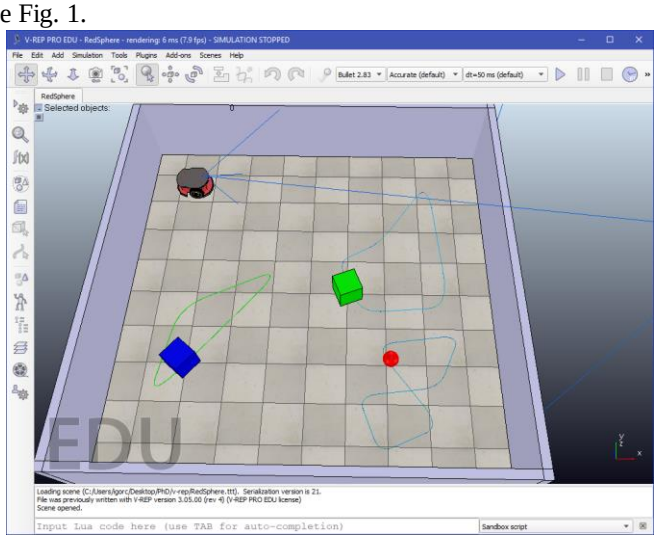


Fig. 1. Overview of the simulator main window.

II. Related work

Stephen James in his Ph.D. dissertation, "3D Simulated Robot Manipulation Using Deep Reinforcement Learning" [7], describes the use of the V-REP simulator with the reinforcement learning algorithm. The algorithm in question is the deep Q learning algorithm. The paper shows, among other things, the effects of the sparse reward on the training process.

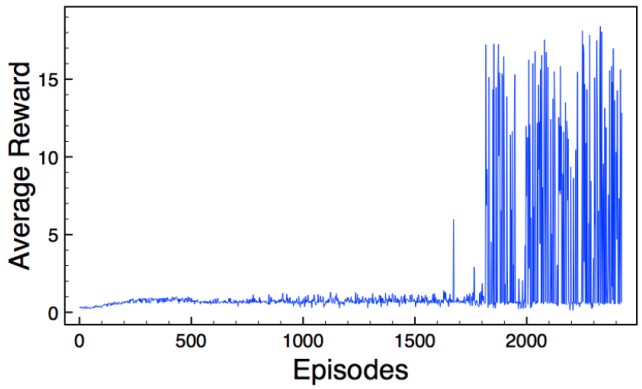


Fig. 2. Average reward by episode in the deep Q learning algorithm. (Taken from [7].)

It can clearly be seen in the Fig. 2 that it takes a lot of time for the agent to find the solution and to start optimizing it. The Fig. 2 shows that only after nearly 1800 episodes the agent finds the expected solution. Taking into account that one episode is the complete simulation run it is easy to conclude that the time needed to find the solution can be a limiting factor. Considering that each simulation run can last up to couple of minutes. The paper further states that the performance and delay of the V-REP simulator proved to be the limiting factor. Hence it was decided that the custom simulation was to be developed in the Unity 3D engine.

Although the paper states that the Python bindings were used to control the V-REP simulator it does not go into great detail in describing how this was used. The solution in this paper tries to detail the setup of the simulator of the use with the reinforcement learning algorithm.

The team from OpenAI led by John Schulman developed the proximal policy optimization algorithm [5] to try to address the issue of the training stability of the reinforcement learning algorithms. In their paper they show that the clipping of the gradients during the gradient descent can have beneficial effects on the stability of the training process. Because reinforcement learning approximates the target of the gradient descent, based on the observed reward, the direction of the target can vary significantly between the iterations. By clipping the gradients a much smoother trajectory can be achieved. Also because the steps taken during the gradient descend are much smaller, samples can be used multiple times. This also gives the better sample efficiency of the algorithm.

The team from Microsoft led by Kaiming He in their paper "Deep Residual Learning for Image Recognition" [6] present the solution to the vanishing gradient problem when training very deep neural networks. The gradient becomes smaller with every layer it passes during the back propagation. After certain number of layers the gradient becomes almost zero and the network becomes impossible to train. To address this issue afore mentioned paper introduces the residual network.

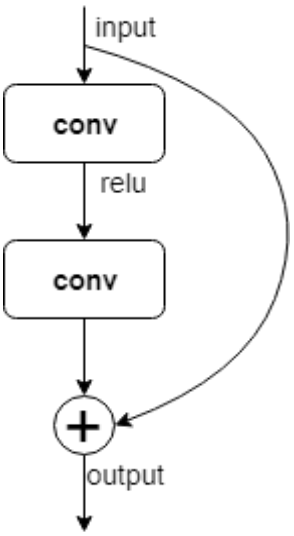


Fig. 3. Diagram of the residual block.

The idea is to divide the convolutional network into blocks. Each block consists of couple of convolutional layers. But the difference is that each block also has its input connected directly to the output as illustrated in the Fig. 3. By building the network like this, the gradients have a way to flow to the deeper layers. On the other hand, all the convolutional layers are still present and the capacity of the network is preserved.

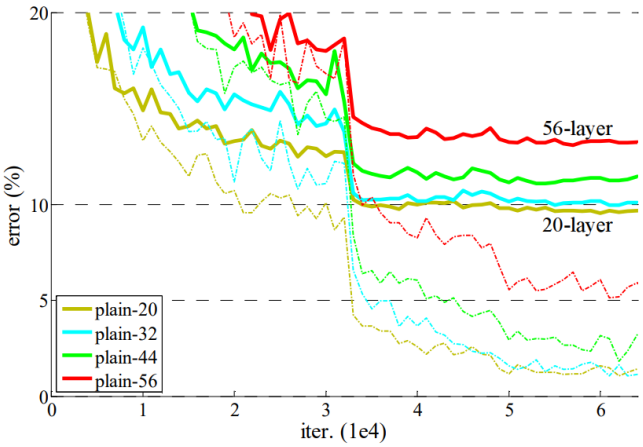


Fig. 4. Convolutional network CIFAR-10 training (dashed lines) and testing (solid lines) error. (Taken from [6].)

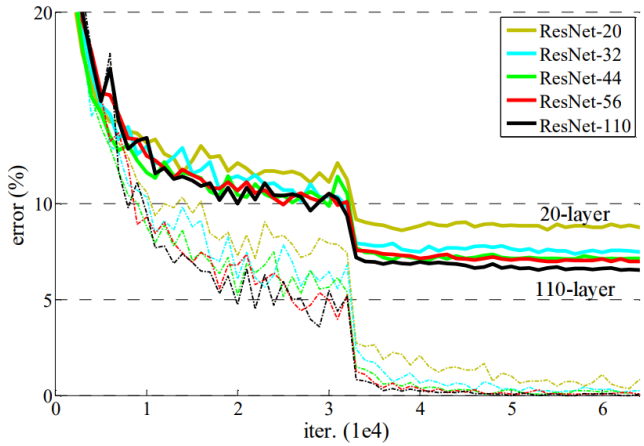


Fig. 5. Residual network CIFAR-10 training (dashed lines) and testing (solid lines) error. (Taken from [6].)

The difference in the learning capacity and the achieved accuracy is illustrated in the Fig. 4 and 5. It can be seen that after around 20 layers of network depth regular convolutional networks start losing accuracy with increased number of layers. In contrast residual networks continue increasing capacity with added depth. Although with the diminished returns after around 56 layers. Suggesting that in order to achieve further improvements the network architecture should be revised.

III. Simulator Setup

When designing the environment for the use with the reinforcement learning algorithm, the inner workings of the algorithm should be considered. Usually the agent will be developed using third party library (i.e. TensorFlow). Because of that the simulator should allow the communication with the external applications. Also, many more efficient algorithms allow for the parallel execution of the replicated agents. Because of the way the reinforcement learning algorithms separate experience recording and learning tasks and the former is much faster than the latter, the simulator should allow for the simulation to be paused and resumed.

In particular the simulator should allow external application to run and stop the simulation, read the data from the environment (sensors and environment variables), write to the actuators and control the execution of the simulation step by step.

Coppelia Robotics V-REP simulator covers all the functionality that was needed for the experiment presented in this paper. They provide the Python bindings for the communication with the external applications. There is no need for any setup on the side of the simulator. All the settings can be provided through Python bindings and the command line parameters. However the V-REP simulator can be used to build the simulation. Although objects in the simulation can be instantiated from the Python code, it is preferable to use the graphical interface to design the scene.

On the other end, the Python language was decided to be used for the solution presented in this paper. Coppelia Robotics provides the library with Python bindings. Three files are needed to be copied to the project working directory. Namely "remoteApi.dll", "vrepConst.py" and "vrep.py" which can all be found in the V-REP installation directory.

Everything else is done through code. Considering how reinforcement learning algorithms run multiple parallel actor tasks, it was decided that it was necessary to have the functionality to start, stop and reset the simulator from the algorithm. The idea being that the learning algorithm should be able to spawn as many workers as needed and control them to collect the experiences (data points). Using Python "sub-process" library it is possible to execute console commands to run the simulator with the command line arguments. Namely "-h" flag that makes the V-REP run headless (without rendering the window) and "-gREMOTEAPISERVERSERVICE_<PORT>_FALSE_TRUE" flag that sets the desired port for the TCP communication, enables or disables remote debugging and triggers. Since it is expected that multiple instances of the simulator will be running in parallel, each of them needs a unique TCP port to communicate with the agent. Also the path to the file containing the simulation scene should be provided and it will be loaded automatically when the simulator process starts. By running the simulator in this way it is possible to run multiple instances locally or remotely even with different simulation scenes.

After the simulator has been started the agent needs to establish a TCP connection to the server running inside the simulator. The Python bindings provide this functionality. Namely the function "simxStart" establishes the connection to the provided IP address at the provided port. After the connection has been made the simulation should be configured to run in the synchronous mode using the "simxSynchronous" function. This allows the external application to run the simulation one step at the time. This is done using the "simxSynchronousTrigger" function, which when called, progresses the simulation to the next step. Then the application can start executing the simulation by calling the "simxStartSimulation" function.

When the simulation is over, the application should first call "simxStopSimulation" function to stop executing the simulation, and only then "simxFinish" to disconnect the TCP connection. After the simulation is closed the process running the simulation should be terminated. During the experiment presented in this paper it was determined that the best way to implement simulation reset functionality is to stop the simulation, terminate the process and start it again.

When the simulation is running it is possible to get the handles to all the objects in the simulation scene, which can then be used to read and write all the relevant variables in the simulation. Detailed explanation of controlling the objects (robots in this example) in the simulation is beyond the scope of this paper. However during the experimentation it was determined that when reading or writing simulation variables "simx_opmode_oneshot" operation mode should be used even though the documentation does not suggest doing this [10], because the other modes either include blocking calls or buffering of the values. Since the simulation is running one step at the time there is no need neither for blocking nor for buffering.

IV. Residual Network Benchmark

With the simulator set up, it was possible to start implementing the intelligent agent for controlling the robot in the scene. The first step is almost always designing the

neural network. Since the reinforcement learning does not have the wide spread practical use because of the training instability, majority of the solutions are proofs of concepts for the learning algorithms. Almost all of them use very simple neural networks because the capacity of the network is not of the great importance. However it has been shown that using the deeper networks can have beneficial effect on transferring the learning between the tasks as described in [4]. As it was stated earlier [6], using the residual neural networks allows for the training of deeper architectures and offers better training stability.

The idea behind this paper was to benchmark the stability of training residual neural network in comparison to the classical convolutional neural network. Since the experiment should test only the stability of the network devoid of the influence of the algorithm, it was decided that simpler (and more stable) supervised learning algorithm should be used.

The human operator would control the robot and the issued commands together with the readings from the sensors on the robot would be recorded. These would then be used as training data for the supervised learning algorithm. Details of this methodology can be found in the paper [11] "Autonomous car driving - one possible implementation using machine learning algorithm".

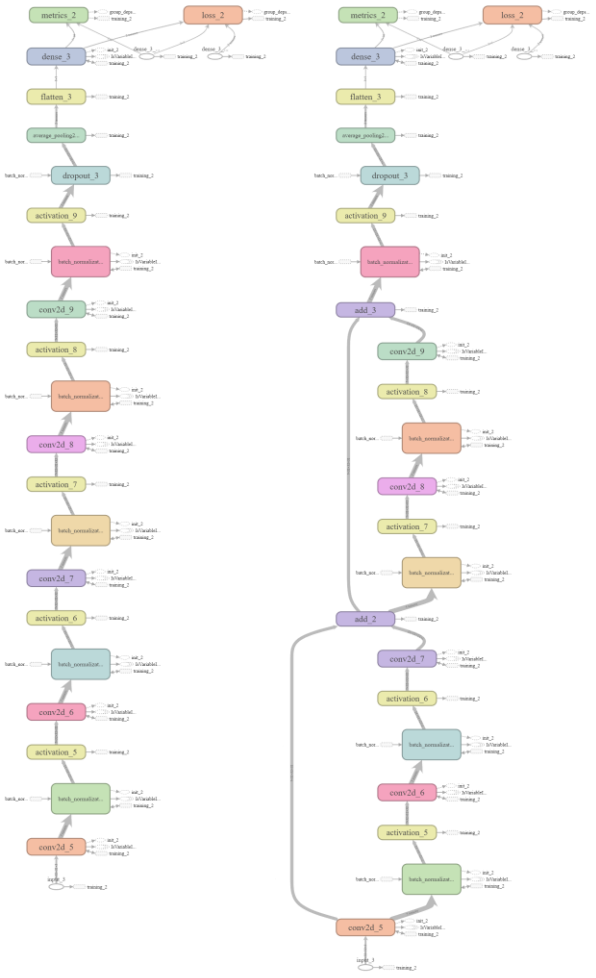


Fig. 6. Diagram of convolutional (left) and residual (right) neural networks 2 residual blocks deep.

With the data collected, two neural networks were constructed, the only difference between them being that one had the skip connections of the residual architecture, illustrated in the Fig. 6. The networks were first trained without the residual blocks to test the validity of the architecture. Then the residual blocks were added one by one and after each step the networks were trained and results recorded. Each residual block consisted of two batch normalization layers and two convolutional layers. In addition to those layers the residual network also had the input of each block connected directly to the output of that block. The testing was done up until 12 blocks, giving 25 convolutional layers in total. During the testing it was decided that going beyond 12 residual blocks was impractical due to hardware limitations.

The training was done with the same dataset and the same number of iterations (epochs) each time. Taking into consideration that the true potential of the residual architecture is in training the networks with more than 50 layers, it can be seen in Fig. 7 and Fig. 8 that the difference exists even with shallower networks.

The experiment was repeated four times. Two times on the local machine and two times on the Google Colaboratory platform [12] to avoid the effects of hardware limitations on the testing results. The results were then averaged for both platforms.

Fig. 7. Values of the loss on the testing dataset.

The Fig. 7 shows the expected decline in the loss value with the increase of the network depth. Also it can be observed that the rate of decline is considerably smoother for the residual network than the convolutional one. This can be seen in particular with the network depths of 5 and 6 where the convolutional network simply failed to converge to an optimal solution. The results shown in the Fig. 7 were obtained on the testing dataset (from the standard 20:20:60 testing: validation: training division of the dataset) after the networks were trained.

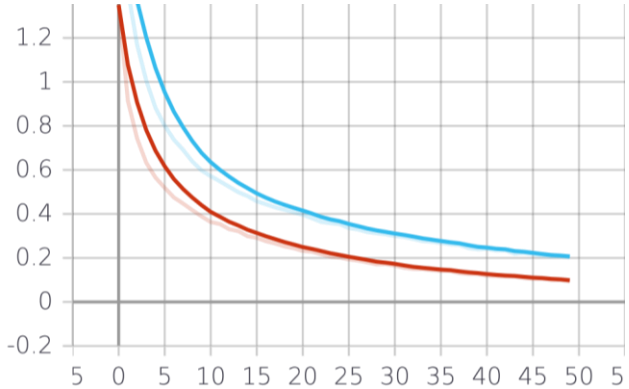


Fig. 8. Values of the loss on the training dataset recorded every time step for the convolutional (blue) and residual (red) networks with the depth of 12 residual blocks.

The Fig. 8 and Fig. 9 show the results of the loss metric on the training and validation dataset respectively on the per epoch basis. It can be seen that the residual network (drawn in red in both charts) converges faster, much more steadily and to a lover optimal value.

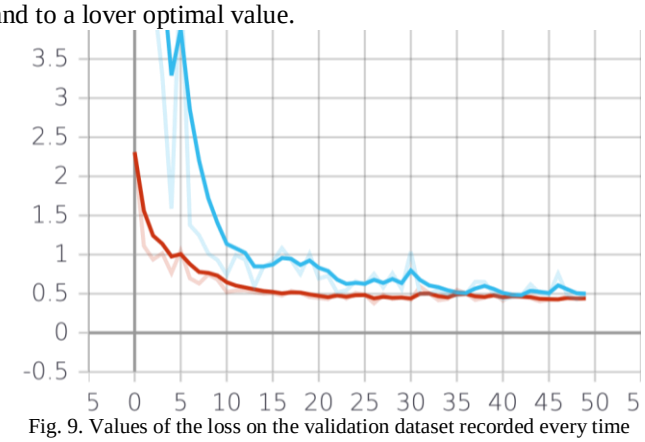


Fig. 9. Values of the loss on the validation dataset recorded every time step for the convolutional (blue) and residual (red) networks with the depth of 12 residual blocks

To calculate the loss value the categorical crossentropy function was used. It is a standard classification loss function. It is used when there is more than 2 categories. The formula for calculating categorical crossentropy is

$$H = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} 1_{y_i \in C_c} \log p[y_i \in C_c].$$

Where N is the number of samples in the dataset, C is the number of classes, $1_{y_i \in C_c}$ is 1 if the i-th sample belongs to the c-th class or 0 otherwise and p is the predicted probability of i-th sample belonging to the c-th class.

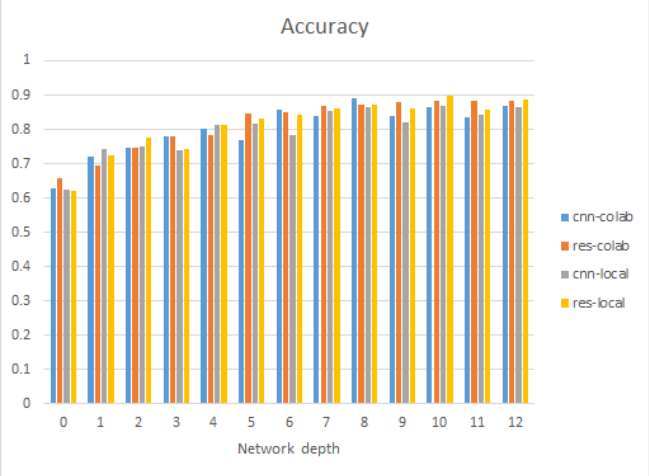


Fig. 10. Values of the accuracy on the testing dataset.

The Fig. 10 shows the achieved accuracy on the testing dataset. The numbers here show the expected correspondence with the loss values shown in the Fig. 7. Again the examples with the depth of 5 and 6 blocks show the unexpectedly low values. The Fig. 10. also clearly shows that for the simple classification tasks with networks of up to 25 layers deep there is no great benefit in using the residual networks. Which agrees with the results in the "Deep Residual Learning for Image Recognition" paper [6]. However better stability is achieved which does have beneficial effects on the stability of the reinforcement learning algorithms.

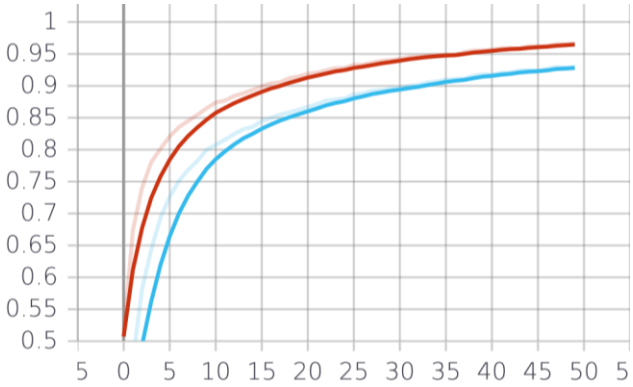


Fig. 11. Values of the accuracy on the training dataset recorded every time step for the convolutional (blue) and residual (red) networks with the depth of 12 residual blocks.

The Fig. 11 and 12 again, expectedly, show faster and more stable convergence to the higher values.

The accuracy metric shown in the Fig. 10, Fig. 11 and Fig. 12 is the categorical accuracy. It is calculated as a simple percentage of accurate classifications averaged across classes. Again, it is the most commonly used metric for the classification tasks when multiple classes exist.

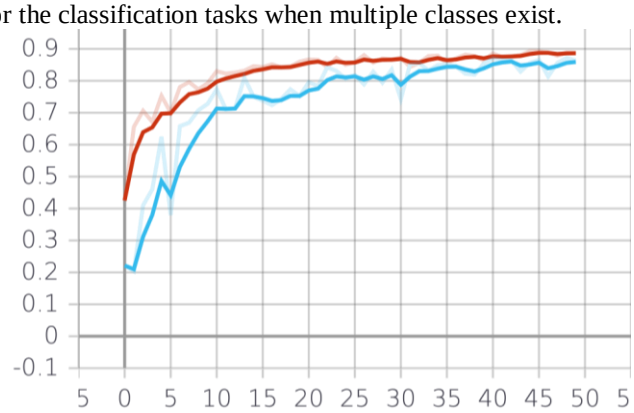


Fig. 12. Values of the accuracy on the validation dataset recorded every time step for the convolutional (blue) and residual (red) networks with the depth of 12 residual blocks.

V. CONCLUSION

It is a matter of time before the reinforcement learning sees more widespread practical use. But for that to happen the algorithms have to mature and become more stable and therefore more predictable, but so do the network architectures. As it was shown in this paper it can be achieved by using the residual neural network instead of the classical convolutional network. This should alleviate some instability form the reinforcement learning process and make it somewhat more stable and repeatable.

Also the majority of the more widely used simulators are currently not designed or very well documented for the use with the reinforcement learning agents. New simulators are being developed particularly for this use and the existing ones are being modified, but currently there is still a lot more work to be done. However, a lot of simple environments, usually in the form of the game console emulators, exist. But those are only usable for the proof of concept benchmarks while designing the algorithms.

ACKNOWLEDGMENT

The research in the paper is funded by the Serbian Ministry of Education Science and technological development under the grants TR-35003 and III-44008.

REFERENCES

[1] Richard S. Sutton, Andrew G. Barto, “Reinforcement Learning” in *Reinforcement Learning: An Introduction*, London, Country: Eng, 2017, ch. 1, sec. 1, pp. 1 – 4.

[2] Shin Ishii, Wako Yoshida, Junichiro Yoshimoto, “Control of exploitation–exploration meta-parameter in reinforcement learning”, *Neural Networks*, vol. 15, no. 4-6, pp. 665-687, Jun.-Jul. 2002.

[3] Berkenkamp, Felix and Turchetta, Matteo and Schoellig, Angela and Krause, Andreas, “Safe Model-based Reinforcement Learning with Stability Guarantees”, Advances in Neural Information Processing Systems, Long Beach, USA, vol. 30, pp. 908-918, Dec. 4-9, 2017.

[4] Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Volodymir Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, Shane Legg, Koray Kavukcuoglu, “IMPALA: Scalable Distributed Deep-RL with Importance Weighted Actor-Learner Architectures”, *CoRR*, vol. 1802-01561, Jun. 2018.

[5] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, Oleg Klimov, “Proximal Policy Optimization Algorithms”, *CoRR*, vol. 1707.06347, Aug. 2017.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, “Deep Residual Learning for Image Recognition”, The IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, vol. 1, pp. 771-778, Jun. 26.-Jul. 1., 2016.

[7] Stephen James, “3D Simulated Robot Manipulation Using Deep Reinforcement Learning”, Ph.D. dissertation, Department of Computing, Imperial College London, London, UK, 2016.

[8] Aaron Staranowicz, Gian Luca Mariottini, “A Survey and Comparison of Commercial and Open-Source Robotic Simulator Software”, International Conference on Pervasive Technologies Related to Assistive Environments, Crete, Greece, vol. 1, pp. 56-63, Jun. 6-8, 2011.

[9] Lucas Nogueira, “Comparative Analysis Between Gazebo and V-REP Robotic Simulators”, *SICA*, vol. 2014, no. 1, pp. 5-10, 2014.

[10] E. Rohmer, S. P. N. Singh, M. Freese, "V-REP: a Versatile and Scalable Robot Simulation Framework", The International Conference on Intelligent Robots and Systems, Tokyo, Japan, vol. 1, pp. 1321-1326, Nov. 3-7, 2013.

[11] Igor Ciganović, Aleksandar Pluškoski, Miloš D. Jovanović, “Autonomous car driving - one possible implementation using machine learning algorithm”, 5th International Conference on Electrical, Electronic and Computing Engineering, Palić, Serbia, vol. 1, pp. 1016-1021, Jun. 11-14, 2018.

[12] Google, “Colaboratory: Frequently Asked Questions.”, 2018, Accessed: Apr. 20, 2019. [Online]. Available: https://research.google.com/colaboratory/faq.htm

ROS as a Rapid Prototyping Platform for LIDAR Based Stopping Distance Monitor

Marko Dragojević, Momčilo Krunić, Ninoslav Jovanov and Nemanja Lukić

***Abstract*—Automotive industry is on the rise in last couple of years, and it's most notable goal is producing fully autonomous vehicles. To achieve this goal, complex software systems are introduced. Development of such systems can be very costly, and possible mistakes made in early stages of system design can be hard to amend. That's why different prototyping methods are used to solve potential problems in early stages of development. Different rapid prototyping platforms are used to achieve this. In this paper we will present one prototype solution for obstacle detection application which uses LiDAR to sense vehicle's environment and perform those detections, and also take appropriate actions to warn driver about possible collision. Technologies used for this solution involve, C++ as a language of choice, ROS and Autoware as a framework and GTest/ROStest which are utilized for verification and validation.**

***Index Terms*— Autonomous vehicles; Autoware; ROS; LiDAR; Rapid prototyping; Adaptive AUTOSAR;**

I. Introduction

Autonomous vehicles are slowly but steadily becoming reality. Automotive industry is working towards achieving this goal from day to day. Autonomous vehicles will drastically reduce number of accidents, but also improve quality of life and social welfare in near future. Diverse alliances (such as AAM (Alliance of Automobile Manufactureres) or AUTOSAR Consortium), made of OEMs (Original Equipment Manufacturers), electronics and semiconductors suppliers and technology suppliers, are formed in order to achieve this goal. Autonomous vehicles are exceedingly complex fusions of software systems, electrical systems and mechanical systems. Complexity of software systems begins to grow exponentially with recent introduction of autonomous vehicles. That is why special attention is given to software domain of a modern vehicle. Development of such complex software systems is severely expensive, and mistakes made in early design phases of software systems can prove to be quite difficult to amend. To avoid such mistakes, prototyping takes significant part in automotive software's lifecycle. To make transition from prototype software to actual automotive grade software, different type of prototyping platforms are used. Majority of such platforms are proprietary. Furthermore, different companies have their own prototyping platforms, which makes potential cooperation of different companies complicated. Open source prototyping platforms could potentially solve this problem. One such platform is used as a base for this solution. Autoware [1] is ROS (Robot operating systems) [2] based rapid prototyping platform. It's abundant set of features can be used to produce different kinds of application level software for autonomous vehicles. Furthermore, Autoware platform has interfaces provided to application software similar to those in Adaptive AUTOSAR [3] platform, which presents new platform for Autonomous vehicles and is successor of Classic AUTOSAR platform used in traditional automotive software applications. Despite those similarities, these two platform have their differences where main one is that Autoware is open-source based platform, while Adaptive platform is strictly proprietary and OEM specific, even though platform's specifications are made public by AUTOSAR consortium. Additionally Adaptive platform covers some additional functionalities, which are not present in Autoware, such as execution management and vehicle's diagnostics. In addition to that, underlying implementations may differ severely as Adaptive platform specifies different underlying communicational protocols than ones used in Autoware platform.

Prototype application presented in this paper is used for obstacle detection based on data acquired from LiDAR scans and vehicle stopping distance monitoring based on current vehicle's velocity and distance to detected obstacle.

The rest of this paper is organized as follows: Section 2 describes Autoware platform and related work. Section 3 presents the detail of implemented application. Section 4 describes measurements of applications performances, and comparison to some other similar applications. Section 5 concludes this paper and introduces some possible future works.

II. Related work

Autonomous vehicles as complex systems can be abstracted into sensing, computing and actuation modules. Sensing devices such as LiDARs and cameras allow vehicle to acquire data, which is later used to percept it's environment. Actuation modules handle actual vehicle controls, typically steering wheel and pedals (throttle, break). Computation modules represent the actual "brain" of the vehicle and it

Marko Dragojević is with the RT-RK Institute for Computer Based Systems,Novi Sad and Faculty of Technical Sciences, University of Novi Sad, Serbia (e-mail: marko.dragojevic@rt-rk.com).

Momčilo Krunić is with the RT-RK Institute for Computer Based Systems, Novi Sad and Faculty of Technical Sciences, University of Novi Sad, Serbia (e-mail: momcilo.krunic@rt-rk.com).

Ninoslav Jovanov is with the RT-RK Institute for Computer Based Systems, Novi Sad and Faculty of Technical Sciences, University of Novi Sad, Serbia (e-mail: ninoslav.jovanov@rt-rk.com).

Nemanja Lukić is with the RT-RK Institute for Computer Based Systems, Novi Sad and Faculty of Technical Sciences, University of Novi Sad, Serbia (e-mail: nemanja.lukic@rt-rk.com).

usually involves submodules used for perception of the vehicle's environment, submodules used for planning, and submodules used for decision making based on data acquired from sensing modules. Autoware software model is shown in Figure 1.

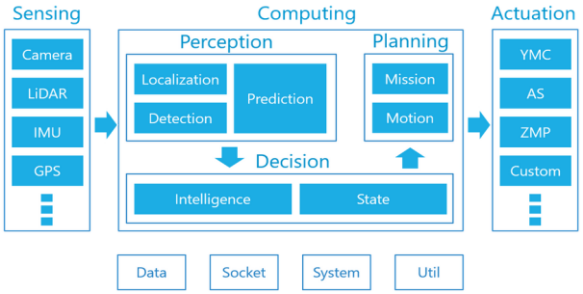


Figure 1. Autoware software stack

Autoware provides this rich set of modules and it can be utilized as an admirable framework for development of autonomous driving applications. Sensors record environmental information, and that information serve as an input to the computing modules. Artificial intelligence plays significant role in perception of vehicle's environment in form of object detection algorithms. Decision portion of the data flow usually takes the form of state machine. And planning portion uses information about detected objects and information acquired form decision making cluster and utilizes it to plan the optimal actions that needs to be executed. More detailed description of Autoware platform can be found in [4].

As already mentioned, Autoware prototyping platform is based on ROS which is a framework for developing robot based applications. ROS is designed to provide means for developing of modular applications in distributed manner. In ROS software is abstracted as nodes and topics. Nodes represent software components that provide some kind of functionality (e.g. Processing LiDAR data to generate 3D map of given environment). Topics, on the other hand serve as a communicational interfaces, which allows nodes to exchange messages. This communicational paradigm is known as Publisher/Subscriber design. When a node publishes message on certain topic, all nodes that are subscribed to that topic get notified about existence of a new message, and the data stored in message becomes available to them. Similar communicational pattern is also used in Adaptive AUTOSAR platform and that is why Autoware, as ROS based platform, could prove to be excellent candidate for prototyping applications, which will later be integrated into Adaptive software stack. ROS also provides rich set of visualization tools such as RViz and data driven simulation tool ROSBAG, which gives user possibility to record and playback ROS traffic on specified topics that was previously recorded. Autoware also utilizes abundant set of well known open-source libraries for different kind of processing. Point Cloud Library (PCL) [5] that is used for processing LiDAR scans and 3D mapping. Beside this, computer vision library for image processing, OpenCV [6] is also used. Furthermore, deep learning frameworks are also utilized e.g. Caffe [7]. Notable part of calculations performed during processing relies heavily on GPU computation frameworks such as CUDA [8] which is developed by NVIDIA.

Utilizing this rich set of software packages provided by Autoware, our prototype application is developed. By using Autoware we ensure certain compatibility of our prototype application with future application as a part of Adaptive AUTOSAR stack. Application should enrich functionality of Autoware platform with additional module that is capable of monitoring vehicle stopping distance based on its velocity and warn or even take an action if there is LiDAR detection in the region of interest. In case of this solution, region of interest represents the lane in front of the vehicle. LiDAR is chosen as a main sensing tool because of it's advantages over cameras, which underperform in extreme conditions, such as situations when weather is poor. To the best of our knowledge, such functionality is not provided in current Autoware stack. This solution could be combined with existing solutions, such as ACC (Adaptive Cruise Control) [9] or FCD (Forward Collision Detection) [10] [11]. Additionally, one existing solution [12] tackles similar problem, which involves pedestrians detection based on inputs from multiple sensors and AEB (Autonomous Emergency Braking) actions. In addition to LiDAR based obstacle detection and emergency brake actions, [13] tackles additional complex problem of time delays which are present in communicational link of teleoperated vehicles. Unlike these solutions, our current solution does not takes those additional actions into account but instead focuses on stopping distance monitoring and providing appropriate warnings. Additionally, outputs from our application could be used as one of the inputs to other systems which can take breaking actions.

III. SOLUTION

In order to provide functionality of stopping distance monitoring, information about vehicle's velocity and LiDAR data acquired from Autoware environment is processed. Top level architecture of solution is shown in Figure 2.

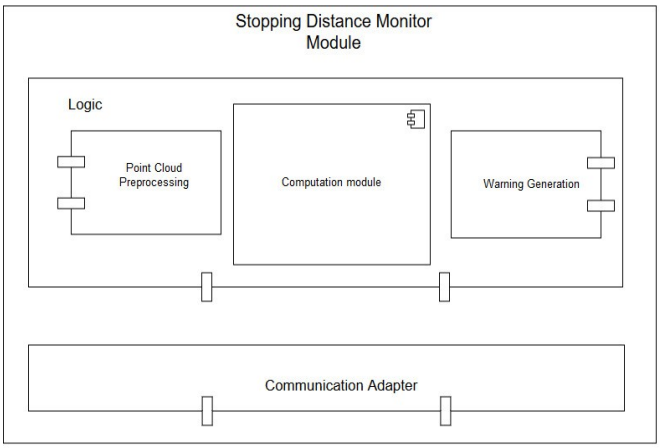


Figure 2. Stopping Distance Module

Proposed solution is organized in several submodules, where each of them provides certain type of functionality. This kind of organization grants us modularity, and loose coupling of our components, which makes it easier to modify each of the submodules, if such action is needed. As this solution is actually a prototype of Adaptive AUTOSAR application, modifications of submodules shall be necessary, when final specification of Adaptive AUTOSAR platform becomes available.

Communication Adapter submodule represents a abstraction of communicational module which allows this application to communicate with its environment. In this case, environment implies to other ROS nodes within Autoware platform, where these nodes provide different kinds of inputs for our application (node), such as point cloud data and vehicle speed estimation. Application can use functionally of Communication Adapter to acquire mentioned data and other relevant information, without actually relying on any of the ROS specific functionalities. In order to use this application in Adaptive AUTOSAR context, only submodule which should be modified is Communication Adapter, with the rest of application unchanged.

After point cloud data is acquired, it's passed to the Point Cloud Preprocessing submodule. Point Cloud Preprocessor will then filter acquired point cloud data, to isolate ROI (Region of interest) and remove other needless detections. Such preprocessing is done by utilizing PCL. By reducing size of point cloud, we relieve computation submodule of additional intense processing. Also, Point Cloud Preprocessor module isolates usage of PCL, so if PCL is not available as a part of Adaptive AUTOSAR platform or different point cloud processing library needs to be used, only this submodule should be modified.

Computation submodule does the work of finding the nearest cluster of points within the preprocessed point cloud, which represents the actual ROI and marks that cluster as a detected obstacle. Organizing detected points in clusters is done with placing each of the detected points into appropriate cells of logical grid, which corresponds to the ROI. After all points from ROI are assigned to the appropriate cell, each cell's point count is checked and if it passes certain threshold, location of that cell is considered as a location of detected obstacle. Furthermore, this submodule keeps track of current vehicle speed, which it acquires from other ROS nodes within Autoware platform, through the Communication Adapter as a medium. Latest received value of vehicle is used to calculate stopping distance in addition to vehicle's maximum deceleration, which is specified as a parameter and distance from the vehicle to the nearest grid cell with point count that surpasses the threshold.

Calculated stopping distance is passed to the Warning Generation submodule, which generates visualization markers which are passed to the ROS visualization tool. Generated markers contain information about free space in front of the vehicle, and the actual length of the stopping distance. If distance to detected obstacle is smaller then calculated stopping distance, warning is generated. Warning Generation submodule utilizes functionality Communication Adapter submodule, to send those generated markers to it's environment. In current prototyping phase of application development, only visualization markers are generated, but in later more matured versions of application as a part of Adaptive stack, this submodule could be modified to generate different types of warnings, or event request adequate actions.

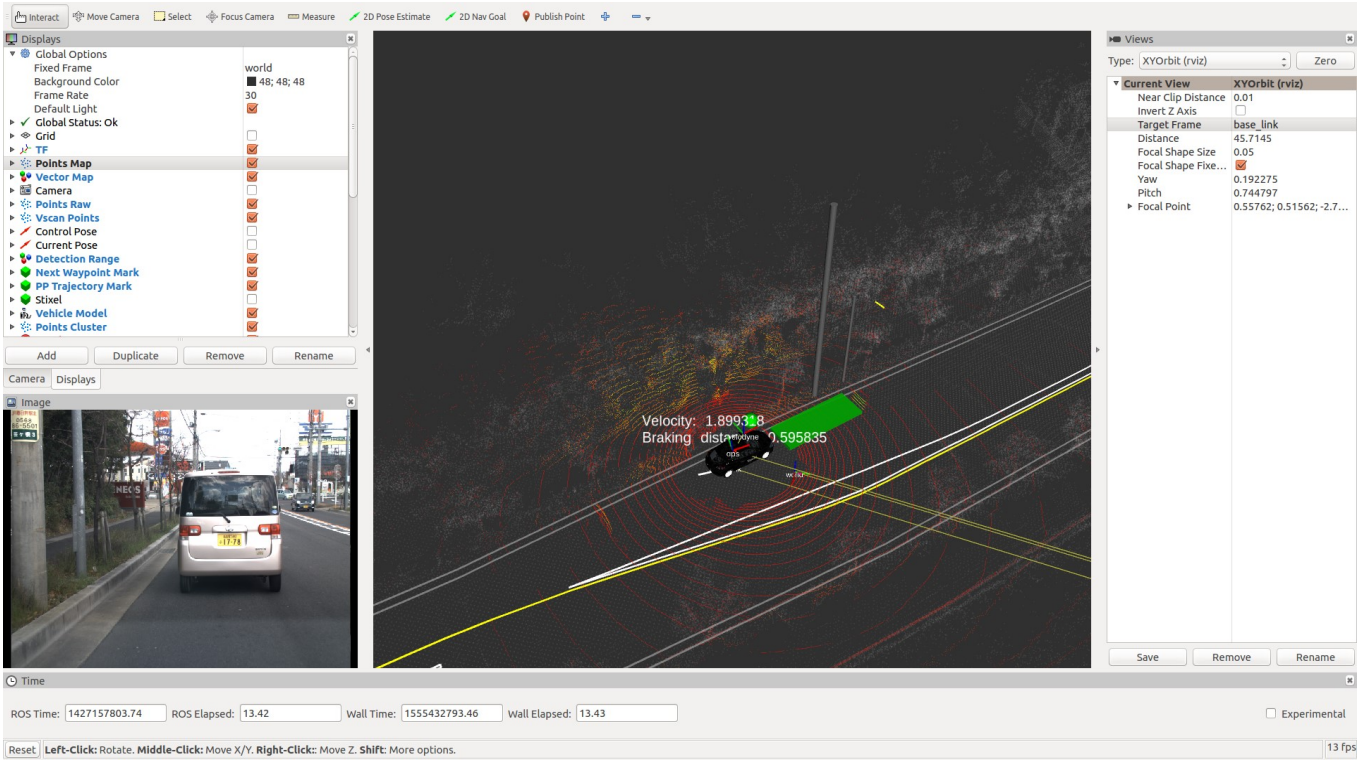


Figure 3. Autoware evaluation environment

IV. EVALUATION

As already previously mentioned, this application is prototype for future application which will be part of Adaptive AUTOSAR stack. Evaluation environment used during prototyping is shown in Figure 3.

Visualization tool RViz was especially useful during development of this solution, because that tool is capable of displaying different types of visualization markers, such as green one in front of the vehicle in Fig. 3 which represents free space in front of the vehicle. In bottom left side of Fig. 3 feed from frontal camera is shown, which gives us an option to visually confirm object detections made by Computing submodule. Data such as internal representation of estimated vehicle's linear velocity, feed from frontal camera or point cloud data is acquired from the set of nodes within Autoware environment. These nodes acquire raw data from ROSBAG file which was recorded during real-life test drive of a test vehicle that had different kinds of sensors attached. These ROSBAG files can later be used to replay acquired sensor data, providing us with valuable simulation scenarios, that gives us opportunity to develop applications similar to one described in this paper.

Described prototype application is developed by utilizing TDD (Test Driven Development) [14] [15] methodology, which grants us with code coverage. Considering this in synergy with CI (Continuous Integration) systems like Jenkins [16] and Autoware's already established support for Docker [17] containerization, we get complete platform that enables us for rapid prototyping, and reduces actual cost of traditionally expensive automotive grade software development. Furthermore, it allows us to focus on providing more complex functionality of software systems rather than focusing on development processes.

Proper functionality of prototype application was verified by means of unit testing. Each of the submodules within application was properly tested, by writing ROStest/GTest test cases which cover different possible scenarios. Beside unit testing, comprehensive testing on integration level was also performed with a goal of ensuring quality of software. Proper integration of our prototype application within the Autoware platform was necessary in order to test its actual functionality. In order to achieve this goal, additional testing ROS nodes were developed to simulate the inputs for our application and to acquire, but also verify its output. With this amount of test cases, we were granted the flexibility of changing functionality of each and everyone of submodules, without concern of accidently breaking down existing functionality. Unit testing results for one of the test cases is provided in Table 1. Execution times displayed in Table 1 represent test cases execution times and do not refer to execution time of application's functionality.

Name	Execution time	Verdict
TestAlphaCalculation	258 ms	Pass
NonEmptyGrid	355 ms	Pass
GenerateROINonEmptyGrid	347 ms	Pass
GenerateROIEmptyGrid	457 ms	Pass
CheckROISizeParam	509 ms	Pass
FindClosestPointValid	253 ms	Pass
FindClosestPointInvalid	305 ms	Pass
CheckClosestPoint	242 ms	Pass
StoppingDistanceCalculation	263 ms	Pass
CalculateVelocityAverage	240 ms	Pass
FilterTestOffset	267 ms	Pass
FilterTestNoOffset	235 ms	Pass
FilterInvalidTestInput	369 ms	Pass
FilterValidTestInput	235 ms	Pass
PointCounting	369 ms	Pass
PublishROIInvalidInput	233 ms	Pass
PublishROIValidInput	371 ms	Pass

Table 1 – Testing results

With that in mind, it is safe to say that we have achieved our goal of establishing a system which would grant us an ability to do rapid prototyping of our application in its early stages of design.

V. CONCLUSION

In this paper, we have shown that open-source prototyping platform, such as Autoware, can be used in rapid prototyping of autonomous vehicles applications. With utilizing Autoware, we were granted a chance of early feasibility check of our application thus reducing chance of inducing flaws in design of our application. Furthermore, new functionality is introduced in form of stopping distance monitoring application based on vehicle's velocity and LiDAR detections, that can warn driver or even, in later stages of development, take appropriate action in form of braking.

As this application prototype was developed by considering some of the applicable requirements defined in ISO26262 [18] standard such as providing full vertical and horizontal traceability, providing specification for software architecture and also integration and testing strategy, final Adaptive application could be more easily adapted to satisfy safety standard to given extent.

ACKNOWLEDGMENT

This work was partially supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia, under grant number: III_044009_2.

REFERENCES

[1] "Autoware: Open-source software for urban autonomous driving," https://github.com/CPFL/Autoware, accessed: 2019-04-01.

[2] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Y. Ng, "ROS: an open-source Robot Operating System,"

[3] Simon Fürst, Markus Bechter, "AUTOSAR for Connected and Autonomous Vehicles: The AUTOSAR Adaptive Platform".

[4] Shinpei Kato, Shota Tokunaga, Yuya Maruyama, Seiya Maeda, Manato Hirabayashi, Yuki Kitsukawa, Abraham Monrroy, Tomohito Ando, Yusuke Fujii and Takuya Azumi "Autoware on Board: Enabling Autonomous Vehicles with Embedded Systems"

[5] "Point Cloud Library (PCL)," http://pointclouds.org/, accessed: 2019-04-02.

[6] "OpenCV," http://opencv.org/, accessed: 2019-03-30.

[7] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, “Caffe: Convolutional architecture for fast feature embedding,” arXiv preprint arXiv:1408.5093, 2014.
[8] “Compute Unified Device Architecture (CUDA),” https://developer.nvidia.com/cudazone, accessed: 2018-02-14.
[9] Junmin Wang, R. Rajamani, “Adaptive cruise control system design and its impact on highway traffic flow”
[10] Zhipeng Di, Dongzhi He, “Forward Collision Warning System based on vehicle detection and tracking”
[11] E. Dagan, G.P. Stein, O. Mano, Amnon Shashua, “Forward collision warning with single camera”
[12] Hyuck-Kee Lee, Seong-Geun Shin, Dong-Soo Kwon “Design of emergency braking algorithm for pedestrian protection based on multi-sensor fusion”
[13] Johannes Wallner, Tito Tang, Markus Lienkamp “Development of an Emergency Braking System for Teleoperated Vehicles Based on Lidar Sensor Data”
[14] Hussan Munir, Krzystzof Wnuk, Kai Petersen, Misagh Moayyed, “An Experimental Evaluation of Test Driven Development vs. Test-Last Development with Industry Professionals”
[15] Burak Turhan, Lucas Layman, Madeline Diep, Forrest Shull, “How Effective is Test Driven Development”
[16] “Jenkins”, https://jenkins.io/doc/, accessed: 2019-04-06.
[17] “Docker”, https://www.docker.com/, accessed: 2019-04-06.
[18] “ISO 26262-1:2011”, https://www.iso.org/standard/43464.html, accessed 2019-04-07

Never-Ending Ontology Learning Approach Applied to Biomolecular Function Prediction

Nenad Petrović and Milorad Tošić

***Abstract*— Biomolecular function prediction is an important task in biomedical sciences. However, most of the existing solutions for automated function prediction are based on structural sequence similarity methods that do not always give satisfying results, as structural similarity does not necessarily lead to functional similarity. In this paper, we propose a semantic approach for biomolecular function prediction based on never-ending learning paradigm applied to ontology learning and knowledge extraction. PubMed biomedical literature repository is used for this purpose with objective to refine the results returned by BLAST that is performing the calculation of similarity using local sequence alignment. For the refinement of results, the molecule similarity on semantic level is also considered. The implementation of ontology learning and knowledge extraction modules is presented, while the prediction-related module is still a prototype in progress.**

***Index Terms*—bioinformatics; ontology learning; text mining.**

I. Introduction

Every day, organizations all over the world generate enormous amount of data in textual format about various topics – including emails, articles, books to reports and experiment results. Nowadays, the storage capacity of computers is more than enough to enable storage of all the content without deleting anything. However, one of associated challenges is ability to manage the stored information considering mutual relations [1], as it does not only ease the information retrieval but also provides ability to generate new knowledge using reasoning and inference mechanisms. A possible approach to organize information in such way in computer science is to use ontologies.

Ontologies are formal and explicit specifications in form of concepts and relations of shared conceptualizations within a particular domain of interest [2]. They are used as metadata schemas, providing a controlled vocabulary of concepts, each with explicitly defined and machine-processable semantics [3]. In this sense, they govern the way the corresponding knowledge bases are populated [4]. However, the creation of ontologies requires additional efforts leading to an initiative to automatize the process of ontology creation. Ontology learning is the process of deriving high-level concepts and relations from information in order to construct an ontology in automatic or semi-automatic way [1, 4].

Amount of available biomedical literature, texts and experimental data is increasing at exponential rate. The same applies to the number of biomolecules (proteins and genes) being identified and their structure determined, due to advances in high-throughput sequencing techniques [5]. Furthermore, it is needed to extract useful knowledge instead of simple information from these large data volumes in order to answer critical question, such as what do all these newly discovered biomolecules do. However, it is impossible to keep up with the data volume growth rate by only performing the manual annotation of discovered molecules. For that reason, the scientists have been turning to sophisticated computational methods when it comes to annotation of huge volume of molecular sequence and structure data [5]. Most of the current molecule function prediction systems are based on sequence similarity approach (homologous transfer). Despite the fact that it seems intuitive, this approach might not be always the right choice, as structural similarity does not always lead to functional similarity [5,6,7]. Moreover, the structure of biomolecules inside cell is far from static and many functions are associated with the cellular environment. Therefore, combination of heterogeneous data about molecules (publications, experimental results) has become a promising direction for function prediction [6,7].

In this paper, we propose a semantic approach to biomolecular function prediction based on never-ending learning paradigm applied to ontology learning and knowledge extraction. For this purpose, PubMed[1] biomedical literature repository is used. The main goal is to refine the results returned by BLAST[2] that is performing the calculation of similarity using local sequence alignment. For the refinement of results, similarity of the molecules at semantic level is considered. A proof-of-concept prototype is presented to illustrate the system capabilities.

One of the solutions that are using heterogeneous sources to enhance biomolecule function prediction [6] combines information from structure and sequence homologies with protein–protein interaction networks for improved protein function predictions. Moreover, the Extended Similarity Group (ESG) [8] for protein annotation prediction iteratively searches the homology space around the queried protein and draws conclusion about protein function based on annotations of similarly annotated proteins using gene ontology. In contrast to solution presented in [8] that relies on pre-defined Gene Ontology[3] annotations, approach proposed in this paper builds upon framework for automated ontology learning and knowledge extraction, with more flexibility and possibility to improve the prediction in time by analyzing new biomedical publications on a daily basis by adoption of the never-ending learning approach.

Nenad Petrović is with the Faculty of Electronic Engineering, University of Niš, Aleksandra Medvedeva 14, 18000 Niš, Serbia (e-mail: nenad.petrovic@elfak.ni.ac.rs).

Milorad Tošić is with the Faculty of Electronic Engineering, University of Niš, Aleksandra Medvedeva 14, 18000 Niš, Serbia (e-mail: milorad.tosic@elfak.ni.ac.rs).

[1] https://www.ncbi.nlm.nih.gov/pubmed/
[2] https://blast.ncbi.nlm.nih.gov/Blast.cgi
[3] http://geneontology.org/

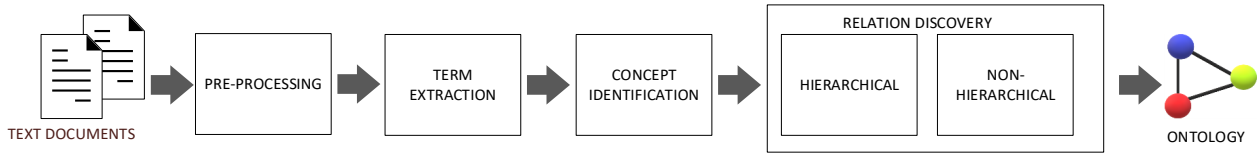


Fig. 1. Ontology learning process

II. Background and Related Work

In this section, the work that our research is build upon is presented and described how it contributes the overall solution.

A. Ontology learning

Ontology learning is a research area focused on automated discovery and construction of ontologies using corpus of textual sources [1]. Ontology learning process consists of identifying terms, concepts and relations between them within the text, in order to construct an ontology [1,4]. The process is illustrated in Fig. 1.

First, the documents need to be pre-processed in order to make sure they are in the right format, so they can be further used within the ontology learning system. After that, the term extraction is performed. Term extraction is a part of ontology learning process whose objective is to obtain terms that could be considered as linguistic realizations of everything important and relevant to the domain [1,4]. Concept is a set of one or many terms which denote an abstract idea or mental picture of a group or class or objects [1]. However, not all the detected terms are suitable to be chosen as concepts. We consider that a corpus of documents on the specific topic consists several knowledge abstractions that are regularly found several times within the text. Therefore, it is needed to identify these terms for the particular topic. Furthermore, it is needed to find the relations between the concepts that describe some kind of bond between them. There are two types of relations: taxonomic (hierarchical) and non-taxonomic (non-hierarchical) [1,4]. The main task related to taxonomic relations is to construct hierarchies by discovering *is-a* relations (hypernym/hyponym) between the concepts [1,4]. Non-taxonomic relations are other interactions between the domain concepts, such as meronymy (*part-of*), possession (*has-a*), roles, attributes and causality [4]. In general, the discovery of non-taxonomic relations relies on the analysis of syntactic structure and dependencies. Verbs are often good candidates when it comes to identification of non-taxonomic relations [4].

In context of this paper, the ontology learning process is considered for construction of ontology schemas for the knowledge base about proteins based on public repository of biomedical literature PubMed with purpose of providing means for prediction of gene functions based on semantic technology. Furthermore, a variation of ontology learning is used in order to populate the semantic triple store by extracting the knowledge from the text documents about given topic.

B. tf-idf

As it was mentioned previously, it is important to find the terms that are most relevant to the specific topic, as they could be candidates for important domain concepts. For this purpose, we decide to use the tf-idf technique.

Term frequency- inverse document frequency (tf-idf) is a numerical statistic often used in information retrieval and text mining to determine the importance of word within to a document within the collection or corpus [9]. The value of tf–idf increases proportionally to the number of times that the observed word appears in the document and is offset by the number of documents in the corpus that contain the word, which makes up for the fact that some words appear more frequently in general. Typically, the tf-idf consists of two factors: term frequency (tf) and inverse document frequency (idf).

Term frequency is the number of times a word appears in a document, divided by the total number of words in that document:

$$tf(t) = \frac{\text{Number of times term t appears in a document}}{\text{Total number of terms in the document}}$$

Inverse document frequency (idf) is computed as the logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears:

$$idf(t) = ln\frac{\text{Total number of documents}}{\text{Number of documents with term t in it}}$$

C. Ontology alignment

Alignment represents correspondence between entities of two ontologies [10]. It is produced by ontology matcher using ontologies as input.

There is a distinction between two types of matching: schema-based and instance-based [11]. A schema-based matcher takes various aspects of the concepts and relations within the ontologies and uses some similarity measure to determine correspondences. Instance-based matcher takes the instances that belong to the concepts in the ontologies and compares them to discover similarity. On the other side, there is also a distinction between element-level and structure-level matching [11]. While element-level matcher compares properties of the particular concept or relation, a structure-level matcher compares the structure of the ontologies to find similarities. However, these matchers can also be combined.

In scope of this paper, we use instance-level, property-based ontology similarity matcher to find the molecules that have similar semantic descriptions obtained from document corpus during the instance-level ontology extraction. Similarity on semantic level is taken into account as indicator of potential molecule function similarity in combination with molecule-level sequence homology. Moreover, schema-level ontology alignment is used for evaluation of the created ontologies. For implementation, we use Alignment API[4] presented in [10].

[4] http://alignapi.gforge.inria.fr/

D. Never-ending learning

Never-ending language learner is an intelligent computer agent that runs continuously with purpose of reading information from the web in order to populate its knowledge base, while it learns to perform this task better every day [12, 13]. Most machine learning systems learn to perform some function based on statistical analysis of a single data set. On the other side, never-ending learning paradigm is inspired by the fact that humans learn many different functions and obtain various types of knowledge over years of accumulated diverse experiences, using extensive background knowledge learned from previous experiences to guide subsequent learning [12, 13].

A never-ending learning problem is defined as £=(L, C) [4]. L is an ordered pair that consists of a set $L = \{L_i\}$ of learning tasks, where the i^{th} learning task $L_i = (T_i, P_i, E_i)$ is to improve the agent's performance, as measured by performance metric P_i, on a given performance task T_i, through a given type of experience E_i. Furthermore, it includes a set of coupling constraints $C = \{(\varphi_k, V_k)\}$ among the solutions to these learning tasks, where φ_k is a real-valued function over two or more learning tasks that specifies the degree of satisfaction of the constraint, and V_k is a vector of indices over learning tasks, specifying the arguments to φ_k. Each performance task is a pair $T_i \equiv (X_i, Y_i)$ defining the domain and range of a function which has to be learned $f_i^* : X_i \rightarrow Y_i$. The performance metric $P_i : f \rightarrow R$ defines the optimal learned function for the i^{th} learning task defined as $f_i^* \equiv argmax_{i \in Fi} P_i(f)$, where F_i is the set of all possible functions from X_i to Y_i. As time passes, the goal is to increase the quality of learned functions (as measured by the individual performance metrics $P_1 \ldots P_n$) and the degree to which the coupling constraints C are satisfied.

In our system, never-ending learning approach is being adopted for both the ontology learning process and biomolecule function prediction. In the first case, an ontology evaluation metric is used as a performance metric, while in the second case prediction success rate is considered for that purpose. Therefore, as time passes, it is expected that system is able to collect more knowledge that enables to perform better prediction by extracting knowledge from larger number of documents describing the biomolecule structure and features. The imposed constraints are related to the number of texts taken as input. First, the theoretical texts about biomolecules are analyzed to build domain ontologies. After that, the information from biomedical literature is extracted and knowledge base populated. Once the system extracts knowledge from a given number of publications and experimental results, it is possible to perform prediction. The prediction performance is measured every day and system continuously explores more and more texts every day in order to extend its knowledge about both the ontology schema and concrete instances that could improve the quality of prediction. The whole process is illustrated in Fig. 2.

E. Sequence homology and homology-based transfer

Sequence alignment is a method of arranging the sequences of DNA, RNA, or protein in order to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences.

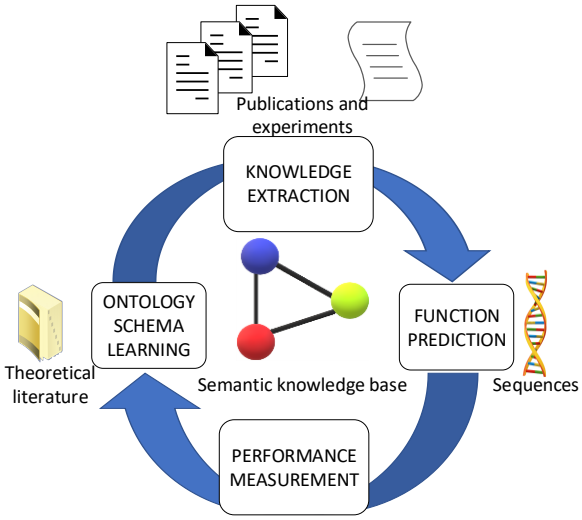


Fig. 2. Applying the never-ending learning approach to biomolecule function prediction

The discovery of sequence homology to a known molecule provides the first ideas about the function of one that has just been discovered [14]. This the most common method of computational function prediction, known as homology-based transfer. The biological rationale behind the homology-based transfer is that if there are two sequences with high degree of similarity, then it is assumed that they have possibly evolved from a common ancestor, so they could have similar functions [5].

Sequence similarity measures can be either global or local. Global similarity algorithms attempt to align every residue in each sequence with purpose of performing the overall alignment of two sequences. This method is most useful when the sequences involved are similar and of roughly equal size. On the other side, local similarity measures seek relatively conserved subsequences, and a single comparison may return several distinct subsequence alignments. Local similarity measures are generally preferred for database search, as cDNAs could be compared with partially sequenced genes and more useful in case of dissimilar sequences that are suspected to contain regions of similarity or similar segments within the wider context.

Basic Local Alignment Search Tool (BLAST) [14] is one of the most commonly used software platforms for sequence similarity which can be used either via web interface or as a stand-alone tool to compare the user's provided sequence against the database of known sequences. There are several types of BLAST to compare all combinations of nucleotide or protein queries with nucleotide or protein databases. BLAST performs comparisons between pairs of sequences, looking for regions of local similarity. The output of BLAST is a list of molecules ordered by descending similarity score between the queried sequence and those that were found similar within the sequence database. However, the sequence that is most similar to the query is not necessarily the one that is identical in function.

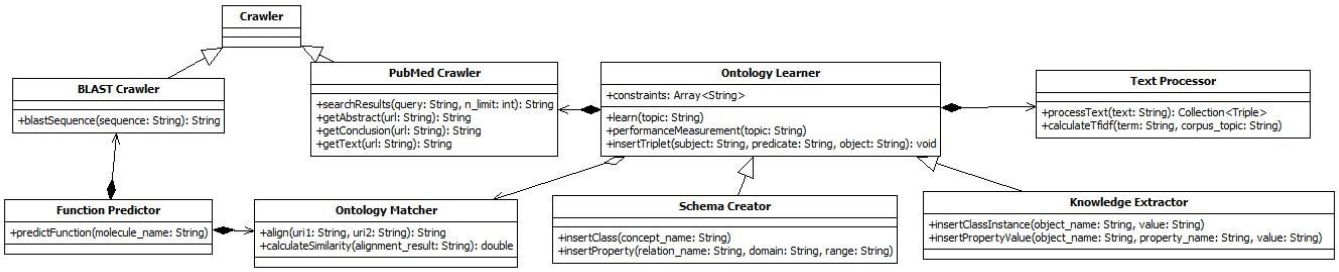


Fig. 4. UML class diagram of the system prototype

Within the scope of this paper, homology-based transfer is considered as an auxiliary factor for function prediction of newly discovered sequence, but not the only one, as our approach is also involving techniques related to semantic similarity of sequences, based on the instance-level ontology matching.

III. Architecture and Implementation Overview

In this section, the architecture overview of the implemented system is given (illustrated in Fig. 3.).

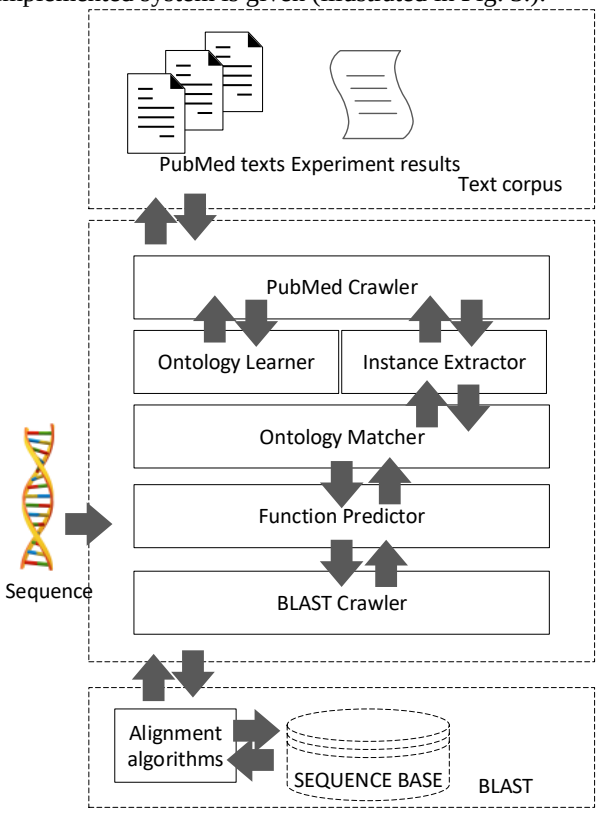


Fig. 3. Gene function prediction system based on never-ending ontology learning approach

The implemented prototype is a Java application with REST API, accessed via web browser. There are three main processes in our system: ontology learning, knowledge extraction and gene function prediction.

First, the ontology learning is performed. Various textual inputs can be used for this purpose (scientific texts, experiment results etc,). In our implementation, the PubMed crawler extracts abstracts and contents of text documents on a given topic in order to construct corresponding ontology schemas according to the algorithm given in Listing 1.

When it comes to text processing, the triplets in form *(subject, predicate, object)* are extracted from each text. For this purpose, we use Stanford CoreNLP Natural Language Processing Toolkit [15], that transforms the given text into a collection of triplets.

Input: topic name
Output: ontology
Steps:
1. Retrieve top n documents from PubMed about the given topic;
2. for each of the retrieved documents
3. get triplets from text;
4. calculate tf-idf of terms;
5. end for each
6. find most relevant concepts about the topic based on tf-idf of terms
7. for each noun
8. insert class(noun);
9. end for each
10. for each verb
11. if(is-a) then
12. insertProperty(subject, is subclass of, object)
13. else
14. insertProperty(domain, verb, range);
15. end if;
16. end for each
17. end

Listing. 1. Pseudo-code of the ontology learning algorithm

Moreover, the most relevant nouns are extracted according to tf-idf and they become classes within the ontology schema. After that, the properties are identified. The most important verbs become properties – either describing taxonomy or other types of relations between the discovered classes.

Once the ontologies are constructed, the knowledge extraction is performed. In this process, the knowledge base is populated by triplets that are instances of the previously constructed ontologies for various molecules. This can phase can be described as instance-level ontology learning and the algorithm is quite similar to one presented in Listing 1., starting from line 7, with additional steps of class and relation matching with the existing schema. The processes of ontology learning and knowledge extraction are alternating according to the never-ending learning approach. The more time passes, the more knowledge is held by the system, so it is expected to perform much better predictions based on this knowledge.

Finally, it is possible to perform the prediction of gene function based on its sequence and knowledge extracted from PubMed literature, according to the algorithm given in Listing 2.

In what follows, the gene function prediction algorithm is described. The pre-condition is that the knowledge about the given molecule is extracted (if there is literature on the topic). After that, the BLAST crawler queries the sequence database with a given input sequence. As output, all the sequence homologs are returned. Among them, there are potential candidates whose functions might be similar to the molecule that contains the queried sequence. This is where the ontology learning and knowledge extraction plays role. Furthermore, the instance-level ontology alignment of the

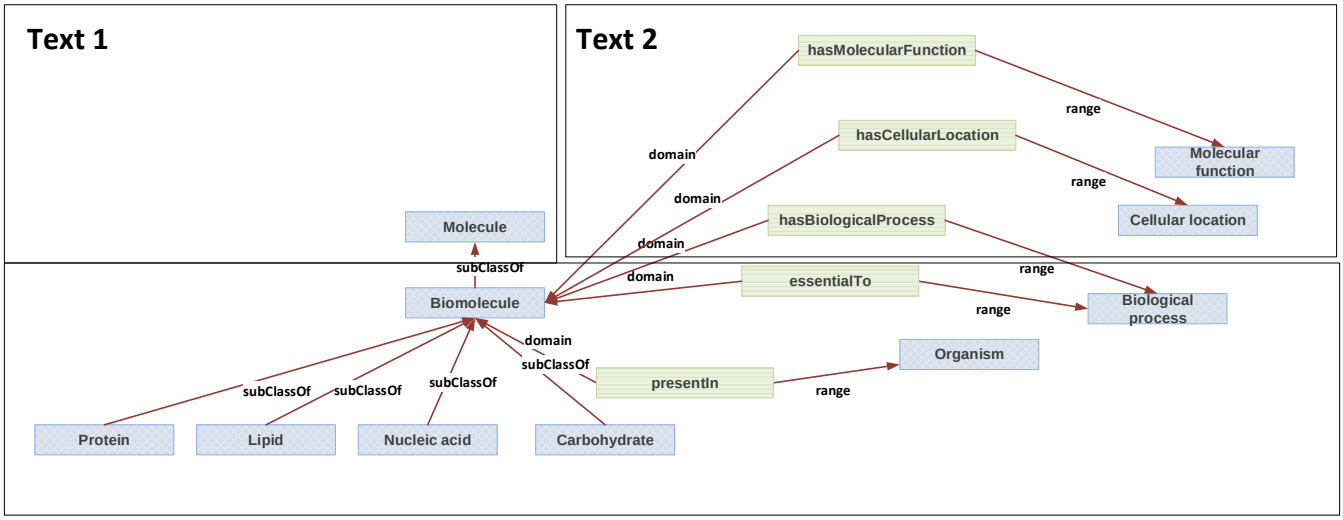


Fig. 5. Ontology schema learning example

knowledge about the input molecule is performed against the knowledge of the homologs. The one that is semantically most similar is the one whose function is transferred to the new sequence. In Fig. 4, a detailed UML class diagram of the implemented system prototype is given.

Input: molecule name, molecule sequence
Output: function prediction
Steps:
1. extract semantic knowledge about molecule from texts
2. candidates=blast(molecule sequence, top-n)
3. max=0;
4. for each candidate
5. similarity=ontologyAlign(candidate, molecule name) ;
6. If(similarity>max) then
7. max=similarity;
8. end if;
9. end for
10. retrieve functions of candidate molecule with max similarity to input molecule
11. return functions prediction;

Listing 2. Pseudo-code of gene function prediction algorithm

As it can be seen in Fig. 4, two central classes within the system are *Ontology Learner* and *Function Predictor*. *Ontology Learner* is further split into two classes: *Schema Creator*, that is responsible for insertion of classes and properties into the ontology schema and *Knowledge Extractor* whose role is to populate the triple store with the knowledge about the particular topic (property values and detected relations between relevant classes) following the structure of previously constructed schema. Both of them contain *Text Processor* and *PubMed Crawler* objects. *Text Processor* is based on Stanford CoreNLP and transforms the text into a collection of triplets. Moreover, it is able to calculate tf-idf of each term within the document corpus. *PubMed Crawler* is able to get the n most relevant texts about the given topics and extract their abstracts and contents. These texts are used for both ontology learning and knowledge extraction. It uses *Ontology matcher* for alignment performance measures according to the pre-defined set of constraints for different knowledge base sizes and number of documents processed within the corpus. On the other side, *Function Predictor* contains *Ontology Matcher* that aligns two ontology instances and is able to return the numeric value assigned to the achieved alignment. Moreover, *Function Predictor* contains *BLAST Crawler* that is responsible for molecule segment alignments used for creation of the initial set of candidate molecules for gene function prediction. The final output of the system is a string that holds the list of predicted functions for a given sequence.

IV. Ontology Learning and Knowledge Extraction Scenario

In this section, we present a simple scenario illustrating how the implemented prototype works when it comes to ontology learning and knowledge extraction. Let us assume that we have two text from theoretical biomedical literature that are analyzed about biomolecules.

The first one is taken from Wikipedia:

„Biomolecule is a molecule that is present in organisms, essential to some typically biological process. Biomolecules include large macromolecules such as proteins, carbohydrates, lipids, and nucleic acids.“[5]

And, the second is an excerpt from [5]:

„There are three important functional aspects of biomolecules: molecular function, biological process and cellular location“.

Once these two texts pass through our system, the ontology about biomolecules is being built, as shown in Fig. 5. As it can be seen, the information from both texts is taken into account to form the ontology schema. However, redundant relations and concepts might appear, such as *hasBiologicalProcess* and *essentialTo* that both express the same relation between biomolecule and biological process. Therefore, it is needed to eliminate duplicate concepts and relations.

Let us consider the following text about protein kinase: *“Protein kinase is found in human, animals, bacteria and plants. Kinase has molecular function to modify other proteins. It is involved in biological process of cellular pathway regulation”.*

After text analysis, the knowledge base is populated according to the ontology schema, as illustrated in Listing 3.

[5] https://en.wikipedia.org/wiki/Biomolecule

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:j.0="http://www.w3.org/1999/02/22-rdf-syntax-ns#/"
    xmlns:j.1="http://www.tasorone.com/tsc/resources/bionel/" >
  <rdf:Description rdf:about="http://www.example.com/proteins/Kinase">
    <j.0:type rdf:resource="http://www.tasorone.com/tsc/resources/bionel/Protein"/>
    <j.1:hasBiologicalProcess>Cellular pathway regulation</j.1:hasBiologicalProcess>
    <j.1:hasMolecularFunction>Modify other proteins</j.1:hasMolecularFunction>
    <j.1:presentIn>Bacteria</j.1:presentIn>
    <j.1:presentIn>Plant</j.1:presentIn>
    <j.1:presentIn>Animal</j.1:presentIn>
    <j.1:presentIn>Human</j.1:presentIn>
  </rdf:Description>
</rdf:RDF>
```

Listing 3. Protein kinase knowledge base population example

For evaluation of the achieved results, a gold standard-based approach was used [16]. The learned ontologies were compared with a previously created reference ontologies referred to as gold standard. After that, the ontology alignment was performed between the learned ontology and gold standard. According to the results, the average matching for ten ontologies about biomolecules created using texts from Wikipedia and PubMed was around 78%. It was noticed that the ontology learning system struggled with texts containing complex language constructions with many ambiguities.

V. Conclusion and Future Work

In this paper, we have proposed an architecture of a system for biomolecule function prediction based on publicly available biomedical literature and experiment results adopting the never-ending ontology learning and knowledge extraction approach. As a proof-of-concept, we have implemented a prototype in Java programming language.

According to the initial testing, the approach gives satisfactory results when it comes to ontology learning and also seems promising for refinement of BLAST similarity score results in combination with knowledge extracted from PubMed text summaries. However, as this is still a work in progress, we plan to further improve the ontology learning system and perform accurate evaluation of molecule function prediction performance in near future. Furthermore, we would like to experiment with various similarity techniques based on expert knowledge from biomedical domain in order to determine which kind of alignment on semantic level leads to better prediction results. It could be also beneficial to include various text sources from the web (beside PubMed) to construct more complete knowledge base. Moreover, we would like to adopt the similar approach to respond other challenges in biomedical sciences, such as diagnosis and predictive medicine [17]. Finally, the idea is to extend the proposed approach to sensor data and integrate it with our framework for ontology-based coordination of autonomous robots [18].

Acknowledgment

This work has received funding from the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Grant Agreement No 645220, project RAWFIE (Road-, Air- and Water- based Future Internet Experimentation).

References

[1] J. I. Toledo-Alvarado, A. Guzman-Arenas, G. L. Martinez-Luna, "Automatic Building of an Ontology from a Corpus of Text Documents Using Data Mining Tools", *Journal of Applied Research and Technology Vol. 10 No. 3 June 2012*, pp. 398-404, 2012.

[2] T. Gruber, "A translation approach to portable ontology specifications", Knowledge Acquisition - Special issue: Current issues in knowledge modeling, Vol. 5, No. 2, June, pp. 199-220, 1993.

[3] A. Maedche and S. Staab, "Ontology Learning for the Semantic Web", *IEEE Intelligent Systems vol. 16 (2)*, pp. 72-79, 2001.

[4] W. Wilson, W. Liu and M. Bennamoun, "Ontology Learning from Text: A Look Back and into the Future", ACM Computing Surveys Vol. 44 No. 4 August 2012, Article 20, pp. 1-36, 2012.

[5] I. Friedberg, "Automated protein function prediction--the genomic challenge", *Briefings in Bioinformatics, 7(3)*, pp. 225–242, 2006.

[6] C. Zhang, P. Freddolino and Y. Zhang, "COFACTOR: improved protein function prediction by combining structure, sequence and protein-protein interaction information*", Nucleic Acids Research 2017 vol. 4*, pp. 291-299, 2017.

[7] M. Chitale, T. Hawkins and D. Kihara, "Automated Prediction of Function from Sequence*", Prediction of Protien Structures, Functions and Interactions*, John Wiley and Sons, pp. 63-85., 2009.

[8] I. K. Khan, Q. Wei, M. Chitale and D. Kihara, "PFP/ESG: automated protein function prediction servers enhanced with Gene Ontology visualization tool*", Bioinformatics, 31(2)*, pp. 271–272, 2014.

[9] A. Kao and S. Poteet, *Natural Language Processing and Text Mining*, Springer, 2006.

[10] J. David et al., "The Alignment API 4.0", *Semantic Web*, 2 (2011), pp. 3–10, 2010.

[11] M. L. Caliusco and G. Stegmayer, "Semantic Web Technologies and Artificial Neural Networks for Intelligent Web Knowledge Source Discovery", *Emergent Web Intelligence: Advanced Semantic Technologies*, pp. 17-36, 2010.

[12] A. Carlson et al., "Toward an Architecture for Never-Ending Language Learning", AAAI'10 Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, pp. 1306-1313, 2010.

[13] T. Mitchell et al., "Never-ending Learning", *Communications of the ACM, 61(5)*, pp. 103-115., 2018.

[14] S. F. Altschul et al., "Basic local alignment search tool". Journal 74 Keanean Journal of Science Vol. 2 2013 of Molecular Biology 215, pp. 403–410, 1990.

[15] C. Manning et al., "The Stanford CoreNLP Natural Language Processing Toolkit", pp. 1-6, 2014.

[16] J. Raad and C. Cruz, "A Survey on Ontology Evaluation Methods", 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, pp. 1-8, 2015.

[17] D. Dragulescu and A. Albu, "Medical Predictions System", *Acta Polytechnica Hungarica vol. 4, no. 3*, pp. 89-101, 2007.

[18] V. Nejkovic, N. Petrovic, N. Milosevic, M. Tosic, "The SCOR Ontologies Framework for Robotics Testbed", 2018 26th Telecommunication Forum (TELFOR), Belgrade, pp. 1-4, 2018. https://doi.org/10.1109/telfor.2018.8611841

Upravljanje redundantnom robotskom rukom s višestrukim pogonima i pokretanjem sajlama

Aleksandar Rodić, Miloš Jovanović i Ilija Stevanović

***Sadržaj*—U radu se razmatraju aspekti projektovanja inteligentnog upravljanja redundantnom robotskom rukom pokretanom sajlama s višestrukim servo-pogonima. Na početku, biće prikazana konstrukcija robotske ruke redundantne kinematske strukture s 7 stepeni slobode kretanja, čiji je prototip mehanizma u fazi pripremem za izradu. Takođe, biće dat matematički model višestruko aktuirane robotske ruke na bazi kojeg je urađena simulacija sistema i sprovedena inkrementalna poboljšanja mehaničke strukture u fazi razvoja prototipa robotskog mehanizma. U radu se sintetizuje fuzzy kontroler robotske ruke, na višem hijerarhijskom nivou, s zadatkom upravljanja višestrukim pogonima I pokretanjem mehanizma posredstvom sajli. Verifikacja performansi sintetizovanog inteligentnog kontrolera urađena je korišćenjem simulacionih eksperimenta. Rezultati će biti analizirani i prikazani u poslednjem odeljku rada. Budući pravci istraživanja i razvoja biće istaknuti na kraju rada kao i neki aspekti primene redundantne robotske ruke s višestrukim pogonom.**

***Index Terms*—Laka robotska ruka, višestruki pogoni, redundantni mehanizam, pokretanje sajlama, inteligentno upravljanje**

I. UVOD

Rad se bavi problemima inteligentnog upravljanja višestruko aktuiranog robotskog mehanizma veličine ljudske ruke, s pogonom uz pomoć sajli i višestrukim pogonima. Cilj istraživanja i razvoja u ovom radu je postizanje performansi robotske ruke koje se mogu porediti s manipulativnim sposobnostima i veštinama ljudske ruke kao biološkog organa. Čovekova ruka sastoji se od skeletnog sistema koji joj daje neophodnu čvrstinu, mišića koji ostvaruju kretanje, nervnog sistema koji omogućava osetljivost na spoljašnje pobude i sistema krvotoka koji obezbeđuje ishranu tkiva i dovođenje energije do mišićnih vlakana. Anatomski posmatrano, ruka s šakom se pokreće uz pomoć više od 40 mišića raspoređenih u ramenom pojasu, nadlaktici, podlaktici i u samoj šaci. Njihovom sadejstvom, ostvaruje se veliki broj raznovrsnih kretanja. Velika redundansa u broju mišićnih vlakana podrazumeva da pojedini stepeni slobode kretanja budu ostvareni sadejstvom 3-4 nezavisna mišića. Na taj način, mišići ruke superponiraju svoje mišićne tonuse tako da generišu pokrete ruke. Takođe, u slučaju povrede ili pucanja mišićnog vlakna, ruka ne ostaje nepokretna već nastavlja da obavlja svoje funkcije u oslabljenom kapacitetu. Može se reći da je priroda za vreme evolutivnog razvoja ljudskih bića stvorila vrlo suptilan anatomski sistem koji minimizuje rizike od potpunog gubitka funkcionalnosti upravo korišćenjem bioloških redundansi u sistemu. U ovom radu, cilj nam je bio da sintetizujemo jedan inteligentni robotski kontroler robotske ruke koji će urediti aktivnosti redundantnog pogona mehanizma ruke u cilju postizanja boljih dinamičkih performansi mehanizma u poređenju s konvencionalnim robotskim pogonima tipa "jedan zglob – jedan servo-motor".

Kod konvencionalnih robotskih mehanizama, u slučaju oštećenja i/ili otkaza jednog od aktuatora, robot trajno gubi sposobnost kretanja u tom pravcu (oko partikularne ose). Po našem mišljenju, redundantnost strukture i višestrukost pogona značano mogu unaprediti performanse sistema ali zato otežavaju (komplikuju) njegovo upravljanje.

Prosečna težina ruke odraslog čoveka je 4-5 kg i zavisi od konstitucije i pola osobe. Istovremeno, ljudska ruka u opštem slučaju, može podići teret od 3-4 kg u odručenom položaju. U proseku, čovek taj teret može držati rukom u odručenom položaju ne više od jednog minuta. Nakon toga, ljudski mišići se zamare, ulaze u zasićenje i moraju se relaksirati da bi povratili svoju funkciju. U robotici, sposobnost da robot podigne teret, lakši ili teži od svoje sopstvene težine, se izražava jedinstvenim indeksom koji se naziva "maseni odnos" (engl. mass-payload ratio). Ovaj indikator predstavlja fizički odnos mase robota i maksimalnog tereta kojeg robot može da podigne bez preopterećenja ili oštećenja strukture. Kod ljudi, ovaj indeks je približno jednak jedinici. Ljudska ruka može kratkotrajno podići teret koji je teži od sopstvene težine. Nasuprot, industrijski roboti imaju značajno veću vrednost pomenutog masenog indeksa od jedinice. Razlog za to je strog zahtev da se robotska struktura industrijskih robota učini što robusnijom i krućom. Kada govorimo o preciznosti ljudske ruke, ona iznosi ne više od 1-1,5 mm. Ljudska šaka ima ulogu da obezbedi veću tačnost (recimo manju od 0,2 mm) dok je zadatak ruke (do zgloba šake) da obezbedi tzv. prenosno kretanje i da dovede šaku u određenu poziciju [1].

II. STANJE U OBLASTI

Poslednjih godina bilo je uspešnih pokušaja da se razviju različite lake robotske ruke [2]-[8]. Mnogi od njih, implementirali su način za pokretanje zglobova potezanjem sajli, po uzoru na pokretanje ruku tetivama. Glavne poteškoće kod ovih sistema pokretanim sajlama odnose se na elastičnosti sistema i unutrašnja trenja a to dovodi do problema s preciznošću i ponovljivošću robota. Drugi problem odnosi se na mehaničku robusnost robotske ruke. Jedna od interesantnih

Aleksandar Rodić iz Instituta "Mihajlo Pupin" Beograd, Centar za robotiku, Volgina 15, 11060 Beograd, Republika Srbija (e-mail: aleksandar.rodic@pupin.rs).

Miloš Jovanović iz Instituta "Mihajlo Pupin" Beograd, Centar za robotiku, Volgina 15, 11060 Beograd, Republika Srbija (e-mail: milos.jovanovic@pupin.rs).

Ilija Stevanović iz Instituta "Mihajlo Pupin" Beograd, Centar za robotiku, Volgina 15, 11060 Beograd, Republika Srbija (e-mail: ilija.stevanovic@pupin.rs).

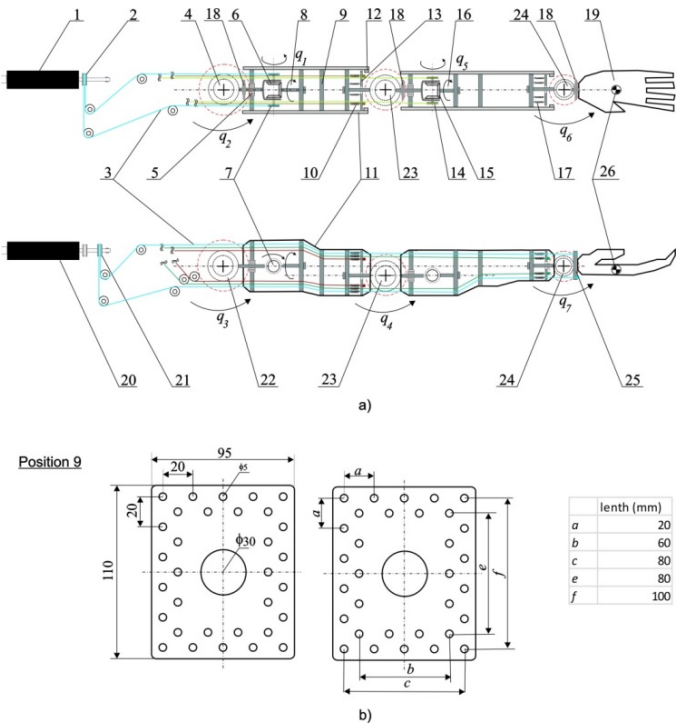
konstrukcija u zadnje vreme je AMBIDEX robotska ruka [2] koja je nastala kao rezultat istraživanja i razvoja uzajamne saradnje i koegzistencije biološko-tehnološkog para čovek-robot. Robotska ruka AMBIDEX koristi inovativno rešenje pogona uz pomoć sajli koje čini interakciju s čovekom bezbednijom nego obično. Sa samo 2,6 kg, ova ruka teži manje od prosečne ljudske ruke muške, odrasle osobe. AMBIDEX može razviti maksimalnu brzinu od 5 m/s i može podići do 3 kg tereta. Zato što se AMBIDEX robotska ruka može upravljati na istovetan način kao industrijski robot, ona može imati veliki broj primena, od jednostavog podizanja tereta pa do izvođenja složenih zadataka koji zahtevaju preciznu manipulaciju i kooperaciju s drugim subjektima. AMBIDEX podržava velike brzine, bežičnu komunikaciju, upravljanje u realnom vremenu s daljine korišćenjem malog kašnjenja i velikog protoka 5G mreže. Inovativno rešenje, koje će biti predstavljeno u ovom radu, uprošćava mehaničku strukturu robotske ruke, povećava robusnost sistema zadržavajući u isto vreme preciznost i ponovljivost sličnu biološkom modelu.

III. KONSTRUKCIJA

Robotska ruka je tako konstruisana da ispuni zadate tehničke zahteve koji su komentarisani u Odeljku-I ovog rada. Aktuatori robotske ruke su izmešteni iz zglobova u njegovu bazu da bi se smanjila težina mehanizma. Uobičajeno, masa servo-motora, odgovarajućih reduktora i kućišta (mehaničkog rama) predstavljaju najteže elemente robotske ruke. Ako pretpostavimo da robotska ruka predstavlja deo jednog dvoručnog robotskog sistema, onda je pogodno da servo-motori budu smešteni u osnovi sistema. U specijalnom slučaju, robotsku osnovu čini torzo (trup) dvoručnog robota. Pokretanje robotske ruke, u poređenju s konvencionalnim industrijskim robotima gde su servo-motori smešteni u zglobovima mehanizma, izvodi se sinhronizovanim povlačenjem i otpuštanjem žica (sajli) napravljenih od nerastegljivog materijala (čelična žica, Kevlar i sl.). Pogonske sajle su jednim krajem prebačene (vezane) preko motorizovanih koturača, postavljenih na izlazna vratila reduktora pogonskih motora, dok su svojim drugim krajem vezane za mesta na robotskim segmentima, kao što je prikazano na Sl. 1a.

U biološkom svetu, tetive kao nastavci mišića su povezani u odgovarajućim tačkama za kosti ruke. Segmenti robotske ruke, da bi bili što lakši i istovremeno imali odgovarajuću mehaničku čvrstoću, su izrađeni od Al-ploča (poput sendviča) s dodatnim ramom između ploča (Sl. 1a i 2) koji dodatno ukrućuje strukturu. Unutrašnjost segmenta se koristi za prolaz poteznih sajlica, električnih provodnika i vodova senzorskih signala. U unutrašnjosti segmenta robotske ruke (Sl. 1a, detalj 11) postavljene su i fiksirane za zid, odgovarajuće sprovodne ploče (Sl. 1, detalj 9). Sprovodne ploče (nalik na „kapije“) imaju male otvore raspoređene po obimu u dve paralelene vrste na rastojanju 20mm jedni od drugih (Sl. 1b). Male rupice, u sprovodnim pločama pokretačkih sajli, dozvoljavaju provlačenje uskih silikonskih cevčica kroz koje su provučene nerastegljive sajlice (poput ručnih kočnica kod bicikla).

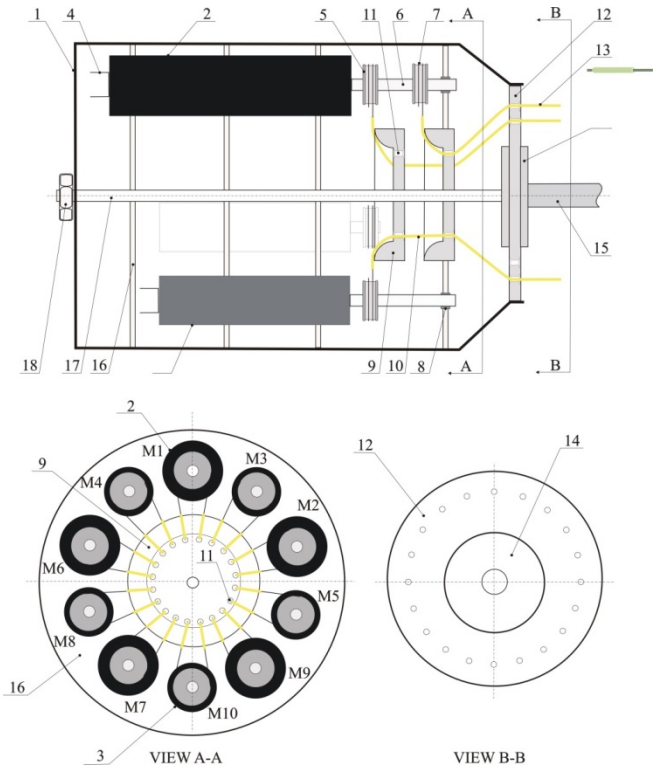
Višestruko aktuirani zglobovi robotske ruke konstruisani su kao sverni (rotacija oko 3 ose) ili dupli cilindrični zglobovi (Sl. 2, detalji 2 i 3). Zglobovi su konstruisani kao kardanski zglobovi kao što je ilustrovano na Sl. 2. Robotska ruka ima redundantnu kinematsku strukturu s 7 mehaničkih stepeni slobode kretanja. Ose zglobova 22 i 23 (Sl. 1a) se ne poklapaju i međusobno su pomerene za 20 mm jedna u odnosu na drugu. Ose zglobova 22 i 24 su kolinerane (Sl. 1a). Nepodudarnost osa zglobova (ramena i lakta) obezbeđuje smanjenje opterećenja pogonskih servo-motora koji pokreću zglob lakta. Sinhronizovanim povlačenjem i otpuštanjem nerastegljivih sajli, od strane servo-motora posatvljenih u bazi robota, generišu kretanje robotske ruke u zglobovima ramena, lakta i zgloba šake (Sl. 1a). Mehanička snaga servo-motora, uz pomoć motorizovanih doboša, se prenosi na sajle i dalje do tačaka vezivanja na segmentima robota. Pronacija i supinacija nadlaktice se ostvaruje pomoću koničnog zupčastog para. Linearno kretanje poteznih sajlica se pretvara u obrtno kretanje koničnog zupčastog para smeštenog neposredno iza ramenog zgloba (Sl. 1). Pogonski servo-motori mehanizma robota su smešteni u bazi robota kao što je prikazano na Sl. 2. DC-motori s planetranim reduktorima su poređani u krug i montirani su u cilindrično kućište kako bi zauzeli što manje prostora. Motorizovani doboši su postavljeni na izlazna vratila servo-motora. Jedan kraj nerastegljive sajle je namotan preko doboša. Drugi kraj sajle je ankerisan za određene tačke na segmentu robotske ruke. Na ovaj način, obrtno kretanje servo-motora je preneseno na linerno kretanje sajlice kako je to prikazano na šemi na Sl. 3.



Slika 1. Viši nivo prikaza žičanog pogona, višestruko aktuirane robotske ruke: a) struktura mehanizma i princip pogona; b) konstrukcija vođica (prolaznih ploča) namenjenih za sprovođenje poteznih sajli. Značenje pojedinih pozicija na slici: 1-DC-motor; 2-motorizovani kotur; 3- potezne sajlice; 4-sverni zglob u ramenu; 5-kuglični ležaj; 6-konični zupčanik; 7-kotur; 8-pronacija/supinacija nadlaktice; 9-vođica; 10-opruge; 11-Al-kućište; 12-graničnik elastičnog modula; 13-završna ploča prvog segmenta; 14-kotur; 15-konični zupčanik; 16-pronacija/supinacija podlaktice ; 17-opruge; 18-senzor momenta u zglobu (ramena, lakta, zgloba šake), 19-šaka; 20-DC-motor; 21-motorizovani doboš; 22-rameni zglob; 23-zglob lakta; 24-zglob šake; 25-prirubnica šake; 26- cenatar mase šake.

IV. MATEMATIČKI MODEL

Matematički model robotske ruke uključuje model mehanizma robotske ruke i model pogona. Konstruisani sistem je višestruko aktuiran pošto ima veći broj aktuatora nego što robot ima zglobova.



Slika 2. Konfiguracija pogonskog modula robotske ruke. Princip prenosa snage od servo-motora do zglobova robota: M1, M2 – savijanje/istezanje nadlaktice; M3, M4 – savijanje/istezanje podlaktice; M5 – savijanje/istezanje robotske šake; M6, M7 – odručenje/priručenje nadlaktice, M8 – odručenje/priručenje robotske šake; M9 – pronacija/supinacija nadlaktice and M10 - pronacija/supinacija podlaktice. Značenje pojedinih pozicija na grafičkoj prezentaciji je: 1-Al-kućište, 2-servo-motor Tipa A; 3-servo-motor Tipa B; 4-krajevi električnog konektora; 5-motorizovani doboš; 6-izlazno vratilo (DC-motor and planetarni reduktor); 7-nerastegljive sajle; 8-uležištenje; 9-sakupljač sajli i konduktor; 10-silikonska cevčica; 11-rupa; 12-izlazna konudktivna ploča; 13-sajla; 14-prirubnica ramenog zgloba; 15-istezanje zgloba; 16-okvir kućišta; 17-navojno vreteno kućišta; 18-navrtka.

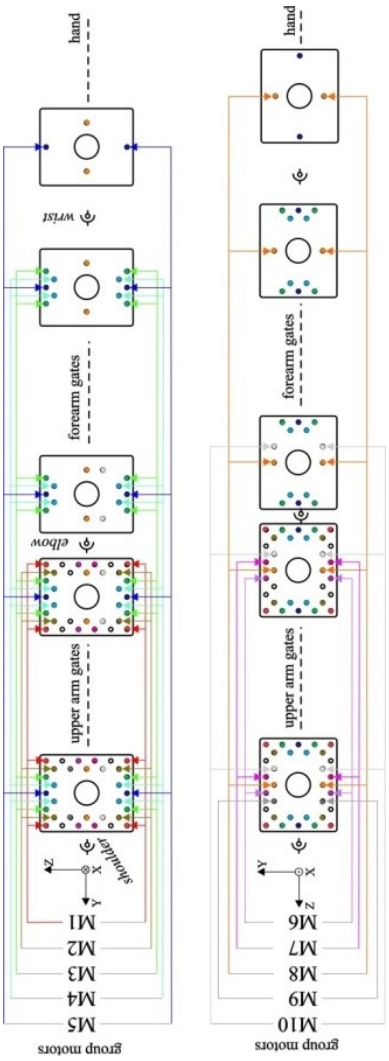
Model robota uključuje kinematski i dinamički model mehanizma. Kinematski model mehanizma opisan je jednačinama koje dovode u vezu unutrašnje koordinate robota iz prostora zglobova i spoljašnje koordinate iz prostora zadatka. Kinematski model se definiše sledećim relacijama:

$$s = f(q, p) \tag{1}$$

$$\dot{s} = J(q, p)\dot{q} \tag{2}$$

$$\ddot{s} = J(q, p)\ddot{q} + \frac{\partial J(q, \dot{q}p)}{\partial q}\dot{q}^2 \tag{3}$$

gde je: $q = [q_1\ q_2 \dots q_7]^T$ vektor unutrašnjih koordinata zglobova robota; $s = [x\ y\ z\ \varphi\ \theta\ \psi]^T$ je vektor spoljašnjih koordinata koje opisuju poziciju i orjentaciju završnog organa robota u prostoru zadatka (Kartezijanske koordinate *x, y, z* za poziciju i odgovarajući Ojlerovi uglovi *φ, θ, ψ* za orjentaciju);



Slika 3. Šema aktuacije robotske ruke – segmenta nadlaktice, podlaktice i šake. Potezne sajlice i odgovarajući otvori kao provodne vođice u robotskoj ruci.

$J(q, p)$ je Jakobijeva matrica dimenzija 6 x 7; Kinematski i dinamički parametri mehanizma su definisani vektorom *p*.

Dinamički model mehanizma robota je izražen u vektorskoj formi diferencijalnom jednačinom (4):

$$\tau = H(q, p)\ddot{q} + C(q, \dot{q}, p)\dot{q} + G(q, p) + \cdots + J^T(q, p)S + J^T(q, p)F \tag{4}$$

gde je $\tau = [\tau_1\ \tau_2 \dots \tau_7]^T$ vektor momenata u zglobovima robota; $H(q, p)$ je matrica inercije; $C(q, \dot{q}, p)$ je vektor Koriolisovih i centrifugalnih sila/momenata; $G(q, p)$ je vector gravitacionih sila i momenata; $J(q, p)$ je Jakobijeva matrica mehanizma; S je vektor elastičnih sila i momenata koji potiču od niza opruga i elastične strukture mehanizma; *F* je vektor spoljašnjih sila i momenata koje dejstvuju na mehanizam i *p* je vektor geometrijskih i dinamičkih parametara robotskog mechanizma. Pogonski momenti u zglobovima robota se ostvaruju na indirektan način, sinhronizovanim povlačenjem

sajli (Sl. 1, detalj 3). Sajle su jednim krajem povezane na pogonske motore koji su postavljeni u bazi. Drugi kraj kabla je fiksiran za mehaničku strukturu odgovarajućeg segmenta robota (Sl. 3). Pogonski momenti u zglobovima robotske ruke generisani su posredstvom odgovarajućih momenata na izlaznim vratilima servo-motora (Sl. 2). U relacijama (5)-(7) date su relacije koje dovode u vezu pogonske momente u zglobovima robotske ruke τ_j i odgovarajuće momente na izlaznim osovinama reduktora servo-motora.

$$\tau_1 = T_9, \tau_5 = T_{10} \tag{5}$$

$$\tau_2 = T_6 + T_7, \tau_6 = T_8 \tag{6}$$

$$\tau_3 = T_1 + T_2, \tau_4 = T_3 + T_4, \tau_7 = T_5 \tag{7}$$

U primeru koji se razmatra u ovom radu koristi se 10 servo-motora koji aktiviraju 7 zglobova robota i to: 3 u ramenu, 2 u laktu i 2 u zglobu šake robotske ruke. Model pogonskog motora uobičajeno se predstavlja sledećim relacijama:

$$L_r \frac{di_r}{dt} + R_r i_r + C_E \dot{q} = u \tag{8}$$

$$-C_M i_r + J_r \ddot{q} + B_C \dot{q} = -T_l \tag{9}$$

gde je: $L_r\ (H)$ induktivnost motora; $i_r\ (A)$ struja u rotoru; otpornost rotora $R_r\ (\Omega)$; konstanta proporcionalnosti elektromotorne sile $C_E\ (V/(\frac{rad}{s}))$; ugao obrtanja izlaznog vratila motora $q\ (rad)$; napon u rotoru $u\ (V)$; konstanta proprocionalnosti momenta $C_M\ (Nm/A)$; moment inercije motora J_M redukovan na izlazno vratilo $J_r\ (kg\,m^2)$; koeficijent viskoznog trenja redukovan na izlazno vratilo motora $B_C\left(Nm/(\frac{rad}{s})\right)$ i momenti od spoljašnjeg opterećenja $T_l\ (Nm)$. Konstante, C_E i C_M, koeficijent B_C i moment inercije J_r su određeni iz sledećih relacija: $C_E = C_e N_v$; $C_M = C_m N_m$; $B_C = B_c N_v N_m$; $J_r = J_m N_v N_m$ gde su N_v i N_m prenosni odnosi za moment i brzinu, redom, a C_e, C_m i B_c su uzeti iz kataloga proizvođača.

V. FUZZY KONTROLER

Momenti u zglobovima robotske ruke se ostvaruju dejstvom servo-motora postavljenih u osnovi robta (Sl. 1 i 2). U ovom slučaju, bazu robota predstavlja komplet od deset DC motora smeštenih u ramenom pojasu torzoa dvoručnog robotskog mehanizma koji će biti razvijen kao krajnji cilj projekta. Momenti na izlaznim vratilima motora se prenose do odgovarajućih zglobova robota posredstvom niza vođica (otvora u pločicama - kapijama) kroz koje prolaze potezne sajlice (Sl. 1). Na ovaj način, sinergijsko dejstvo dva motora se koristi da pokrene jedan stepen slobode kretanja mehanizma u pojedinom zglobu ruke. Ključno je pitanje kako izabrati pravu strategiju upravljanja redundantnim brojem pogona robotske ruke da bi se ostvarilo željeno kretanje mehanizma. U ovom radu biće prodiskutovana primena tri različite strategije aktuacije motora ali one nisu jedine koje je moguće definisati.

- Strategija ravnomerne raspodele opterećenja u zglobovima robota;
- Strategija kontrolisanog rada motora u nominalnom režimu, i
- Strategija raspodeljenog opterećenja motora među zglobovima.

Upravljački algoritam, koji ispunjava prvo navedenu strategiju ravnomerene raspodele opterećenja na servo-motorima koji pokreću pojedinačan zglob, obezbeđuje generisanje jednakih momeneta na izlaznim vratilima motora koji zajednički pokreću isti zglob ruke. Na bazi simulacionih rezultata u ovom radu, potvrđeno je da je ova strategija upravljnja najbolja u smislu postizanja najbolje energestke efikasnosti sistema.

Drugi predloženi algoritam, obezbeđuje rad motora u nominalnom režimu što generiše željene momente u zglobovima robota. To podrazumeva da neki od više motora, koji deluju sinergijski, ne mora biti aktivan kako bi ostali radili u nominalnim režimima.

Treći po redu upravljački algoritam, odgovoran za ostvarivanje postavljene strategije upravljanja, obezbeđuje da se za isti tip pokreta ruke (npr. priručenje/odručenje, savijanje/opružanje, itd.) opterećenje na motorima (više njih) tako raspoređuje da se mogu po potrebi koristiti i motori koji pokreću druge zglobove i ne pripadaju istoj grupi motra dodeljenih pojedinim zglobovima. Recimo, za savijanje zgloba u ramenu može se delimično koristiti aktivnost motora koji inače savija zglob u laktu ruke i na taj način „pomaže“ motorima ramena koji su realno najopterećeniji.

Inteligentni kontroler kretanja robotske ruke može kombinovati sve tri strategije od slučaja do slučaja u cilju postizanja boljih performansi sistema. Na sličan način funkcionišu i biološki sistemi gde se aktivnost mišića menja i prilagođava različitim uslovima manipulativog zadatka i opterećenja.

Na ovom mestu biće prikazano kako se sintetizuje fuzzy kontroler jedne prekobrojno pogonjene robotske ruke. Fuzzy kontroler ima četiri ulaza (ulazne promenljive) i dva izlaza. Ulazne promenljive fuzzy sistema su označene kao: „*Jtorq-gradient*“, „*M1-torq*“, „*M2-torq*“ i „*torq-deficit*“. Pomenute promenljive su definisane sledećim relacijama (10)-(13).

$$grad(\tau_k) = \frac{\Delta\tau_k}{\tau_k^{max}} \tag{10}$$

$$amp(T_j) = \frac{|\tau_j|}{\tau_j^{max}} \tag{11}$$

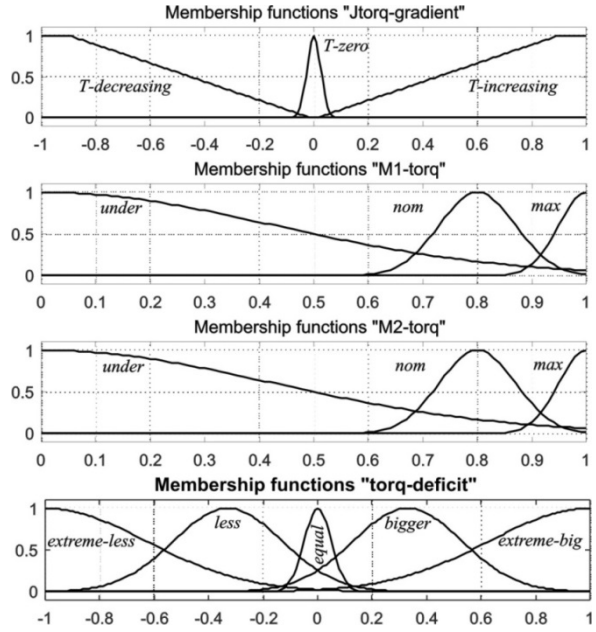
$$amp(T_{j+1}) = \frac{|\tau_{j+1}|}{\tau_{j+1}^{max}} \tag{12}$$

$$\Delta T_j^k = \frac{\tau_k - \tau_j}{\tau_j^{max} + \tau_{j+1}^{max}} \tag{13}$$

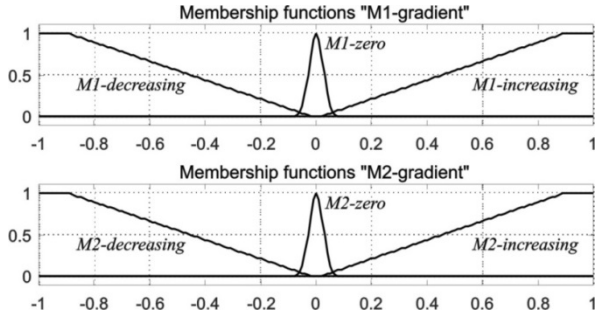
gde je: $grad(\tau)$ relativni gradijent momenta opterećenja τ_k koje dejstvuje u k-om zglobu robota; $amp(T_j)$ je relativni priraštaj amplitude momenta tzv. vodećeg „master“ servo-motora koji pokreće k-ti zglob robota; $amp(T_{j+1})$ je relativni priraštaj amplitude momenta pomoćnog (dopunskog) servo-

motora koji pokreće *k*-ti zglob ruke; ΔT_j^k je odstupanje (deficit/suficit) pogonskog momenta motora od odgovarajućeg momenta opterećenja u *k*-om zglobu robota. Izlazne promenljive fuzzy kontrolera T_j^k i T_{j+1}^k predstavljaju vrednosti pogonskih momenata generisanih na paru servo-motora M_j i M_{j+1} koji sinergijski pokreću k-ti zglob robotske ruke. U relacijama (5)-(7) dati su momenti opterećenja robotskog mehanizma τ generisani posredstvom odgovarajućih pokretačkih servo-motora $T_1 \cdots T_{10}$ svih zglobova mehanizma.

Struktura fuzzy kontrolera sastoji se od funkcija članstva čiji su oblici dati na Sl. 4 i 5.



Slika. 4. Ulazne promenljive i funkcije članstva projektovanog fuzzy kontrolera.



Slika 5. Izlazne promenljive i funkcije članstva projektovanog fuzzy kontrolera.

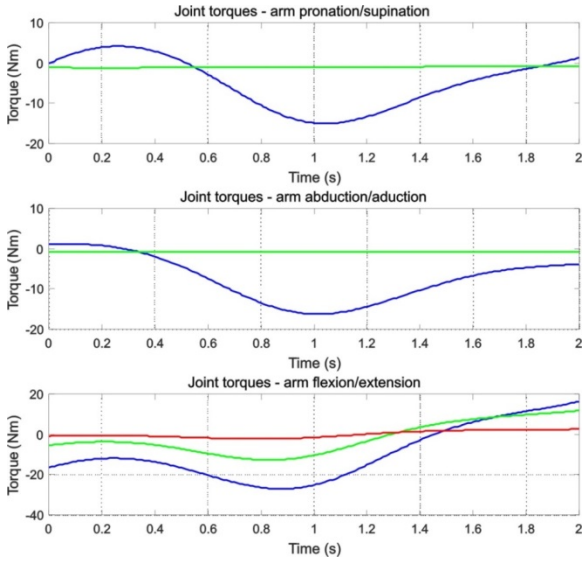
Logički deo kontrolera čini 11 fuzzy pravila koja uzimaju u obzir ranije definisane ulazne i izlazne promenljive. Uvedena fuzzy pravila materijalizuju stanja sistema koja ispunjavaju odgovarajuće postavljene kriterijume funkcionisanja. To su: a) kriterijum nominalnog režima rada koji vodi računa da motori rade uvek u optimalnom modu rada, ili b) kriterijum ravnomerne raspodele opterećenja među motorima. Na Sl. 6 prikazani su rezultati primene inteligentnog robotskog kontrolera, na najvišem nivou upravljanja, za slučaj ispunjenja kriterijuma a). U radu se razmatra slučaj kada tri para servo-motora (M6, M7), (M1, M2) i (M3, M4) rade u „kooperativnom“ režimu rada odn. sadejstvu. Za taj slučaj se u ovom radu prikazuju odgovarajući simulacioni rezultati. Uočava se da glavni (master) motori M6, M1 i M3 pretežno rade u nominalnom režimu rada (s približno 80% snage) kada je njihova efikasnost najveća. Istovremeno, pomoćni, dopunski motori M7, M2 i M4 (Sl. 2 i 3) su manje opterećeni. Zadatak pomoćnih motora je da nadomeste eventualni deficit u snazi glavnih motora. Zajedno, glavni i pomoćni motori obezbeđuju sprovođenje željenog kretanja zglobova robotskog mehanizma.

Fuzzy pravila koja obezbeđuju realizaciju postavljenog niza kriterijumskih funkcija su:

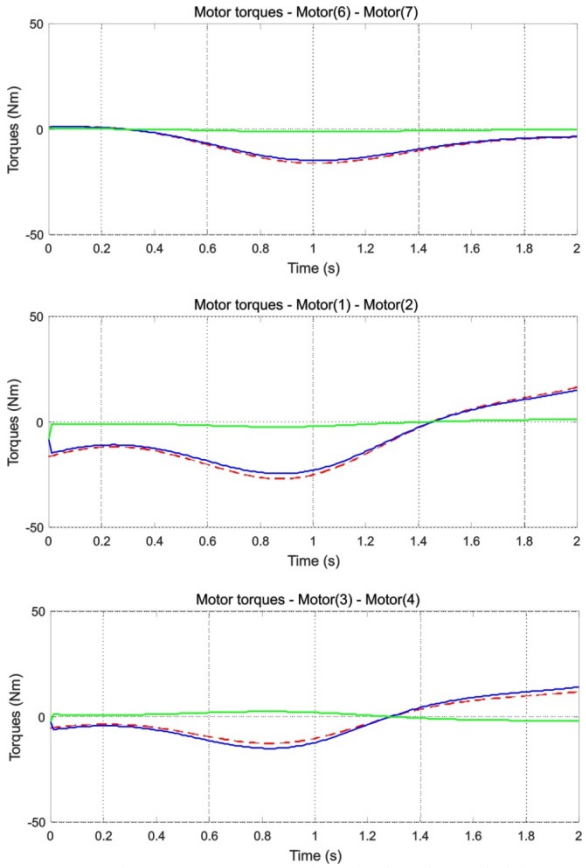
1. If (Jtorq-gradient is T-zero) then (M1-gradient is M1-zero) and (M2-gradient is M2-zero) (1)
2. If (Jtorq-gradient is T-increasing) and (M1-torq is not nom) then (M1-gradient is M1-increasing) (1)
3. If (Jtorq-gradient is T-decreasing) and (M1-torq is not nom) then (M1-gradient is M1-decreasing) (1)
4. If (M1-torq is max) then (M1-gradient is M1-zero) (1)
5. If (M2-torq is max) then (M2-gradient is M2-zero) (1)
6. If (M1-torq is not max) and (M2-torq is not max) and (torq-deficit is extreme-big) then (M1-gradient is M1-increasing) and (M2-gradient is M2-increasing) (1)
7. If (M1-torq is not max) and (M2-torq is not max) and (torq-deficit is extreme-less) then (M1-gradient is M1-decreasing) and (M2-gradient is M2-decreasing) (1)
8. If (Jtorq-gradient is T-increasing) and (M1-torq is nom) and (M2-torq is under) then (M1-gradient is M1-zero) and (M2-gradient is M2-increasing) (1)
9. If (Jtorq-gradient is T-decreasing) and (M1-torq is nom) and (M2-torq is under) then (M1-gradient is M1-zero) and (M2-gradient is M2-decreasing) (1)
10. If (Jtorq-gradient is T-increasing) and (M1-torq is nom) and (M2-torq is nom) then (M1-gradient is M1-increasing) and (M2-gradient is M2-increasing) (1)
11. If (Jtorq-gradient is T-decreasing) and (M1-torq is nom) and (M2-torq is nom) then (M1-gradient is M1-decreasing) and (M2-gradient is M2-decreasing) (1)

VI. SIMULACIONI REZULTATI

S ciljem da se odrede momenti opterećenja u zglobovima robota i da se sračunaju pogonski momenti u servo-motorima, urađeni su odgovarajući simulacioni eksperimenti. Zadat je robotu jedan tipičan manipulativni zadatak gde se od robota zahtevalo da napravi pokret u vertikalnoj ravni protivno sili gravitacije i da pri tom nosi teret u hvataljci težine 3 kilograma. To predstavlja jedan od najzahtevnijih robotskih zadataka, u smislu opterećenja, koja se javljaju u zglobovima mehanizma. Model robota, definisan relacijama (1)-(4) je iskorišćen za simulaciju kinematskih i dinamičkih efekata. Simulacioni rezultati su prikazani na Sl. 6 i 7.



Slika. 6. Momenti opterećenja u zglobovima robota za vreme savijanja ruke u vertikalnoj ravni suprotno ubrzanju gravitacije.



Slika 7. Momenti servo-motora generisani tako da savladaju momente opterećenja u zglobovima robotskog mehanizma (u primeru robotskog zadataka koji je obrađen u tekućem simulacionom primeru važećem i na Sl. 6). Grafik prikazan isprekidanom linijom predstavlja momente u zglobovima robotske ruke.

VII. ZAKLJUČAK

Višestruko pogonjen robotski sistem ima značajne prednosti nad konvencionalno aktuiranim sistemima (jedan zglob – jedan motor). Prednosti se ogledaju u sledećem. Kada prekobrojno aktuiran sistem ima kvar na pojedinom servo-motoru, preostali pogoni preuzimaju teret (odgovornost) da omoguće da sistem i dalje regularno radi čak i pri smanjenom kapacitetu snage. Druga prednost se odnosi na mogućnost da se kod prekobrojno pogonjenih sistema mogu na pogodan način primeniti različiti optimizacioni, inteligentni algoritmi upravljanja koji minimizuju potrošnju energije, koordiniraju radom servo-motora tako što ih teraju da funkcionišu u nominalnom režimu rada tako da ih štede i produžavaju radni vek. Treća prednost ovih sistema odnosi se na kompaktnost konstrukcije i racionalizaciju prostora potrebnog za smeštanje pogonskih motora u mehanizmu. Predložena je i potvrđena simulacionim eksperimentima upotreba većeg broja slabijih pogona (redundantsnost u broju motora) nasuprot manjem broju, po dimenzijama većih i snažnijih motora. Ako smo svesni činjenice da se kod bioloških sistema koristi veći broj mišića za pokretanje jednog zgloba onda se pristup konstruisanju robotske ruke u ovom radu smatra opravdanim i u saglasnosti s poznatim bio-kibernetskim principima funkcionisanja živih bića.

ZAHVALNOST

Ovaj članak je rezultat rada na istraživačkom projektu "Istraživanje i razvoj ambijentalno-inteligentnih robotskih sistema antropomorfne strukture" pod brojem TR-35003, 2011-2019 koje finansira Ministarstvo prosvete, nauke i tehnološkog razvoja Republike Srbije.

REFERENCE

[1] Rodić, A., Miloradović, B., Popić, S., Urukalo, Dj.: On developing lightweight robot arm of anthropomorphic characteristics. In book: New Trends in Medical and Service Robots. Book 3, Series: Mechanisms and Machine Science, Springer Publishing House, Vol. 38, Eds. Bleuler, H.; Pisla, D.; Rodic, A.; Bouri, M; Mondada, F; (Eds.), ISBN: 978-3-319-23831-9, Book ID: 332595 _1_En, (2015).

[2] AMBIDEX Cable-Driven Robot Arm, https://www.naverlabs.com/en/storyDetail/12, last accessed 2019/02/06

[3] Yang, G., Lin, W., Kurbanhusen, M. S, Pham, C., B., Yeo, S. H.: Kinematic design of a 7-DOF cable-driven humanoid arm: a solution-in-nature approach. Journal IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM), pp. 24-28, Monterey, CA, (2005)

[4] Lum, G.Z., Mustafa, S.K., Lim, H.R., Lim, W.B., Yang. G., Yeo, S.H.: Design and motion control of a cable-driven dexterous robotic arm. In proceedings 2010 IEEE Conference on Sustainable Utilization and Development in Engineering and Technology. DOI: 10.1109/STUDENT.2010.5686997, Petaling Jaya, Malaysia, (2010)

[5] Lens, T., Kirchhoff, J., Stryk, von O.: Dynamic modeling of elastic tendon actuators with tendon slackening. In: 12th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2012). DOI: 10.1109/HUMANOIDS.2012.6651608, Osaka, Japan, (2012)

[6] Palli, G., Borghesan, G., Melchiorri, C.: Modeling, Identification, and Control of Tendon-Based Actuation Systems. In: IEEE Transactions on Robotics (Volume: 28, Issue: 2 , April 2012). pp. 277 – 290, DOI: 10.1109/TRO.2011.2171610, (2012)

[7] Rost, A., Verl, A.: The QuadHelix-Drive - an improved rope actuator for robotic applications. In: 2010 IEEE International Conference on Robotics and Automation. 3-7 May 2010, ISSN: 1050-4729, DOI: 10.1109/ROBOT.2010.5509764, Anchorage, AK, USA, (2010)

[8] Potkonjak, V., Svetozarevic, B., Jovanović, K., Holland, O.: The Puller-Follower Control of Compliant and Noncompliant Antagonistic Tendon Drives in Robotic Systems. International Journal of Advanced Robotic Systems, pp. 143-155, DOI: 10.5772/10690, (2011)

[9] P.I. Corke, "Robotics, Vision & Control", Springer 2017, ISBN 978-3-319-54413-7, last accessed 2019/02/06

Аутоматизација система за наводњавање применом Интернета ствари

Владо Крунић, *ПМФ, Универзитет у Бањој Луци,* Момчило Крунић, *ФТН, Универзитет у Новом Саду,* Предраг Ранитовић, *Висока пословна школа струковних студија Нови Сад*

***Апстракт*— Интернет ствари (IoT – Internet of Things) је довео до великих промена свакодневног живота људи кроз примену у градовима, енергетици, пословању, образовању, медицини, индустрији, пољопривреди и другим областима. Интернет ствари омогућује повезивање већег броја корисника, уређаја, сервиса и апликација на Интернет, при чему се сваком уређају додељује јединствен идентификатор. Међусобно повезани уређаји и апликације размењују податке и прослеђују их удаљеним серверима, да би им крајњи корисници приступали по потреби путем мобилних и/или веб апликација. Примена IoT у пољопривреди кроз пројектовану инфраструктуру са одређеним сензорима, актуаторима, микроконтролерима и микрорачунарима омогућује се стални надзор и аутоматско покретање одговарајућих процеса. Кључне активности, које се ефикасно могу аутоматизовати применом IoT, се односе на процесе наводњавања. Тема рада је модел аутоматизације система за наводњавање применом IoT.**

***Кључне речи* - *И*нтернет ствари; Дизел генератор; Систем за наводњавање; Сензор; Актуатор.**

I. Захтеви корисника

Интернет свари (IoT – Internet of Things) представља концепт који омогућава објекима из реалног света, који имају рачунарску и сензорску подршку, да се међусобно повежу и да приступе апликацијама и сервисима на Интернету. IoT треба да, посредством Интернета, повеже уређаје и да омогући корисницима приступ и размену информација. Једна од области са већом потенцијалном употребом IoT је пољопривреда где се и убудућности може очекивати још шира примена [1], [2].

Циљ овог рада је да се представи IoT концепт на коме је заснован систем за даљинско надгледање и контролу процеса наводњавања. Систем за наводњавање који треба аутоматизовати применом Интернета ствари се састоји из следећих уређаја и инфраструктурне опреме:

- Дизел генератор електричне енергије, трофазни;
- Пумпа за наводњавање за бунар, трофазна;
- Фреквентни регулатор, трофазни;
- Разводни ормар са главним прекидачем, осигурачима и одговарајућом инсталацијом;
- Разводни систем са три кугласте славине за воду са прикључцима пречника 50 mm;
- Систем цеви за наводњавање три парцеле;

Први и основни захтев корисника се може изразити кроз реченицу: Дизел генератор електричне енегрије (у даљем тексту Агрегат) треба реконструисати са доградњом одговарајућих уређаја који ће омогућити његово стартовање, надзор и гашење на даљину.

Други захтев се односи на даљинско стартовање пумпе, надзор њеног рада и гашење по истеку подешеног временског периода или на захтев.

Трећи захтев подразумева постављање сензора за влажност на одговарајућим местима и праћење параметара влажности на три локације које се независно наводњавају.

Осим наведена три основна захтева, потребно је унапредити разводни систем уградњом електомагнетних вентила који ће се селективно отварати/затварати према потреби, односно захтеву апликације.

Такође, треба поставити систем видео надзора са циљем повећања безбедности и заштите система за наводњавање, као и објекта за смештај главне опреме.

II. Анализа опреме и захтева

Први корак у реализацији захтева корисника се односи на анализу структуре и функција расположиве опреме и сагледавање могућности њеног унапређења са циљем стварања услова за испуњење постављених захтева. Полазна тачка је први захтев корисника и прва ставка наведене расположиве опреме. Дакле, анализираћемо могућности реконструкције и доградње расположивог дизел генератора – Агрегата (Слика 1). Агрегат је марке Willager, номиналне снаге 5,5 kW.

Сл. 1. Дизел генератор електричне енергије марке Willager, ознаке VGD5500-0. Мотор се стартује преко контакт браве електропокретача (анласера) повезаног на акумулатор 12V, капацитета 20 Ah. Контакт брава има три положаја (0 – нема нема контакта, 1- спреман за палење и 2 – покретање анласера).

Агрегат који је представљен располаже са дизел мотором снаге 7,4 kW, који постиже брзину од 3600 rpm. Пали се електропокретачем или ручно, потезањем стартног ужета. Палење, као и гашење дизел агрегата је много сложеније у односу на бензинске агрегате.

Агрегат који се користи у систему за наводњавање има техничке карактеристике које могу да задовоље

корисничке захтеве везане за аутоматизацију (Табела I).

ТАБЕЛА I
КАРАКТЕРИСТИКЕ АГРЕГАТА

Тип мотора	Четворотактни, ваздушно хлађени
Запремина мотора	418 cm³
Снага мотора	7.4 kW (3600 rpm)
Тип стартера	Реверзибилни / Електро-старт
Номинална снага kW	5 kW
Максимална снага kW	5.5 kW
Јачина ел. струје	8.3 A
Номинални напон	230 V / 400 V
Утичнице	1 × 230 V / 1 × 400 V
Време рада	≈ 9 h 30 min.
Гориво	Дизел
Маса	120 kg
Запремина резервоара	15 l

Наводимо два проблема која су идентификована код решавања аутоматизације:

А. Дизел агрегати имају ручицу којом се отвара довод горива непосредно пре палења. Агрегат се гаси тако што се затвори довод горива. Затварање довода горива је аутоматизовано померањем контакт кључа из позиције 1 на позицију 0. Потребно је осмислити решење које омогућује даљинско отварање довода горива.

Б. Дизел агрегати се теже пале у зимским условима, па је некад потребно да се више пута стартују док се не упале. Приликом ручног стартовања треба претходно повући ручицу за декомпресију ради лакшег почетног покретања мотора. Декомпресија је пожељна и код палења преко стартера (анласера), ради смањења почетног обртног момента. Треба наћи решење да се ручица за декомпресију аутоматски повлачи.

Сценарио палења Агрегата је следећи: 1 – Отвара се славина за гориво повлачењем ручице. 2 – Повлачи се ручица за декомпресију мотора. 3 – Кључ стартне брава се из позиције 0, помера на позицију 1, а затим се помера на позицију 2. На позицији 2 кључ се задржава од 2 до 4 секунде, а затим се враћа на позицију 1. Ако се мотор не упали, посутпак се понавља.

После палења (или непосредно пре палења) Агрегата, проверавају се линије за наводњавање и евентуално се подешавају за наводњавање одређене парцеле. Избор парцеле је условљен процентом влажности.

После стабилизације рада мотора у периоду од 5 до 10 s, трофазни прекидач фреквентног регулатора се поставља у контакт позицију чиме се пумпа за воду. Агрегат се гаси постављањем прекидача стартне браве у нулту позицију чиме се активира механизам за затварање довода горива.

Пумпу за воду коју покреће Агрегат, спада у групу посебних пумпи намењених за транспорт воде из бунара. То је такозвана потапајуће пумпе за бунаре, Pedrollo 4SR8/13, са уљним мотором предвиђеним за дуг временски рад. Услови рада пумпе су предвиђени за бунаре који имају значајне залихе воде (Табела II).

Снага пумпе (2,2 kW) је усклађена према капацитету бунара и укупној површини парцела који треба наводњавати (2,7 ha) технологијом „кап по кап“ [7].

III. НАДОГРАДЊА СИСТЕМА ЗА НАВОДЊАВАЊЕ

Треба подсетити да Агрегат поседује техничке карактеристике које су у складу са потребама напајања пумпе, уз нешто резерве снаге што је пожељно ради повећане потрошње код стартовања пумпе као и напајања додатне опреме. Између Агрегата и пумпе је смештен фреквентни регулатор који регулише рад пумпе и оптимизује потрошњу горива. Пумпа је повезана на разводни систем са три кугласте слине којима се према потреби отварају или затварају линије за наводњавање три посебне парцеле.

Резултати претходне анализе расположиве опреме и захтева корисника упућују на неколико техничких захвата који подразумевају пре свега доградњу механизама који ће омогућити аутоматизацију појединих операција. На самом почетку анализе проблема наведеног под А. намећу се два решења:

1. Даљинско померање ручице славине за отварање довода горива се може извести уградњом вентила одговарајуће снаге са елетромагнетним актуатором повезаним на акумулатор 12 VDC.

2. Друго решење је доградња електромеханичког уређаја непосредно поред ручице славине за ручно затварање довода горива, који ће моћи да помера ручицу славине. Свако од решења има предности и недостатке.

Наш избор је решење са уградњом електромагнетног актуатора (Слика 2), због неколико разлога:

- Постојећа линија довода горива остаје у оригиналној изведби, што је увек пожељно због будућег одржавања;
- Доградња електромеханичког уређаја који покреће ручицу постојеће славине је јефтиније него набавка новог електромагнетног вентила и његова уградња;
- Електромеханички уређај је довољно поуздан и његово одржавање је једноставније него одржавање електромагнетног вентила.
- Решење опције избора палења Ручно/Аутоматско, је једноставно (реконструкција Агрегата подразумева да постоји опција избора палења Ручно/Аутоматско).

ТАБЕЛА II
КАРАКТЕРИСТИКЕ ПУМПЕ

Модел	4SR8/13
Тип пумпе	Потапајућа бунарска 4"
Потис	2"
Снага	2200 W
Напон	380 V
Напор	85 - 30 m
Проток	40 ~ 200 L/min
Потапање	Макс. 100 m
Температура воде	Макс. +35°C
Температура околине	Макс. +35°C
Димензије (Ø x H)	98mm x 1043mm
Тежина	19,2 kg

Анализом је установљено да треба осмислити и направити електромеханички механизам за померање ручице славине за отварање линије за довод горива.

Механизам који је направљен, тестиран и који функционише без грешке се састоји од електромагнета

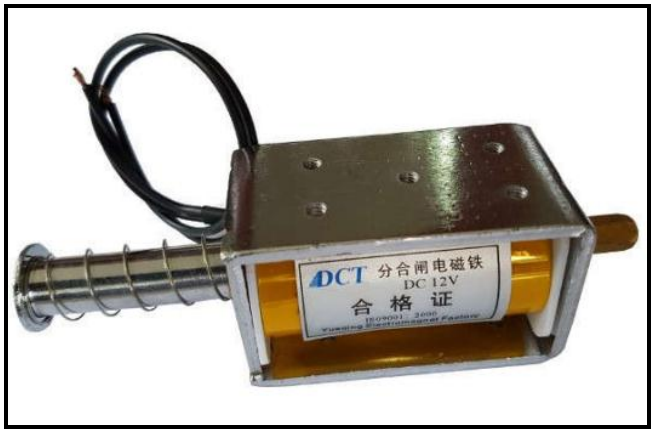
одговарајућих карактеристика, носача електромагнета постављеног на кућиште Агрегата и полуге која повезује електромагнет и ручицу славине на линији за довод горива. Карактеристике електромагнета су у складу са захтевима који дефинишу момент силе за затварање славине на линији за довод горива и дужину лука кружног кретања ручице славине (Табела III).

ТАБЕЛА III
КАРАКТЕРИСТИКЕ ЕЛЕКТРОМАГНЕТА

Модел	DCT 150
Добављач	DOSSY
Напон	12VDC
Ход	30 mm
Сила (ход 0)	2 kg
Сила (ход 30 mm)	0.15 kg
Тип	Push

Осим описаног механизма за аутоматско отварање ручне славине на линији за довод горива, потребно је осмислити и направити и механизам за декомпресију Агрегата. Решење уоченог проблема (наведеног под Б. у анализи опреме и захтева) је скоро исто као претходно описано решење аутоматизације ручног вентила на линији за довод горива, осим разлике у начину фиксирања електромагнета на кућиште Агрегата.

Механизам за аутоматску декомпресију се разликује у конструкцији носача електромагнета у односу на механизам за отварање ручне славине за гориво. Електромагнетни актуатори се не разликују с обзиром на врло сличне техничке захтеве за конструкцију наведених механизама. Осим тога, оба електромагнетна актуатора се покрећу помоћу акумулатора 12 VDC.



Сл. 2. Електромагнет DCT150 12VDC је коришћен као актуатор у уређају за аутоматско покретање ручног вентила за гориво. Исти такав електромагнет се користи као актуатор који повлачи полугу за декомпресију. Поузданост је експериментално потврђена оптерећењем електромагнета двоструко већим од захтеваног.

Претходно описани механизми за аутоматско отварање довода горива и аутоматску декомпресију су тестирани у радним условима тако што су активирани затварањем струјног кола где је извор напајања акумулатор Агрегата. Радни услови за тестирање уграђених електромагнетних механизама су остварени у режиму ручног управљања.

Наредни корак који је предузет у процесу решавања захтева корисника се односи на анализу електричних кола чија се стања мењају променом позиција контакт кључа стартне браве Агрегата. Идеја аутоматизације дефинисане корисничким захтевима је заснована на идентификацији свих могућих стања електричних кола која се отварају или затварају различитим позицијама контакт кључа у стартној брави.

IV. ЛОГИКА СТАРТОВАЊА АГРЕГАТА

Основна идеја која се односи на аутоматизацију Агрегата, који се пали помоћу контакт кључа стартне браве са три могућа положаја (0, 1 и 2), је заснована на симулацији позиционирања контакт кључа. У ту сврху, треба пресећи све каблове који су повезани на стартну браву и направити „Т“ везу која ће омогућити опцију Ручно/Аутоматско палење.

Када се изабере опција Ручно палење, враћамо се на оригинално повезаивање и прекидање веза које воде до система за аутоматско палење. Избор опције Аутоматско палење прекида све везе до контакт браве и повезује каблове који воде до система за аутоматско палење. Следи идентификација електричних кола и опис логике система ручног палења Агрегата.

Идентификација стања електричних кола је заснована на анализи и мерењу стања контакта која се реализују променом позиција контакт кључа стартне браве, узимајући у обзир и смер окретања стартног кључа. Са стартном бравом је повезано пет каблова (црни, жути, сиви, бели и црвени), којима се успостављају различита електрична кола дефинисана шемом електро инсталације система палења Агрегата.

Стартни кључ има три могућа положаја (Положај 0, Положај 1 и Положај 2) преко којих се успостављају или раскидају везе између каблова према опису који следи:

Положај 0 је безконтактни (црна и жута жица спојене, док су све друге растављене).

Положај 1 је контактни (када се окрене кључ у смеру кретања казаљки на сату на прву попозицију (прекида се веза црне и жуте жице и спајају се сива и бела жица)

Положај 2 је контакт за електропокретач (анласер). Овај положај се добије када се савлада опруга кретањем стартног кључа у смеру казаљки на сату полазећи из позиције Положај 1. У овом положају се црвена жица спаја са белом и сивом жицом на кратко време од 2-3 секунде, када треба да се упали Агрегат. После палења пустимо стартни кључ да га опруга врати у Положај 1 где остаје док све док Агрегат ради.

Ако се Агрегат не упали после 2-3 секунде, кључ стартне браве треба вратити у Положај 1 и поново стартовати електропокретач. Гашење агрегата се изводи враћањем контакт кључа стартне браве на Положај 0, чиме се активира затварање довода горива.

Претходна анализа стартне браве је довољна за дефинисање управљачке логике аутоматског стартовања Агрегата, што подразумева испуњавање првог захтева корисника. У наставку је представљена управљачка логика аутоматског система за наводњавање.

Логика аутоматског палења Агрегата треба да буде идентична ручном палењу које смо описали. Мобилна IoT апликација eWeLink, која омогућава комуникацију,

подешавање и управљање Sonoff паметним уређајима, је подешена према захтевима управљачке процедуре система наводњавања. Структура управљачког модула садржи следеће нивое:

Највиши ниво чини eWeLink апликација (интерфејс између корисника и управљачког хардвера), затим следи слој који обухвата Sonoff IoT уређаје, наредни слој (испод Sonoff uređaja) се састоји од електронске плочице и два електромагнета (актуатори славине на доводу горива и вентила за декомпресију). Карактеристике датих електромагнета су наведене у претходном поглављу у оквиру описа електромеханичких уређаја који су уграђени у Агрегат. Остали део опреме за аутоматско палење обухватају три уређаја:

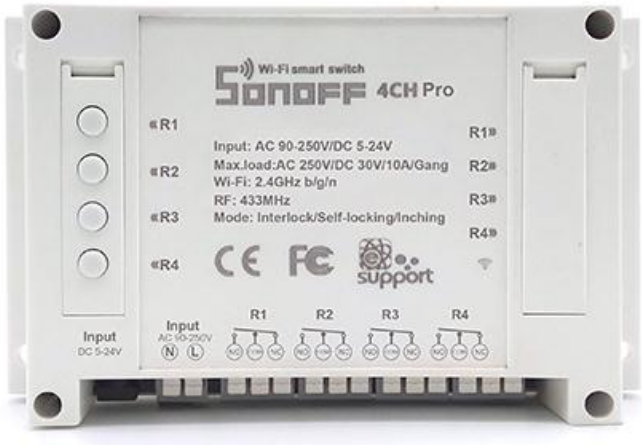
1. Електронска плочица RST-SW/2R12V/2R220V

Електронска плочица садржи RST склопку са 3 релеја 12V/16A, којом се покреће фреквентни регулатор за пумпу. Трофазна склопка се активира сигналом 12 VDC помоћу R3 релеја Sonoff 4CH Pro микроконтролера, са спецификацијом представљеном у наставку. Осим тога на плочици се налази и 4 релеја за управљачку логику: 2 релеја 12V/16A за стартне контакте (RS1 и RS2), 1 релеј 220V/16A намењен за раскидање струјног кола електропокретача ако је агрегат упаљен (RE) и 1 релеј 220V/16A (RP), који служи за покретање пуме за наводњавање помоћу SonoffTH10/TH16 прекидача.

2. Sonoff 4CH Pro – микроконтролер (Слика 3)

Спецификација:

Радни мод: inching/interlock/self-locking mode
Подесиво време: 0.25-4s
Напајање: 5-24V DC
Макс. струја: 10A
Макс. снага: 2200W
Wireless стандард: Wi-Fi 2.4GHz b/g/n, 433MHz RF
Заштита: WEP/WPA-PSK/WPA2-PSK



Сл. 3. Sonoff 4CH Pro је микрокомтролер са 4 релеја која се могу подешавати на апликативном нивоу према захтевима управљачке логике. Техничке карактеристике су у складу са потребама система за аутоматско наводњавање.

3. SonoffTH10/TH16 - паметни прекидач (Слика 4)

Спецификација:

Напајање: 90V~250V AC
Макс. струја:16A
Wireless стандард: WiFi 2.4GHz b/g/n
Заштита: WEP/WPA-PSK/WPA2-PSK
Оперативна температура: 0℃~40℃
Оперативна влажност: 5%-95%
Материјал: FR-ABS

Сл. 4. SonoffTH10/TH16 је паметни прекидач са једним релејем који се користи за покретање потапајуће пумпе. Осим тога, овај прекидач има и сензоре за мерење влажности и температуре који се користе за аутоматско управљање процесом наводњавања.

V. Мобилна апликација

Интерфејс мобилне апликације eWeLink садржи четири дугмета за палење Агрегата и видео надзора и једно дугме за покретање пумпе система за наводњавање. Прва три дугмета (Дугме 1, Дугме 2, Дугме 3) су непосредно везана редом за релеје микроконтролера Sonoff 4CH Pro, са ознакама R1, R2 и R3, респективно, којима се активирају одређени уређаји система за наводњавање. Дугме 4 мобилне апликације eWeLink је везано за четврти релеј микроконтролера Sonoff 4CH Pro, са ознаком R4, којим се активира видео надзор објекта са опремом система за наводњавање.

Пето дугме апликацие eWeLink је везано за паметни прекидач SonoffTH10/TH16, који је предвиђен за покретање потапајуће пумпе активирањем трофазног прекидача уграђеног у фреквентни регулатор, посредно преко релеја RP 220V/16A који се налази на електронској плочици RST-SW/2R12V/2R220V.

Дугме 1 (КОНТАКТ) доводи до стања које је адекватно стању Положај 1 код ручног палења (раскида се веза црне и жуте жице и спајају се сива и бела жица. Ово се дешава на електронској плочи где се истовремено напајају релеји RS1 и RS2. RS1 релеј раскида нормално затворено коло - раздваја црну и жуту жицу, док RS2 релеј спаја нормално отворено коло - спаја сиву и белу жицу. Поновни клик на Дугме 1 (гашење) враћа претходно стање адекватно Положај 0 ручног палења (Слика 5).

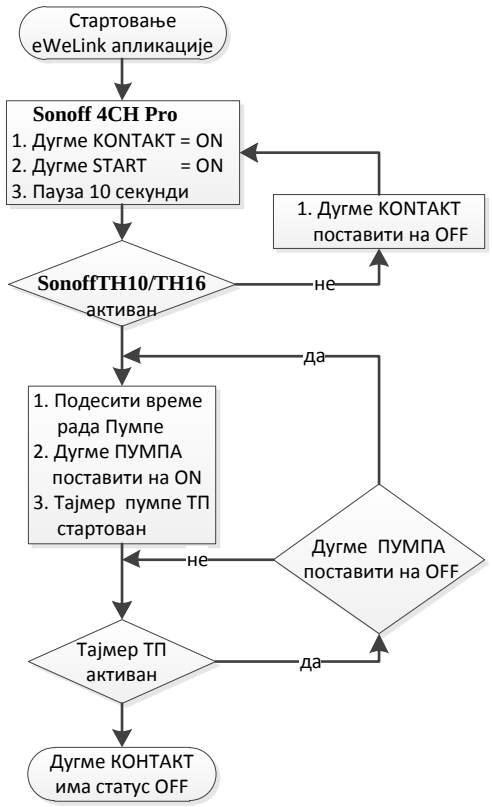
Дугме 2 – (СТАРТ) долазимо до стања које је адекватно стању Положај 2 код ручног палења. У овом положају, који следи после клика које активира Дугме 1, долази до спајања црвене и сиве жице које траје 2-4 секунде, зависно од сетовања (претходно су спојене сива и бела жица када смо активирали Дугме 1).

Црвена жица се спаја са сивом преко R2 релеја Sonoff 4CH Pro који се напаја кликом на Дугме 2 у апликацији

мобилног телефона. Црвена жица из стартне браве се спаја са средњим контактом који је заједнички за нормално отворен и нормални затворен излаз.

Са нормално отвореног контакта црвена жица одлази до електронске плоче на релеј RA3-220V/16A за укидање контакта електропокретача, ако је упаљен агрегат. Црвена жица пролази кроз нормално затворен контакт до сиве жице на шини у разводном орману. Када се агрегат упали раскида се веза која спаја црвену и сиву жицу.

Дугме 3 (ПУМПА) пали фреквентни регулатор преко кога се покреће трофазна пумпа. Ово дугме се активира увек после палења агрегата. Провера да ли је агрегат упаљен се ради тако што се погледа да ли је уређај SonoffTH10/TH16 активан (прослеђује информацију о портошњи струје).



Сл. 5. Блок дијаграм даљинског покретања система за наводњавање применом IoT апликације eWeLink повезане са описаном опремом за управљање Sonoff 4CH Pro и SonoffTH10/TH16.

VI. Интернет комуникација

Стигли смо до слоја који треба да нам омогући комуникацију преко Интернета. Кључну улогу у Интернет комуникацији има рутер намењен мобилним корисницима, као и корисницима стационираним у руралним подручјима. Наравно, подразумевано је да на локацији где желимо да приступамо Интернету постоји довољно јак сигнал мобилног оператера чије услуге намеравамо да користимо. Из богате понуде различитих комуникационих уређаја издвојили смо Huawei E5573 4G Mobile Wi-Fi рутер који је робусан, поуздан и једноставан за инсталацију. Намењен је за кориснике у подручјима где није доступна кабловска комуникација.

Спецификација рутера:
Димензије (mm): 94 x 58 x 15mm
Тежина (g): 75.00
Модем: 4G
Капацитет батерије (mAh): 1500
Меморијско проширење (GB): 32GB
Брзина: до 150Mbps download, до 50Mbps upload
LTE категорија: LTE CAT4
LTE фреквенције: (800/900/1800/2100/2600MHz)
WLAN standards: 802.11a/b/g/n
5 GHz WLAN подршка: Да
Заштита (енкрипција): WEP, WPA, WPA2

Непрекидно напајање је обезбеђено преко квалитетног пуњача за мобилне телефоне који се користе у аутомобилима. Комуникација је обезбеђена преко мобилног оператера који има сигнал довољно јак на локацији где је постављен рутер са SIM картицом.

Важно је нагласити да је систем комуникације потпуно аутономан са сталним напајањем преко акумулатора Агрегата који се допуњава, сваком приликом када је агрегат упаљен током процеса наводњавања. Интернет је доступан у сваком тренутку, што је потребан услов за функционисање управљачког система заснованог на IoT.

Слике 6, 7 и 8 представљају локацију примене IoT апликације и детаље опреме управљачког система.

Сл. 6. Пољопривредна парцела 2,7 ha са засадом малине у Бечејском атару са детаљем вентилације шахта у коме је смештен систем за наводњавање и опрема управљачког система. У белој кутији је смештен рутер који мора бити вани због сигнала мобилне мреже.

Сл. 7. Део шахта са Агрегатом где се види механизам за декомпресију са електромагнетним актуатором DCT150 12VDC. Са леве стране се види детаљ изолованог цевовоа за издувне гасове којим се гасови одводе ван просторије са опремом.

Сл. 8. Фреквентни регулатор за пумпу, паметни прекидач SonoffTH10/TH16 , микроконтролер Sonoff4CHPro и део инсталације у управљачком ормару система за наводњавање.

VII. ZAKLJUČAK

Пројекат аутоматизације система за наводњавање применом Интернета ствари, који је представљен у раду, је практично реализован на пољопривредној парцели у Бечејском атару. Систем наводњавања је „кап по кап“ типа, покрива парцелу од 2,7 ha, снабдева се бунарском водом и у експлоатацији је већ годину дана. Корисници управљају наводњавањем из Новог Сада и до сада нису имали примедби. Недостаци који су се појавили у току имплементације су отклоњени током тестирања, које је обављено пре пуштања унапређеног система у рад.

Аутори су предложили оригинално решење даљинског стартовања Агрегата са дизел горивом. Техничко решење обухвата доградњу два механизма са електромагнетним линеарним актуаторима:

- механизам за отварања ручне славине на линији за довод горива и
- механизам за отварање вентила за декомпресију.

Поред наведеног, инсталацији система за даљинско стартовања Агрегата је претходила реконструкција електро-инсталације Агрегата и разводног ормара са уградњом склопке за алтернативни избор стартовања Агрегата (ручно/аутоматско).

У припремне техничке захвате спада и обезбеђивање сталне Интернет везе са аутономним напајањем.

На крају, требало је на тржишту расположиве опреме одабрати поуздану опрему са перформансама које су у складу са захтевима корисника у које спада и економска оправданост аутоматизације система за наводњавање.

Идеја о унапређењу постојећих техничких система, као што је систем за наводњавање који је био тема рада, и њихова адаптација да се створе услови за примену савремених информационих технологија, као што је Интернет ствари, је кључни допринос рада. Осим тога, анализа система за стартовање Агрегата помоћу кључа стартне браве и идентификација електричних кола која се затварају/раскидају позиционирањем стартног кључа, је модел који може послужити у сличним ситуацијама које се односе на аутоматизацију техничких система.

Даље унапређење система за наводњавање се односи на уградњу аутоматских вентила за селективно наводњавање које је пожељно када се на различитим парцелама гаје различите врсте засада. Осим тога, треба поставити мрежу сензора који ће давати константне информације о различитим параметрима на одређеним парцелама [3] [4].

Са наведеном опремом се могу направити регулациони кругови и потпуно аутоматизовати систем наводњавања, који тренутно захтева мануелно подешавање линија за заливање. Постојећи систем треба интегрисати са новим функционалностима кроз јединствену веб апликацију [5], [6], [8] .

LITERATURA

[1] D. Marković, R. Koprivica, U. Pešović, S. Ranđić, „Application of IoT in monitoring and controlling agricultural production“, Acta Agriculturae Serbica, Vol. XX, 40 145-153, 2015.

[2] J. Zhao, J. Zhang, Y. Feng, J. Guo: „The study and application of the IOT technology in agriculture“, Computer Science and IT (ICCSIT), 3rd IEEE International Conference on, 2: 462-465, 2010.

[3] V. C. Patil, K. A. Al-Gaadi, D. P. Biradar, M. Rangaswamy, “Internet of things (Iot) and cloud computing for agriculture: An overview”, Proceedings of AIPA, India, 292-296, 2012.

[4] N. Gondchawar, R. S. Kawitkar, „IoT based Smart Agriculture“, International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 6, June 2016.

[5] Q. Wang, A. Terzis and A. Szalay, “A Novel Soil Measuring Wireless Sensor Network”, IEEE Transactions on Instrumentation and Measurement, pp. 412–415, 2010.

[6] O. Mirabella, M. Brischetto, “A Hybrid Wired/Wireless Networking Infrastructure for Greenhouse Management”, IEEE transactions on instrumentation and measurement, vol. 60, no. 2, pp 398-407, 2011.

[7] http://elvissabac.com/proizvod/pedrollo-4sr-8-13/

[8] http://www.icent.hr/wp-content/uploads/2018/10/2.-Podnar-%C5%BDarko_AGRO-ARCA.pdf

ABSTRACT

Internet of Things (IoT) has led to major changes in the daily lives of people through application in cities, energy, business, education, medicine, industry, agriculture and other fields. The Internet of things allows the connection of a large number of users, devices, services and applications to the Internet, whereby each device is assigned a unique identifier. Interconnected devices and applications share data and forward them to remote servers so that end-users can access them as needed via mobile and / or web applications. The application of IoT in agriculture through the projected infrastructure with appropriate sensors, actuators, microcontrollers and microcomputers enables continuous monitoring and automatic start of the appropriate processes. Key activities, that can be efficiently automated using IoT, refer to irrigation processes. The theme of the paper is the model of automation of irrigation systems using IoT.

Automation of irrigation systems using the Internet of Things

Vlado Krunic, Momcilo Krunic, Predrag Ranitovic

Data Acquisition, Collection and Storage in Smart Home Solutions

Sandra Ivanović, Neven Jović, Marija Antić, Ištvan Papp

***Abstract* — In this paper, one solution for smart home IoT data collection and storage is presented. In the proposed solution, MQTT protocol is used to report changes in the system, while RabbitMQ, MongoDB, and NodeJS are used on cloud side to capture, process and store data. Processed data is stored in views that can be easily retrieved and presented on the client side. The proposed solution does not introduce significant deployment and maintenance costs, as it does not introduce technologies that are not otherwise already present in the system. Performance of the proposed solution is explored, and it is shown that the rate of changes that can be captured satisfies the needs of the medium sized IoT system.**

***Index Terms*—data acquisition; smart home automation; big data, IoT**

I. Introduction

Recently, we have been witnessing the growing popularity of various IoT systems and applications, such as fitness companions, daily activity trackers, navigation services, etc. One of the IoT areas which is of particular interest to consumers is the field of home automation [1], [2]. Home automation (HA) solutions allow users to connect multiple sensing and actuating devices into one system, to monitor their state, perform actions, or setup automatic execution of some tasks. Depending on the HA system, single networking technology can be used to connect devices [3], [4], or multiple different technologies, such as Zigbee, Z-Wave and ONVIF can work together [5-7].

HA systems generate large volumes of sensory data, even when the users are not interacting with the system. Additionally, user commands change states of actuating devices, and these changes can also be tracked. The information about the way users interact with the system can be of interest for further analysis because once it is properly operated, it can serve to improve the system itself, customer experience, as well as to users to have better insight in their system usage. This can lead to a more efficient smart home system and cost minimization.

Sandra Ivanović is with Faculty of Technical Sciences, University of Novi Sad, Serbia (e-mail: sandra.ivanovic@rt-rk.uns.ac.rs).

Neven Jović is with with the RT-RK Research Institute, Narodnog fronta 23a, 21000 Novi Sad, Serbia (e-mail: neven.jovic@rt-rk.com). .

Marija Antić is with Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia (e-mail: marija.antic@rt-rk.uns.ac.rs).

Istvan Papp is with the Oblo Living, Narodnog fronta 21a, 21000 Novi Sad, Serbia (e-mail: istvan.papp@obloliving.com).

Capturing and storing large volumes of IoT data is a challenging task, as the number of changes that should be tracked grows with the number of users and the number of devices they connect into their smart home. Typically, NoSQL databases are used to store IoT data [8], [9]. Data is aggregated and transformed periodically, in order to form batch views, tailored for particular applications. If needed, the difference between the last calculated batch view and current system state is calculated using real-time processing. Such architecture is called *lambda architecture* [10]. Although Cassandra or HDFS are usually the best choice for large scale data collection and processing systems, some of the research indicates that MongoDB can also be used in small or medium-sized solutions [11], [12].

In this paper, the focus is on user data acquisition, processing and storage for the purpose of the existing home automation system [5-7]. The goal is to build a cost-effective cloud module, which can store information about changes in HA system network (when devices are added to or removed from the HA system, or failures of devices happen), as well as track changes of particular sensors and actuators in the system. Since MongoDB is already used by the existing system's cloud, data collection modules will also use this database as the primary data storage.

First, the architecture of the existing HA system is presented, and it is explained how changes of device states are reported to HA cloud. Then, the implemented data collection cloud modules are described. Finally, the performance of the solution is analyzed, in order to determine its limitations.

II. HA System architecture

In this section, the architecture of the target HA system is described, the way devices are represented in the system, as well as the way system components communicate with each other.

Architecture of the system is presented in Fig. 1. HA gateway connects end devices using Zigbee, Z-Wave, IP and ONVIF protocols. All devices are controlled by the HA gateway, which is acting as the controller/coordinator of different PAN networks. HA gateway is connected to HA cloud, in order to provide remote control of the system using mobile client applications. In addition to remote control, HA cloud provides user management and system administration services, alarm notifications, voice control, etc. Components of the system communicate using MQTT protocol [13].

HA gateway provides an abstraction layer, allowing for all devices to be represented in the same way within the system,

regardless of their underlying communication technology [5], [6]. Every device is represented by the list of supported services and service properties. Value of each property reflects the physical state of the device (light on/off, dim level, current temperature, etc.). Devices report changes of their physical properties to the HA gateway, using the appropriate IoT communication protocol (Zigbee, Z-Wave). The HA gateway maps every change to the device representation in the HA system, and informs other components of the HA system about the changes. To distribute information about device property changes, MQTT protocol is used. Gateway publishes the information about property change to the *event* topic.

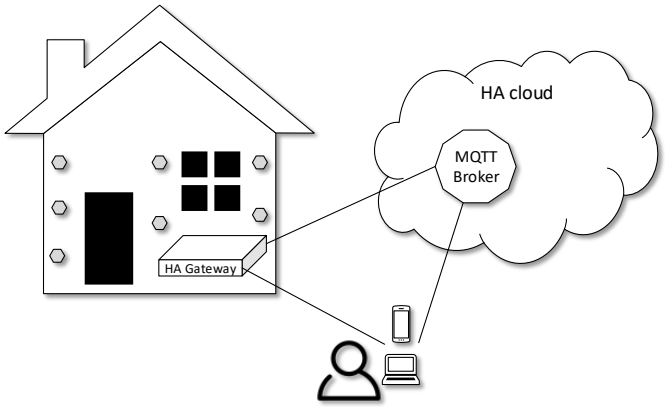


Figure 1 - Home automation system architecture overview

MQTT messages published to the *event* topic contain information about the ID of the device in the HA system, name and new value of the property that has changed, and name of the affected service – Table I.

TABLE I
DEVICE STATUS CHANGE EVENT PAYLOAD

Parameter Name	Description	Example
Device ID	Unique ID of the device within the system, assigned by the HA gateway.	12345
Service name	Name of the service	Alert Service
Property name	Name of the service property	Alert Start
Property value	Value of the property	True/False
Timestamp	Time when change happened	2019-04-03T09:52:08+02:00

III. Data Collection Service

In this section, details of the implementation of data collection module within HA cloud are presented. The main goal is to collect historical data about device property changes in the system. Additionally, for the purpose of network issues debugging, changes in the HA network are monitored, and snapshots of the local network topology are stored.

The architecture of the implemented data collection service is presented in Fig. 2. Device history sub-module monitors changes of physical device properties in the system. Network history sub-module is responsible for storing snapshots of network topology. In order to manage potential bursts of events reported by HA gateways in the system, all relevant MQTT messages are first queued using RabbitMQ [14], and then further processed, in order to store them in the appropriate MongoDB collection.

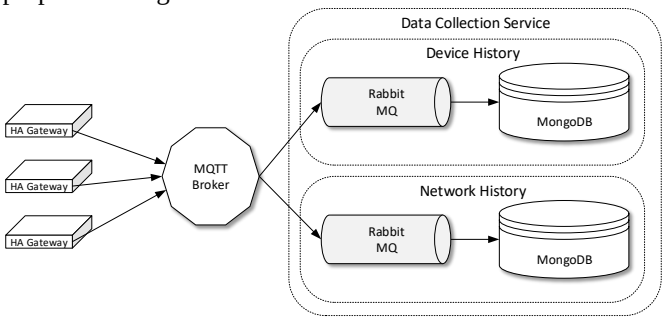


Figure 2 – Data collection service architecture

A. *Device History*

Device history module is subscribed to MQTT messages (events) reporting device state changes. As already said, device change event message contains information about the ID of the device, and modified property of some service. Depending on the service and property in question, data is stored in dedicated MongoDB collection.

All data is aggregated by the hour, i.e. every entry in the database contains the list of changes for one particular device ID, during one hour. This data is available for users to see in the web client application, once they log in into their account. Also, for some of devices, such as plugs, current state can be monitored in real-time. Graphs are shown in Fig. 3, Fig. 4 and Fig. 5.

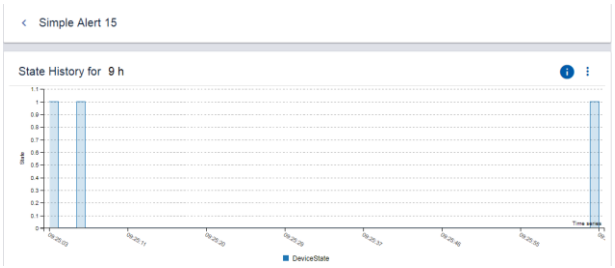


Figure 3 - Node history for the last hour

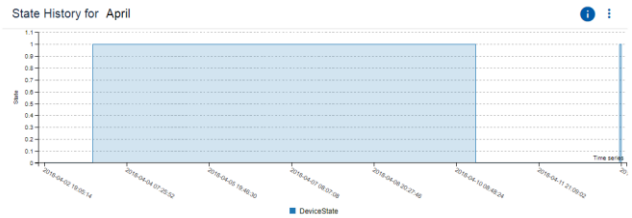


Figure 4 - Node history for the last month

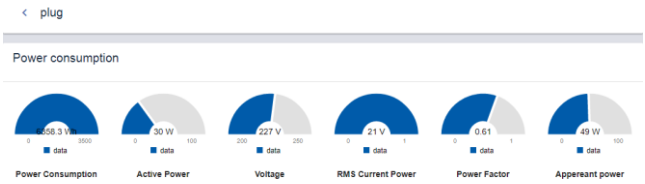


Figure 5 – Real-time changes

B. Network History

Network history sub-module provides an overview of local mesh network and device routing. This information is of particular interest for debugging purposes, if there are issues in device connectivity. Network history listens to events announcing devices that are added or removed from the system. These events cause network topology changes, which are referred to as *major changes*. Whenever a major change occurs, network history module requests detailed information about network topology from HA gateway, and stores it as a snapshot in the database. Users can browse through snapshots, as presented in Fig. 6 and Fig. 7.

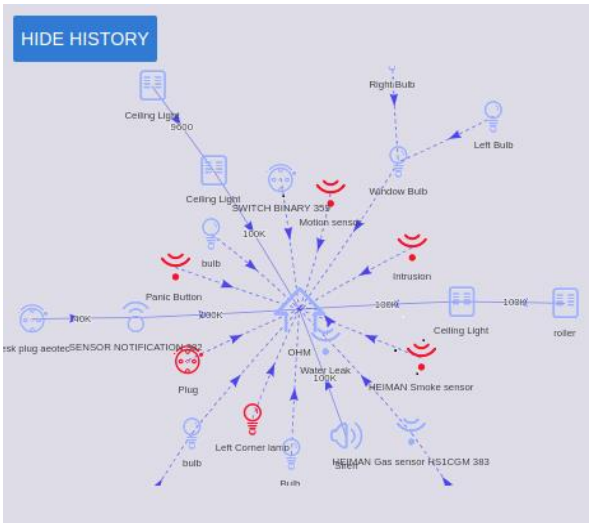


Figure 6 - Network history graph

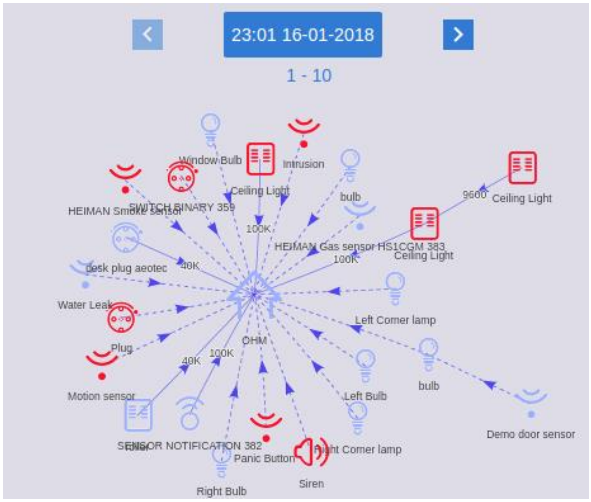


Figure 7 - Network history graph changes

IV. System Performance

In order to test the performances of the data collection module, we monitored consume time and number of messages in a queue. For the testing purposes, a lot of events had to be sent in a relatively short time period. Since it was not realistic to use real consumer data to perform these tests, we had to accomplish this by using virtual HA gateway models. Virtual HA gateway models are created solely for system testing. They are written in NodeJS and fully configurable. Virtual gateways can connect to cloud MQTT broker and generate asynchronous events, reporting the changes in the system. Parameters in the performed tests were the same for the each test. Configured values were:

- Number of virtual gateway models
- Number of different or same devices per model
- Frequency of device property change

HA virtual gateways support the MQTT communication, as well as the real one, and all the other feature that the HA gateway has. Regarding the HA Cloud architecture, it is consisted of two processor cores virtual machines, with 2 GB RAM Memory. Each service within the HA cloud setup is duplicated and event load balancing was implemented in a master slave manner. In case of an error, only one instance is affected, while the other one remains running. To improve system efficiency, database and MQRabbit are separated as virtual machines as well. Total number of the machines is 23.

As virtualization environment, ProxMox platform was used [15]. ProxMox is an open source solution for virtual server management, based on QEMU/KVM and LXC technology. It enabled us to cluster the machines, handle the instances and connection between them. Separating the micro services and scheduling them across various machines was done based on their load. Since node history is very highly loaded, it is an instance of its own. Setup architecture overview is given in the Figure 8.

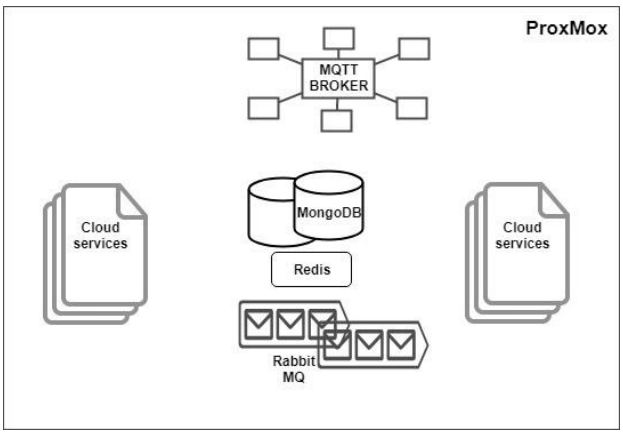


Figure 8 - ProxMox architecture overview

Total number of virtual HA gateways in setup was 600. Each one had around 30 devices.

RabbitMQ performance

When performing load tests, we monitored RabbitMQ queues, in order to detect potential limits in event load and processing. For each instance of Node History it was possible to vary the number of queues. The other variable parameter was the prefetch size. Consumer prefetch number is limit number of unacknowledged messages per channel / customer.

Firstly, we observed CPU on broker. On Figure 9 and Figure 10, we see CPU load when no events from virtual gateways had been fired. From the moment events started, we can see CPU load went above 80 percent. On Figure 11 we can see message rate in queue.

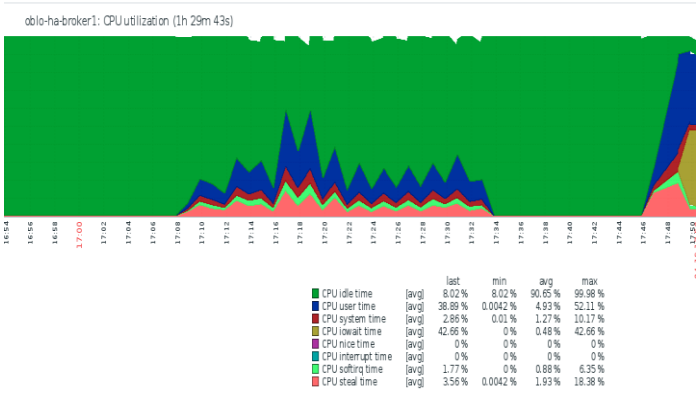


Figure 9 – CPU load

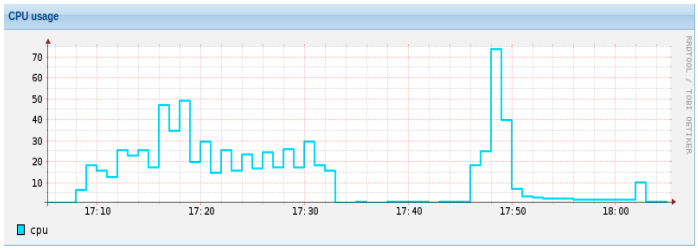


Figure 10 - CPU load graph

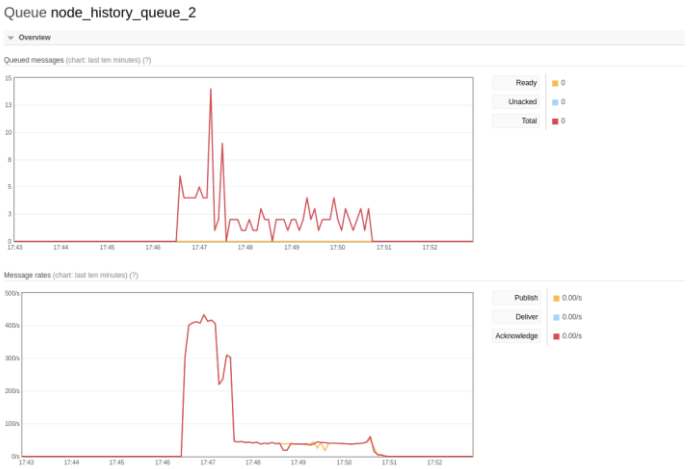


Figure 11 - RabbitMQ queue state

Secondly, we observed RabbitMQ queues and consume / publish ratio. Monitoring was done in RabbitMQ Console.

1. In the first scenario, number of queues was one, with prefetch parameter set to one as well. In Figure 12 we see that number of messages in queue went from 0 to around 60 000 in about five minutes. Also, on the lower graph we see that, in this case, consume is about 50% slower than publish rate. Therefore, this hasn't proven to be the optimum solution.

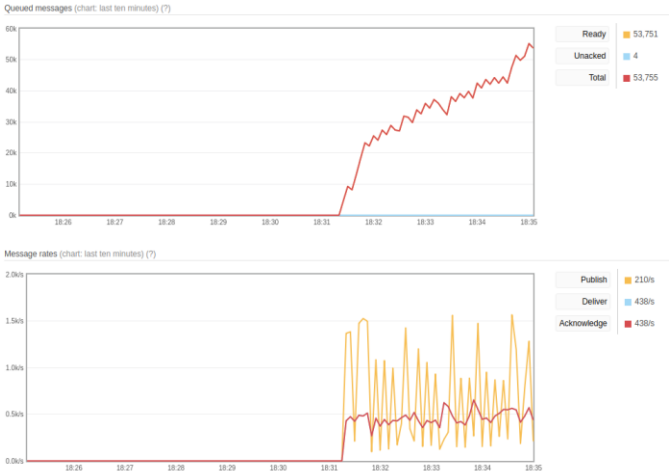


Figure 12 - One queue with prefetch parameter set to one

2. In the second scenario, number of queues was still one, with prefetch parameter set to 90. In Figure 13 we see that consume and publish rate are nearly the same, with publish being a little bit higher. Also, there are a few peaks of published messages in queue. Still, this result is much better than the previous one.

Figure 13 - One queue with prefetch parameter set to 90

3. In third scenario we tried having two queues, with prefetch parameter set to one again. In this case, we can see on Figure 14 that there was a publish peek at first that lasted for about 30 seconds, after which publish and consume rate mostly stabilized, but still publishing was faster than consuming. Graph from other queue, shown on Figure 15, show us that publish and consume rate were in fairly the same ratio as in the first queue.

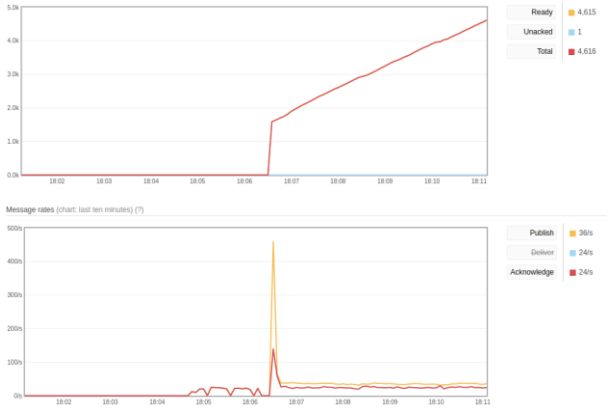


Figure 14 - Two queues with prefetch parameter set to one - First queue

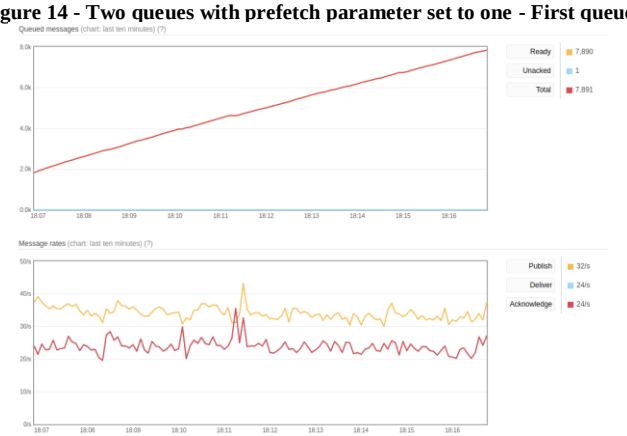


Figure 15 - Two queues with prefetch parameter set to one - Second queue

4. Fourth scenario was setting up two queues with prefetch parameter set to 90. This has proven to be the ideal case, since, as we can see on Figure 16, publish and consume rate are the same. There was only one short peak at the begging of publishing, but after that, queue was low on messages, since they were immediately consumed.

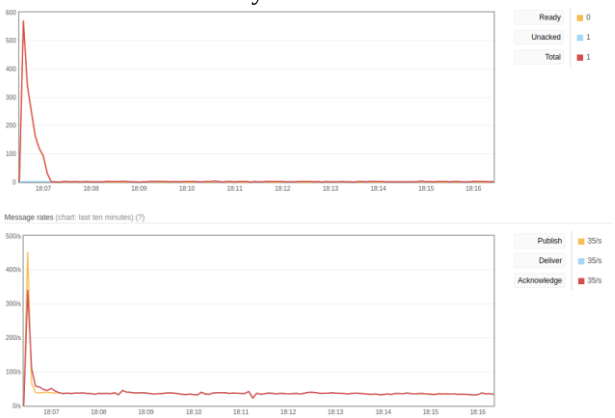


Figure 16 - Two queues with prefetch parameter set to 90

V. Conclusion

One implementation of the data collection service was presented, which is based only on the technologies that are already used in the existing HA cloud. Testing was done for the number of 600 gateway devices, simulated with virtual gateways. Event frequency was around 600 events per second. Variables in the testing process were number of queues as well as prefetch parameter. Results show us that rising the prefetch parameter improves consume rate massively.

Also, more queues help the balancing of the events, therefore we get better response time, but in combination with higher prefetch value it gives the best results.

Further tests will be done on a larger number of gateways and higher frequency rates. Also, we will test the system response times when MongoDB is replaced with Cassandra database. Another direction we would like to go for is replacing RabbitMQ with Kafka.

Acknowledgment

This work was partially supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia, under grant number: TR32014.

References

[1] C. Gomez, J. Paradells, “Wireless home automation networks: A survey of architectures and technologies”, *IEEE Communications Magazine*, Vol. 48 (6), 2010.

[2] B. Brush, B. Lee, R. Mahajan, S. Agarwal, S. Saroiu, C. Dixon, “Home automation in the wild: challenges and opportunities”, *Proc. of SIGCHI Conference on Human Factors in Computing Systems*, 2011.

[3] K. Gill, S. H. Yang, F. Yao, X. Lu, “A zigbee-based home automation system”, *IEEE Transactions on Consumer Electronics*, Vol. 55 (2), 2009.

[4] R. Piyare, M. Tazil, "Bluetooth based home automation system using cell phone", *Proc. of IEEE ISCE*, 2011.
[5] M. Tucic, R. Pavlovic, I. Papp and Dj. Saric, "Networking layer for unifying distributed smart home entities", *Proc of IEEE Telfor*, 2014
[6] V. Moravcevic, M. Tucic, R. Pavlovic and A. Majdak, "An approach for uniform representation and control of ZigBee devices in home automation software", *Proc. of IEEE 2015 ICCE– Berlin*, 2015
[7] M.Sekulic, I. Lazarevic, M.Bjelica and V. Pekovic, "Asynchronous application programming interface library for distributed home automation software", *Proc. of IEEE 2015 ICCE– Berlin*, 2015.
[8] H. L. Truong, S. Dustdar, "Principles for Engineering IoT Cloud Systems", *IEEE Cloud Computing*, Vol. 2 (2), 2015.
[9] H. Cai, B. Xu, L. Jiang, A. Vasilakos, "IoT-based Big Data Storage Systems in Cloud Computing: Perspectives and Challenges", *IEEE Internet of Things Journal*, Vol. 4 (1), 2016.
[10] N. Marz, J. Warren, "*Big Data: Principles and best practices of scalable real-time data systems*", Manning Publications Co., 2015.
[11] J. S. van der Veen, B. van der Waaij, R. J. Meijer, "Sensor Data Storage Performance: SQL or NoSQL, Physical or Virtual", *Proc. of IEEE Conference on Cloud Computing*, 2012.
[12] Y. S. Kang, I. H. Park, J. Rhee, Y. H. Lee, "MongoDB-based Repository Design for IoTgenerated RFID/Sensor Big Data", *IEEE Sensors Journal*, Vol. 16 (2), 2015.
[13] *MQTT Version 3.1.1*, OASIS standard, 2014, Available [Online]: http://docs.oasis-open.org/mqtt/mqtt/v3.1.1/os/mqtt-v3.1.1-os.pdf
[14] David Dossot, "*RabbitMQ Essentials"*, Packt Publishing, 2015.
[15] Rik Goldman, "*Learning Proxmox VE"*, Packt Publishing, 2016.
[16] Josiah L. Carlson, "*Redis in Action"*, Manning Publications Co., 2013

Using Online Weather Data to Improve Smart Home User Experience

Milica Matić, Milan Tucić, Marija Antić, Roman Pavlović

***Abstract* – In this paper we present one solution for integrating online geolocation and weather services with the existing smart home system. The goal is to improve smart home user experience, by enabling automatic system reaction to different astronomical data parameters (sunrise, sunset, moon phase), or changes in weather forecast. The existing smart home system is extended to hold information about user location, and to monitor online weather data for this location. Details of the implementation and communication between components of the system are presented and functional and load testing of the implemented service are evaluated.**

Index Terms – smart home; weather; geolocation; cloud; gateway

I. INTRODUCTION

The progress of computer and networking technologies has enabled the development of various smart devices, which can be controlled and monitored remotely. These devices are grouped into Internet of Things (IoT) networks in which they are communicating with each other, in order to convey status reports, network control messages and commands [1]. IoT technologies can be applied to both industrial and consumer use-cases. Smart home systems represent one of the growing areas of IoT consumer applications [2] – [4]. In smart home solutions, various devices (sensors and actuators) can connect to automate daily tasks in the household, and enable remote control and monitoring of the system by end users.

Typically, actions in the smart home system are executed by the user command from client application, but they can also be automated by a set of user-defined rules. Rule execution can be triggered at a certain time during the day, or as a reaction to values measured by some sensors in the system [5]. In this paper, we explore the possibility of extending smart home user experience by adding online weather data information to the system. Measurements of current sunlight and temperature have previously been used to control power consumption in smart homes, and effort has been made to design algorithms to control this process automatically [6] – [8]. However, in these scenarios user has no control over the actual algorithm implementation. The idea behind this work is to represent weather data in a form of a virtual "weather device", according to the unified device model in the existing smart home system [9]. This way, users can monitor current weather information, as well as set rules which use weather information as triggers. For example, the system can be set to turn off all lights after sunrise, or turn them on after sunset. The information about sunrise and sunset is retrieved from the online weather service, based on the smart home location. Also, the system can be set up in a way that allows notifications to be sent to users, whenever weather data changes. For example, user can be notified if it is raining outside or if the storm is close, and similar. Additional rules can be configured by the user, in order to protect the home from bad weather conditions, or setup heating or cooling depending on the outside temperature changes.

As already said, online weather data will be used in the existing smart home system. Therefore, the architecture of the target smart home system is first presented, and communication between system components is explained. Then, details of the integration with location and weather services are provided. Finally, functional and load testing of the newly implemented module is performed.

II. SMART HOME SYSTEM ARCHITECTURE

In this section, the architecture of the existing smart home system is briefly explained – Fig. 1.

The existing home automation system [5], [9] allows users to connect end devices (nodes) which use Zigbee, Z-Wave, ONVIF, and IP to communicate with each other. The central component of the system is the home automation gateway, which acts as a bridge between different IoT networks integrated into a system. It creates a level of abstraction, so that all nodes in the system are represented in the same way, by a set of their functionalities and characteristics [9]. This allows applying the same control logics to all of the nodes, while being agnostic of the actual communication protocol components are using. With this abstraction, gateway becomes bridge for communication between cloud and end devices, in the system [10].

User application allows user to manage and control their smart home system. They provide intuitive UI to all gateway functionalities allowing user fine-grade control over gateway while in the same time hiding complexities of underlying technologies.

Remote access to the system is provided by the home

Milica Matić is with Faculty of Technical Sciences, University of Novi Sad, Serbia (e-mail: milica.matic@rt-rk.uns.ac.rs).

Milan Tucić is with with Oblo Living, Narodnog fronta 21a, 21000 Novi Sad, Serbia (e-mail: milan.tucic@obloliving.com).

Marija Antić is with Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia (e-mail: marija.antic@rt-rk.uns.ac.rs).

Roman Pavlović is with Oblo Living, Narodnog fronta 21a, 21000 Novi Sad, Serbia (e-mail: roman.pavlovic@obloliving.com).

automation cloud, as well as a number of advanced features, such as voice control and integration with different third party services. Cloud has a micro-service architecture, i.e. it consists of multiple clearly decoupled services, each being in charge of different functionalities. For communication between different cloud micro-services HTTP protocol is used. Also, HTTP REST APIs are used to integrate with various third party services, such as video surveillance systems, voice control (Amazon Alexa, Google Home), etc.

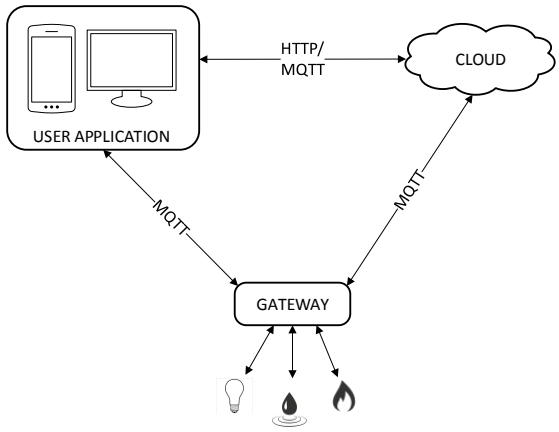


Fig. 1. Smart home system architecture

A. *Communication between smart home system components*

Main components of the home system communicate using MQTT (Message Queuing Telemetry Transport) protocol [11]. A universal messaging model is applied to all system components, i.e. the MQTT topic structure and payloads are defined on system level. MQTT broker defines topics that client is allowed to send messages on, based on the client type (gateway, application, or cloud micro-service).

There are three different message types exchanged between the system components:

1) Request messages – used to send commands or queries
2) Response messages – used to respond to requests
3) Event messages – used to asynchronously report changes in the system (device property changes, alarms, etc.)

In this paper, the communication between gateway and cloud is of interest. Topics that MQTT messages are published to contain information about the following:

- gateway id – unique gateway identifier
- user id – unique identifier of user connected to specific gateway
- service name – cloud service name that gateway is sending message to
- type – message type (request, response, event)

Regardless of their type, all MQTT messages contain the message ID and a variable number of parameters - TABLE 1.

TABLE 1
MQTT MESSAGE PAYLOAD

Field name	Required	Description
ID	yes	Unique message identifier
Parameters	no	Message parameters

The actual parameters depend on the message type. Common parameters for all request messages are listed in TABLE 2. Every request message contains information about its sender and type (defining if it's a command or a query), as well as the name of the message. Common parameters for all response messages are listed in TABLE 3. As a response to a command or query, status code is returned, which may be accompanied by additional information in the description. Finally, event messages must carry the event name, but there can also be additional information within the optional event fields - TABLE 4.

TABLE 2
REQUEST MESSAGE PARAMETERS

Field name	Required	Description
Sender	yes	Message sender
Type	yes	Message type (command or query)
Name	yes	Message name

TABLE 3
RESPONSE MESSAGE PARAMETERS

Field name	Required	Description
Code	yes	Is request processed successfully
Description	no	Additional information

TABLE 4
EVENT MESSAGE PARAMETERS

Field name	Required	Description
Name	yes	Event name

III. WEATHER AND GEOLOCATION INTEGRATION

In this chapter, the details of integration with third party geolocation and weather services are presented. On gateway side, the weather information is incorporated into a virtual weather device, while on cloud side two micro-services are implemented, which communicate with the selected geolocation and weather APIs. Gateway communication model is extended to support new cloud. User application is customized for a better user experience by allowing user to monitor information obtained from online weather service and to act upon it.

In Fig. 2, a high-level architecture of the solution is presented. First, during first time run and system setup, or when the home address is changed, geolocation micro-service within home automation cloud is notified by the application. This micro-service then maps the address provided by the user

to the location information (longitude and latitude) of the city the provided address belongs to. Location information is then used by the weather micro-service, as a parameter required by the third party weather service API, to fetch weather forecast and astronomical data for the location in question. Changes in weather data are reported to gateway and applications using MQTT event messages. In order to avoid frequent communication with third party weather service, information about current weather is cached in Redis [12]. Cached data is set to expire within a predefined period of time, and after this time has elapsed it will be automatically deleted from cache. Whenever a gateway requests information about weather in its location, the lookup in weather cache is first performed. If there is no entry in weather cache, the fresh weather data is fetched from the third party weather service, new entry in weather cache is formed, and all gateways in the same location are informed about new weather parameters.

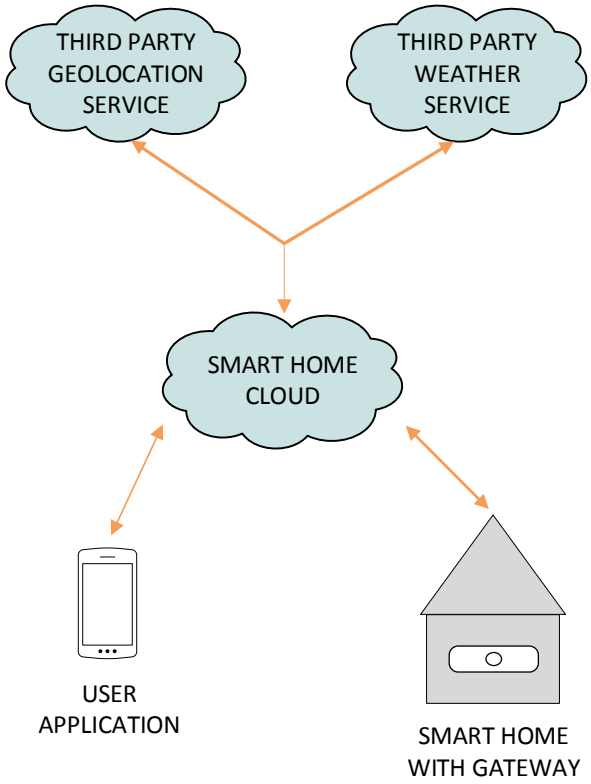


Fig. 2. High level system architecture

A. *Geolocation micro-service*

In order to be able to fetch relevant weather data from online weather service, home automation cloud has to be aware of the gateway location. Therefore, end user has to allow the smart home system to use their personal information, i.e. home address. There are different techniques to obtain location information:

1. GPS location – method used to detect exact position of the user's mobile device
2. Mobile cell information – using triangulation algorithm based on cell tower information where cell towers in exact locations are well known
3. WiFi network information – Internet services constantly update WiFi network information from routers all over the world. These services match GPS data and available WiFi networks information sent from mobile phones thus calculating precise location of each WiFi router
4. IP information – user's location is fetched based on the device IP address
5. Providing location via smart home application – user manually enters postal address of his household.

Techniques 1 to 4 are already used by mobile phones and browsers to estimate the location of the user during the home automation system setup. If the automatic estimation of the address is not precise enough, user can always provide the correct address manually, using the application.

Geolocation micro-service is in charge of mapping the provided address to the latitude and longitude. It communicates with third party online service in order to geocode the given address [13]. Once address is geocoded and latitude and longitude values are obtained from third party online service, unique identifier (location ID) representing specific location is created and stored in the database, containing information about the location it refers to. Since there is no need to request weather information for exact latitude and longitude of every household in the system, only city-level weather information is of interest. Therefore, all gateways in the same city are mapped to the same location ID, and information from the geolocation location ID entry in the database is afterwards used in communication with third party weather service. Also, once the gateway address is mapped to the location identifier, the gateway is notified about the assigned location ID, in order to correctly subscribe to the relevant weather information.

In case where user applications provide address information without user interference, usually they provide geolocation micro-service with latitude and longitude values. On condition that this data is provided to it, geolocation micro-service will do reverse geocoding [14] of provided data in order to save physical address to database. In this case, third party geolocation is called, as it is for geocoding purposes.

Figure Fig. 3 represents communication taken between home automation gateway, geolocation micro-service within home automation cloud and third party geolocation service in order to connect home automation gateway with physical address.

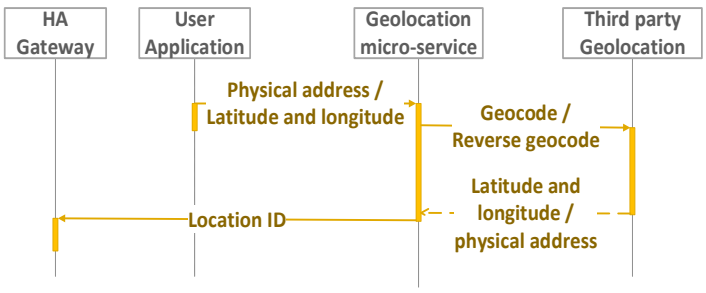


Fig. 3. Obtaining location information from third party geolocation service

B. *Weather micro-service and weather device*

When weather service is enabled for the particular gateway, information retrieved from it will be presented as properties of a virtual weather device. This virtual device is listed in the user application among other physical devices. Virtual weather device abstracts services and properties given in the TABLE 5.

TABLE 5
WEATHER DEVICE SERVICES

Service name	Description
Address	Physical address of the gateway
Astronomy time	Sunrise and sunset information
Location	Gateway latitude and longitude
Weather configuration	Weather data refresh interval
Weather	Weather information
Wind	Wind information

Users can configure how often weather information should be updated for their virtual device. After the specified time period, gateway will query the cloud for updated weather data.

Gateway asks cloud for new data in three cases:

1. Gateway has come online - if gateway was offline, first time it comes online, weather data should be checked
2. Gateway's location has changed – in case user has changed gateway's address, weather data has to be updated accordingly
3. Refresh period time has passed and data needs to be updated.

The communication between the gateway, weather micro-service, weather cache and third party weather service is depicted in Fig. 4.

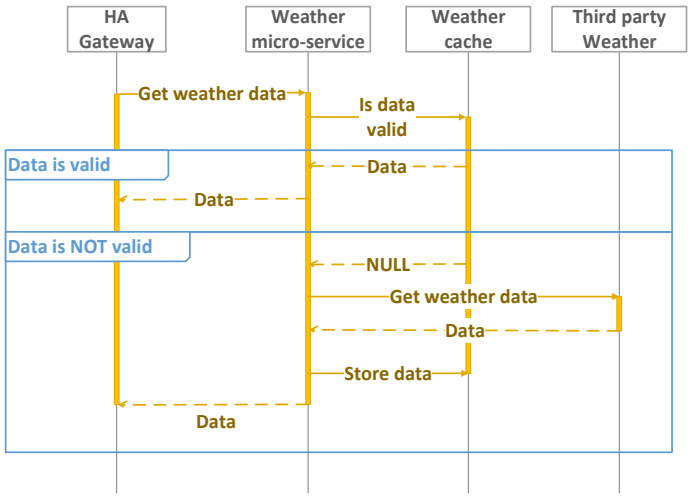


Fig. 4. Obtaining weather information from third party weather service

Gateway sends MQTT request to home automation cloud, to fetch new weather data. Cloud first checks the weather cache for information. If cached data is valid, cloud will fetch data from cache and send it to the gateway as a response message. However, if data in cache is not valid, cloud has to contact third party online weather service to fetch new data. Once data is obtained from the third party weather service, response is sent to gateway. In order to minimize the number of query requests sent by gateways in the same location, notification by MQTT event messages is added. Hence, after sending the response to the gateway which asked for updated weather information, cloud will notify all gateways in the same location, by sending them updated weather data in an MQTT event message.

Whenever a new location (location for which weather data was not obtained before) is added to the system, weather information has to be obtained for it. New location is saved to the database with location ID created by geolocation micro-service. Then, weather micro-service will fetch data from online weather service and store this information in the weather cache.

C. *Messaging model extension*

In order to support new geolocation and weather service functionalities, MQTT messaging model had to be extended to support new topics and messages.

New topic was added to allow communication between gateways and geolocation micro-service. It is used for providing information about new postal address of the gateway. Also, two new topic groups are added for communication between gateways and weather micro-service. One topic group is used for direct communication between one specific gateway and the cloud, in cases when the gateway requests weather data and cloud sends response to it. Second topic group is used for cloud to send events about weather data changes to groups of gateways in the same location. For every location, a new topic is formed.

IV. EXPERIMENTAL EVALUATION

In order to fully test provided solution, we had to use gateways located in different world areas. This was done by using virtual gateway model. Virtual gateway model is a software gateway model created for purpose of system testing. It is implemented in NodeJS and can be configured to support various end devices and rules. Software model can be assigned to a user in the system, same as the physical gateway. After linking the user profile with the virtual model of the gateway, user is able to control it in the similar way as he would control real physical device. Virtual gateway model supports communication using MQTT protocol, as well. Existing model is extended to support geolocation and weather services and to test communication between cloud and gateway. Various locations were added to the system, and cloud was tested for obtaining data for proper locations and serving it to the proper gateways.

Home automation cloud architecture is presented on the Fig. 5. Every block in the figure represents one virtual machine with two processor cores and 2 GB of RAM memory. These virtual machines are part of ProxMox virtual environment [15]. Geolocation micro-service is located in one of three machines with home automation micro-services. As figure shows, geolocation micro-service and Redis database

are placed in separate virtual machines which reduce the load on the individual machine.

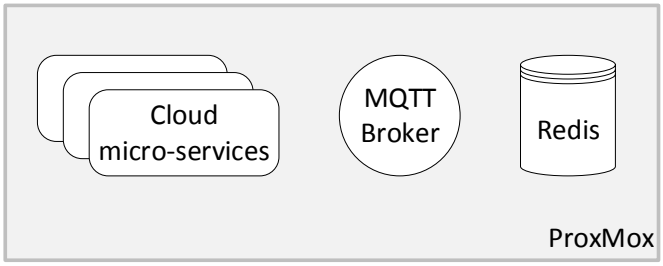


Fig. 5. Home automation cloud architecture

First, we test the performance of Redis weather cache, depending on the number of different location IDs in the system. We observe memory consumption and response times when gateway requests information, as well as CPU load on Redis instance.

Further, we vary the frequency of weather data update on the gateways, in order to determine its impact on the system performance.

For testing purposes 4000 virtual gateways, all located in different locations, were configured to obtain weather data in the same time. Weather cache was cleared before tests were started. Additionally, we tested system where 4000 virtual devices were located in the same location, but frequency of weather data update was changing.

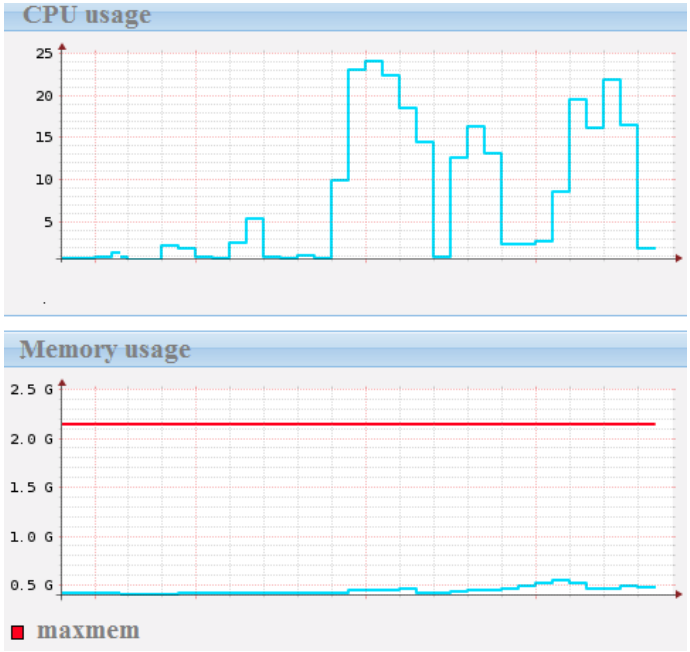


Fig. 6 Average CPU and memory usages on Redis machine

Average tests results are depicted in Fig. 6. Results show that CPU usage does not exceed the value of 20%, and memory consumption is beneath 1 GB after 8000 different weather information is stored in Redis.

Finally, we test the end-to-end performance when using real physical gateway and third party geolocation and weather services. Results are presented in the TABLE 6. Typically, it takes 910,89 ms to map the address to the location ID if data is obtained from online third party service, apropos 39,43 ms if data is gathered from database. Fetching weather data from weather cache requires 14,45 ms, on average. However, if data is obtained from online third party weather service, it takes 222,11 ms to get the data, on average.

TABLE 6
EXPERIMENTAL RESULTS

Test case	Average time in ms
Get existing location ID from database	39,43
Get location from third party online service	910,89
Get weather data from cache	14,45
Get weather data from third party online service	222,11

V. CONCLUSION

One approach to using weather data within the smart home system was presented. This approach enables users to monitor weather information for their household location, as well as set automatic rules that can be executed if weather or astronomical data changes in some way of interest. Future work will be focused on better location analysis, as bigger cities can have multiple localities, and obtaining weather data for these localities could provide better and more accurate results.

ACKNOWLEDGMENT

This work was partially supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia, under grant number: III_044009_2.

REFERENCES

[1] K. J. Singh, D S. Kapoor, “Create Your Own Internet of Things: A survey of IoT platforms”, *IEEE Consumer Electronics Magazine*, Vol. 6 (2), 2017.

[2] B. Brush, B. Lee, R. Mahajan, S. Agarwal, S. Saroiu, C. Dixon, “Home automation in the wild: challenges and opportunities”, *Proc. of SIGCHI Conference on Human Factors in Computing Systems*, 2011.

[3] S. H. Park, S. H. Won, J. B. Lee, S. W. Kim, “Smart home – digitally engineered domestic life”, Personal and Ubiquitous Computing Journal, Vol. 7 (3-4), 2003.

[4] Moataz Soliman, Tobi Abiodun, Tarek Hamouda, Jiehan Zhou, Chung-Horng Lung, “Smart Home: Integrating Internet of Things with Web Services and Cloud Computing”, 2013 IEEE 5th International Conference on Cloud Computing Technology and Science

[5] I. Lazarević, M. Sekulić, M. Savić, V. Mihić, “Modular home automation software with uniform cross component interaction based on services”, *Proc. of IEEE ICCE-Berlin*, 2015.

[6] H. Tiescher, G. Verbic, “Towards a Smart Home Energy Management System – A Dynamic Programming Approach”, *IEEE PES Innovative Smart Grid Technologies*, 2011.

[7] T. Zhu, A. Mishra, D. Irwin, N. Sharma, P. Shenoy, D. Towsley, “The Case for Efficient Renewable Energy Management in Smart Homes”, *Proc. of ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*, 2011.

[8] M. Khan, B. N. Silva, K. Han, "Internet of Things Based Energy Aware Smart Home Control System", *IEEE Access*, Vol. 4, 2016.
[9] M. Pandurov, I. Lazarević, R. Pavlović, N. Smiljković, "Unified device access in home automation environment", *Proc. of TELFOR*, 2014.
[10] M. Tucić, V. Moravčević, G. Velikić, Đ. Sarić, V. Mihić, "Device abstraction and virtualization: Concept of device in device", 2015 IEEE 5th International Conference on Consumer Electronics - Berlin (ICCE-Berlin)
[11] V. Karagiannis, P. Chatzimisios, F. Vazquez-Gallego, J. Alonso-Zarate, "A Survey on Application Layer Protocols for the Internet of Things", Transaction on IoT and Cloud Computing, vol. 3, no. 1, pp. 11-17, 2015
[12] V. Abramova, J. Bernandino, P. Furtado, "Experimental evaluation of NoSQL databases", International Journal of Database Managament Systems, vol. 6, no. 3, June 2014
[13] X Ge, Google LLC, Address geocoding – US Patent 6,934,634, 2005
[14] M. Zarem, E. Vuillermet, J. DeAguiar, TeleCommunication System Inc, Intelligent reverse geocoding – US Patent 8,731,585, 2014
[15] B.R. Chang, H.-F. Tsai, C.-M. Chen, "Evaluation of Virtual Machine Performance and Virtualized Consolidation Ratio in Cloud Computing System", Journal of Information Hiding and Multimedia Signal Processing, vol. 4, num. 3, July 2013

Industrial Fog Computing Platform and System Testing Through GUI

Rade Tišma, Ivan Velikić and Velibor Mihić, *Member, IEEE*

***Abstract*—Fog Computing is next step in supporting industrial IoT. It offers solutions for harsh demands in modern industrial environment. Manipulation with enormous data generated on daily level in industry and enabled almost real time analysis with safeness always on mind, is supported by Fog Computing model. This paper shortly presents industrial IoT platform based on Fog Computing architecture. Basic functionality of available GUI components is described with some quick turn on system testing through available interfaces for users.**

***Keywords*—Fog Computing, Edge Computing, Industrial Internet of Things, Cloud Computing.**

I. Introduction

In the world of industry (machines, sensors and actuators), which is connected on different levels and strives to be even more connected, integrated to look like one automated system, and where need for flexible structure emerges to support incoming rapid business demands, Fog Computing comes to scene to offer a solution. Fog Computing (or Edge Computing, also commonly used synonym), term firstly used by Cisco company, shifts storing and processing power to the edge of network, close to source [2]. The massive amount of data that industrial IoT (*Internet of Things*) devices are generating presents problem for exclusive use of cloud computing model [2]. Cloud architecture relies on internet links and their throughput is the bottleneck for IoT traffic. Also, internet links are unstable and variable. Transferring a lot of data is not the only biggest challenge for industrial IoT, real time response is also very important factor. Real time reaction, system reaction in real time based on data accessible in milliseconds (analysed in almost real time) is imperative for modern industrial systems. The security of sensitive data can be on highest level simply because there is no need to distribute all data over internet [1]. Only partial data can be sent over internet (in raw form or analysed), which is useful for high level analysis. Openness of this model allows user to choose, between other things, which data are going to be sent and in what form. An implementation of Fog Computing architecture, Nerve platform [3], provides openness when it comes to connectivity with multiple IoT cloud solutions by using protocols AMQP, MQTT, REST/JSON and OPC UA. Beside capability to integrate with existing infrastructure of user, there is also ability to connect with internal cloud-based component called fogSM (*Fog System Manager*) [3]. FogSM is central component that manages fogNodes (edge devices) and underlying applications. FogNode [3] is a network edge device that combines the capabilities of a computer system, network system and storage system in purpose-built hardware for full life cycle management of IoT edge devices. It hosts both real time and non-real time applications on the same node. FogOS (*Fog Operating System*) [3] is software stack that runs on fogNode, based on CentoOS Linux distribution.

Rade Tišma is with RT-RK Institute for Computer Based Systems, Jovana Dučića 23A, 78000 Banjaluka, Bosnia and Herzegovina (e-mail: rade.tisma@rt-rk.com).

Ivan Velikić is with the RT-RK Institute for Computer Based Systems, University of Novi Sad, Narodnog Fronta 23, 21000 Novi Sad, Serbia (e-mail: ivan.velikic@rt-rk.com).

Velibor Mihić is with the RT-RK Institute for Computer Based Systems, University of Novi Sad, Narodnog Fronta 23, 21000 Novi Sad, Serbia (e-mail: velibor.mihic@rt-rk.com).

II. FogOS – Software Stack

FogOS is software stack based on CentOS Linux distribution that supports secure virtualization of Linux and Windows based virtual machines, Docker container environment management, Windows application hosting (within Windows virtual machines), and open interfaces for fieldbus device communication. FogOS provides complete fog computing solutions along with fogSM. CentOS serves only to support all other applications and services that are part of Nerve software and rely on it. Most important thing that CentOS supports is integrated virtual machine – adminvm. FogSM is part of fogOS and it represents cloud-based component for managing fogNodes and applications on them [3]. FogOS enables storing, visualization and processing of collected data using customized open source and proprietary software. KVM (*Kernel-based Virtual Machine*) hypervisor (part of fogOS) runs VM (*Virtual Machines*) in real time, with user software, and allows user to integrate existing software into Fog Computing platform. Figure 1. illustrates virtual machines deployed on fogNode by user and real time OS.

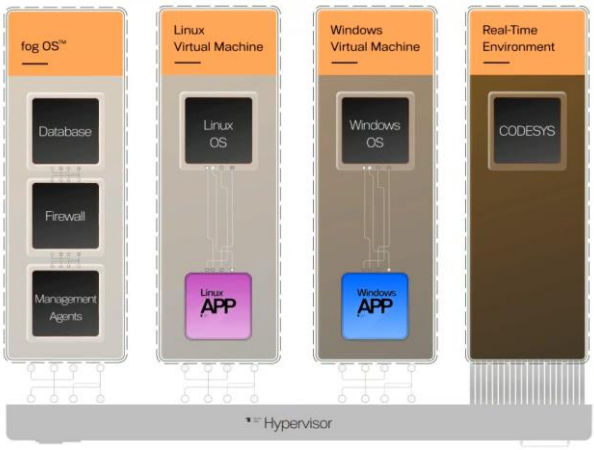


Fig. 1. Users virtual machines and Real Time OS with Codesys.

For the system users on the local level (near to a fogNode) there is a local UI (*User Interface*). Local web server allows users to interact through local UI with fogNode. To host a web server and to manage Docker containers (deployed by user) an integrated virtual machine (adminvm) is used. Local UI provides network related settings for fogNode, fogSM connection settings (IP address field), user password changing, and displaying some information for running system such as: CPU usage, RAM usage, IP address, software version, etc. Analysis and visualisation of gathered data is available with Grafana [4] – a general purpose dashboard and graph composer, which runs as a web application [3]. The collected data are stored locally using open source TSDB (*Time Series Databases*) – influxDB [5]. At the same time, it is possible to send data to fogSM, where the user can view data in different forms. Figure 2. shows WAN interface settings.

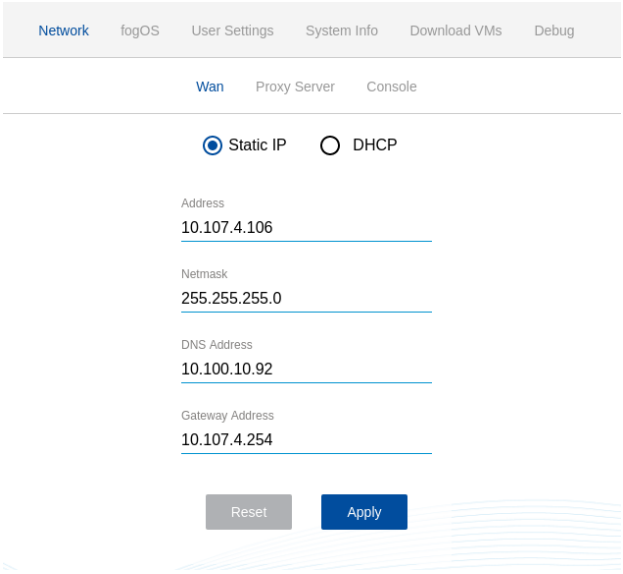


Fig. 2. Wan interface settings for local UI.

System testing of local UI can't be done using only interface itself. Testing certain functionality requires doing checkups on underlying system by using CLI (*Command Line Interface*) tools. Changing network configuration in local UI can be verified with CLI tools on the system, like ip utility tool. The Local UI also provides feature to download raw VM images and image definition file of a virtual machine from the fogNode to the development machine. This feature supports scenarios where a VM is locally modified by the user and then stored, either to be rolled out to other fogNodes or to be used as backup. Testing VM download on local UI requires background check on the system. For this purpose qemu-img tool [6] is used to verify downloaded images. Local UI provides a few features more. It gives a user ability to configure network interfaces according to internal network structure, to download changed VMs (for backup purpose or distribution). Also it allows setting connection with cloud component and getting some system information that can be very useful to the end user.

III. FOGSM - SYSTEM MANAGER

FogSM is embedded cloud-based component of fogOS that manages fogNodes, application provisioning, and deploying over cloud portal. It can be deployed on commercial cloud services such as AWS (*Amazon Web Services*) and Azure. Or at customer's premise server. In both cases it's necessary to have fully qualified hostname of the fogSM instance to connect to the cloud portal. Figure 3. shows fogSM cloud portal interface.

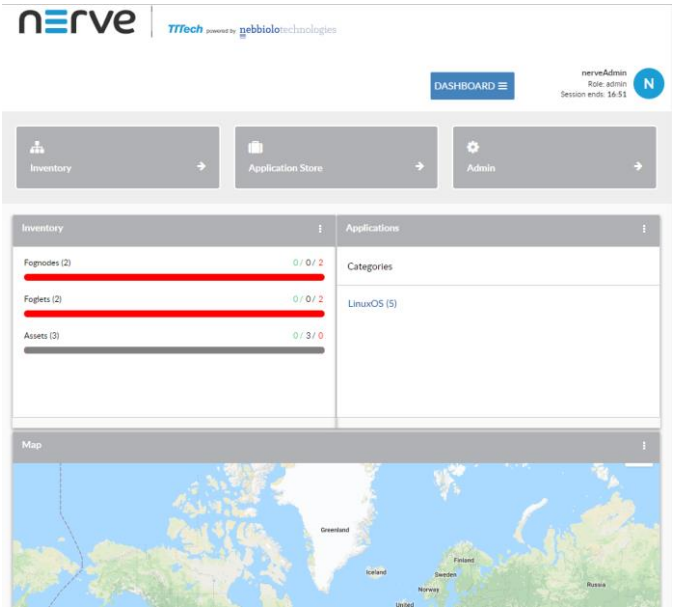


Fig. 3. Home page of fogSM portal interface.

Interface menu offers user management – RBAC (*Role Based Access Control*), fogNode management, and

application management. Main menu consists the following six items:

A) Dashboard
B) Inventory
C) Config Store
D) Application Store
E) Datastream
F) User

A. Dashboard

Dashboard is the default view for logged user. It gives a quick overview of currently connected fogNodes, notifications for user actions, provisioned applications, and geographical location of fogNodes on map.
Testing dashboard is pretty much straightforward because it requires testing correctness of links to specific page (e.g. Inventory) and comparison of data with data on particular page.

B. Inventory

Inventory page allows users to manage fogNodes and software on them. Opening this page reveals topology browser that hierarchically organize the devices (fogNode, assets, virtual assets, software components, virtual machines, containers). Selecting any topology node provides information that is applicable to that node and its children (descendants) [3]. Information that is displayed is node type dependent. Basic info for each node consists of connection status, version information and network related information. It's also possible to see system information such as CPU type, amount of free RAM, storage capacity and network interfaces (drivers in use). Grafana [4] allows users to view data collected on the particular fogNode through fogSM database. User can create, edit, move and delete nodes. The different topology nodes in the hierarchy are arbitrary and user defined to enable customization of device organization. Each fogNode has its own ID, in order to be uniquely identified on fogSM. When it comes to testing inventory functionality it is mostly based on using interface itself. Only for testing topology tree regeneration (e.g. removing fogNode from it) and getting it back after restarting fogNode, it's necessary to read fogNode ID directly from the system (using CLI tool like cat) in order to check presence of reattached fogNode that was previously removed.

C. Config Store

It allows users to store application configurations. The configs can be a text-based config files such as json, xml, yaml etc or it could be a binary config files. The configurations can be added or deleted from the config store. Each config is identified by a unique key. Once the config item is created, it can be applied to a Docker or a Windows application [3]. Configs can be added or deleted from the CONFIG STORE section. For a text configs, appropriate type such as json, yaml can be selected to enable syntax highlighting. For testing applied config files it is necessary to check Windows or Docker application on fogNode.

D. Application Store

Application hosting feature provides the function of on-boarding applications into the application store and then deploying them on the fogNodes. Application hosting comprises of an application store and an associated set of operations, which are application on-boarding, application authorization, application deployment, application policy management, and Role Based Access Control management [3]. Different application categories are used to scope the list of applications currently available for deployment. Application categories are used to group a set of related applications.
Testing application store functionality, to be more specific, application deployment, demands verification on fogNode. For everything else, interface itself is sufficient.

E. Datastream

Datastream offers to user ability to create different stream pipes with analysis tools using simple intuitive drag&drop technique for stream pipe creation. User can save, edit and delete created stream pipe. Deployment of created&saved pipe is not restricted to fogSM, it can be deployed on any connected fogNode. Grafana is used for testing created data streams, both on fogSM and fogNode.

F. User

User page gives access to notifications (for each user action there is response from system - notification), information for system version, RBAC management (role creation or editing, operation and resource managing, user bindings) and for administrator account user and tenant management. User page also provides some system related settings, like allowing user to choose between Google Map or Open Street Map use for fogNode location. Some under development features can be enabled (for testing purpose) and a few minor things. Testing user page functionality is mainly done over existing interface because there are no fogNode related settings or anything else that involves communication with fogNode.

IV. Conclusion

Nerve (Fog Computing) platform for industrial IoT is complex software stack that includes a lot of customized or configured open source solutions chained to work together. GUI (*Graphical User Interface*) components offer good interaction with underlying system and allows user to have better insight of entire working process. Openness is a key factor for integration into existing software environment and at the same time it's biggest strength for this Fog Computing implementation. Comparing Nerve to other solutions is not very easy task. For each platform there is another approach and also different set of protocols that are used to leverage specific hardware – i.e. gateway. Most of platforms are built with specific hardware and that is not the case with Nerve. Nerve comes with purpose built hardware (fogNode) but it's designed to be hardware independent Fog Computing platform. Connection to several higher-level systems (clouds) via different protocols (if not available in the base product, the

function is easily expandable) and integrated Codesys, which enables accessing data, as well as taking over the control, are features that are not present on another platforms - at least not in most of them. Another thing that differentiate Nerve from other similar platforms is installation of additional applications (Docker and Windows) and virtual machines specific for particular customer.

When it comes to system testing through available GUI components it brings chalenges in entire process. Since GUI components are web applications (fogSM and local UI) it's natural to automate them using Selenim and Selenium bindings for Pyhon, Java, Javascript or other suitable programming language. Difference in automating tests for usual web applications and Nerve components (fogSM and local UI) is in existing hardware that is mostly managed through GUI. So, additional checks are required in order to get valid test results. Formerly means that operation or action taken in GUI, in majority of test cases, needs to be validated on fogNode (on CentOS host or adminvm) using some of Linux CLI tools available on distribution. This is one of main reasons why running automated system test suite takes so much time, and sometimes implementing particular automated test case can be really chalenging.

ACKNOWLEDGMENT

This work was partially supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia, under grant number: III_044009_2.

REFERENCES

[1] Amir M. Rahmani,Pasi Liljeberg, Jürgo-Sören Preden, Axel Jantsch "Fog Computing Fundamentals in the Internet-of-Things," in *Fog Computing in the Internet of Things – Intelligence at the Edge*, Gewerbestrasse, Switzerland: Springer, 2018.

[2] Zaigham Mahmood, "Fog Computing in the IoT Environment: Principles, Features and Models," in *Fog Computing Concepts, Frameworks and Technologies*, Gewerbestrasse, Switzerland: Springer, 2018.

[3] *https://docs.nebbiolo.io/fogOS2.1.0/*

[4] *http://www.grafana.com*

[5] *http://www.influxdb.com*

[6] *http://www.qemu.org*

One solution of vehicle control software based on camera in ROS environment

Maksim Egelja, Nikola Teslić, *Member, IEEE, Research Institute RT-RK,*

Nemanja Lukić, *Member, IEEE, Faculty of Technical Sciences, University of Novi Sad,*

Zvonimir Kaprocki, *Research Institute RT-RK Osijek,*

***Abstract*—Software in automotive industry today is a very popular branch and brings many new challenges to the world of engineering. Recently, autonomous driving became one of the biggest challenges of this field. Goal is to develop a system capable of controlling vehicle almost completely independently.**

Work presented in this paper defines several modern autonomous driving algorithms and simulate them on modern platform. These algorithms are working with sensors attached to vehicle, such as camera, radar, lidar, etc. and include functionalities such as keep lane, controlling vehicle when "Stop" sign is detected and adapting vehicle speed when speed limit sign occurs in front of camera sensor.

***Index terms*—Autonomous driving; ROS; vehicle control based on camera**

I. Introduction

Recently, autonomous driving became one of the most popular software industry branches since engineers are working on providing highly automated driving systems, while over the past few years more focus is given on providing driving assistance. There are five levels of autonomous driving defined by experts and each of them describes the extent to which a car takes over tasks and responsibilities from its driver, and how a car and driver interact. Short explanation of each autonomy level is presented below.

Level 1 - Driver Assistance: systems on this level only support the driver, they do not take any control of the vehicle.

Level 2 - Partly Automated Driving: vehicle can control itself but driver remains responsible for control.

Level 3 - Highly Automated Driving: autonomous systems are taking control for the extended periods of time under certain conditions.

Level 4 - Fully Automated Driving: complete control of the vehicle by the system. Driver must remain able to drive.

Level 5 - Full Automation: completely autonomous driving, people are passengers only.

Vehicle needs many sensors due to its capability of sensing nearby environment. In figure 1 there is an illustration of most used sensors and their field of detection in autonomous car.

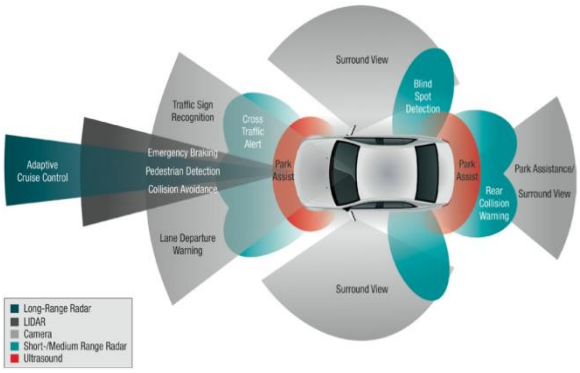


Fig. 1. Autonomous car with sensors

Today, there are commercial vehicles with integrated systems of autonomy levels 1 and 2. Further levels are bringing more challenges to the world of engineering. Due to the fact that there are many sensors attached to the vehicle such as lidar, full range and short range radars, multiple camera, ultra sonic sensors, there is a lot of data to collect and process in real time. Having this in mind engineers need huge computing power, special hardware units with very powerful processors, which are able to execute very complex algorithms. These include functionalities such as: trajectory planning, path finding, object fusion, etc, and they are mostly based on machine learning and neural networks.

Most components such as image processing, which are used for autonomous driving, are already solved as individual problems. Those components are highly important for this engineering field, but fusion of them still needs to be defined, in order to construct fully automated vehicle. In order to achieve this goal, engineers need to combine a great deal of research from different areas to produce highly automated vehicle pilot. Paper presents definition, implementation and

This work was partially supported by the Ministry of Science and Technological Development of Serbia under the project No. III 44009-2, year 2019

Maksim Egelja and Nikola Teslic are with the Research Institute RT-RK, 23a Narodnog fronta, 21000 Novi Sad, Serbia (e-mail: maksim.egelja@rt-rk.com, nikola.teslic@rt-rk.com).

Nemanja Lukić is with the Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obardovića 6, 21000 Novi Sad, Serbia (e-mail: nemanja.lukic@rt-rk.com).

Zvonimir Kaprocki is with the Research Institute RT-RK Osijek, 10b Cara Hadrijana, 31000 Osijek, Croatia (e-mail: zvonimir.kaprocki@rt-rk.com).

integration of such autonomous driving algorithms in simulated environment.

The rest of material is organized as follows. At the beginning, section II describes software libraries and platform used in this research, while section III describes system architecture, where its sub sections present more detailed description of each system block. Section IV gives the experimental results of this work, whereas section V presents a conclusion of the demonstrated work and proposes some of its possible future improvements.

II. Platform and Software Libraries

Goal of this work was to simulate real part of an autonomous driving system. ROS [1] (Robot Operating System) was used to simulate independent hardware units such as camera sensor or system control unit. Today, ROS is a very popular software system used for robot control, where the self driving vehicle can be defined as robot. Hardware units are simulated using ROS nodes where nodes are processes that are completely independent and can communicate with each other through specified topics.

Other part of this investigation was the OpenCV [2] (Open Computer Vision) library. All of the image processing was done using this open source library while complete work was realized using modern C++ programming language. Finally, units of this work are verified and tested using *gtest* [3] framework.

III. System Architecture

As explained before, ROS nodes were used to simulate hardware units in autonomous driving system. Schematic view of the system is shown in figure below.

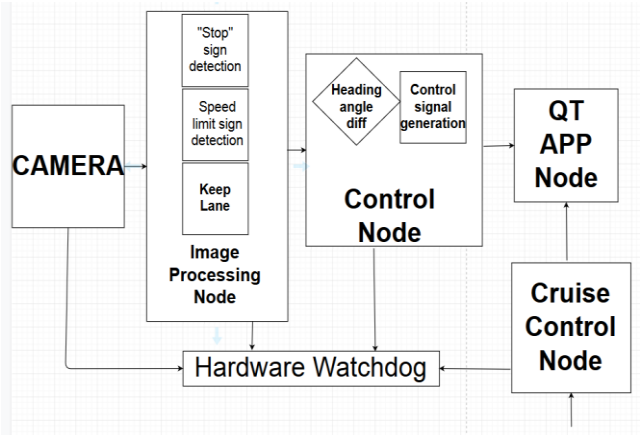


Fig. 2. Illustration of simulated hardware and signals between them

Camera node simulates camera sensor data and it sends raw pictures to image processing node. Due to the fact that camera needs to be calibrated, the reason why and the exact process of camera calibration is described in section III.A. Image processing node is in charge of all image processing algorithms in the system. It uses HAAR [4] feature based cascade classifier, which will be described in one of the following sections. It also communicates with camera sensor and control unit. Control node represents control unit of an autonomous driving system in a car. It can communicate with image processing and QT [5] application nodes. At the end, QT application node represents car's brakes, steering wheel and acceleration control units. It is possible to see how these systems work though GUI (Graphic User Interface). Besides recently mentioned nodes, there is a cruise control node which is in charge of collecting wanted speed data from the driver. There is a node that simulates hardware watch dog as well. This unit is capable of shutting down the whole system if something goes wrong. In further text, more detailed explanation of each node is given.

A. *Camera calibration process*

Camera needs to be calibrated due to measuring the size of an object in world units such as centimeters, or determine the location of camera in the scene. Camera calibration needs to be done since, according to this research, distance to "Stop" sign needs to be determined in order to properly stop the vehicle.

Today, in automotive industry, the most commonly used camera sensors are pinhole cameras. Pinhole camera model assumes that there is no lens in front of camera sensor and no distortion can occur in ideal case. In real world, there is a lens which introduces two types of distortion: radial and tangential. Radial occurs because of different lens thickness around the edges while tangential occurs because lens and camera sensor itself are not perfectly parallel. In order to remove all the image imperfections caused by lens and camera sensor, calibration is performed, which covers intrinsic and extrinsic camera parameters. Parameters are determined using matrix arithmetic and DTL (Direct Linear Transformation).

OpenCV currently supports three types of patterns used for camera calibration

- Classical black-white chessboard
- Symmetrical circle pattern
- Asymmetrical circle pattern

In this investigation, classical black-white chessboard pattern is used to calibrate the camera. For most of the patterns two snapshots are sufficient, while circular requires more snapshots in order to successfully calibrate the camera.

B. *Haar feature-based cascade classifier*

HAAR cascade classifier is machine learning based approach, where a cascade function is trained using many positive and negative samples. It is then used to detect particular objects in various images. This technique needs to extract features from the input images which are defined as value obtained by subtracting sum of pixels under black rectangle.

In order to calculate all needed features by using possible sizes and positions of kernels, big amount of calculation needs to be done in real time. In order to increase performance, HAAR introduces an integral picture, which uses an operation involving only surrounding four pixels for each pixel processed. Few types of features are shown in picture below.

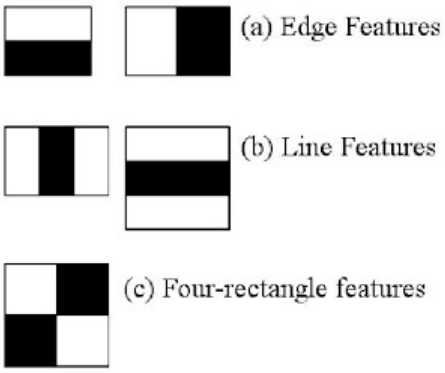


Fig. 3. Illustration of HAAR features used in cascade classifiers

C. *Image processing node*

This node firstly crops the image received from camera, and then extracts ROI (Region Of Interest). After cropping, frame is transferred to three HAAR feature-based cascade classifiers, trained for detecting "Stop" and speed limit signs and traffic light. Camera calibration process is necessary due to the fact that "Stop" sign's distance from the car needs to be calculated. Traffic lights are checked only if they signal red or green light. When speed limit sign is detected and cropped from a frame, this fraction of image is transferred to an OCR [6] (Optical Character Reading) module which is specialized to recognize digits only.

Lane detection is done by specifying ROI to road part only. After that, filtering by color is implemented, which finds white and yellow parts of the frame. In the next step picture is blurred and canny [7] edge detection process is performed. This feature assumes that camera is in the center of a car and predicts the angle needed to turn the steering wheel to keep the car in the current lane. After all detections and calculations, data is transferred to control node.

D. *Control node*

Control node is in charge of the whole system, by generating vehicle control signals and sending those inside different ROS topics. This unit exchanges information with image processing node and cruise control node. It collects the necessary data needed for vehicle control signals generation. Image processing node sends the information about detection of "Stop" sign with calculated distance, number read from speed limit sign and predicted angle of steering wheel turn to control unit. When information arrives, this part of the system is responsible for checking if distance to "Stop" sign is sufficient in order to stop the vehicle. If this is the case, node produces breaking signal.

When predicted angle is received, calculation of difference between current and predicted angle of steering wheel is performed. If this difference is smaller than predefined threshold, signal for steering wheel rotation is generated. This is done to prevent sudden rotations.

Also this node is receiving maximum allowed speed signal from image processing unit and driver desired speed from cruise control node. It combines those two signals and sets vehicle speed to maximum allowed if driver set speed is greater than limit. This can be overwritten by the driver if same, greater speed, is required twice. On top of all mentioned features, this part of the system has watchdog timer function as well, which is required to turn off autonomous driving system if something goes wrong with camera sensor or the image processing node.

E. *QT Application*

This part of the system simulates vehicle itself. It interacts with user and shows which control signals are generated and sent from control node. Starting complete system, entering wanted vehicle speed and monitoring of all control signals can be done through application. Application is able to show processed frame which is used for keep lane function. The screenshot of application behavior is presented in picture below.

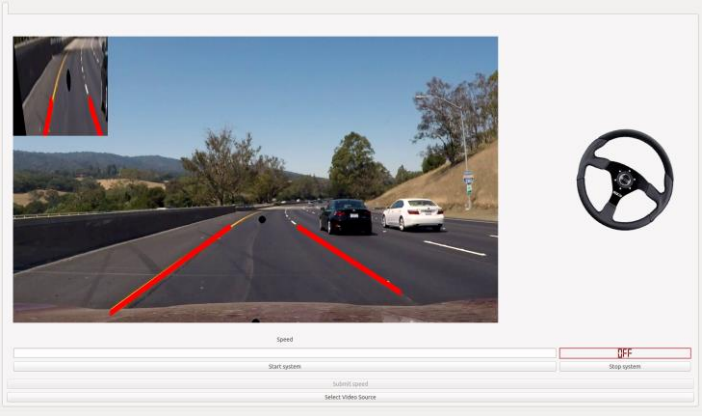


Fig. 4. Application screen shot with graphic user interface

F. *Watchdog node*

Watchdog is an electronic timer that is used to detect and recover devices from malfunctions. They are commonly used in embedded systems, such as this one. During normal operation of the system, watchdog timer is reset. In case that timer is not reset due to hardware fault or program error, this node's timer will elapse and generate timeout signal. When timeout signal is generated this node will try to recover the whole system to a working state if it is possible. Importance of this part of the system is high due to safety procedures in automotive industry.

G. *Cruise control node*

Modern cars cannot be imagined without cruise control function. This part of the system is in charge of collecting user defined speed. When user enters desired speed, this node collects this information and transfers it to control node, which will in return decide the actual driving speed. Also through this node, cruise control function of vehicle is being turned on or off, according to the driver's desires.

IV. RESULTS

Testing of system is done in multiple ways. Speed limit and "Stop" signs detection processes are tested automatically

using test scripts. System has its own log file system which is used to dump detected data inside. Script is run automatically when prerecorded video ends, to check if all detections are done correctly.

Keep lane function is tested visually by monitoring prerecorded camera video and calculated steering wheel angle. Unit and integration testing is done in this research too. It is done by using *google test* and *google mock* library. Mocking the objects is designed with use of DI (Dependency Injection) design pattern. In figure, below output of testing script used to test system is presented.

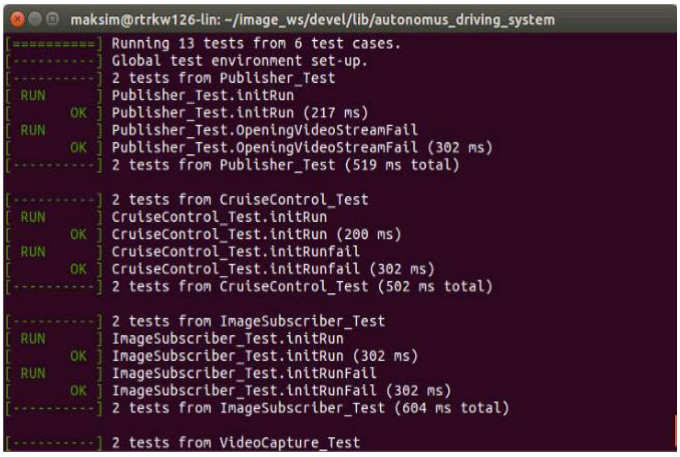


Fig. 5. Output of testing script used to check system functionality

Code coverage is measured using *gcov* [8] tool. In this work, due to very strict safety procedures which are very important in automotive industry, memory management is tested using *Valgrind* [9] tool. Static code analysis is performed due to high safety standards. Research done in this work is checked with MISRA 2004 guidelines using PC-lint [10] tool.

V. Conclusion

Contribution of this work is successful implementation of previously defined algorithms used in self driving cars. This research shows how combination of multiple independent building blocks are used to make a self driving system. Demonstration includes definition and development of algorithms for vehicle control using camera sensor. Algorithms shown in this investigation are:

- Keep lane function where software components are keeping a vehicle in current lane.
- Detection of "Stop" sign and its distance calculation due to stopping car in appropriate distance to this sign.
- Speed limit sign recognition and reading with intelligent vehicle speed control using neural networks.
- Traffic light detection with appropriate vehicle control according to traffic lights

The fact that this system uses only one camera sensor, proves that there are many possibilities for future improvements. More sensors can be added to the system, such as lidar and radar, then a fusion of sensors outputs could be made to get a better view of vehicle environment. By selection of ROS as the base platform of this work, program code is platform independent. Next step could be transferring the whole system to other platforms such as *AUTOSAR* framework.

References

[1] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, A. Y. Ng "Ros: an open-source robot operating system", ICRA Workshop on Open Source Software, vol. 3, no. 3.2, pp. 5, June 2009

[2] V. Pisarevsky "Opencv object detection: theory and practice" Slide Presentation, Intel Corporation, Software and Solutions Group, June 2007

[3] Jeganathan Swaminathan "Test-Driven Development" in "Mastering C++ Programming", Birmingham, United Kingdom, Packt Publishing, September 2017

[4] Paul Viola, Michael Jones "Rapid Object Detection using Boosted Cascade of Simple Features", IEEE Conf. on Computer Vision and Pattern Recognition, December 2001.

[5] Jasmin Blanchette, Mark Summerfield "C++ GUI Qt4 Programming" (the second edition), Electronic Industry Press, Beijing, 2008.

[6] R. Smith "An overview of the Tesseract OCR engine", Ninth International Conference on Document Analysis and Recognition, vol. 2, pp. 629-633, September 2007

[7] Zhao Xu, Xu Baojie, Wu Guoxin "Canny edge detection based on OpenCV", 13th IEEE International Conference on Electronic Measurement & Instruments, October 2017

[8] Holger Blasum, Frank Gorgen, Jurgen Urban "Gcov on an embedded system", GCC for Research in Embedded and Parallel Systems, September 2007

[9] N. Nethercote, J. Seward "Valgrind: A program supervision framework", Electronic Notes in Theoretical Computer Science, vol. 89, no. 2, pp. 44-46, vol. 89, October 2003

[10] James Gimpel "Software That Checks Software: The Impact of PC-lint", IEEE Software, pp. 15-19, vol. 31, February 2014

Marko Nerandžić, Petar Milanović, Gardelito Hew A Kee, Ilija Popadić, *Member, IEEE*, Miroslav Perić, *Member, IEEE*

Administration tool for multi-sensor imaging system

***Abstract*— Rapid development in imaging technologies as well as computer science bring both software and hardware improvements to systems used in military and surveillance fields, but there is a constant trade-off between adding new software features and lowering the system's security. Every new feature implemented in such system can potentially be a weak point for individuals with malicious intent. Particularly in multi-sensor imaging systems, initialization process of system itself is the weakest point, when no IP address is acquired, but system is powered on. Also, worth mentioning is the necessary use of static IP addresses, to avoid any DHCP Offer or other magic packets. However, there is a need for remote administration, especially since these types of systems are often found on distant locations. This paper presents a solution applied in such system and gives comments about which programming language should be used, recommended tools for enabling remote administration itself, as well as security shortcomings of such approach.**

***Index Terms* — System administration, multi-sensor imaging, Java, ONVIF protocol, remote system configuration.**

I. Introduction

Modern surveillance systems consist of multiple imaging sensors, cameras capable of different wavelength capturing, which work in tandem to provide best image available for further processing[1]. Picture acquired in this way is suitable for image algorithms, such as object tracking, motion detection, etc. Illustration of such system, and of the potential use-case scenario for service application presented in this paper is shown below (Fig. 1). Camera systems, based on Nvidia Jetson TX2 embedded computer board, communicate internally through a private local area network, with access to a wide area network ensured through a gateway. Operator's console on which the administration tool presented in this paper is running, is connected through Open Network Video Interface Forum - ONVIF protocol[2], which ensures encrypted data transfer as well as secure instruction set exchange.

Marko Nerandžić is with the Vlatacom Institute of High Technologies, Blvd. Milutina Milankovića 5, 11070 Belgrade, Serbia (e-mail: marko.nerandzic@vlatacom.com).

Petar Milanović is with the Vlatacom Institute of High Technologies, Blvd. Milutina Milankovića 5, 11070 Belgrade, Serbia (e-mail: petar.milanovic@vlatacom.com).

Gardelito Hew A Kee is with the European University of Belgrade, Carigradska 28, 11000 Belgrade, Serbia.

Dr Ilija Popadić is with the Vlatacom Institute of High Technologies, Blvd. Milutina Milankovića 5, 11070 Belgrade, Serbia (e-mail: ilija.popadic@vlatacom.com).

Dr Miroslav Perić is with the Vlatacom Institute of High Technologies, Blvd. Milutina Milankovića 5, 11070 Belgrade, Serbia (e-mail: miroslav.peric@vlatacom.com).

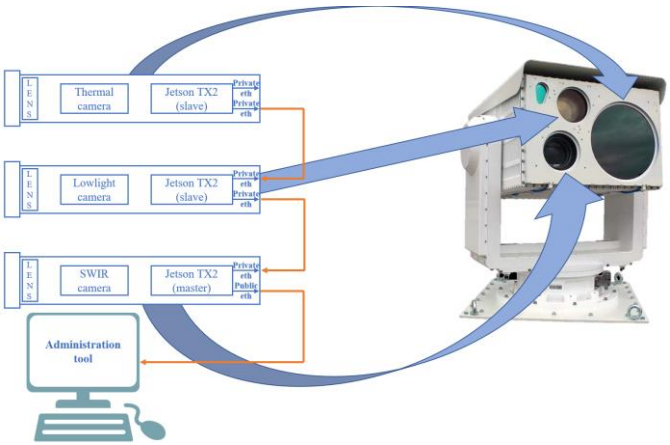


Fig. 1 Illustration of multi-sensor imaging system

With the swift development of both software and hardware, and quick adoption rate of new technologies which widen the abilities of traditional systems, there is an increasing demand for applications which serve to observe remote systems as well as log their activity. This phenomenon is especially pronounced in the security and military fields, where system failure can cause major problems. That is why it is important to have reliable software which can monitor system health, change basic parameters crucial for system functioning and observe any changes to the environment. This type of application faces a difficult trade-off between functionality and too much access to the system and its' resources. On one hand, the greater the number of tools available to user, the more possibilities there are for remote operation, which lessen the chances for obligatory site visits, but on the other hand, it leaves more room for potential harm-doing software and persons with malicious intent[3]. Having that in mind, it is important to define all needed options this type of application should have, and build it with only those operations available, which is why other existing solutions may not be acceptable.

Another thing worth considering is computer knowledge of applications' users, operating systems which will be used to run it and situations where it will be used. For example, it is not recommended to build an internet-dependant application when users have limited or no access to it.

One more thing to keep in mind is user's knowledge of English language, which is sometimes taken for granted, but it is not always the case that the user is fluid with all the technical terms and naming standards. Which is why intuitive software and placement of elements within the Graphical User Interface – GUI is very important[4]. Also, the language used in these types of applications should be as simple as possible, and use of whole sentences should be avoided in all cases except informational screens and logs.

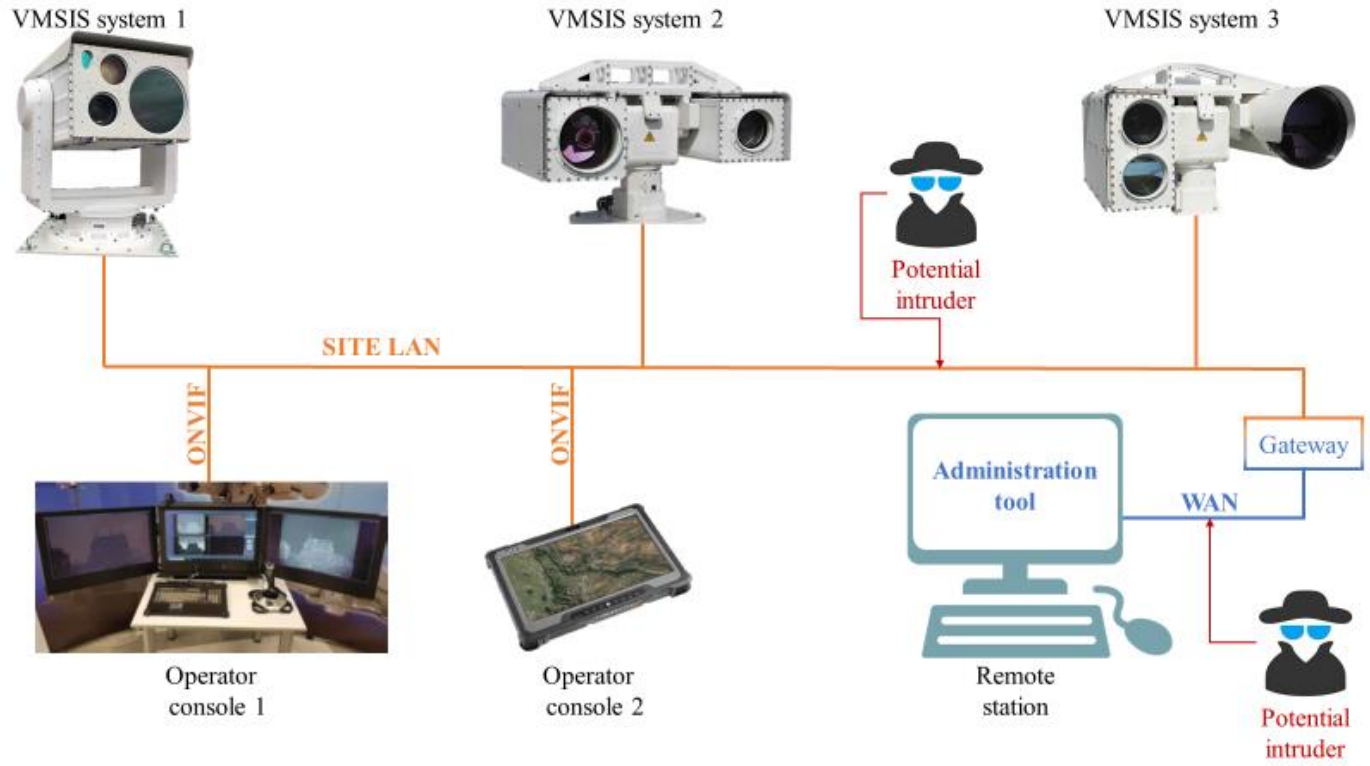


Fig. 2 Network structure on surveillance site

These types of systems are most vulnerable upon power on. At that point, if the system has no IP address assigned to it, potential intruder can manipulate it by giving the system an IP address of his choice, leaving it completely exposed. VMSIS is a server-type device with a static IP address and problems can arise when irregular address is used or assigned one is forgotten. Unlike SOHO networks, simple factory reset can't be used, because of system complexity. Once the power on sequence is finished, security measurements are passed on to the communication protocol of higher level, in this case ONVIF protocol. Network structure on surveillance site, with potential weak spots, is shown above (Fig. 2).

Advantages of this protocol compared to other ones, for example SNMP, is that it is built from ground-up with surveillance and video systems in mind. Since ONVIF is an open industry standard, its adoption rate is measured in hundreds of IP security manufacturers, leading to great interoperability among IP security devices. Also, having a universal protocol, which all manufacturers can use, benefits the developers as well, since software written for one compliant device will work on others without any alterations.

The paper is organized as follows. The Section II describes the methods used in order to achieve remote administration, as well as tools implemented in order to provide the features of administration tools described in this paper. In the Section III system components are discussed, and system prerequisites for running the administration tool presented in this paper are listed. The section IV shows field testing results of the administration tool presented in this paper. The Section V presents conclusions and future work in this research area.

II. System Description

Most multi-sensor camera systems consist of modules which communicate with each other privately, and have at least one public interface that is used for external communications and surveillance. That public interface is used for all purposes of service application presented in this paper. This particular service application is built in Java programming language, since there are end-users both on Windows as well as Linux operating systems[5], so compatibility with both platforms was a strong requirement. Also, Java offers visual elements which can be easily combined to form an intuitive User Interface – UI. Another positive of Java is its' ability to use remote resources and communicate seamlessly with devices which provide information for system functioning. For the most part, this application uses Java Secure Channel – Jsch[6], which is a pure Java implementation of SSH2 (Secure Shell 2)[7], to provide the needed tools to the end-user. This type of approach supports secure remote login, secure file transfer, secure TCP/IP as well as X11 forwarding. Also, there is support for key exchange like diffie-hellman-group-exchange-sha1, diffie-hellman-group-exchange-sha256, etc. Further, one of the features is connection through HTTP Proxy, as well as SOCKS5 proxy. This implementation gives the ability for port forwarding, stream forwarding and SSH File Transport Protocol, among other protocols. Everything mentioned allows the integration of all needed services which are essential to monitor these types of systems. Those include,

changing the network parameters of the remote host, such as device IP address, subnet mask, and default gateway, modifying both system and hardware time, downloading logs collected on the machine, both ones generated by operating system itself, syslogs, as well as those generated by proprietary applications which run on the system.

Another functionality implemented in the service application presented in this paper is remote host finder. This is especially important for systems designed to work only with static IP addresses, since some host may lose its' set IP address due to some system error, or even get a dynamically assigned IP address, by a DHCP server connected to the system, accidentally by an operator or on purpose by a potential intruder. So, it is very important to have a way to find IP address of the system which is to be configured or checked for errors. Since all operations on the remote host rely on knowing the host's IP address, host finder functionality is implemented in such a way that it finds and lists all the connected systems to the same network on which the operator's machine, running this service application, is connected.

For all purposes of finding the lost or forgotten IP address, as well as other IP manipulations, java.inet Java packet is used. It contains various tools for manipulation with Ethernet packets. One of those is a sub-packet called InetAddress, and it contains necessary classes and functions for checking if there is a valid route to the given host or not. To speed up the search for connected hosts, multi-threading is implemented, about which will be more discussion in the System Components section.

Use of ONVIF protocol enables the administration tool presented in this paper to communicate with the multi-sensor imaging system, without compromising the system's security.

III. System Components

A. Java Runtime environment

The basic component for running the service application on the host machine is a Java Runtime environment installed and set up. This environment is a free download, available on Java's owner website, Oracle. It is available for all popular systems, Microsoft's Windows, Apple's macOS, as well as most Linux distributions, such as Canonicle's Ubuntu, CentOS, Debian and others. Installation process is simplified and is a linear procedure, requiring minimal effort from the user, following only the steps listed by the installer. The multi-camera sensor system doesn't have to have Java Runtime installed, which is a plus, as there is no need to modify an already running system. That said, service application presented in this paper can operate with any other type of system, just some of the functionalities won't be available. For example, host finder will work in any given network and system configuration. Functionalities such as changing the target IP address, the system clock, log downloading as well as ping, are built in such a way, so they will retain their usability as long as the target is running a Linux-based operating system.

B. Secure log-in

Since this type of system and service application is intended to be used in surveillance and military systems, security is an important part of its' functionality. Therefore, before user can access any of the application's tools, secure login is required. Parameters of this login are changeable only before the first use of the application itself, leaving less room for any potential attacks and password cracking. Credentials for application are explicitly given to the user, and are not distributed.

After successful environment setup and application starting, user will be greeted with the log-in screen as shown below (Fig. 3).

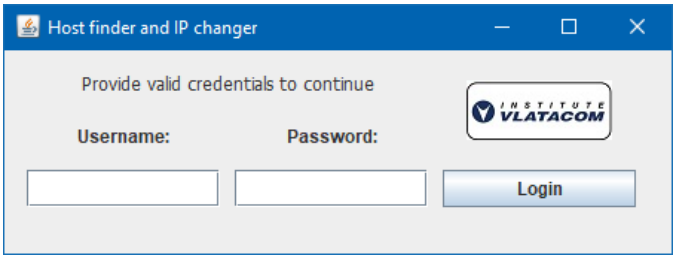


Fig. 3 Log-in screen

In case of valid credentials, user will be given access to the service application itself. On the other hand, if not, user will be informed about invalid combination of username and password, prompting for another try.

C. Service application main screen

Once the login is successful, user will be shown the main screen of the service application itself. It is worth mentioning that all functionalities provided by the service application are documented and described in detail in the manual which is supplied to the users alongside the service application. However, the form and presentation of the tools is composed in such a way that it tries to minimize the need for the manual.

Once opened, service application starts with the host finder tool on its' main screen. In the left part of the application, options for finding hosts are listed, while the right side is populated with some predefined, popular, subnets, as well as a button which leads to the part of the application that provides the user with options for changes on the system, but more on that later.

As shown below (Fig. 4), user has the ability to search only the predefined subnet, or just a given range of addresses, as well as conduct a full search of the whole network. In order to speed up this procedure, every query for an active host is started as a separate thread[8], up to a maximum of 65025 (255*255) threads. This is done to shorten the procedure, since every host has a time out of 200ms to respond, after which the request is discarded meaning that no host is connected on the given IP address. If a host replies in the meantime, it is registered as active and shown in the text area on the bottom of the service application.

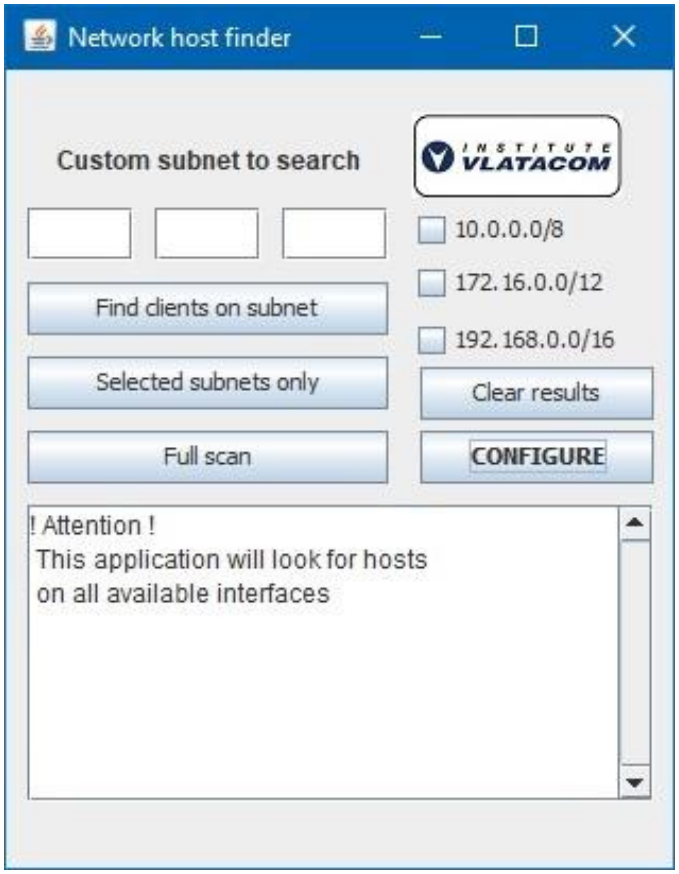


Fig. 4 Network host finder

Multi-threading may not be necessary when small networks are in use, but combination of using multiple threads, as well as setting the timeout limit to an empirically tested number, speeds up the procedure dramatically. The timeout itself can be set to a smaller or bigger number depending on what can be expected of the network and connected hosts.

The polling for active hosts is done by sending ICMP (ping) packets. While most tools rely on using ARP packets, solution presented in this paper doesn't implement that method, as some network firewalls may interpret sending that many ARP packets from one host machine as an attempt to destabilize the network, or as a flooding attack, and disconnect the host. On the other hand, ICMP packets are small, predefined, and are allowed in almost all networks, which makes them ideal for this purpose. Upon reply from the host connected to the network, its' IP address will be shown in the main window of the application. Order of IP addresses will be such that it matches the order of received replies, meaning that the hosts with smaller latency through the network will be on the top, and more distant ones, with bigger propagation latency, near the bottom.

D. Remote system connection

Once operator has the IP address of the system which needs to be configured, connection to that system has to be established in order to proceed to the device administration. Connection to the system is done through the connection panel, which is shown below (Fig. 5).

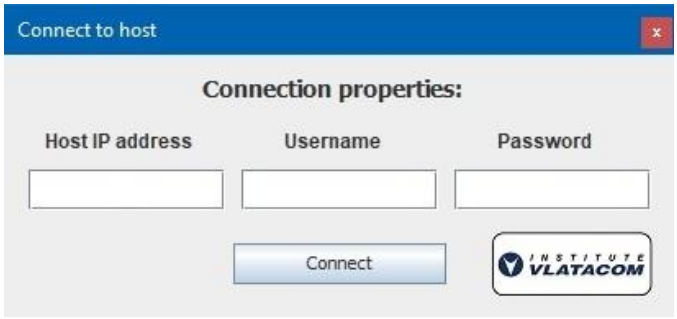


Fig. 5 Connection panel

System parameters used to connect to the device are the same as the operator would use to connect using a standard SSH connection[9], or in other words, IP address of the system which is to be administrated, username of an account on the machine, as well as password for that account.

E. Device administration

Upon successful connection to the system, device administration panel will open. This part of the application contains main tools for changing the system parameters as well as checking the system's health. It is not recommended to leave system parameters susceptible to change all the time the system is powered on, so a security feature is implemented which disables the ability to change the IP address as well as other network parameters after 5 minutes upon powering on. However, there is a way for this to be overridden, by using the button in the top left corner, which allows the user to change all system parameters, but only after providing a predefined administrator password. By doing so, one more security measurement is added through adding another step of authentication. The image below (Fig. 6) shows the state of the service application once the system has been powered on for more than 5 minutes and administration privileges are not granted.

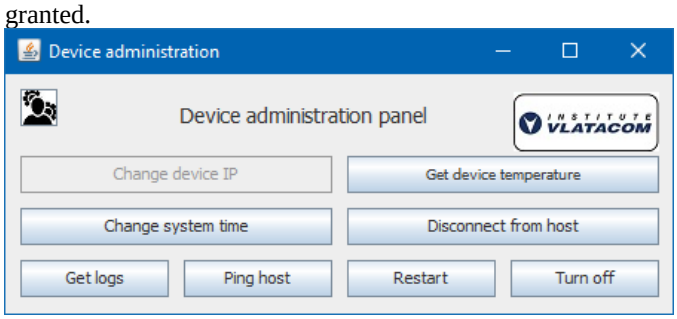


Fig. 6 Device administration panel

As already mentioned, button Change device IP will allow the user to make needed network adjustments, such as changing the IP address of the machine, setting the new gateway address, as well as changing the subnet. After successful parameters change, user will be disconnected, so the new parameters can be set on the remote host.

Another feature enabled by service application presented in this paper is the ability to change system software and hardware time. Before changing, the user will be informed about current time. The mentioned feature, along with its' GUI element is shown below (Fig. 7).

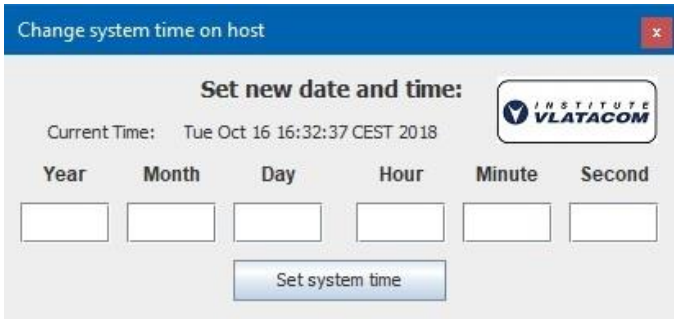


Fig. 7 System time modification panel

For any system intended to be operative for longer periods of time, it is essential to keep log of its' activity, which can be extremely helpful to diagnose a cause of an eventual system failure, or even if everything is running correctly, observe the system state. With logging implemented, tool for downloading those logs from the remote host is found in the lower left corner of the administration panel and once used, the user will be presented with a progress bar of the downloads. Among the Ubuntu's syslog files, log files generated by a proprietary logging tool are also downloaded.

Checking the connection between the host running the service application and the remote host can be done by pinging the remote host using the implemented tool. This feature presents the user with latency present in the network, as well as time needed for the remote host to be reached.

System health can be checked in various ways, depending on the implementation of the system, there are various parameters which can best represent its' state. In this implementation, system temperature is of special interest, that is why, solution presented in this paper has the ability to display to the user, temperature measured by the system.

Among other features, the operator has the ability to remotely reset the system, or power it off. It can be useful in some conditions where system isn't working properly to have a way to restart it remotely, without having to go on site.

F. Additional improvements

Adding the ability of downloading the recordings captured by the multi-sensor imaging system would be the next logical step in improving the administration tool. Also, addition of an informational file which would contain detailed information about the system, and the ability to download it remotely would see benefits in diagnostic speed, as the operator could easily identify components and prepare for needed repairs without having to go on site and look for what parts are installed.

IV. FIELD TESTING

Administration tool presented in this paper has already been field tested in the capital city of United Arab Emirates, Abu Dhabi. Fig. 8 shows the naval surveillance site in the bottom left corner of the figure, network connection used, operator console room with administration tool running on operator machine, on the bottom right in the image, and an image from the multi-sensor imaging system, VMSIS, in the upper portion of the picture. Note that even in the simplified setup, device is 10m above ground and factory reset button isn't accessible.

Fig. 8 Administration tool on surveillance site field testing

V. CONCLUSION

Modern surveillance systems are improving by day, and those improvements bring lots of additional features, but increase the system's overall complexity. One of those improvements comes from the ability to have multiple cameras, or multi-sensor camera system which can extrapolate signals captured by all of the cameras and then combine them to create a picture of higher quality. Image generated in that way can be better used for further processing, whether it be to detect an object in the image, track an object, or something entirely different. But all those improvements come at a cost of increasing system intricacy, since every imaging channel requires a subsystem which has a built-in processor, and an image processing hardware. All of those components have a chance of failing, that is why the need for administration tools is increasing.

REFERENCES

[1] Hamid Aghajan, Andrea Cavallaro, *Multi-Camera Networks: Principles and Applications*, Massachusetts, Cambridge, United States of America: Academic Press, 2009.

[2] Johan Adolfsson, et al. December 2016. "ONVIF Application Programmer's Guide." *ONVIF*. https://www.onvif.org/wp-content/uploads/2016/12/ONVIF_WG-APG-Application_Programmers_Guide-1.pdf

[3] Daniel J. Barrett, Richard E. Silverman, Robert G. Byrnes, *Linux Security Cookbook*, California, Sebastopol, United States of America: O'Reilly Media, Inc., 2003.

[4] Jonathan Anderson, John McRee, Robb Wilson, The EffectiveUI Team, *Effective UI: The Art of Building Great User Experience in Software*, California, Sebastopol, United States of America: O'Reilly Media, Inc., 2010.

[5] C.L. Sabharwal. 1998. "Java, Java, Java." *IEEE.*

[6] Dr. Atsuhiko Yamanaka. 2018. "Java Secure Channel Project Documentation." *JSCH Project.* November 22. http://www.jcraft.com/jsch/examples/README

[7] *The Secure Shell (SSH) Protocol Architecture*, RFC 4251, January 206, IETF.

[8] Vamsi Krishna Myalapalli, Sunitha Geloth. 2000. " High performance JAVA programming." *IEEE.*

[9] Daniel Barrett, Richard Silverman, Robert Byrnes, 2nd edition, *SSH, The Secure Shell: The Definitive Guide*, California, Sebastopol, United States of America: O'Reilly Media, Inc., 2009.

One solution of DTV simulator for PC platform

Đorđe Glišić, Aleksandar Šuka, Aleksandar Plahćinski, Miodrag Đukić

***Abstract* — User requirements of digital TV receivers grows increasingly, so does the complexity of embedded development process. Current STB devices range from cheap low memory device to multi-core device with performance comparable to PC or mobile phones. Low cost devices present a specific challenge in development, as the stability and performance is more in focus than on high performance devices. Limitations in the development of software for digital television showed the need for the development of tools for analysis and testing. This paper presents one solution to overcome limitations of target platform. DTV software stack is taken and adjusted to be used on PC to get sophisticated development tools accessible.**

***Index terms* — simulator; digital TV receiver; software in digital television; Middleware; Middleware Testing Environment; hardware abstraction layer; Inter-process communication; SDL library; ffmpeg library; wxWidgets library.**

I. Introduction

Embedded devices in recent years show sudden increase in terms of performance, so *DTV* receivers no longer serve only for reproduction of television signals, yet they come with a set of additional functionalities, the implementation of which in time becomes more and more complex. Usually the development of software for embedded devices is slower than the PC platform. The main challenges are the lack of tools to detect programming errors, general tools for software development, as well as lack of hardware resources. Idea was to develop tool that could give us ability to use debugging tools at runtime, code coverage analysis and to automate testing of DTV stack as black box and in development.

Similar work was done in commercial DTV software, were aim was mostly toward supporting development of application part, but not around simulating whole DTV stack. One example of that is Observatory [1], that supports minimal functionality of DTV and focuses around UI development. Second more close solution working only on Linux platform is DVBCore [2], it supports DVB-T, DVB-C, DVB-S, and it has similar requirements for building on Linux. It uses older versions of underlaying libaries, limiting capabilities in UI presentation and reproduction, and does not provide version for Windows operating system.

The second chapter gives a general structure of software for *DTV* receiver, the aim is to create a sense of layering and modularity, which facilitates the implementation of the simulator.

The third chapter describes modules that had to be implemented on DTV stack in hardware abstraction layer for target platform. For simulation of hardware devices like tuner, demultiplexer, decoder we used libraries from open source community. For decoding and demultiplexing of content from transport stream we used *ffmpeg* library[3]. For audio and video reproduction, graphical presentation, as well as to send and receive user-generated events with keyboard we used *Simple DirectMedia Layer – SDL* library [4].

Chapter four describes second part of this solution, graphical environment that is used to control simulator from outside and to generate commands and receive simulator states. Environment is possible to send a number of test commands, to trigger a series of user and system activities. Communication between PC simulator and test environment is done using TCP/IP protocol.

II. Software for DTV receiver

Architecture of DTV software can be divided into: operating system layer and DTV drivers layer, hardware abstraction layer (*HAL*), *Middleware*, *Middleware* abstraction layer API *(MAL API)* and DTV application. Applications can be different from application running using build-in UI engines to external clients like Web browser or application written in Java (in case of Android OS) as depicted in Fig. 1.

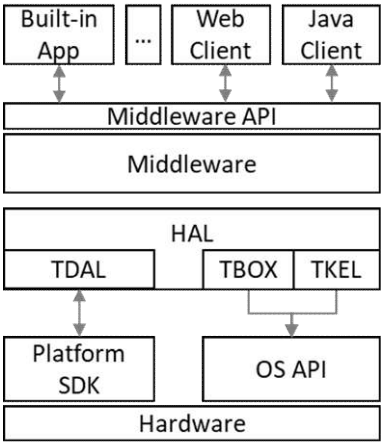


Fig. 1 DTV application architectural overview

A. Middleware

The *Middleware* layer is central part of the DTV stack, it is responsible for parsing DVB information, service management, access control, support for personal video recording (PVR), electronic programming guide (EPG), control reminders, controlling descrambling and decoding of

Aleksandar Šuka – RT-RK Institute for Computer Based Systems, Novi Sad, Srbija, (e-mail: Aleksandar.Suka@rt-tk.com)
Đorđe Glišić – RT-RK Institute for Computer Based Systems, Novi Sad, Srbija, (e-mail: Djordje.Glisic@rt-tk.com)
Aleksandar Plahćanski – RT-RK Institute for Computer Based Systems, Novi Sad, Srbija, (e-mail: Aleksandar.Plahcanski@rt-tk.com)
Miodrag Đukić – RT-RK Institute for Computer Based Systems, Novi Sad, Srbija, (e-mail: Miodrag.Djukic@rt-tk.com)

multimedia content and so on. Middleware controls data flow for multimedia content, from one or more modules in HAL layer. Middleware has defined API as seen in Pic. 1, that is used by user application or other clients, depending on the actual architecture and platform. The role of the user application layer is to control the display of graphic elements, and to define the behavior of the application that interact with the user. Key point for reusability of the middleware is that its code is platform independent, in a sense that it does not assume particular architecture or uses any platform specific features. All system or driver dependencies are encapsulated through HAL layer.

B. Hardware abstraction layer

Hardware Abstraction Layer (HAL), is layer on which *Middleware* relies. This layer has strict predefined API serves to hides physical architecture of the *DTV* receiver, and needs to be implemented for every new targeted platform.

HAL consists of three independent units: *Thin Kernel Encapsulation Layer (TKEL)*, a layer for functional abstraction of real time operating system, *Tool Box (TBOX)* module provides I/O capability over serial connection (UART, TCP/IP etc.), and *Thin Driver Adaptation Layer (TDAL)* that serves as an abstraction layer for drivers which consists of many modules, each block of physical architecture of the *DTV* receiver has corresponding *TDAL* controller.

III. Adjustment of DTV Software to Targeted PC Platform

As part of the *DTV* receiver software adjustment to target platform it was necessary to implement hardware abstraction layer. This chapter describes the details of implementation used to port DTV stack to the *Windows* and *Linux* platform. It was selected to use built-in application as DTV application with stack as in Fig. 2.

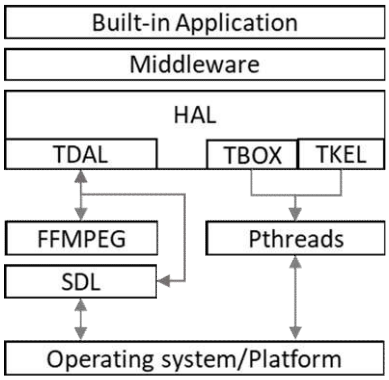


Fig. 2 Architectural overview of PC simulator for DTV

A. Adjustment of TKEL module for targeted platform

TKEL module implements functions that are specific to the used operating system, provides *API* for: semaphores, critical sections, creation of software threads, communication between program threads, memory allocation and deallocation, timers.

In this paper we used programing interface from *Pthreads* library to implement *TKEL* module parts related to the creation, synchronization and communication between threads - tasks. *Pthreads* is *POSIX* standard directly supported on *Linux* platforms, unlike the *Windows* operating system, which in this case it was necessary to include *Minimalist GNU for Windows* (*MinGW*) solution. As a result, we got code which does not depend directly on the programming interface provided by the operating system, and accordingly it is not necessary to maintain separate *Windows* and *Linux* version as shown in Fig. 2.

Exchange of data between tasks in the operating system adaptation module is provided by using message queues. As queues are simple structures, there is no difference between this implementation and the implementation of the DTV receiver. Attention is paid only to synchronization, depending on the system calls typical for the targeted platform, for which was used the same module as described in the section for synchronization between tasks.

On DTV receiver timing is achieved using special thread that is blocked, and signaling to that thread takes place every hardware cycle and every cycle of that thread is one unit of time. On the targeted platform, the part of hardware that performs signaling is switched by system signaling.

B. Adjustment of TBOX module for targeted platform

The main role of *TBOX* is to enable communication of the *Middleware* with the system environment for the purpose of performing debugging output via the serial interface. *TBOX* allows multi-level printing, such as: critical prints, warning, monitoring hierarchy calls, general printouts.

Adjusting *TBOX* module comes down to the implementation of *TBOX* as a tool to print to standard output or to output for errors.

C. Adjustment of TDAL module for targeted platform

TDAL is driver abstraction layer, which allows the use of physical architecture of *DTV* receiver by the *Middleware*, without needing to know the details of the implementation of that architecture. *TDAL API* is the most comprehensive *API* and most complex for porting to targeted platform. *TDAL* is modularly organized, and so that in most cases one *TDAL* module corresponds to one block of the physical architecture, one module can be viewed as the driver handler from the point of *Middleware*.

In the following section are given solutions of how some of the modules from *TDAL* are implemented.

1) Adjusting the block that controls demodulator

Block that controls demodulator and set the reception at a specific frequency *TDAL_DMD* [8] is responsible for managing the physical network module. Some of the roles are: setting up reception at a certain frequency, performs search on a given frequency band, notifies other modules of the important events related to the reception, provides access and control parameters of signal quality.

Within the simulator input data are represented as a transport stream in a digital format. The process of signal demodulation has been done already in the process of recording a transport stream. Complete transport stream that would be obtained as an output from the demodulator is in the file. In addition to the file that contains *MPEG-TS* data, there is a configuration file that contains all the information necessary to describe the signals that are obtained by the process of demodulation. *MPEG-TS* data that are transmitted at different frequencies are stored in separate files, when adding a new *MPEG-TS* file it is necessary to modify the configuration file. Each line in the configuration file specifies *MPEG-TS* in the following format: the frequency at which *MPEG-TS* is being broadcasted

expressed in Hz, path to *MPEG-TS* file, signal quality which is set in the range of 0 to 100, the signal strength, which is applied in the range from 0 to 100, speed at which data is transferred expressed in number of bits per second, comments in the configuration file begin with * and they do not interpret.

During initialization of the block that controls demodulator, the configuration file is being parsed. Parsed information are stored in the structure and later used to set the reception at a specific frequency – *tuning*. Setting the reception at a specific frequency for us means positioning to the next *MPEG-TS* file. Signals the library *ffmpeg* which first deinitializes old contexts for formatted input and output, and then initialize the new contexts. After the opening of the new context library closes previous, and then opens new file. When opening the file, functions to read the file and positioning within the file are provided. It is also necessary that the local structures for the management of the demodulator is filled with information from the current *MPEG-TS* file, after which the system is ready to fetch new data.

2) Adjusting the block that controls demultiplexer

The role of the *TDAL_DMX* [7] module is to manage the demultiplexer to separate the audio/video components, *Packet Elementary Stream (PES)* packages and filters *Service Information (SI)* sections [5].

TDAL_DMX module allows the creation and deletion of channels, which are used for the delivery of data by type of channel. The channels may be of different types: video stream, audio stream, PCR stream, teletext stream, stream for subtitle, stream for sections etc. Each channel is dependent on the correct one *packet identifier – PID*. Within the channel, *TDAL_DMX* allows usage of filters that additionally filters data from transport stream. Two modes of filtration are supported: positive filter (client requires filtering according to certain criteria, *TableId*, *ExtensionId*, *VersionNumber*) and negative filter (client requires fetching sections that do not contain specific criteria).

How hardware support for demultiplexing on *PC* platform does not exist, we have implemented demultiplexer in the TDAL_DMX module. Functions, for reading and positioning within *MPEG-TS* file that we registered to *ffmpeg* library, shall be fetching data. The data in this case are the packets. Special demultiplexing thread is reading data and applies filtering, first on the basis of *PID* value, and then on the basis of defined filter, if one exists. Given that *ffmpeg*, itself, may differ audio and video packets from other *PSI* packets, we used that ability so in demultiplexer thread goes only data related to *PSI* packets. *PSI* packets are after filtering grouped into sections, that are by *callback* function passed to the *Middleware*.

3) Adjusting the block that decodes audio and video packets from elementary stream

The role of the *TDAL_AV* [11] module is to fully control the process of decoding video and audio *PES* packets. Within *DTV* receiver the physical audio and video decoders are controlled directly. On PC platform decoding function performs library *ffmpeg*, to display content we used *SDL* library. Below it will be explained how the audio and video decoding process is implemented, forming video frames and displaying of synchronized audio and video content.

Block that reads *MPEG-TS* file is implemented using *ffmpeg* library. As we stated in part for demultiplexing, library can distinguish audio and video packets, so depending on the type of packet, package is sent to the block for video processing or into the block for audio processing.

Block for video processing takes sent video package, decodes it, forms a frame and synchronizes its display. Then *SDL* library is informed to allocate space for a picture and based on frame creates video texture for displaying. Showing is realized by the timer that signals *callback* function, that joins part with graphic and part with video texture and refreshes the display.

In the audio processing block is defined callback function, in which are audio packets collected, decoded and synchronized. This function is passed to library *SDL* when audio and video content are being played.

4) Adjusting the block that controls graphical rendering

TDAL_GFX [9] serves as a loop which is used by the application to display a graphical user interface. Some functionalities of module for graphical rendering are: graphical regions management, color palettes management, transparency level management, bitmaps manipulation and drawing, filling a rectangular region with color, blit command to combine multiple bitmap and multiple regions into a single final image.

The graphics module used in the simulator relies on library SDL. The library provides the concept of texture and the concept of region, as the *Middleware* expects graphic layer and windows within the graphic layer, we used texture as a graphic layer, and regions represent windows within a layer. Rendering graphics comes down to going through all regions and bliting with texture.

Initialization of graphic layer boils down to creation of graphic texture and initialization of ten free regions. When the *Middleware* is prompted to create a new region, *TDAL_GFX* finds the next available region, creates texture for a specific region and returns to *Middleware* handler for specific region. All drawings are related to a texture region. Using *SDL* library primitives can be rendered, for complex images are first created bitmaps, which are then glued to the texture Blit function we implemented so that, in addition to the gluing bitmap on the texture of region, it can perform sticking color to bitmap, as well as some memory content on bitmap.

The content of the graphic texture is copied to the display, together with the content of video texture described in the section referring to the video decoding. Refresh the display gets the final image.

5) Adjusting the permanent memory management block

TDAL_FLA [12] module is responsible for providing access to *flash* permanent memory. *Flash* memory is divided into blocks, predefined size.

In the simulator the file is opened during the initialization of the permanent memory management block. File is mapped with buffer, and further manipulation with permanent memory comes down to operations with buffer.

6) Adjusting the block for receiving user events

In the case of *DTV* receiver for receiving and processing user events is responsible *TDAL_KDB* [10] module. For each user action on the remote control, *TDAL_KDB* module generates an event that can be received and processed in order to carry out certain actions. *TDAL_KDB* module supports the following events on the keys of the remote control: key is pressed, key released and key is hold. Physical codes are translated into *Middleware* codes using configuration files.

Instead of an infrared remote control as an input device to generate a user event parameters are utilized with the keyboard. To receive events generated by the keyboard we used structure and program interface offered by *SDL* library. For event handling *SDL* library offers a collection of events. For the purpose of the simulator we used fields that relates to keyboard events. The structure for the events from the keyboard has a field type, which can have values pressed or released which allows us to distinguish events key is pressed or released key. Key identifier is read from the field symbol keys which are then mapped to the corresponding code that is sent to the *Middleware*.

The mechanisms for receiving user control event distinguish between two threads. The first thread receives user events over the functions of *SDL* library and sends it to another thread over a common buffer, while another thread processes event and calls the appropriate *callback* registered by the *Middleware*. Running PC simulator is depicted in Fig. 3.

Fig. 3 PC simulator at execution time (Windows version)

7) Adjusting the block for playing multimedia content

The main job of the *TDAL_MP* [13] is to control audio/video playback of multimedia content from local hard disk, extern flash memory or extern hard disk connected via USB plug. Module allows to play, pause, stop, rewind or fast forward reproduction.

Within *DTV* receiver, drivers capable of decoding specific multimedia coding are already present on board. In case of PC simulator, we used feature of *ffmpeg* which distinct audio and video packets, provides access to a single package, the fields and content, as well as content decoding. It has support for a large number of audio and video codecs, as well as container formats. In simulator extern storage memory is simulated with directory on PC platform local hard disk.

After opening multimedia file, there are three threads that control: reading data, decoding of audio and video packets and performing their reproduction. Functions for file reading and positioning within the file are provided to *ffmpeg,* after that using those functions *ffmpeg* reads packets from elementary stream and decodes them. For displaying video content and playing audio content, is again used *SDL* library.

It was necessary to implement function used to manipulate playing of content. Rewind was enable by function which changes position in file. Fast forward function skips parts of file periodically so that frames are displayed with longer time intervals in between, different sets of arguments for speed increase or decrease distance between presented chunks.

8) Adjusting the block for conditional access

Conditional Access System (CAS) is a mechanism by which television services can be accessed if the relevant provisions of the term of use defined by the digital television provider are met. The need for a conditional system occurs whenever the commercial content is transmitted over to the user by unsafe networks and is used to protect the content of television services transported by satellite, cable or terrestrial systems for the transport of digital content.

Within the elementary stream each packet has a header, that contains information relevant to the transport protocol itself and a segment containing useful data, for example, audio and video. Scrambling takes place at the packet level and only segment containing useful information is encrypted, header remains the same. Main task of *CAS* is to process *Entitlement Control Messages – ECM* and *Entitlement Management Messages – EMM*, descramble encrypted packets using *Control Words – CW* located inside *ECM* and responds to messages to request transported over *EMM*.

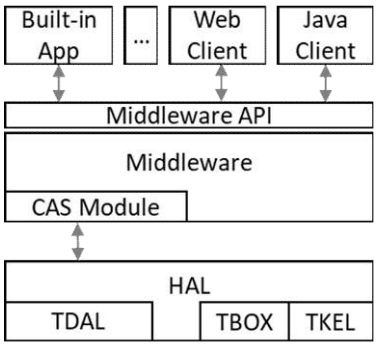


Fig. 4 Conditional access system module

Since library for descrambling is kept in privacy by *Conditional Access Provider*, when adjusting *CAS* to PC platform we had to bypass all the details concerning scrambling and descrambling of content obtained through the transport stream. Existence of descrambling block is abstracted so testing is possible only on streams that are not scrambled. Module is part of middleware Fig. 4, and is only module modified that is not part of HAL. Testing of *CAS* and its importance will be explained in following chapter.

IV. Middleware Testing Environment

In order to assure satisfying test results it is necessary to have a number of test procedures established at various levels. Testing is which, in addition to the software itself, checks all accompanying components and features. Testing is a process that involves a large number of activities and is closely related to software development itself. Bearing in mind that the errors are relatively frequent, it is necessary to introduce a quality control system that will serve for early detection of errors and their removal.

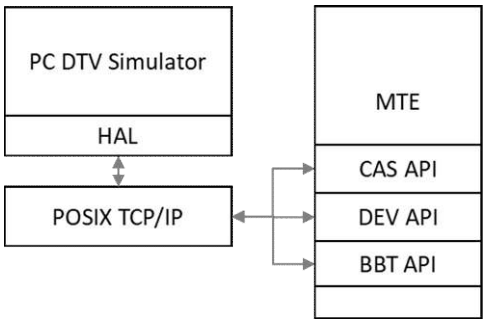


Fig. 5 Communication between PC simulator and MTE

The idea is to develop special modules with a graphical environment that would facilitate work. Their existence in the long run saves a lot of time that would otherwise be spent on writing tests, which ultimately promotes the development of the product itself. Also, in some situations, when several different parties work on development of different modules of the same product, their testing may require the cooperation of several parties. In this case, having a tool that would be standardized during testing can further promote the development itself.

Environment used for the purpose of testing middleware is the *Middleware Testing Environment – MTE* as given in Fig. 5. Graphical environment is developed in Python using wxWidget library, and consists of window that mimics *Remote Controller* and *CAS Middleware Testing Environment – CAS MTE* window.

From user point of view, using keyboard is not practical. Purpose of *Remote Controller* in *MTE* is to enable interaction between user and PC simulator. Beside obvious functionality which is sending events to *TDAL_KBD* module, *Remote Controller* supports *Logging* mechanism as well as environment for automatic testing. It allows to search and filter output from PC simulator using *Logging* mechanism. It is also possible to make automated tests in the form of a series of commands that are generated individually in the *MTE Environment*, send it all together via generated communication stream and execute sequentially on the simulator. Control over digital content is also possible from MTE, PC simulator and MTE can communicate sources as tuple (file name, frequency, signal quality, and bit rate) as seen in Fig. 6.

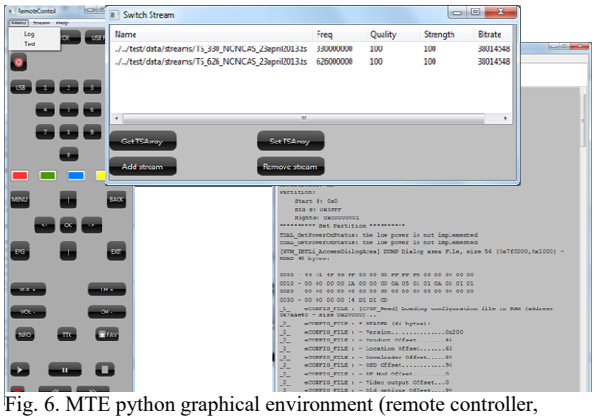


Fig. 6. MTE python graphical environment (remote controller, logger, stream picker)

Part of MTE is dedicated to CAS simulation. It is implemented to send commands that will simulate actual CAS system. On PC simulator side there was need to implement CAS module that would receive commands and process them. It was not possible to use any proprietary CAS libraries as they are not available for PC. Data exchanged simulates use cases for real data that would come through stream in *STB* environment from proprietary CAS. Some of use cases are: access permissions to certain services, fingerprint security, popups, smart card states, email messages.

Simulator for DTV and MTE are two different applications which are also written in two different languages C and Python. Reason behind splitting this was to minimize performance and memory impact on DTV software in execution and to move all the user interaction away from simulator. Communication was realized through TCP/IP protocol using sockets. In simulator HAL layer, which is used to work with sockets, is using *Winsock* library in case of *Windows OS* and *Berkeley sockets* library in case of *Linux OS* [14].

With given MTE environment following features have been tested and checked for functionality flaws:

Table I
Group of tests that were executed using MTE

Test group	Result
Switching services	*PASS*
Electronic programming guide	*PASS*
Personal video recording	*PASS*
User settings	*PASS*

Automated testing reduces amount of time developer has to do manual testing, and removes possibility of missing test cases. Currently test cases prepared for this work take about 10 minutes to execute, where developer will take more than half hour. On the other hand, with implantation of CAS module it is possible to test scenarios that would otherwise require additional resources.

V. Conclusion

With this work we have shown that it can make a fully functional environment for the development of digital TV applications supported by *Windows* and *Linux* platforms. This would mean that all the solutions in the *Middleware*, which does not depend on the hardware performances, can be developed on PC platform, which exceeds the unavailability of hardware and software complexity in the process of software development for embedded systems. Also, *MTE Environment* is designed so that it can easily be adapted for different projects and leaves a lot of space for its further development.

From perspective of development, all possible IDE's that have analytic capabilities like Eclipse, NetBeans, Visual Studio and many others can be attached to PC simulator for debugging. On MTE side, it is also possible to attach IDEs for debugging python execution. It has been tested for both PC simulator and MTE to be run in debugging mode simultaneously.

From perspective of quality assurance, automatic testing black box testing [15] was implemented to help validate changes done in development. Aim of the paper was to expose work that had to be done to create working DTV simulator, and not so much about testing, so it was not given in detail.

Further directions of development PC simulator can go toward adding support for descrambling of protected digital content, adding support for IPTV, adding support for multiple outputs (SD and HD) and supporting for web clients. On MTE side, development should focus toward

improving automated testing, DTV content generation and run-time modification that would allow testing all possible scenarios. This gives the possibility of simulating the operation of the entire television network.

ACKNOWLEDGEMENT

This work was partially funded by the Ministry of Science and Technology of the Republic of Serbia, the number of technological development project: III_044009_2.

REFERENCES

[1] Observatory (by Cabot), http://www.web4595.vs.speednames.com/products/eclipse.html; accessed May 2019
[2] DVBCore (by DTVKit), https://dtvkit.org/wiki/, accessed May 2019
[3] FFmpeg, https://www.ffmpeg.org/, 2019
[4] Simple DirectMediaLayer, http://wikie, 2019
[5] ETSI - European Telecommunications Standards Institute, *Digital Video Broadcasting (DVB), Specification for Service Information (SI) in DVB systems*, 2014.
[6] W. Fisher, *Digital Video and Audio Broadcasting Technology*, 2008
[7] Iwedia, *TDAL_DMX Technical Specifications*, 2009.
[8] Iwedia, *TDAL_DMD Technical Specifications*, 2009.
[9] Iwedia, *TDAL_GFX Technical Specifications*, 2009.
[10] Iwedia, *TDAL_KBD Technical Specifications*, 2009.
[11] Iwedia, *TDAL_AV Technical Specifications*, 2009.
[12] Iwedia, *TDAL_FLA Techical Specifications*, 2009.
[13] Iwedia, *TDAL_MP Techical Specifications*, 2009
[14] G. Miljkovic, "DTV Linux Device Abstraction for Embedded Systems", *ISCE*, ISBN:978-1-4244-6673-3, 2010
[15] T. Tarkan, "User-driven Automatic Test-case Generation for DTV/STB Reliable Functional Verification"*; IEEE Transaction on Consumer Electronics,* vol.58, no.2, pp. 587-595, ISBN: ISSN:0098-3063, 2012

Reproduction of high quality object-based audio content using GStreamer multimedia framework

Srđan Šuvakov, Jelena Kovačević, *Member, IEEE*, Dejan Bokan and Andrej Popović

***Abstract* — The new generation of audio technology introduces a significant increase in audio processing demands from playback devices. The implementation of such technologies for embedded devices represents a challenging task. By using an already existing multimedia framework, we can speed up the development process, but most of the state of the art multimedia frameworks do not support new types of audio stream formats such as object based audio. This paper describes the process of extending the GStreamer multimedia framework with support for object-based audio stream playback. A new GStreamer plugin for rendering audio objects, has been created. The plugin receives and prepares the object-based input stream, invokes the rendering function and after the rendering is finished, the plugin sends the rendered data to the next element in the GStreamer pipeline. This paper also presents a method for audio objects and metadata transfer between GStreamer. The plugin is evaluated on an ARM based SoC device running Linux OS.**

***Index Terms*— Object based audio coding, Audio object rendering, GStreamer, Embedded, Audio systems**

I. Introduction

Over the past few years, we have witnessed a significant progress in developing new technologies for audio content processing, distribution and delivery. Most of these advancements are aimed towards enabling lifelike, scalable and interactive audio experiences to consumers across a wide range of devices and applications. Some of the key properties of modern audio content is spatial realism and immersion. Immersive audio content provides an accurate representation of the spatial attributes of an audio program while providing user experiences that engage our sensory system in a more natural way – allowing sensory systems to process information in a way comparable to how listeners experience the natural world by providing more realistic and natural auditory cues to the listener. [1]. Modern audio technologies allow extending existing mix practices, sound mixers can use music, effects and dialogue in new ways. This allows content creators with the ability to deliver significantly improved, personalized and immersive experience to consumers as shown in [2].

In order to achieve the immersive experience, modern audio technologies usually combine several elements for producing an audio scene, such as: object based audio representation, enhanced height surround sound, and advanced virtualization algorithms. Introducing these new technologies significantly increased the complexity of the tasks performed by audio playback devices. Additionally, these new technologies are being introduced not only as part of high end audio devices, but also in different home audio systems, in-vehicle infotainment systems, mobile devices etc. Thus real time implementation of these technologies has become an important and ongoing topic.

Modern multimedia and infotainment systems are commonly based on System on Chip (SoC) platforms [3]. The audio DSP cores found in most SoCs lack the processing power required to execute new generation complex audio processing technologies on their own. There are two effective solutions to this problem. One is simply increasing the number of audio DSP cores in the SoC or adding additional Audio DSP chips to the device. This solution does not resolve the problem of adding support for modern audio technologies within the existing audio device designs. The other solution to enable support for complex audio processing technologies is to use the application processor for audio processing. This would enable new audio technologies to be implemented within already existing hardware.

In most SoC based multimedia systems application cores are used to execute operating system and user interface tasks. Providing support for multimedia tasks within the OS is usually performed using a multimedia framework. One of the commonly used frameworks is GStreamer [4]. GStreamer is an often used framework when it comes to developing real time media processing and playback solutions for SoCs. GStreamer is an open source project developed by the GNU community. It is written in programming language C, and it supports multiple OS (Windows, Linux, Android, iOS, Mac OS X). Examples of GStreamer being used for playback, recording and streaming in different multimedia systems can be seen in [5], [6] and [7]. GStreamer framework is designed to make it easy to write applications that handle audio or video or both. The pipeline design is made to have little overhead above what the applied processing modules induce. This makes GStreamer a good framework for designing even high-end audio applications which put high demands on latency. GStreamer framework enables easier porting of existing solutions to other SoCs that support the GStreamer framework. Although GStreamer contains support for an extensive set of plugins and tools for streaming, decoding, formatting and processing traditional channel-based audio content, there is currently no support for any form of object based audio representation.

In this paper, GStreamer's functionality is extended with support for object based audio stream playback. An existing

Srđan Šuvakov, RT-RK Institute for Computer Based Systems LLC, Novi Sad, Serbia (e-mail: srdjan.suvakov@rt-rk.com).

Jelena Kovačević, Faculty of Technical Sciences Novi Sad, Trg Dositeja Obradovića 6, Novi Sad, Serbia (e-mail: jelena.kovacevic@rt-rk.uns.ac.rs).

Dejan Bokan, Faculty of Technical Sciences Novi Sad, Trg Dositeja Obradovića 6, Novi Sad, Serbia (e-mail: dejan.bokan@rt-rk.uns.ac.rs).

Andrej Popović, RT-RK Institute for Computer Based Systems LLC, Novi Sad, Serbia (e-mail: andrej.popovic@rt-rk.com).

audio object rendering module is used to create a new GStremaer plugin. Additionaly, support for a new type of stream and metadata handling is added to GStreamer framework.

The object based audio renderer was built as a static library. To run the object based audio renderer within GStreamer a new plugin was created. This plugin would serve as an interface between GStreamer and the object based audio renderer, the plugin receives and prepares the input stream, and after the rendering is done the plugin sends the rendered data to the next element in the GStreamer pipeline. The plugin was implemented on *Raspberry pi 2 model b* as it includes an ARM based chip, which are very prevalent in multimedia SoCs.

II. Object Audio Coding

There are two approaches to 3d audio scene reproduction, channel based and object based [8]. With channel based audio the number and position of the speaker setup must be predefined. This means that any audio content has to be mastered for any speaker setup on which we want to reproduce it on, trying to play audio content mastered for one speaker setup on a different one may result in loss of audio content. Object audio coding has no such issue [9]. Every object in object based audio coding contains audio content for reproduction and its proprietary metadata. This metadata contains additional information that is used to accurately reproduce the audio scene, such as the position of the audio content, its gain as well as its trajectory, if its position changes over time. Thanks to metadata, object based audio systems offer more flexibility when it comes to speaker setups. This also means that the complexity of audio reproduction devices increases, as it's now up to the end device to not only decode, but accurately render the audio scene as well. The added complexity also makes it more difficult to implement object based technologies in existing hardware.

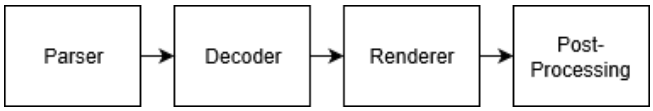


Figure 1 Data processing order within object audio processing systems

There are many different approaches to object audio content distribution and reproduction (such as MPEG-H[10], Dolby Atmos[11] etc.). While different in many ways, the basic principles are similar. Each approach needs to simultaneously unpack objects and metadata, ensure their synchronization and rendering. Accurate rendering requires a pre prepared model of the reproduction room. The creation and presentation of that model is not covered in this paper as that process differs greatly for each object based audio technology. The post-processing of object based audio does not differentiate much from post-processing within channel based audio technologies. Modules dedicated to rendering object audio content receive audio objects that contain raw audio content along with its proprietary metadata from the input stream. The module then renders the audio content from the audio objects according to metadata. Depending on the object audio technology, and its use case, the rendered audio content is then either sent to a post-processing module or towards an audio output, such as speakers. The object rendering module in most object audio technologies is situated between the decoding module and the post-processing module, as is shown on figure 1.

III. GStreamer Framework

The GStreamer framework provides a pipeline mechanism that allows a complex audio system to be divided into several major functionalities and be developed independently of one another. A *bin* in GStreamer is container for a collection of elements. As bins are GStreamer elements themselves, they can be controller as elements, thereby abstracting a lot of complexity of the developed application. For example, you can change the state of all of the elements within one bin by changing the state of the bin. A *pipeline* is a top-level bin. It provides a data buss and manages data flow and synchronization between its elements. Once started, pipelines run in a separate thread until stopped. GStreamer also allows for a pipeline to be split into multiple threads, for example when both video and audio needs to be decoded simultaneously. An example of one pipeline instance can be seen in figure 2.

Plugins are a core building block for the GStreamer framework. Each plugin has one specific function, such as decoding, reading data from a file, outputting to a file or sound card etc. By linking plugins together a pipeline that can do a specific task is created, for example a pipeline for media playback. Each plugin has input and output ports called "Pads". These pads are used to negotiate links and data flow between plugins in GStreamer. Each pad has specific data handling capabilities, or *caps* for short, and will only allow data flow between plugins with compatible pads. In addition to providing a pipeline mechanism, GStreamer also provides an extensive set of tool that cover media type handling, synchronization between pipeline elements and a vast library of existing plugins that provide multimedia processing functionalities. These functionalities include a collection of parsing, decoding and post processing algorithms implementations for both video and audio content. Additionally, there are plugins for handling input and output transport streams, which allow receiving multimedia content from different sources (file, disc, network stream, etc.) and directing audio to a specified output. [4]

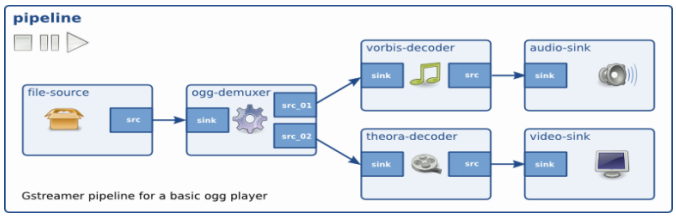


Figure 2 Gstreamer pipeline example

GStreamer offers base classes that provide base functionality for some types of plugins. These classes provide everything required for a plugin to function within the GStreamer framework, meaning that they only need to be expanded with the functionalities that the plugin itself is designed to perform.

IV. Plugin Implementation

As it was stated in the introduction, an object rendering module from an already existing object based audio solution was used

as the base for the created object rendering plugin (ORP) in GStreamer. The rendering module is given in form of a prebuilt C library. The ORP plugin serves as an interface between GStreamer and the rendering module itself, it receives and prepares all the necessary data for the rendering module, invokes the rendering function and afterward it sends the rendered data forward along the pipeline. *BaseTransform* was used as the base class for the ORP. *BaseTransform* is commonly used as a base class for plugins that have raw audio content on both input and output pads. Plugins which extend the *BaseTransform* usually modify only audio content, preserving the format, but it also contains a capability to change audio format (number of channels, sample format, sample rate etc.). This makes is a suitable base class for a module that would transform input objects into a channel based stream that could be reproduced on a variety of speaker configurations. In order to create ORP, *BaseTransform* class is extended and the following functionalities were implemented:

- Caps negotiation and transformation
- Bitmask formation
- Metadata handling
- Rendering

A. *Caps negotiation and transformation*

Caps negotiation is a process between two connecting GStreamer elements in which the output capabilities of the output pad of one element is compared to the input pad of the other to determine if they are compatible with one another. In most GStreamer elements the output caps are inherited from its input caps. The ORP expects audio objects with metadata on the input and sends raw channel audio containing rendered audio scenes to its output. The output caps of ORP are set based on audio scene model which is loaded during module initialization. Output caps of ORP are not negotiable, which means that input pad of successor plugin must match them. These caps are always in form of raw audio content with the possible format parameters:

- Number of channels: 1 to 23
- Sample rate: 1 to 96000
- Audio format: S32LE - 24 bits in 32 bits, unsigned, little endian
- Channel layout: interleaved, non-interleaved

The input caps of ORP are set based on output caps of previous plugin in the pipeline. GStreamer currently does not have ability for identifying object based audio stream. A new identifier for this type of stream is introduced, and named: "Audio/object-audio", the possible audio parameters for object audio content are the same as with raw audio content, with the exception that the number of channels now represents the number of objects being transmitted. Each decoder which has the ability to decode object based audio needs to be modified to support this caps identifier, and to emit the extracted metadata with the audio content.

B. *Bitmask formation*

Since the audio channel bitmask of the object rendering module that was used in the ORP differs from GStreamer's bitmask, the two needed to be aligned. The object rendering module requires the used channels mask to be in the form of a string. For example for a stereo stream that uses the left and right speaker, the string that is sent to the ORP is "LR". The string is then parsed and list of GstAudioChannelPositions is filled. The list of audio channel positions is then turned into a mask by calling the gst_audio_channel_positions_to_mask() function. Alongside the GstAudioChannelPositions list, the function also requires the number of channels. The bitmask created in such a manner is then set as the output channel bitmask by calling the gst_value_set_bitmask().

C. *Metadata handling*

Object audio relies on metadata to provide information needed for accurate reproduction of the audio scene. GStreamer framework contains a mechanism to transfer some additional information about the stream which is referred to as "metadata". But in GStreamer terms, metadata can contain only a small number predefined fields with the information on the properties of the stream (such as its format, type etc.) and as such cannot be used to transfer the metadata needed for object audio rendering. This problem was solved by creating an additional buffer for object audio metadata which is sent alongside the audio content. Specifically, a new buffer is created that encompasses all the metadata required for one frame to be rendered. That buffer is then bound to the main buffer containing audio data that needs to be rendered, and that is achieved using the using GStreamer's ability bind the life expectancy of one buffer to another. This binding is performed using *gst_buffer_add_parent_buffer_meta()* mechanism, which adds a reference to the metadata buffer, as shown in figure 3, and prevents the parent buffer to be returned to the buffer pool while the child buffer is still in use. Additionally a class containing the new metadata structure as well as its functions needed to be made. The functions contained in this class provide functionalities such as initialization, filling and registering the metadata structure with its implementation name. Registering a metadata structure is required so that it can retrieved from the metadata buffer later by using its implementation name.

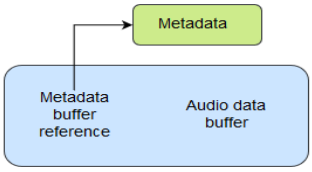


Figure 3 Binding the metadata buffer to the audio data buffer

D. *Rendering*

While the rendering itself is done within the rendering module, as the interface between GStreamer and the module, the task of the ORP is to handle the data prior to, and after the rendering is done. When a new input buffer is received on the ORP input pad, a function called gst_renderer_transform() would be called. The gst_renderer_transform() contains all the rendering functionality such as buffering, rendering and preparing the output stream. After a buffer is received from the input stream, its metadata buffer is extracted using the gst_buffer_get_parent_buffer_meta() function. It is possible that one audio data frame contains more than one metadata frame. After the metadata is extracted, the metadata frames are counted. The entire content of audio data is placed in the input FIFO structure. For each metadata frame, the rendering function is invoked with the FIFO structure as audio input. The rendering module renders audio samples covered by one

metadata frame and increases the FIFO pointer. After rendering all the samples covered by one metadata frame, the rendered audio is placed into the output buffer. This process is repeated while there are still unprocessed metadata frames and enough audio data in the FIFO buffer for the rendering of current metadata frame. After all metadata frames are processed, the output buffer is resized to fit the number of samples rendered, and sent towards the next element within the pipeline for further processing. In case there are samples left in the FIFO buffer after all the metadata frames were rendered, this data is preserved for the next rendering cycle. As GStreamer has the ability to change the size of the buffer transferred between elements of its pipeline, in case that the amount of samples that were rendered is less than the set size of the output buffer the object renderer plugin will resize the buffer to fit the actual size required to transmit all the rendered samples. One example of dataflow in ORP is shown on the figure 4.

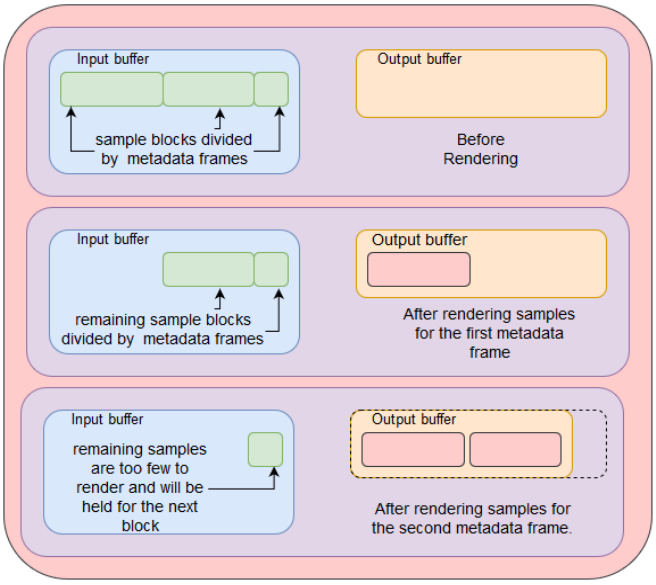


Figure 4 Visual representation of the rendering process for one block of samples

V. TESTING AND RESULTS

The end goal for the ORP was running it in embedded devices. For this purpose the testing was done on Raspberry Pi 2, as it is a good example of a Linux embedded device with GStreamer support. The full test environment is given in table 1.

As previously stated, the rendering module expects objects and their metadata as its input stream. This would normally be done within the decoding or parsing modules, or even the combination of the two. As neither of the two have a GStreamer implementation that supports object audio processing, for the purpose of testing the rendering module, a separate module was created that will load pre prepared objects with their metadata, pack them into a buffer and send them to the rendering module. The module was named "Loadfile" and uses basesrc class template. As its input, this module expects an XML file. This XML file contains the location and names of all the object files and their metadata on disk. The audio object were pre prepared and each object was stored as a separate uncompressed WAV file. Metadata files for all objects were stored in a single file.

Table 1 Test environment

Platform	Raspberry Pi 2 model B
SOC	Broadcom BCM2836
Instruction set	ARMv7-A (32-bit)
CPU	4× Cortex-A7 900 MHz
OS	Raspbian stretch 1.4.6
GStreamer version	Gstreamer1.0
Audio output:	Analog out, HDMI

A. *File I/O test*

The ORP was tested using a pipeline created using GStreamer's gst-launch-1.0 tool, which builds and runs GStreamer pipelines from elements passed as its arguments. The pipeline created by gst-launch-1.0 can be seen in figure 5. The resulting file would then be compared with the output file of the reference object rendering module. This output file was created by running the reference module binary, built for the raspberry pi.

Figure 5 ORP file I/O test pipeline

B. *Real-time test*

The gst-launch-1.0 also enabled the pipeline to be executed in real time. This is done by setting the last element of the pipeline to alsasink. The Alsasnik plugin uses system audio output, such as analog output or HDMI output. The HDMI output was recorded using RT-AG[12]. RT-AG is a digital audio grabber and player with support for both analog and digital audio playback and recording. RT-AG supports recording of up to fourteen channels, and recording up to eight channels at frequencies of up to 192kHz and a depth of 32 bits per sample. The test setup is shown in figure 7. The RT-AG connects to the PC using the Ethernet port to ensure a lossless transmission of high quality audio content. Real time test were executed using a script from the pc. The script would copy test stream files to the Raspberry Pi and run the pipeline shown in figure 6 using gst-launch-1.0. The stream would then be sent through the HDMI to RT-AG . RT-AG would record the stream and send it to the pc through the Ethernet port. The stream would then be saved on the PC as WAV file. This file can then be compared to the output file of the reference module generated on the PC.

Figure 6 ORP real-time test pipeline

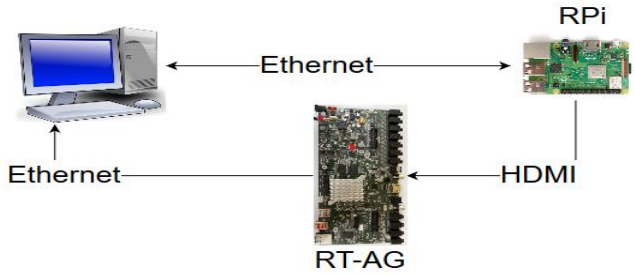


Figure 7 Raspberry Pi 2 connected to RT-AG

C. System resource utilization

After it was determined that the outputs of the reference module and the implemented module were bit exact, the CPU usage of the module was calculated. This was done with the help of the Linux *Perf* tool [13]. Perf tool was used to get performance information of the module such as the total amount of cycles that the implementation used and the running time of the implementation. Perf was also used to calculate how much of CPU time was used for the rendering module itself, as Perf was counting this information for the whole audio chain, including the Loadfile, converter and sink. The final calculation was expressed in MIPS (Millions of instructions per second) and was done by using the following formula:

$$MIPS = \frac{nr\_cycles * \frac{R}{100}}{1000000 * total\_time}$$

Where "R" is the percentage of CPU utilization of the rendering module itself, divided by 100 to normalize it. Table 2 contains the MIPS calculations using the test from the certification package of the object based technology used as the base for the GStreamer plugin. The results from Table 2 show that the rendering time, as well as the processor usage represented in MIPS, doesn't linearly scale up with the amount of objects nor the stream duration. As the object count increased the amount of time it took the module to render an audio scene did increase, while the amount of MIPS the module required didn't change. For example we see that the tests three through five have the lowest amount of MIPS required, even though they are the longest streams with the largest object count. That is due to the stream only having general speech in it that would sound off on which audio channel, and by proxy speaker, it should be positioned. In contrast, the shortest stream from test eight with a relatively low amount of objects has the highest MIPS count. That test used a complex stream that contains a full audio scene that consisted out of ambient sounds as well as foreground noise, fully utilizing the rendering capabilities of the rendering module. On the other hand the rendering time for tests three through five, all of which contain 11 object, compared to their duration was around 80% of the stream duration, while the rendering time of test eight, which contains 3 objects, was 61%. But, these test have shown that, even in the worst case scenario that happened in test eight, that required 383.986 MIPS to run, or 42.66% of a single CPU core, the Raspberry Pi 2 is able to run it on a single core with no issues.

VI. Conclusion

This paper aimed to present GStreamer as a viable option when it comes to the development of new generation audio technology solutions for the embedded devices. The CPU usage alongside the real time tests confirm that GStreamer is able to render objects in real time. With further testing that includes implementing the complete audio reproduction chain within GStreamer as well as utilizing GStreamer's ability to split the pipeline into multiple threads to run the decoder and renderer simultaneously, GStreamer could become a very viable option for embedded media streaming devices that would eliminate the need to add additional dedicated DSP cores. This would also have the added benefit of using GStreamer is lowering the development time for new technologies as GStreamer can reuse the reference code, only requiring an interface to be able to execute it. Most DSP based solutions on the other hand require the whole reference code to be ported and adapted to fit their architecture.

Table 2 MIPS calculation results

Test	Object count	Output channels	Stream duration [s]	Render time [s]	CPU usage [mips]
test_1	1	8	2	1,071	344,640
test_2	1	8	2	1,03	331,969
test_3	11	2	30	23,141	176,272
test_4	11	8	30	23,921	172,595
test_5	11	8	30	23,479	178,689
test_6	5	6	2	1,24	312,563
test_7	5	6	2	1,23	281,501
test_8	3	8	1	0,61	383,986
test_9	1	8	2	1,009	284,937

Acknowledgement

This work was partially supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia. Under Grant number TR32029.

References

[1] J. C. Riedmiller and N. Tsingos. "Recent Advancements in Audio – How a Paradigm Shift in Audio Spatial Representation & Delivery Will Change the Future of Consumer Audio Experiences". Spring Technical Forum Proceedings, Chikago, May 2015.

[2] S. Mehta, T. Orders and J.Riedmiller. "Recepies for Creating and Delivering Next-Generation Broadcast Audio". SMPTE Annual Technical Conference & Exhibition, Hollywood, California. 2015.

[3] S.-J. Chen, G.-H. Lin, P.-A. Hsiung, and Y.-H. Hu, Hardware Software Co-Design of a Multimedia SOC Platform. Springer Science & Business Media, 2009.

[4] *Gstreamer – open source multimedia framework*, https://gstreamer.freedesktop.org/documentation/ [accesed: April 2019.]

[5] Haibin Zhang, Hui Li, Dan Wu, Husheng Yuan, Tao Sun, Peng Yi, Hongchao Hu i Bingqiang Wang. *The design and implementation of an embedded high definition player*. The 2nd International Conference on Computer and Automation Engineering *(ICCAE)*, Singapore, 2010.

[6] Wang H., Hao F., Zhu C., Rodrigues J.J.P.C., Yang L.T. (2012) An Android Multimedia Framework Based on Gstreamer. In: Rodrigues J.J.P.C., Zhou L., Chen M., Kailas A. (eds) Green Communications and Networking. GreeNets 2011. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 51. Springer, Berlin, Heidelberg

[7] D. Darling, C. Maupin, & B. Singh. "GStreamer on Texas Instruments OMAP35x Processors". Proceedings of the Linux Symposium, Montreal, Quebec Canada, pp. 69-78, July 13th–17th, 2009.

[8] G. Potard, "3D-audio object oriented coding", PhD thesis, School of Electrical, Computer and Telecommunications Engineering, University of Wollongong, 2006.

[9] "Multichannel sound technology in homeand broadcasting applications", International Telecommunication Union, Report ITU-R BS.2159-4 (05 /2012), Geneva, Switzerland 2012

[10] Herre, J. and et al. "MPEG-H Audio—The New Standard for Universal Spatial/3D Audio Coding". Journal of the Audio Engineering Society, (2015), 62(12), pp.821-830.

[11] "Dolby Atmos® for the Home Theater". (2016). [PDF] Available at: https://www.dolby.com/us/en/technologies/dolby-atmos/dolby-atmos-for-the-home-theater.pdf [Accessed 18 Apr. 2019].

[12] N. Pekez, J. Kovaĉević, M. Beljulji, "The technical solution of the software architecture of RT-AG digital grabber/player device", International conference for Electronics, Telecommunications, Computers, Automatic Control and Nuclear Engineering (IcEtran) , Kladovo, Serbia, June, 2017

[13] *perf: Linux profiling with performance counters*, https://perf.wiki.kernel.org/, [accessed: April 2019.]

Spectral Analysis of Male and Female Speech Signals

Omar Zelmati, Boban Bondžulić, Milenko Andrić, Dimitrije Bujaković

***Abstract*—In this paper a spectral analysis is applied on a male and female speech audio database. The effect of unsounded audio signal parts on the spectrum rate is studied and it is shown that silent parts disturb strongly the spectral analysis and these parts should be deleted. A comparison between different spectra is made based on correlation. For the case of the spectrums that origins from the same speaker, it is shown that these spectrums are strongly correlated, while a significant correlation between spectrums of the same speaker's gender is highlighted. Finally, the effect of audio signal duration on the spectrums correlation is discussed. The obtained results are very promising and can be used in several fields of speech signal analysis such as speaker recognition and speaker gender identification.**

***Index Terms*— Spectral analysis, Speech signal, Correlation.**

I. Introduction

ANALYSIS of speech signal has various applications such as speaker identification, automatic speech recognition, speaker gender recognition, speech enhancement, etc. In recent years, many researches have been carried out in order to take some advantages of the spectral analysis benefits over the speech time analysis.

As it is stated in [1], spectral analysis (known also "Fourier representation") is often used to highlight certain properties of the speech signal that may be hidden or less obvious if the signal is represented in time domain. These properties are extracted from a spectrum using different techniques and, furthermore, they can be exploited in various methods according to the application and the purpose of the research. Various approaches and implementations in order to extract features of spectrum have been studied in literature [2], such as Discrete Fourier Transform (DFT), and its fast implementations: Fast Fourier Transform (FFT) and Short-Time Fourier Transform (STFT).

Regardless to the signal type, Fourier representation aims to decompose it into its frequency components [3]. For the special case of speaker recognition, numerous signal decomposition techniques based on DFT are proposed. Moreover, some alternatives such as non-harmonic bases, aperiodic functions and data-driven bases derived from independent component analysis have been discussed, and their effectiveness has been placed in evidence [4-6]. However, because of its simplicity and efficiency, DFT is used in practice and usually only the magnitude spectrum is considered, based on the belief that phase has little perceptual importance [7]. The overall shape of the DFT magnitude spectrum, called spectral envelope contains information on the resonance properties of the vocal tract and has been found to be the most informative part of the spectrum for purpose of speaker recognition [1].

Dynamic spectral features (spectral transition) as well as instantaneous spectral features play an important role in human speech perception [8] and many researches on feature extraction for isolated word recognition have investing the effectiveness of using spectral features. In [9] it is shown that Mel Frequency Cepstral Coefficients (MFCCs) are the robust features for isolated word recognition. In this research these coefficients are calculated in six steps: pre-emphasis, applying Hamming windowing, frame blocking, FFT, Log Mel spectrum calculation and applying Discrete Cosine Transform. Furthermore, MFCCs and other measures like pitch, log energy and mel-band energy, have been fixed as based features for emotion recognition using speech signal in [10].

Linear Prediction Coding (LPC) method is applied for speaker gender recognition in [11]. LPC method is such a filter applied on the FFT in order to spectrally flatten the input signal. Beyond its use in gender recognition, this method has been used in speech enhancement and particularly for noise reduction [12].

In this research the spectral analysis of recently formed database of audio signals is performed. Spectral analyses are done regard to the speaker gender (male or female) and as the measure of spectral properties of different speakers the correlation between spectrums of the recorded audio signals is used. These correlations are analyzed for different speakers and for different texts. Beside this, the effects of the audio signals duration are also analyzed through the correlation between male and female speakers.

The rest of paper is organized as follow: in the Section II the used audio signals database is described; in Section III it is applied spectral analysis based on FFT for audio recordings of used database through three analyses: speaker-based correlation, text-based correlation and analysis of the duration of speech signals on the correlation between various speakers.

Omar Zelmati is with the Military Academy, University of Defence in Belgrade, 33 Generala Pavla Jurišića Šturma, 11000 Belgrade, Serbia (e-mail: omarzelmati1991@gmail.com).

Boban Bondžulić is with the Military Academy, University of Defence in Belgrade, 33 Generala Pavla Jurišića Šturma, 11000 Belgrade, Serbia (e-mail: bondzulici@yahoo.com).

Milenko Andrić is with the Military Academy, University of Defence in Belgrade, 33 Generala Pavla Jurišića Šturma, 11000 Belgrade, Serbia (e-mail: andricsmilenko@gmail.com).

Dimitrije Bujaković is with the Military Academy, University of Defence in Belgrade, 33 Generala Pavla Jurišića Šturma, 11000 Belgrade, Serbia (e-mail: dimitrije.bujakovic@va.mod.gov.rs).

In the last part of this paper, results are discussed and some conclusions are highlighted with some direction of further research.

II. Description of Speech Signals Database

For the purposes of the male and female speech signal analysis, an audio recordings database is created with the recording of five already prepared texts in the Serbian language read by five male and five female speakers. As each speaker read five texts, the complete database consists of 50 audio records. The duration of each record is about 30 seconds.

All speakers are aged between 20 and 25 years (students of the Military Academy in Belgrade). In order to guarantee the same environmental conditions of recording, the recording is done in the same place, in a sound-isolated room. All voice recordings were recorded using the SpectraLAB software package on a DELL laptop, with sampling rate f_s = 8 kHz and 16-bit resolution. As input, the microphone of the headphones for the VoIP communication Genius HS04S is used. The sensitivity of used microphone is –60 dB, while its frequency response is within 50 Hz and 20000 Hz.

Audio recordings contain words used in military terminology (gun, pistol, airplane, attack, defense, etc.), so it represents a good basis that in further research they can be used for isolated word recognition or for speaker identification.

III. Speech Signals Spectral Analysis

In this research the FFT of the recorded audio set is applied in order to calculate the correlation between the different obtained spectrums. Although the recording is done in special conditions, the segmentation step is necessary in order to eliminate the effect of the background noise in the spectrum form. All analyzed speech signals are preprocessed due to elimination of the audio signal silent parts. The algorithm for silence removal is proposed in [13] and in this part of the paper is briefly described. The method of silence removal is based on two audio features, the signal energy and the spectral centroid. The applied algorithm consists of four stages:

1) Extraction of the signal energy and spectral centroid from the already decomposed sequences of audio signal;

2) Thresholds estimation for each sequence, where two thresholds are calculated on the base of the extracted features;

3) Application of thresholds criterion on the audio signals sequences, and

4) Speech segments detection based on the threshold criterion and post-processing.

Firstly, the audio signal is divided into non-overlapping frames of the same duration (in this research the frame duration is 50 ms), where $s_i(n)$, n = 0,…, N-1 are the audio samples of the i-th frame of length N. The energy of one frame of audio signal can be calculated using:

$$E_i = \frac{1}{N}\sum_{n=0}^{N-1}\left|s_i(n)\right|^2. \quad (1)$$

This energy is used to detect the silent frames presented in the audio signal, based on the assumption that, if the level of background noise is not very high, the energy of the voice segments is significantly greater than the energy of the silent segments [13]. Silent segments may contain environmental sounds, and for that reason, the measurement of the spectral centroid is performed. The spectral centroid of low energy segments is much smaller, as these noisy sounds tend to have lower frequencies and, as a result, spectral centroid values are lower [13]. If the $X_i(k)$, k = 0,…, N-1 are the DFT coefficients of the i-th frame of length N, the spectral centroid is calculated as:

$$C_i = \frac{\sum_{k=0}^{N-1}(k+1)\left|X_i(k)\right|}{\sum_{k=0}^{N-1}\left|X_i(k)\right|}. \quad (2)$$

After determining all voiced segments of the audio signal, the new-voiced signal is created using concatenation of all voiced segments of the original audio signal. Furthermore, the FFT is applied on each modified signal. Magnitudes of the FFT of an audio signal from created database before and after silence removal are shown in Fig. 1.

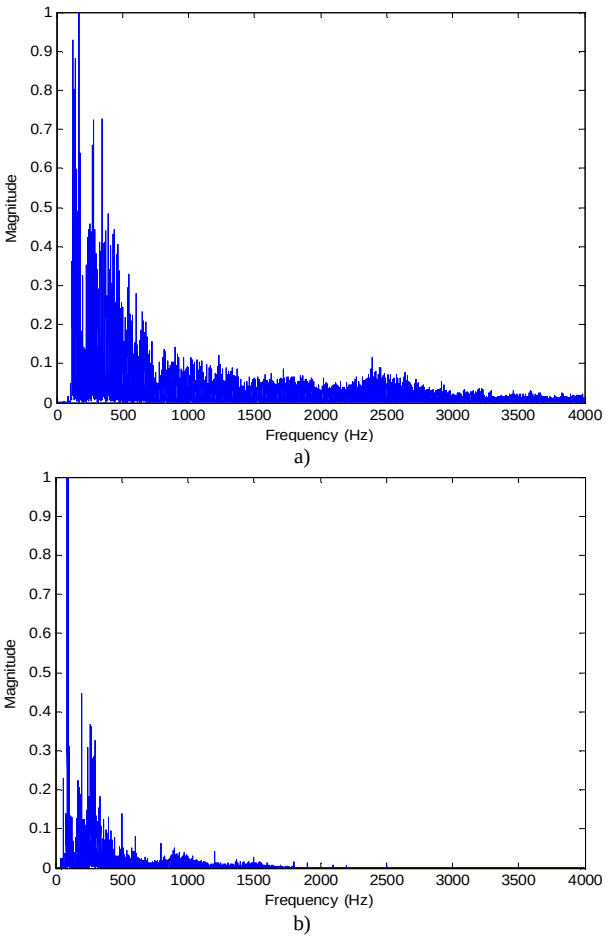


Fig. 1. Normalized spectrum of the audio signal: a) before silence removal, b) after silence removal.

Comparing spectrums before silence removal (Fig. 1.a) and after silence removal (Fig. 1.b), it can be noticed that silent parts strongly affect the spectrum shape. After silence removal, the magnitudes of frequencies above 1800 Hz are suppressed. The correlation between the spectrums of audio signal before and after silence removal is about 20%.

A. Speaker-based Correlation

In the first analysis, the spectrums of audio signals after silence removal of one speaker that speak different texts are compared. The normalized spectrums of audio signals from a one male (speaker 3) and one female (speaker 7) extracted from used database are shown in Fig. 2.

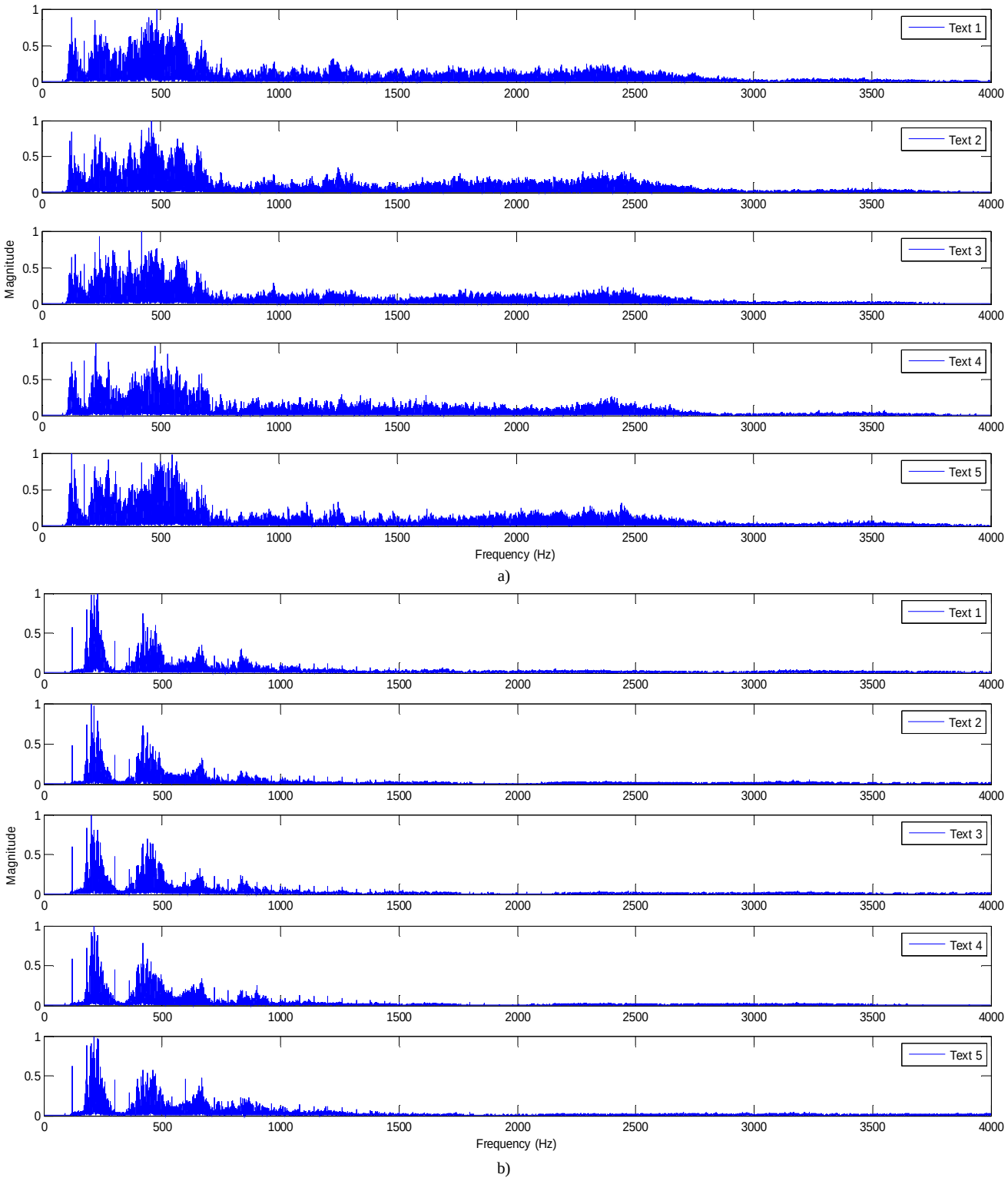


Fig. 2. Spectrums of the different audio speech signals for the same: a) male speaker (speaker 3) and b) female speaker (speaker 7).

From the Fig. 2, it can be noticed that signals origin from the same speaker and different read texts, have the similar spectral shape. Analyzing the spectral properties of male speaker (Fig. 2.a), it can be noticed that spectral components are wider, while for female speaker (Fig. 2.b), spectral components are concentrated around 200 Hz and 400 Hz. In order to quantify spectral similarity, the correlation of spectrum of audio signals produced by male and female speaker is calculated with regard to the spoken texts with and without silence removal. Results of this analysis are shown in Table I (male speaker) and Table II (female speaker).

TABLE I
SPECTRUM BASED CORRELATION FOR THE MALE SPEAKER

	Text 1	**Text 2**	**Text 3**	**Text 4**	**Text 5**
	Without silence removal				
Text 1	1	0.48	0.46	0.41	0.45
Text 2	0.48	1	0.47	0.40	0.43
Text 3	0.46	0.47	1	0.39	0.43
Text 4	0.41	0.40	0.39	1	0.40
Text 5	0.45	0.43	0.43	0.40	1
	With silence removal				
Text 1	1	0.69	0.68	0.66	0.64
Text 2	0.69	1	0.71	0.68	0.67
Text 3	0.68	0.71	1	0.70	0.68
Text 4	0.66	0.68	0.70	1	0.70
Text 5	0.64	0.67	0.68	0.70	1

TABLE II
SPECTRUM BASED CORRELATION FOR THE FEMALE SPEAKER

	Text 1	**Text 2**	**Text 3**	**Text 4**	**Text 5**
	Without silence removal				
Text 1	1	0.37	0.36	0.36	0.32
Text 2	0.37	1	0.37	0.38	0.32
Text 3	0.36	0.37	1	0.37	0.31
Text 4	0.36	0.38	0.37	1	0.32
Text 5	0.32	0.32	0.31	0.32	1
	With silence removal				
Text 1	1	0.70	0.68	0.68	0.65
Text 2	0.70	1	0.70	0.69	0.70
Text 3	0.68	0.70	1	0.71	0.71
Text 4	0.68	0.69	0.71	1	0.72
Text 5	0.65	0.70	0.71	0.72	1

Comparing results of spectral correlation presented in Table I and II, a high correlation can be noticed between the spectrums of the same speaker for different texts read either for male or for female in the case after silence removal while it is much lower for the case without silence removal. The highest correlation value is obtained for the female speaker (72%), while the lowest correlation is obtained for the male speaker (64%). However, the average spectral correlation for the male and for the female speaker is nearly the same: 68% for the male and 69% for the female speaker. From this, it can be concluded that spectral shape after silence removal may be used for feature extraction in order to perform speaker recognition.

B. Text-based Correlation

In order to determine the correlation between the spectrums of the same text uttered by different speakers, the correlation between the spectrums of different speakers is calculated on the text 1 with and without silence removal. The results of spectral correlations are shown in Table III. The part of this table colored in blue represents the correlation between spectrums of five male speakers for the read text, while the part colored in orange represents the correlation between spectrums of all the five female speakers for the same text. The yellow part of Table III is the spectral correlation between males and female speakers.

TABLE III
SPECTRAL CORRELATION OF TEXT 1

		Male speakers					**Female speakers**				
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
		Without silence removal									
Male speakers	**1**	1	0.21	0.36	0.33	0.37	0.23	0.22	0.17	0.21	0.10
	2	0.21	1	0.25	0.23	0.24	0.14	0.10	0.14	0.16	0.08
	3	0.36	0.25	1	0.38	0.33	0.28	0.24	0.29	0.33	0.14
	4	0.33	0.23	0.38	1	0.35	0.22	0.20	0.24	0.26	0.15
	5	0.37	0.24	0.33	0.35	1	0.20	0.18	0.23	0.22	0.19
Female speakers	**6**	0.23	0.14	0.28	0.22	0.20	1	0.31	0.28	0.36	0.14
	7	0.22	0.10	0.24	0.20	0.18	0.31	1	0.27	0.30	0.15
	8	0.17	0.14	0.29	0.24	0.23	0.28	0.27	1	0.37	0.29
	9	0.21	0.16	0.33	0.26	0.22	0.36	0.30	0.37	1	0.21
	10	0.10	0.08	0.14	0.15	0.19	0.14	0.15	0.29	0.21	1
		With silence removal									
Male speakers	**1**	1	0.53	0.54	0.59	0.55	0.44	0.43	0.32	0.38	0.31
	2	0.53	1	0.56	0.61	0.55	0.43	0.35	0.40	0.45	0.36
	3	0.54	0.56	1	0.60	0.50	0.49	0.44	0.43	0.51	0.37
	4	0.59	0.61	0.60	1	0.57	0.50	0.46	0.46	0.51	0.43
	5	0.55	0.55	0.50	0.57	1	0.39	0.34	0.37	0.39	0.42
Female speakers	**6**	0.44	0.43	0.49	0.50	0.39	1	0.63	0.52	0.66	0.44
	7	0.43	0.35	0.44	0.46	0.34	0.63	1	0.50	0.58	0.42
	8	0.32	0.40	0.43	0.46	0.37	0.52	0.50	1	0.58	0.59
	9	0.38	0.45	0.51	0.51	0.39	0.66	0.58	0.58	1	0.51
	10	0.31	0.36	0.37	0.43	0.42	0.44	0.42	0.59	0.51	1

From the Table III, it can be noticed that the average spectral correlation of the different male speakers after applying silence removal is about 56% while for different female speakers is about 54%. The average correlation of the male-female part is about 42%. From this, it can be concluded that spectrums of the same gender are notably correlated, whilst spectrums of different gender are less

correlated. This can be explained by the fact that the energy of the audio origins from female speaker is concentrated around 200 Hz and 400 Hz, while energy for male speaker is distributed on larger frequency range. The average spectral correlation of all male speakers for the case without silence removal is about 30% and for all female speakers it is about 27%. These results confirm that the correlation decreases significantly because of the parts of the signal having a low energy (unsounded parts). These conclusions may be used for feature extraction in order to perform gender recognition using speech signal.

C. *Effect of Audio Signal Duration*

For the purpose of the audio signal duration effect investigation on the spectrum correlation, ten signals were prepared by concatenating all the uttered texts for each speaker. In such manner, each speaker is presented by a larger signal. The spectrum of this signal after silence removal is determined and the correlation between it and original signals is calculated. The results are shown in Table IV.

TABLE IV
CORRELATION BETWEEN SPECTRUMS OF SIGNALS SUM AND ORIGINAL ONES

	Speaker	Text1	Text2	Text3	Text4	Text5
All texts	**1**	0.77	0.75	0.81	0.72	0.70
	2	0.76	0.80	0.78	0.81	0.77
	3	0.81	0.73	0.75	0.76	0.76
	4	0.85	0.84	0.81	0.77	0.72
	5	0.73	0.79	0.67	0.81	0.80
	6	0.82	0.69	0.68	0.66	0.64
	7	0.79	0.78	0.71	0.80	0.68
	8	0.78	0.71	0.76	0.70	0.78
	9	0.81	0.68	0.75	0.79	0.70
	10	0.74	0.82	0.77	0.70	0.83

From the Table IV, it can be noticed that the correlation is enhanced using longer speech signal duration. Analyzing results from the Table I, it can be noticed that the average correlation between the different texts of the speaker 3 (with silence removal) is around 68%. On the other hand, the average correlation of the same speaker calculated based on Table IV is about 76%. From this, it can be concluded that if the audio signal is longer, the spectral correlation after silence removal is higher.

In order to support this conclusion, correlation matrices for each of the five analyzed texts and for all concatenated texts are calculated. From each obtained matrix, the average correlation between spectrums of male speakers (male vs. male), the average correlation between spectrums of female speakers (female vs. female) and the average correlation between speakers of different gender (male vs. female) are calculated. Referring to the Table III the male vs. male average correlation is computed based on the blue part of the table, while the orange part serves to calculate the female vs. female average correlation and the part in yellow is used to calculate the male vs. female correlation. Results of this analysis are presented in Fig. 3.

From the Fig. 3, it can be noticed that for higher duration of analyzed signal the spectral correlation between speakers is higher with silence removal than without it. Beside this, from this figure it can be concluded that spectral correlation after silence removal is higher to the speaker gender, while by comparing audio signals that origin from the speakers of different gender, it is concluded that the spectral correlation is lower. These results can be used for determination of the optimal speech recording duration in data training of speech analysis datasets.

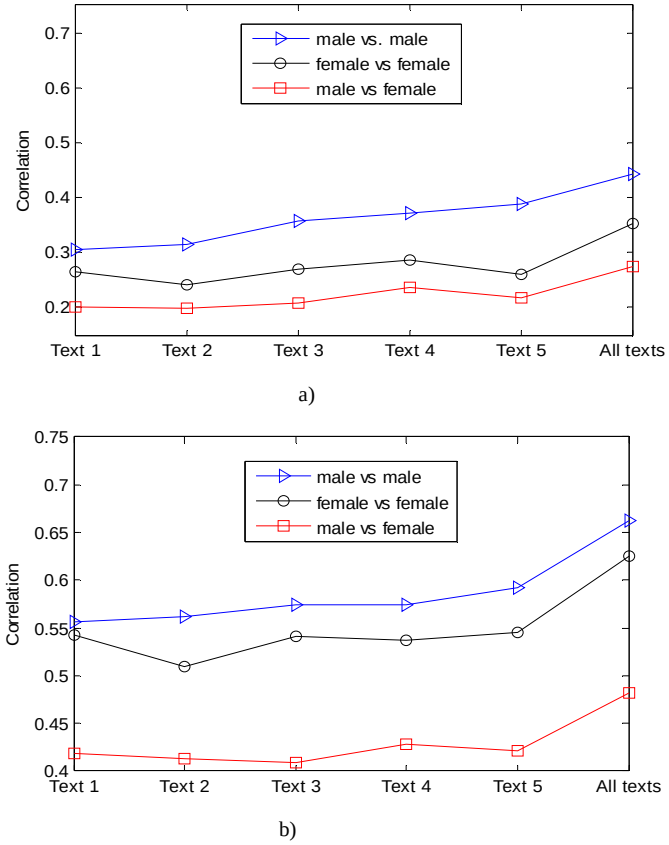


Fig. 3. Average correlation for the different texts and speaker gender: a) without silence removal and b) with silence removal.

IV. CONCLUSION

This paper presented a spectral analysis applied on recently created database that consists of male and female speech samples. Using the spectral correlation measure, in this research it is analyzed the effect of unsounded audio signal parts on the spectrum shape and it is shown that silent parts strongly affects the results of a spectral analysis. Beside this, in this research the spectrums obtained from signals of the same speaker for different uttered texts are compared and it is shown that they are strongly correlated. Moreover, there is a significant correlation between spectrums obtained from speakers of different gender reading the same text. In this research is also analyzed the

effect of the audio signal duration on the spectrum correlation. Obtained results show that for longer duration signal, the spectral correlation is higher. These results may be ground for future researches related to the speech signal analysis. In future research, the problem of spectral-based feature extraction for speaker identification will be considered. Furthermore, the case of background noise, music and distant speakers will be taken into account.

REFERENCES

[1] L. R. Rabiner and R. W. Schafer, *Theory and applications of digital speech processing*, NJ, USA: Pearson, 2011.

[2] A. V. Oppenheim, *Discrete-time signal processing*, ND, India: Pearson Education India, 1999.

[3] L. R. Rabiner and R. W. Schafer, *Digital processing of speech signals*, New Jersey, USA: Prentice-Hall, 1978.

[4] K. Gopalan, T. R. Anderson, and E. J. Cupples, "*A comparison of speaker identification results using features based on cepstrum and Fourier-Bessel expansion*," IEEE Trans. on Speech and Audio Proc, vol. 7, no. 3, pp. *289-294*. Apr. 1999.

[5] B. Imperl, Z. Kai, and B. Horvat, "*A study of harmonic features for the speaker recognition*," Speech Communication, vol. 22, no. 4, pp. 385-402. Feb. 1997.

[6] G.-J. Jang, T.-W. Lee, and Y.-H. Oh, "*Learning statistically efficient features for speaker recognition*," Neurocomputing, vol. 49, no. 4, pp. 329-348. Jun. 2002.

[7] T. Kinnunen and H. Li, "*An overview of text-independent speaker recognition: From features to supervectors*," Speech Communication, vol. 52, no. 1, pp. 12-40. May 2010.

[8] G. Ruske and T. Schotola, "The efficiency of demisyllable segmentation in the recognition of spoken words," in ICASSP'81. IEEE International Conference on Acoustics, Speech, and Signal Processing, NY, USA, vol. 1, pp. 971-974. 04-01-1981.

[9] M. P. Kesarkar, "Feature extraction for speech recognition," M.Tech. Credit Seminar Report, Bombay, India, 2003.

[10] O.-W. Kwon, K. Chan, J. Hao, and T.-W. Lee, "Emotion recognition by speech signals," European Conference on Speech Communication and Technology. Geneva, Switzerland, interspeech. 09-01-2003.

[11] K. Rakesh, S. Dutta, and K. Shama, "*Gender Recognition using speech processing techniques in LABVIEW*," Internat. Journal of Advs. in Engin. & Tech., vol. 1, no. 2, p. 51. Jul. 2011.

[12] M. Hydari, M. R. Karami, and E. Nadernejad, "*Speech Signals Enhancement Using LPC Analysis based on Inverse Fourier Methods*," Contemporary Engineering Sciences, vol. 2, no. 1, pp. 1-15. Jan. 2009.

[13] T. Giannakopoulos, "Study and application of acoustic information for the detection of harmful content, and fusion with visual information," PhD. Dissertation, IT Dept., NKA Univ., Athens, Greece, 2009.

pyHRV: Development and Evaluation of an Open-Source Python Toolbox for Heart Rate Variability (HRV)

Pedro Gomes, Petra Margaritoff and Hugo Silva

***Abstract*—Heart Rate Variability (HRV) is a continuously growing field of interest in research, with an increasing number of new measurements being introduced over the recent decades, followed by complementary software tools. However, mostly closed source HRV tools are available, with high license costs and that prevent source code access for developers, limiting the possibilities of custom application development. Open-source solutions, on the other hand, face different limitations such as reduced functionality, non-validated results, or full support for mainstream programming languages. In this work, we describe a fully open-source Python toolbox named pyHRV for HRV in custom research and development applications, which is validated against a gold standard software. pyHRV computes state-of-the-art Time Domain (TD), Frequency Domain (FD), and Nonlinear (NL) HRV parameters. As for the evaluation, short-term parameters have been computed from 50 Normal-to-Normal Interval (NNI) series of 5 minutes in duration, with long-term parameters being computed from 50 NNI series of 60 minutes in duration. The results have been computed using the pyHRV toolbox and the KUBIOS HRV gold standard software, against which the pyHRV results have been compared. Overall, pyHRV computes a total of 78 HRV parameters (23 TD, 48 FD, 7 NL), from which 12 have achieved identical results as the gold standard software, 38 showing marginal and/or neglectable differences, and 26 showing significant differences, thus requiring further investigation.**

***Index Terms*—Heart Rate Variability, Python, Open-Source**

I. Introduction

Heart Rate Variability (HRV) has been established as a well known and important measure of cardiac and overall health, due to the extensive amount of research conducted over the past decades [1], [3], [4]. This is a continuously evolving research field in which series of temporal, spectral, and non-linear parameters are extracted from Inter-Beat-Interval (IBI) series, i.e. the intervals between successive heart beats, which has strongly benefited from the fast pace of technological advancements in the biomedical field [3], [18]. In addition, the advancements in consumer technology facilitated the access to tools with Heart Rate (HR) extracting capabilities, opening up new fields in which HRV can be found beyond the traditional clinical context, such as in pharmacology, sports science, and virtual reality applications [12], [16], [17], [19].

Many software solutions specialized in HRV have been published, motivated by its increasing research popularity and application fields providing researchers the necessary tools to support their research. However, solutions on both ends of the closed-source and open-source spectrum exhibit different shortcomings in their functionality, usability, or selection of extracted HRV parameters often limiting developers in their work towards new and custom software solutions. For instance, commercially available tools may often be well-established in research communities, but lack the insights in the underlying source code with little to no possibility to extract individual HRV algorithms for further integration possibilities. In addition, high-cost licenses are required to grant full access to all functionality of the software tools to researchers and developers. While open-source tools do not, in most cases, face the same limiting aspects of closed-source solutions, they do not always provide all features required for the computation of the most popular HRV features, have not been validated against gold-standard software, or lack support for mainstream programming languages.

The goal of this work is to validate the results of a newly developed open-source HRV toolbox for the Python programming language named pyHRV, to provide a reliable solution for HRV researchers and developers[1]. This toolbox complies with the HRV guidelines *"Heart Rate Variability: Standard of Measurement, Physiological Interpretation, and Clinical Use"* issued by the Task Force of The European Society of Cardiology and The North American Society of Pacing Electrophysiology in 1996 and newer findings in HRV research [1], [4], [18]. As for the evaluation, series of Normal-to-Normal Interval (NNI) datasets have been computed with the pyHRV toolbox and with KUBIOS HRV, here identified as the gold-standard software, compared, and evaluated to measure the reliability of the newly developed solution [20].

The remainder of the paper is organized as follows. The methods and materials are described in Section II. In Section III, the experimental results are presented to illustrate the performance of the proposed toolbox. Finally, conclusions and future work are discussed in Section IV.

Pedro Gomes is with the Department of Biomedical Engineering, Hamburg University of Applied Sciences, Ulmenliet 20, 21033 Hamburg, Germany (e-mail: pedro.gomes@haw-hamburg.de).

Petra Margaritoff is with the Department of Biomedical Engineering, Hamburg University of Applied Sciences, Ulmenliet 20, 21033 Hamburg, Germany (e-mail: petra.margaritoff@haw-hamburg.de).

Hugo Silva is with the Instituto de Telecomunicações, Instituto Superior Tecnico, Av. Rovisco Pais, n. 1, Torre Norte - Piso 10, 1049 - 001 Lisbon, Portugal (e-mail: hsilva@lx.it.pt).

[1]pyHRV is available on GitHub: https://github.com/PGomes92/pyhrv

II. Methods and Materials

A. HRV Parameters Computable with pyHRV

pyHRV computes TD and FD parameters in accordance with the HRV guidelines and latest findings in HRV research, adding also newer methods for the extraction of NL parameters [1], [3], [4], [6]. In the TD, temporal statistical parameters are derived from the NNI series, Series of Successive NNI Differences (ΔNNI), and HR series. Table I lists all computable TD parameters using pyHRV. In the FD, all spectral parameters are computed from Power Spectral Density (PSD) estimation methods computed from the NNI series. The Fast Fourier Transform (FFT)-based Welch's and the Autoregressive (AR) methods have been implemented according to the HRV guidelines, with the selection of available PSD methods being extended by the Lomb-Scargle method [1], [7]–[9].

Table II lists all computable FD parameters which can be computed over up to 4 frequency bands, the Ultra Low Frequency (ULF), Very Low Frequency (VLF), Low Frequency (LF), and High Frequency (HF) band. Default frequency ranges are defined for the VLF (0.00Hz-0.04Hz), LF (0.04Hz-0.15Hz), and HF (0.15-0.40Hz) bands [3].

The guidelines failed to establish a standard set of nonlinear methods and parameters, due to the lack of scientific backing for the variety of suggested methods at the time [1]. The Poincaré scatter plot is a commonly found visualization method of HRV, from which a set of parameters can be derived: the Standard Deviation (SD) along the minor axis of the scatter plot (SD1), the SD along the major axis (SD2), the SD2/SD1 ratio, and the area of a fitted ellipse with the focal points SD1 and SD2 along the minor and major axis [3], [6]. In addition, the set of NL parameters has been extended by the Sample Entropy (SamPen) and the Detrended Fluctuation Analysis (DFA) [4]. The computable NL parameters are listed in Table III.

TABLE I
pyHRV Time Domain Parameters computed from the NNI series, series of successive NNI differences (ΔNNI), and HR series.

pyHRV Time Domain Parameters		
Parameter	Acronym	Unit
NNI series (min, max, mean)	–	ms
ΔNNI (min, max, mean)	–	ms
HR series (min, max, mean, SD)	–	bpm
Standard Deviation of NNI	$SDNN$	ms
SD of the ΔNNI series	$SDSD$	ms
Root Mean of Squared ΔNNI	$RMSSD$	ms
Mean of the SD of 5 minute segments	$SDNN_{Index}$	ms
SD of the means of 5 minute segments	$SDANN$	ms
NNX (NNI $>$ Xms)[1]	NNX	–
pNNX (% of NNI $>$ Xms)[1]	$pNNX$	–
Triangular Index	TRI	ms
Baseline width of the NNI histogram	$TINN$	ms

[1]additional functions provided for preset thresholds of 50ms and 20ms

B. Open-Source Third-Party Python Packages

The BioSPPy (v.0.5.1) packages *biosppy.signals.ecg.ecg()* function is used in pyHRV to extract R-peak locations from ECG signals, which are required for the computation of the NNI series, from which the HRV parameters are subsequently computed. Different packages have been used for the computation of the different methods of PSD estimations. The *scipy.interpolate.interp1d()* and the *scipy.signal.welch()* functions of the SciPy package (v.1.1.0) are used for the 4Hz interpolation of the NNI series and the computation of the Welch's PSD. Additionally, the *scipy.signal.lombscargle()* method is used for the computation of the Lomb-Scargle method. The AR PSD is computed using the Yule-Walker algorithm of the Spectrum package (v.0.5.2) using the *spectrum.pyule()* function. The NL parameters, the SamPen and the DFA are computed using the *nolds.sampen()* and the *nolds.dfa()* functions from the nolds (NOnLinear measures for Dynamical Systems) (v.0.4.1) package. All functions with plotting capabilities use the Matlplotlib (v.2.2.2) plotting library. Data series and their manipulation are stored and conducted in the NumPy (v.1.15.1) array format and its supported functionalities. Other Python native packages are used through pyHRV for general purpose functions such as data import and export, time stamp creation, and operating system-specific functionalities. All functions are used, configured, and extended with HRV application-specific input parameters, ranges, and methods [1].

TABLE II
pyHRV frequency domain parameters computed using the Welch, Autoregressive, and Lomb-Scargle methods.

pyHRV Frequency Domain Parameters		
Parameter	Acronym	Unit
Peak Frequencies[1]	P_{Peak}	Hz
Absolute Powers[1]	P_{Abs}	ms^2
Logarithmic Powers[1]	P_{Log}	ms^2
Relative Powers[1]	P_{Rel}	%
Normalized Powers[2]	P_{norm}	–
LF/HF Ratio[2]	LF/HF	–

[1]for ULF, VLF, LF, and HF bands
[2]for LF and HF bands only

TABLE III
pyHRV nonlinear parameters.

pyHRV Nonlinear Parameter		
Parameter	Acronym	Unit
SD Along the Minor Axis[1]	$SD1$	ms
SD Along the Major Axis[1]	$SD2$	ms
SD2/SD1 Ratio[1]	$SD2/SD1$	–
Area of the Fitted Ellipse[1]	S	ms^2
Sample Entropy	$SamPen$	–
Detrended Fluctuation Analysis	DFA	$beats$

[1]computed from the Poincaré plot

C. NNI Series Used for the Evaluation Procedure

The NNI series used is the MIT-BIH Normal Sinus Rhythm Database (MIT-BIH) from the PhysioNet database [2]. This database consists of long-term ECG recordings from 18 healthy subjects (5 men, 26 to 40 years and 13 women, 20 to 50 years) with no observable pathological ECG arrhythmias. The NNI series were extracted directly from the PhysioBank ATM (Automated Teller Machine). The use of

the NNI series as input data is preferred over the use of the raw ECG signals to avoid HRV parameter computation using non-identical input data series, as differences could occur due to the different implementations of the R-peak detection algorithms in BioSPPy and KUBIOS HRV. Such differences could ultimately influence the parameter results (pyHRV vs. KUBIOS HRV).

D. NNI Pre-Processing and HRV Parameter Computation

Although the physiological context and significance of the computed HRV parameters are not of interest for the evaluation process, all efforts were made to provide physiologically reasonable data and HRV input parameters. The NNI series of the MIT-BIH database contain NNI outliers of up to approx. 42 seconds in duration. These outliers are caused by visible motion artifacts and/or electrode disconnections during the acquisition rather than having a physiological source, and are visible the original ECG datasets, distorting the NNI series as depicted in the Tachogram shown in Figure 1. This could greatly alter the signal segmentation process in the next steps and the HRV parameter results, reason for which these outliers have been removed by applying a threshold of 1.2s ($\hat{=}$ 50 bpm) as maximum acceptable NNI duration.

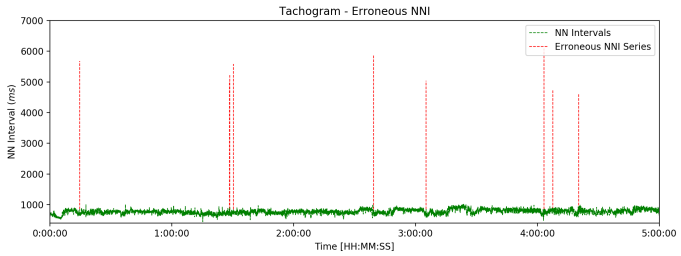


Fig. 1. Tachogram of a NNI series with outliers caused by signal artifacts during the ECG acquisition (in red). These outliers can cause issues during the signal segmentation process of the NNI series pre-processing and eventually alter the results of the HRV parameters (signal source: MIT-BIH, signal 19090 [2]).

After applying the 1.2s threshold, the original NNI series were segmented into multiple blocks of shorter durations, resulting in segments with 5 minute duration for the computation of short-term parameters and 60 minute segments for the computation of the long-term parameters SDNN Index and SDANN. The algorithms computing these long-term parameters contain their own segmentation process, which splits the 60 minute segments into 12 $\times$ 5 minute segments, reason for which additional datasets have been generated to validate these parameters. From all generated segments, 50 x 5 minute and 50 $\times$ 60 minute segments have been randomly selected from which HRV parameters were computed using pyHRV and KUBIOS HRV according to the configurations listed in Table IV.

E. Classification and Validation of the HRV Results

The following statistical taxonomy has been applied to classify the pyHRV results into three categories: *Optimal*, *Acceptable*, and *Divergent*. The mean values $\overline{P_x}$ and $\overline{K_x}$, and the standard deviations (SD) $\sigma(P_x)$ and $\sigma(K_x)$ have been computed for both the pyHRV (P_x) and KUBIOS (K_x) results, with x being one of the computed HRV parameters. In addition, the absolute difference Δ_x between $\overline{P_x}$ and $\overline{K_x}$ values have been computed. The entire pre-processing, parameter computation, and evaluation process is illustrated in Figure 2. Depending on these parameters, the performance of the pyHRV parameters have been classified as follows:

TABLE IV
PYHRV & KUBIOS PARAMETER SETTINGS.

Parameter Settings	
Parameter Specific Settings	Set Value(s)
NNI detrending method	None[k]
HR parameters computed as average of X heart beats	1 beat[k]
VLF Frequency Band	0.00 Hz - 0.04 Hz
LF Frequency Band	0.04 Hz - 0.15 Hz
HF Frequency Band	0.15 Hz - 0.40 Hz
FFT - Interpolation frequency	4 Hz
FFT - Window width	300 s
FFT - Window overlap	50 %
Lomb - Smoothing	None[k]
AR - Model order	16
Sample Entropy - Embedding dimension	2
Sample Entropy - Tolerance	0.2 · SD
DFA - Short term fluctuations	4 - 16 beats
DFA - Long term fluctuations	17 - 64 beats

[k] KUBIOS HRV specific parameter settings only.

Optimal: $\Delta_x = 0$
The computed parameter results are exact. No difference exists between $\overline{P_x}$ and $\overline{K_x}$.

Acceptable: $0 < |\Delta_x| \leq \sigma(K_x)$
The computed parameter results have a difference Δ_x between $\overline{P_x}$ and $\overline{K_x}$ which is $\leq \sigma(K_x)$. Differences within 1 SD of $\overline{K_x}$ are considered non-significant and, therefore, acceptable.

Divergent: $—\Delta_x| > \sigma(K_x)$
The computed parameter results are not identical and have a significant difference Δ_x. These results need further analysis.

III. MAIN RESULTS

A. Time Domain Results

The TD results are presented in Table V. Overall, 10 of the 11 comparable results lie within the optimal and acceptable ranges, with the Triangular Interpolation of the NNI Histogram (TINN) parameter being the only divergent parameter. The TINN's difference between the means is almost twice as high as the KUBIOS mean result with pyHRV's mean value at roughly only a third of the KUBIOS mean ($\sigma(K_{TINN}) = 104.915\ ms$ vs. $|\Delta_{TINN}| = 19.4887\ ms$), thus showing a significant difference. Although KUBIOS does not return the N and M values of the TINN parameter computation, it can be assumed that these parameters are also divergent given that they are the base parameters for the computation of the TINN parameter.

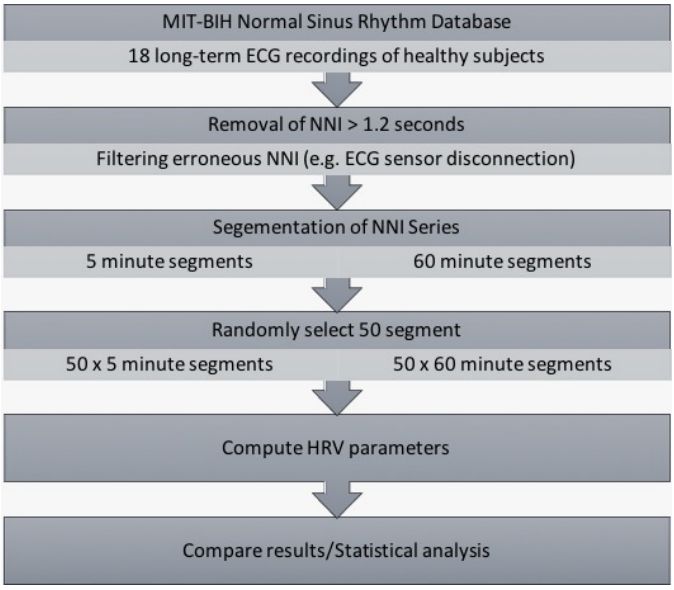


Fig. 2. NNI series pre-processing and evaluation procedure.

The mean of the NNI, minimum HR, SDNN, NN50, and pNN50 parameter results show optimal performance being identical to the KUBIOS results. It can be assumed that the non-compared NN20 and pNN20 parameters also achieve optimal results given that these parameters are computed by the same fundamental function as the NN50 and pNN50 parameters ($tools.time\_domain.nnXX()$). The mean HR ($\sigma(K_{\overline{HR}}) = 13.422\ bpm$ vs. $|\Delta_{\overline{HR}}| = 0.552\ bpm$), SDNN index ($\sigma(K_{SDNNI}) = 17.352\ ms$ vs. $|\Delta_{SDNNI}| = 0.320\ ms$), Standard Deviation of the Mean of NN Intervals in all 5 minute Segments (SDANN) ($\sigma(K_{SDANN}) = 19.331\ ms$ vs. $|\Delta_{SDANN}| = 0.716\ ms$), and the Triangular Index ($\sigma(K_{TRI}) = 3.886$ vs. $|\Delta_{TRI}| = 0.342$) lie within acceptable limits, with the RMSSD ($\sigma(K_{RMSSD}) = 22.611\ ms$ vs. $|\Delta_{RMSSD}| = 0.001\ ms$) being close to the values computed by KUBIOS, with a marginal difference of $0.001\ ms$.

B. Frequency Domain Results

The FD parameters computed from the Welch's method are presented in Table VI, and show overall optimal and acceptable results. pyHRV's peak frequencies in each frequency band are identical to the results obtained with KUBIOS. Differences can be found for the rest of the parameters, although marginal, given that the mean results exist within the acceptable range.

The FD parameters computed from the Lomb-Scargle method are presented in Table VIII. The evaluation results of this method identify half of the parameters within acceptable ranges, with the other half providing divergent results. No optimal results are computed by pyHRV. The Peak Frequencies show acceptable results for the VLF ($\sigma(K_{VLF,Peak}) = 0.009\ Hz$ vs. $|\Delta_{VLF,Peak}| = 0.006\ Hz$) and HF ($\sigma(K_{HF,Peak}) = 0.066\ Hz$ vs. $|\Delta_{HF,Peak}| = 0.063\ Hz$) with the LF results being divergent due to a greater difference than the KUBIOS results ($\sigma(K_{LF,Peak}) = 0.019\ Hz$ vs. $|\Delta_{LF,Peak}| = 0.028\ Hz$). Other acceptable results could be achieved for the Absolute Powers in all Frequency Bands (FB), the Total Power, and the Relative and Logarithmic Powers in LF band. The remaining parameters are divergent, therefore requiring further investigation. It is worth mentioning that no consistency over FB-specific results can be observed. For instance, the Absolute Powers provide acceptable results in the VLF and HF band. However, parameters derived from these parameters (Log. Powers and Relative Powers) show divergent results in these bands. This inconsistency influences additional parameters which depend on the Absolute Powers of the HF band (LF/HF ratio and Normalized Powers), leading to the divergent results.

The evaluation results of the Autoregressive parameters are presented in Table VII. Overall the results of 6 parameters lie within acceptable ranges, with 10 parameter results being divergent due to differences greater than 1 SD of the KUBIOS results. The Peak Frequencies lie within acceptable ranges for every FB. Other acceptable parameter results were computed for the Absolute Powers ($\sigma(K_{VLF,Abs}) = 2666.353\ ms^2$ vs. $|\Delta_{VLF,Abs}| = 645.599\ ms^2$), and Log. Powers ($\sigma(K_{VLF,Log}) = 0.978\ ms^2$ vs. $|\Delta_{VLF,Log}| = 0.742\ ms^2$) of the VLF band, as well as the Relative Powers of the LF frequency band ($\sigma(K_{LF,Rel}) = 14.716\ \%$ vs. $|\Delta_{LF,Rel}| = 6.466\ \%$). However, all remaining results lie outside the 1 SD limits considered to be acceptable. It should be noted that, overall the Absolute Powers, with the exception of the VLF band, are significantly higher and lie multiple SD apart from the KUBIOS results, being almost 4.3 times higher in the LF band ($\sigma(K_{LF,Abs}) = 1137.171\ ms^2$ vs. $|\Delta_{LF,Abs}| = 4858.550\ ms^2$) and 8 times higher in the HF band ($\sigma(K_{HF,Abs}) = 1455.146\ ms^2$ vs. $|\Delta_{HF,Abs}| = 11512.643\ ms^2$).

C. Nonlinear Parameter Results

The evaluation results for the Nonlinear parameters are listed in Table IX. Overall, the 7 comparable parameters provide acceptable results, with the SD2/SD1 ratio showing even optimal results. Marginal differences are present for the SD1 ($\sigma(K_{SD1}) = 16.015\ ms$ vs. $|\Delta_{SD1}| = 0.044\ ms$) and SD2 ($\sigma(K_{SD2}) = 35.962\ ms$ vs. $|\Delta_{SD2}| = 0.123\ ms$) parameters, with the Sample Entropy results achieving almost optimal results ($\sigma(K_{SamPen}) = 0.366$ vs. $|\Delta_{SamPen}| = 0.001$). The highest relative differences are present in the results of the DFA α_1 ($\sigma(K_{\alpha 1}) = 0.240$ vs. $|\Delta_{\alpha 1}| = 0.022$) and α_2 ($\sigma(K_{\alpha 2}) = 0.199$ vs. $|\Delta_{\alpha 2}| = 0.047$) parameters. The area of the fitted ellipse in the Poincaré plot could not be compared, as this parameter is not computed by KUBIOS. However, given that this parameters is computed based on a multiplication of the SD1 and SD2 parameters with π, and considering that these parameters have low relative differences, it can be assumed that it would also lie within acceptable limits.

IV. Conclusion

In the TD, the only concern lies within the computation of the TINN parameters, where significant differences could be identified. A thorough revision of the TINN algorithm is required for future versions of pyHRV. In the FD, no

TABLE V
EVALUATION RESULTS FOR THE TIME DOMAIN PARAMETERS (PYHRV VS. KUBIOS)

Time Domain Evaluation metrics				
Parameter	Unit	pyHRV	KUBIOS	$\lvert\Delta\rvert$
Mean NNI	$[ms]$	836.884 ± 143.429	836.884 ± 143.429	0.000
Min HR	$[bpm]$	57.754 ± 7.508	57.754 ± 7.508	0.000
Max HR	$[bpm]$	91.048 ± 16.342	91.048 ± 16.342	0.000
Mean HR	$[bpm]$	74.505 ± 13.545	73.953 ± 13.422	0.552
SDNN	$[ms]$	65.937 ± 26.291	65.937 ± 26.291	0.000
SDNN Index	$[ms]$	60.018 ± 16.965	60.338 ± 17.354	0.320
SDANN	$[ms]$	41.397 ± 18.281	42.113 ± 19.331	0.716
RMSSD	$[ms]$	42.485 ± 22.611	42.484 ± 22.611	0.001
NN50	$[-]$	58.120 ± 46.789	58.120 ± 46.789	0.000
pNN50	$[\%]$	17.187 ± 15.440	17.187 ± 15.440	0.000
Triangular Index	$[-]$	9.128 ± 3.503	9.470 ± 3.886	0.342
TINN	$[ms]$	110.313 ± 59.100	305.200 ± 104.915	194.887
Optimal		Acceptable		Divergent

Uncomparable parameters: NNI (Min, Max), ΔNNI parameters, NN20, pNN20, TINN N and M
Δ: Difference between pyHRV and KUBIOS mean value

TABLE VI
EVALUATION RESULTS FOR THE PARAMETERS EXTRACTED WITH WELCH'S METHOD (PYHRV VS. KUBIOS)

Welch's Method Evaluation Metrics				
Parameter	Unit	pyHRV	KUBIOS	$\lvert\Delta\rvert$
Peak Frequencies (VLF)	$[Hz]$	0.012 ± 0.009	0.012 ± 0.010	0.000
Peak Frequencies (LF)	$[Hz]$	0.071 ± 0.023	0.071 ± 0.023	0.000
Peak Frequencies (HF)	$[Hz]$	0.226 ± 0.066	0.226 ± 0.066	0.000
Absolute Powers (VLF)	$[ms^2]$	2361.001 ± 2968.427	2325.941 ± 2963.510	35.060
Absolute Powers (LF)	$[ms^2]$	1518.788 ± 1292.658	1510.333 ± 1313.584	8.455
Absolute Powers (HF)	$[ms^2]$	883.304 ± 1561.402	884.503 ± 1568.995	1.199
Log Powers (VLF)	$[log]$	7.192 ± 1.078	7.168 ± 1.088	0.024
Log Powers (LF)	$[log]$	6.954 ± 0.931	6.933 ± 0.953	0.021
Log Powers (HF)	$[log]$	6.115 ± 1.030	6.102 ± 1.050	0.013
LF/HF ratio	$[-]$	3.234 ± 2.241	3.239 ± 2.280	0.005
Total Power	$[ms^2]$	4763.093 ± 4214.227	4722.274 ± 4243.256	40.819
Relative Powers (VLF)	$[\%]$	45.752 ± 20.912	45.594 ± 21.027	0.158
Relative Powers (LF)	$[\%]$	34.643 ± 14.488	34.572 ± 14.607	0.071
Relative Powers (HF)	$[\%]$	19.605 ± 17.882	19.793 ± 18.033	0.188
Normalized Powers (LF)	$[-]$	67.638 ± 19.651	67.381 ± 19.915	0.257
Normalized Powers (HF)	$[-]$	32.362 ± 19.651	32.548 ± 19.894	0.186
Optimal		Acceptable		Divergent

Δ: Difference between pyHRV and KUBIOS mean value

immediate improvements are required for the PSD estimation using the Welch's method, given that the implemented algorithm did correctly identify the peak frequencies of the PSD which are a critical piece of information of the PSD. The remaining 16 extracted parameters show acceptable deviations from the KUBIOS HRV results. The PSD estimation using the Autoregressive method shows acceptable results for 6 out of 16 parameters, with the peak frequencies being also identified with acceptable deviations. The remaining parameters show significant differences in comparison with the KUBIOS HRV results, which can be caused by the selected AR estimation and parameter computation algorithms, in our case on the Yule-Walker algorithm. Higher power estimations can be observed over all computed NNI datasets compared to the KUBIOS HRV estimations. This does greatly alter the absolute powers of the different FB, from which the remaining parameters are extracted and ultimately distorted. For this method, the use of other algorithms or methods can be considered to be further research in order to improve the performance of the implemented algorithms. The Lomb-Scargle PSD estimation achieved 8 parameters lying within acceptable ranges, with the remaining 8 parameters showing significant differences. Other than with the previous PSD estimation methods, this method has shown high sensitivity in the selection of a suitable frequency resolution, which leads to inconsistency and significant changes in its capability of identifying the correct peak frequencies of the PSD. This method failed to accurately identify the peak frequencies of the LF band, critical information of the PSD while identifying these frequencies in the remaining bands with acceptable accuracy. However, this inconsistency can be observed over the results of all the remaining parameters, where the results of different parameters with significant differences can be observed over different FB. With the current version, the selection of a suitable frequency resolution lies

TABLE VII
EVALUATION RESULTS FOR THE PARAMETERS EXTRACTED USING THE AUTOREGRESSIVE METHOD (PYHRV VS. KUBIOS)

Autoregressive Method Evaluation Metrics				
Parameter	Unit	pyHRV	KUBIOS	$\lvert\Delta\rvert$
Peak Frequencies (VLF)	$[Hz]$	0.000 ± 0.000	0.012 ± 0.015	0.012
Peak Frequencies (LF)	$[Hz]$	0.040 ± 0.000	0.045 ± 0.018	0.005
Peak Frequencies (HF)	$[Hz]$	0.155 ± 0.021	0.206 ± 0.071	0.051
Absolute Powers (VLF)	$[ms^2]$	2940.517 ± 62.056	2294.918 ± 2666.343	645.599
Absolute Powers (LF)	$[ms^2]$	6411.443 ± 159.130	1552.893 ± 1137.171	4858.550
Absolute Powers (HF)	$[ms^2]$	12379.985 ± 572.088	867.342 ± 1445.146	11512.643
Log Powers (VLF)	$[log]$	7.986 ± 0.021	7.244 ± 0.978	0.742
Log Powers (LF)	$[log]$	8.766 ± 0.025	7.056 ± 0.847	1.710
Log Powers (HF)	$[log]$	9.423 ± 0.046	6.217 ± 0.903	3.206
LF/HF ratio	$[-]$	0.518 ± 0.015	3.086 ± 1.943	2.568
Total Power	$[ms^2]$	21731.945 ± 748.929	4716.715 ± 3671.023	17015.230
Relative Powers (VLF)	[%]	13.539 ± 0.299	44.497 ± 19.693	30.958
Relative Powers (LF)	[%]	29.515 ± 0.492	35.981 ± 14.719	6.466
Relative Powers (HF)	[%]	56.946 ± 0.766	19.482 ± 17.338	37.464
Normalized Powers (LF)	$[-]$	34.138 ± 0.670	67.972 ± 19.013	33.834
Normalized Powers (HF)	$[-]$	65.862 ± 0.670	31.949 ± 19.006	33.913
Optimal		Acceptable		Divergent

Δ: Difference between pyHRV and KUBIOS mean value

TABLE VIII
EVALUATION RESULTS FOR THE FREQUENCY DOMAIN PARAMETERS EXTRACTED USING THE LOMB-SCARGLE PSD (PYHRV VS. KUBIOS)

Lomb-Scargle Method Evaluation Metrics				
Parameter	Unit	pyHRV	KUBIOS	$\lvert\Delta\rvert$
Peak Frequencies (VLF)	$[Hz]$	0.017 ± 0.010	0.011 ± 0.009	0.006
Peak Frequencies (LF)	$[Hz]$	0.093 ± 0.032	0.065 ± 0.019	0.028
Peak Frequencies (HF)	$[Hz]$	0.290 ± 0.045	0.227 ± 0.066	0.063
Absolute Powers (VLF)	$[ms^2]$	216.763 ± 49.541	2709.483 ± 3700.016	2492.720
Absolute Powers (LF)	$[ms^2]$	608.678 ± 155.304	1508.002 ± 1109.925	899.324
Absolute Powers (HF)	$[ms^2]$	1463.945 ± 267.442	826.661 ± 1454.514	637.284
Log Powers (VLF)	$[log]$	5.354 ± 0.221	7.309 ± 1.038	1.955
Log Powers (LF)	$[log]$	6.380 ± 0.251	7.038 ± 0.806	0.658
Log Powers (HF)	$[log]$	7.271 ± 0.190	6.133 ± 0.911	1.138
LF/HF ratio	$[-]$	0.417 ± 0.077	3.316 ± 2.111	2.899
Total Power	$[ms^2]$	2289.386 ± 414.363	5045.996 ± 4600.503	2756.610
Relative Powers (VLF)	[%]	9.558 ± 1.777	46.894 ± 20.095	37.336
Relative Powers (LF)	[%]	26.422 ± 3.418	34.626 ± 13.803	8.204
Relative Powers (HF)	[%]	64.020 ± 3.843	18.432 ± 17.325	45.588
Normalized Powers (LF)	$[-]$	29.225 ± 3.820	69.102 ± 18.750	39.877
Normalized Powers (HF)	$[-]$	70.775 ± 3.820	30.809 ± 18.725	39.966
Optimal		Acceptable		Divergent

Δ: Difference between pyHRV and KUBIOS mean values

TABLE IX
EVALUATION RESULTS FOR THE NONLINEAR PARAMETERS (PYHRV VS. KUBIOS)

Nonlinear Parameters Evaluation Metrics				
Parameter	Unit	pyHRV	KUBIOS	$\lvert\Delta\rvert$
SD1	$[ms]$	30.041 ± 15.988	30.085 ± 16.015	0.044
SD2	$[ms]$	87.171 ± 35.907	87.294 ± 35.962	0.123
SD2/SD1	$[-]$	3.204 ± 1.306	3.204 ± 1.306	0.000
Sample Entropy	$[-]$	1.332 ± 0.365	1.331 ± 0.366	0.001
DFA - α_1	$[-]$	1.204 ± 0.264	1.182 ± 0.240	0.022
DFA - α_2	$[-]$	0.886 ± 0.215	0.839 ± 0.199	0.047
Optimal		Acceptable		Divergent

Uncomparable parameter: Area of fitted Ellipse (Poincaré)
Δ: Difference between pyHRV and KUBIOS mean value

therefore within the responsibility of the user and the intended application. The comparison of the NL parameters have shown positive results, with 6 out of 7 parameters showing acceptable results, and only marginal differences found when compared to

the KUBIOS results, and the SD2/SD1 ratio achieving optimal results. The functions of this module do not require any further optimization.

In addition to the results discussed in the previous section, which require further revision and improvements in some parameter computation functions of the TD and FD, the developed pyHRV toolbox will experience continued development with the implementation of new HRV methods and features in the future. For example, Björkander et al. [11] have proposed a new geometrical TD parameter named the Differential Index, which is composed based on the NNI histogram from which the geometrical parameters are derived. This parameter is computed from long-term Electrocardiography (ECG) recordings to visualize and extract parameters from NNI series with 10.000 and more intervals without the need for heavy computational algorithms. For the FD, German-Sallo [10] has investigated the use of a Wavelet Package Transform as an alternative method to compute the FFT based PSD estimation method, to measure the power variance and to identify dominant frequencies in a NNI series. His approach has shown good time-frequency resolution and comparable results for the LF/HF ratio, however, it still requires further investigation. It is also worth considering other features for deriving additional data series directly from the ECG signal, such as the ECG Derived Respiration (EDR) algorithm, where respiration data can be derived from the cardiac activity [14]. Although such features do not provide additional HRV parameters, they do provide other interesting measures that are often found in cardiac activity research [12], [13], [15].

REFERENCES

[1] Task Force of the European Society of Cardiology and The North American Society of Pacing and Electrophysiology, "Heart Rate Variability - Standards of measurement, physiological interpretation, and clinical use", Eur Heart J, vol. 17, pp. 354-381, Mar., 1996

[2] A. Goldberger, L. Amaral, L. Glass, J. Hausdorff, P. Ivanov, R. Mark, J. Mietus, G. Moody, C. Peng, H. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals", Circulation, vol. 101, no. 23, pp. 215-220, Jun., 2013

[3] G. Billman, "Heart rate variability - A historical perspective", Front. Physiol., vol. 2, no. 86, pp. 1-13, Nov., 2011

[4] T. Kuusela, "Methodological Aspects of Heart Rate Variability Analysis", in Heart Variability (HRV) Signal Analysis: Clinical Applications, Boca Raton, USA: CRC Press, 2013, ch. 2, pp. 9-42

[5] F. Shaffer, J. P. Ginsberg, "An Overview of Heart Rate Variability Metrics and Norms", Front Public Health, vol. 5, no. 258, Sep., pp. 1-17, Nov., 2017

[6] M. B. Tayel, E. I. Alsaba, "Poincaré Plot for Heart Rate Variability, In: Int J of Medical, Health, Biomedical, Bioengineering and Pharmaceutical Engineering, vol. 9, no. 9, pp. 708-711, 2015

[7] J. D. Scargle, "Studies in astronomical time series analysis II - Statistical aspects of spectral analysis of unevenly spaced data", Astrophys. J., vol. 263, no. 2, pp. 835-853, Jan., 1983

[8] P. Laguna, G. B. Moodz, R. G. Mark, "Power spectral density of unevenly sampled data by least-square analysis: performance and application to heart rate signals", IEEE Trans Biomed Eng, vol. 45, no. 6, pp. 698-715, Jun., 1998

[9] N. R. Lomb, "Least-Squares Frequency Analysis of Unequally Spaced Data", Astrophys. Space Sci., vol. 39, pp. 447-462, Feb., 1998

[10] Z. German-Sallo, "Wavelet based HRV analysis", Procedia Technology, Trgu Mure, Romania, vol. 12, pp. 105-111, 10-11 Oct., 2013

[11] I. Björkander, T. Kahan, M. Ericson, C. Held, L. Forslund, N. Rehnqvist, P. Hjemdahl, "Differential index, a novel graphical method for measurements of heart rate variability", Int J Cardiol, vol. 98, pp. 493-499, 2005

[12] J. Lorenz, P. Gomes, J. Prang, A. Jamshidirad, B. Tolg, "Subjective and Autonomic Responses to a 3D-Virtual Reality Game", Proc. ECGBL, Graz, Austria, 5-7 Oct. 2017, vol. 1, pp. 387-393

[13] B. Tolg, L. Montenegro, S. Steffens, E. Fiedler, L. Lackermeier, A. Leson, J. Voth, P. Gomes, J. Lorenz, "Re-Test Reliability of Subjective and Autonomic Reponses in a VR Performanca-Anxiety Test", Proc. ECGBL, Gratz, Austria, 5-7 Oct. 2017, vol. 1, pp. 671-676

[14] C. Massaroni, A. Nicoloò, D. Lo Presti, M. Sacchetti, S. Silvestri, E. Schena, "Contact-Based Methods for Measuring Respiratory Rate", Sensors, vol. 19, no. 4, pp. 1-47, Feb., 2019

[15] J. Gsior, J. Sacha,2 P. Jeleń, J. Zieliński, J. Przybylski, "Heart Rate and Respiratory Rate Influence on Heart Rate Variability Repeatability: Effects of the Correction for the Prevailing Heart RateEffect of respiration in heart rate variability (HRV) analysis.", Front Physiol., vol. 7, no. 356, pp. 1-11, Aug., 2016

[16] H. Young, D. Benton, "Heart-rate variability: a biomarker to study the influence of nutrition on physiological and psychological health?", Behav Pharmacol, vol. 29, no. 2, pp. 140-151, Apr., 2018

[17] D. Iakovakis, L.Hadjileontiadis, "Standing hypotension prediction based on smartwatch heart rate variability data: a novel approach", Proc MobileHCI, Florence, Italy, vol. 1, pp. 1109-1112, 6-9 Sep. 2016

[18] F. Shaffer, J. P. Ginsberg, "An Overview of Heart Rate Variability Metrics and Norms", Front Physiol., vol. 5, no. 258, pp. 1-17, Sep., 2017

[19] J. Naranjo, B. De la Cruz, E. Sarabia, M. De Hoyo, S. Dominguez-Cobo, "Heart Rate Variability - a Follow-up in Elite Soccer Players Throughout the Season", Int J Sports Med, vol. 36, no. 11, pp. 881-886, Nov., 2015

[20] M. Tarvainen, J. Niskanen, J. Lipponen, P. Karjalainen, "Kubios HRV - Heart rate variability analysis software", Comp Meth Programs Biomed, vol. 113, no. 1, pp. 210-220, Jan., 2014

A Solution of Concurrent Stack on PSTM

Marko Popovic, Branislav Kordic, Miroslav Popovic, Ilija Basicevic

***Abstract*—Nowadays developing concurrent data structures based on foundations of software transactional memory (STM), a.k.a. transactional data structures, is an area of intensive research. In this paper, we developed the concurrent stack on Python STM (CS-PSTM), and verified it using unit and system testing. For system testing we developed the five application workloads, namely: one producer - one consumer, nP producers, nP consumers, nP producers and nC consumers, and nP processes. The CS-PSTM successfully passed all of the unit and the system tests. We also used system tests to estimate CS-PSTM performance. Interestingly, and not unexpectedly, CS-PSTM provides better performance when used by more concurrent processes.**

***Index Terms*— multicore systems; parallel programming; Python; transactional memories; concurrent data structures; concurrent stacks.**

I. Introduction

SINCE the emergence of multicore processors, many efforts have been dedicated to building concurrent data structure libraries (CDSLs) [1, 2], which are a.k.a. thread-safe libraries. Efficient CDSL implementations are available for many programming languages, as C++, Java, C#/.NET, TBB, Python, etc., and are widely adopted [3]. These state-of-the-art CDSLs are implemented using locks and/or atomic instructions such as CAS.

In 1993 Herlihy and Moss introduced the idea of the *transactional memory* (TM) [4], in order to overcome difficulties related to lower-level abstraction based on locks. The concept of TM has the following two main advantages over locks: (1) TM makes parallel programming easier by supporting abstraction and composition, and (2) TM based parallel programs tend to have better performance than traditional *lock* based programs because individual transactions are executed *optimistically* [5].

Software TMs (STMs) enabled research and development of new type of concurrent data structures based on STMs. One notable example of this is development of lock-free (concurrent) data structures using STM in Haskell [6]. Similarly, we in our previous research followed this path to design and implement a concurrent list [7] and concurrent queue [8] based on particular software TM (STM) in Python, named PSTM [9]. Nowadays developing concurrent data structures based on foundations of software transactional memory (STM), a.k.a. transactional data structures, is an area of intensive research [10, 11].

Along these lines of research, this paper makes its main contribution by presenting the solution for a concurrent stack based on PSTM, which we briefly refer to as the concurrent stack on PSTM (CS-PSTM). CS-PSTM is based on the unbounded lock-free stack using CAS [1] (pp. 245, 247).

We verified CS-PSTM using unit and system testing. For unit testing we developed nine test cases, whereas for system testing we developed five application workloads. These application workloads are the following: (1) one producer - one consumer workload, (2) *nP* producers workload, (3) *nP* consumers workload, (4) *nP* producers and *nC* consumers workload, and (5) *nP* processes workload. The CS-PSTM successfully passed all of the unit and the system tests.

We also used system tests to estimate CS-PSTM performance. Interestingly, and not unexpectedly, CS-PSTM provides better performance when used by more concurrent processes, but of course not above the number of available cores (which was four in this research).

One important advantage of the CS-PSTM design is that it may be easily adapted to be used in distributed settings. The key idea in the design of CS-PSTM that enable this adaptation is that stack elements are linked together using t-var identifications, which are strings.

The rest of the paper is organized as follows. Section II introduces the sequential stack, Section III presents the concurrent stack on PSTM (CS-PSTM), Section IV presents the unit and system testing, and Section V presents the final conclusions of this paper.

II. Sequential Stack

In this section we briefly introduce the definition of a sequential stack, which originates from [8] (pp. 245). A *sequential stack Stack<T>* is an ordered sequence of items of a type *T*, which provides the following two methods: (i) the method *push*(*x*) puts an item *x* at the *top* of the queue, (ii) the method *pop*() removes and returns the item from the *top* of the stack.

A sequential stack is implemented as a linked list of nodes. A *node* is a data structure, which comprises the following two fields: (i) the field *item* is the actual item of interest (the variable *x* mentioned above), and (ii) the field *next* is a link to the next node in the list, which is null if that is the last node in the list.

The list consists of regular nodes, where each node contains

Marko Popovic, University of Novi Sad, Faculty of Technical Sciences, Novi Sad, Trg Dositeja Obradovica 6, 21000 Novi Sad, Serbia (e-mail: marko.popovic@rt-rk.uns.ac.rs).

Branislav Kordic, University of Novi Sad, Faculty of Technical Sciences, Novi Sad, Trg Dositeja Obradovica 6, 21000 Novi Sad, Serbia (e-mail: branislav.kordic@rt-rk.uns.ac.rs).

Miroslav Popovic, University of Novi Sad, Faculty of Technical Sciences, Novi Sad, Trg Dositeja Obradovica 6, 21000 Novi Sad, Serbia (e-mail: miroslav.popovic@rt-rk.uns.ac.rs).

Ilija Basicevic, University of Novi Sad, Faculty of Technical Sciences, Novi Sad, Trg Dositeja Obradovica 6, 21000 Novi Sad, Serbia (e-mail: ilija.basicevic@rt-rk.uns.ac.rs).

a stack's item, and the link to the next node. Unlike similar concurrent data structures, such as concurrent lists and queues, there is no need for some special nodes (e.g. sentinels). The list that implements the concurrent stack is organized by a single pointer, *top*, which points to the node at the top of the stack.

Initially, the stack is empty, thus *top* contains the null pointer, see Fig. 1a. The method *push*(*a*) adds the item *a* at the *top* of the stack, see Fig. 1b. The method *pop*() removes and returns the item *a* from the top of the stack, if the stack is not empty; otherwise it indicates that the stack is empty. For example, when the method *pop*() is called on the stack shown in Fig. 1b, it sets *top* to null pointer, and returns the item *a*. The result of this *pop*() is the same as the initial stack shown in Fig. 1a.

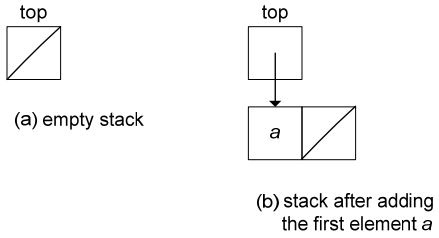


Fig. 1. Stack implemented as a linked-list.

The next section presents the CS-PSTM, which is linearizable to a sequential stack.

III. Concurrent Stack on PSTM

The concurrent queue on PSTM is based on the definition unbounded lock-free stack [8] (pp. 245, 247) and on the previously developed PSTM [3].

A. Short Overview of PSTM

The PSTM is client-server architecture [9]. The client consists of the STM-based application and the set of API functions (defined in the PSTM module) that act as a client's proxy. The PSTM server serves requests created by PSTM API functions and sent over the queue to the server. The queue connects multiple clients (transactions) with a single server. Replies from the server to clients are sent over pipes. A pipe connects the server with a single client.

The PSTM server is the owner of the system *dictionary* that contains current entries for all the t-variables (t-vars) used by transactions. The *dictionary item* is a tuple (*key*, *ver*, *val*), where *key* is the t-var key (aka t-var identification, ID), *ver* is the current t-var version, and the *val* is the current t-var value. The PSTM server serves incoming requests atomically.

The main PSTM API functions are the following:

1. *addVars*(*q*, *keys*)
2. *getVars*(*q*, *keys*)
3. *putVars*(*q*, *vars*)
4. *commitVars*(*q*, *read_write*)

Where *q* is a queue between clients and PSTM server, *keys* is a list of t-variable's IDs, *vars* is a list of dictionary items, and *read_write* is the list <*read*, *write*> where *read* and *write* are lists of items that a transaction reads and writes (updates), respectively. The function *getVars* returns the list of tuples (*exists*, (*ver*, *val*)), where *exists* is True if t-var is found, or False otherwise. Other functions return 'no' if an error occurs, or 'yes' otherwise.

A transaction typically gets its local copies of t-vars using *getVars*, does some processing with them, and finally updates some of them using *commitVars*. New t-vars are added and initialized by *addVars* and *putVars*, respectively.

B. Concept and Data Structures and API Functions

The concurrent stack on PSTM is created by putting the nodes into t-vars and linking them using t-var identifications. The advantage of this solution is that it may be easily adapted for a distributed setting, because t-vars need not be located in the same computer. Next, we describe data structures that are used to implement the concurrent stack on PSTM.

The stack on PSTM and its nodes are represented as Python named tuples *StackOnPSTM* and *Node*, respectively. The elements of the named tuple *StackOnPSTM* are: the name of the stack (*stackName*) and the *top* of the stack t-var identification (*topTvarId*)). The elements of the named tuple *Node* are: the item (*item*) and the t-var identification of the next element in the stack on PSTM (*next*).

Two types of t-vars are used to implement the stack on PSTM. The t-var of the first type contains the t-var identification of the top of the t-vars stack, whereas the t-vars of the second type contain stack's nodes. The t-var *top* is the t-var of the first type, whereas the nodes containing the items are the t-vars of the second type.

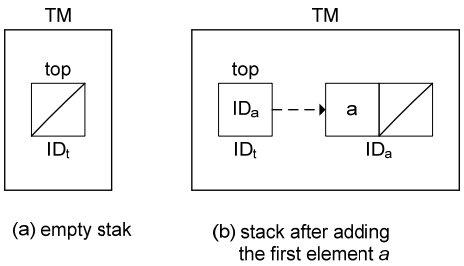


Fig. 2. Stack on PSTM.

Fig 2a shows the empty stack, which consists of the *top* placed into a t-var (in TM) with the identification ID_t. Since the stack is empty, *top* contains the null pointer. Fig 2b shows the stack after the first item *a* is pushed at its top. The item *a* is stored in the t-var with the identification ID_a, and the *top* is updated to point to the node with the item *a*.

The API functions are the following: *createStackOnPSTM*(*q*, *stackName*), *push*(*q*, *S*, *item*), and *pop*(*q*, *S*).

C. The function createStackOnPSTM

At the beggining, the function sets *top* t-var identification, *topTvarId*, by concatenating the stack name parameter, *stackName*, with the string 'topTvarId'.

Then, it uses the PSTM API function *addVars* to add the t-var *topTvarId* to PSTM transactional memory.

After that, it uses the PSTM API function *putVars* to set the identification of the t-var at the top of the stack, to its initial value null pointer, which is encoded as the empty string ''.

At the end, the function returns the instance of the tuple *StackOnPSTM* representing the new stack, which is set to the value (*stackName*, *topTvarId*).

D. The function push

The pseudocode for the function *push* is given in Algorithm 1. At the beginning, the function calls the PSTM API function *createVar* in order to create the new t-var for the new item, and stores the identification of the new t-var in the variable *nodeTvarId* (lines 2 – 3). Then in the endless loop, it calls the supplementary function *tryPush*, which in its turn tries to effectively perform the required push operation (lines 4 – 6). Once the function *tryPush* succeeds to push item on the stack, it returns **True**, and the function *push* returns (line 6).

The function *tryPush* begins by creating the empty read and write lists of t-var PSTM items, *readList* and *writeList*, respectively (lines 8-9). A *t-var PSTM item* is a tuple comprising t-var's identification, its version, and its value [3]. After that, it initializes the read-write list (*readWriteList*) with its elements – *readList* and *writeList* (line 10).

Then it sets the variable *topid* to the identification of the t-var *top* (line 12), and calls the function *getTvarVerAndValue* to get the value (*top*) and the PSTM item (*top_item*) for the t-var *top* (line 13). Next, it unpacks the tuple *top_item* to the variables *tid*, *tver*, and *tval*, holding the *top*'s ID, version, and value, respectively (line 14).

After that, the function creates a new node (*node*) using the argument *item* and links the new node with *top* (line 16). Then, it creates the t-var item for the new node (*node_item*) consisting of the *nodeTvarId*, version value 0, and *node*, and appends it to the write list (lines 18-19).

In order to update the t-var *top*, such that it points to the new node, the function creates the new PSTM item (*new_top_item*) from variables *tid*, *tver*, and *nodeTvarId* (line 21), and then appends *new_top_item* to *writeList* (line 22).

Finally, the function calls the API function *commitVars* to update t-vars as specified by *readWriteList*. If the commit is successful, the function returns **True**; otherwise, it returns **False** (lines 24 – 28).

Algorithm 1. The function *push*

```
01: push(q, S, item)
02:   ret ← createVar(q)
03:   nodeTvarId ← ret[0]
04:   while True
05:     if tryPush(q, S, item, nodeTvarId) then
06:       return

07: tryPush (q, S, item, nodeTvarId)
08:   readList ← []              // initialize read list
09:   writeList ← []             // initialize write list
10:   readWriteList ← [readList, writeList]
11:   // Get the top
12:   topid ← S.tvarIdTop
13:   top, top_item ← getTvarVerAndValue(q, topid)
14:   (tid, tver, tval) ←  top_item
15:   // Create the new node and link with current top
16:   node ← Node(item, top)
17:   // Create node_item and add it to write list
18:   node_item ← (nodeTvarId, 0, node)
19:   writeList.append(node_item)
20:   // Create new_top_item and add it to write list
21:   new_top_item ← (tid, tver, nodeTvarId)
22:   writeList.append(new_top_item)
23:   // Do the Commit
24:   ret ← commitVars(q, readWriteList)
25:   if ret = ['yes'] then
26:     return True
27:   else
28:     return False
```

E. The function pop

The pseudocode for the function *pop* is given in Algorithm 2. The function *pop*, in the endless loop, calls the supplementary function *tryPop*, which in its turn tries to effectively perform the required pop operation (lines 2 – 5). If *tryPop* finds that stack is empty, it returns None. Otherwise, it returns the popped item (*retItem*), which in its turn the function *pop* returns to its caller.

The function *tryPop* begins by creating the empty read and write lists of t-var PSTM items, *readList* and *writeList*, respectively (lines 7 – 8). After that, it initializes the read-write list (*readWriteList*) with its elements – *readList* and *writeList* (line 9).

Then it sets the variable *topid* to the identification of the t-var *top* (line 11), and calls the function *getTvarVerAndValue* to get the value (*top*) and the PSTM item (*top_item*) for the t-var *top* (line 12).

It then checks whether the stack is empty, and if yes, it returns the value None (lines 14-15). If not, it unpacks the tuple *top_item* to the variables *tid*, *tver*, and *tval*, holding the *top*'s ID, version, and value, respectively (line 17).

Then, it gets the *node* and the *node_item* from the top of the stack (line 19). It then unpacks the triple *node_item* to the variables *nodeid*, *nodever*, *nodeval*, respectively (line 20). Then, it tries to set the variable *newTop* to the value of the field *next* of the node at top of the stack (*node*). This is done within the **try-except** construct, because meanwhile *node* may be deleted by another process, in which case the function returns the value None (lines 22-23).

If *newTop* is set successfully, the function adds the *node_item* to the *readList* (line 26).

In order to update the t-var *top*, such that it points to the node after the node that is going to be popped, the function creates the new top PSTM item (*new_top_item*) from variables *tid*, *tver*, and *newTop* (line 28), and then appends *new_top_item* to *writeList* (line 29).

Finally, the function calls the API function *commitVars* to update t-vars as specified by *readWriteList*. If the commit is successful, the function returns *item* (from *node*); otherwise, it returns the value None (lines 31 – 37).

Algorithm 2. The function *pop*

```
01: pop(q, S)
02:   while True
03:     retItem ← tryPop(q, S)
04:     if retItem ≠ None then
05:       return retItem

06: tryPop(q, S)
07:   readList ← []              // initialize read list
08:   writeList ← []             // initialize write list
09:   readWriteList ← [readList, writeList]
10:   // Get the top
11:   topid ← S.tvarIdTop
12:   top, top_item ← getTvarVerAndValue(q, topid)
13:   // If stack is empty, return item None
14:   if top = '':
15:     return None
16:   // Unpack the top_item
17:   (tid, tver, tval) ← top_item
18:   // Get node and node_item from top of the stack
19:   node, node_item ← getTvarVerAndValue(q, top)
20:   (nodeid, nodever, nodeval) ← node_item
21:   // Set newTop to node.next
22:   try newTop ← node.next
23:   except return None
24:   // Prepare read-write list for the commit
25:   // Add node_item to readList
26:   readList.append(node_item)
27:   // Create new item for top and add it to writeList
28:   new_top_item ← (tid, tver, newTop)
29:   writeList.append(new_top_item)
30:   // Do the Commit
31:   ret ← commitVars(q, readWriteList)
32:   if ret = ['yes'] then
33:     // Remove t-var containing node
34:     removeVars(q, [nodeid])
35:     return node.item
36:   else
37:     return None
```

IV. Unit and System Testing

We verified CS-PSTM using unit and system testing. For unit testing we developed nine test cases, whereas for system testing we developed five application workloads. The CS-PSTM successfully passed all of these unit and system tests.

For unit testing we used the following test cases: (1) create single CS-PSTM stack, (2) create multiple CS-PSTM stacks (100 stacks), (3) push a single item on the CS-PSTM stack, (4) enqueue multiple (100) items on the CS-PSTM stack and check its length using the internal function *length*, (5) try to pop a non-existing item from the empty CS-PSTM stack, (6) pop an existing item from the CS-PSTM stack with one item, (7) create a child process and check whether the child process successfully pushed a single item (by checking pop from the stack), (8) create a child process, push a single item on the stack, and check whether the child process successfully dequeues that item, (9) push multiple (100) items on the CS-PSTM stack, pop the first half of the items (50), and check whether the second half of the items (50) remains on the CS-PSTM stack using the internal function *length*.

For system testing we used the following application workloads (which all start from the empty CS-PSTM stack, except NC which starts from the stack with *nItem* items): (1) the one producer - one consumer (OPOC), where producer and consumer exchange *nItems* items over the stack in the two possible modes, namely the sequential mode (where producer and consumer are executed sequentially) and the parallel mode (where producer and consumer are executed in parallel), (2) *nP* producers (NP), where each producer pushes *nItems*/*nP* items onto the stack, (3) *nP* consumers (NC), where each consumer consumes *nItems*/*nP* items from the stack, (4) *nP* producers and *nC* consumers (NPNC), where each producer produces *nItems*/*nP* items onto the stack, and each consumer consumes *nItems*/*nC* items from the stack, and (5) *nP* processes (NPP), where each process produces *nItems* onto the stack and then consumes *nItems* from the stack.

TABLE I. Overview of System Tests

Application Workload		Exe. Time [s]
OPOC	Sequential	0.3619
	Parallel	0.2464
NP	*nP* = 2	0.176
	nP = 3	0.1686
	nP = 4	0.1671
NC	*nP* = 2	0.2152
	nP = 3	0.2189
	nP = 4	0.2137
NPNC	(*nP*, *nC*) = (1, 2)	0.2722
	(*nP*, *nC*) = (1, 3)	0.2695
	(*nP*, *nC*) = (2, 2)	0.2718
	(*nP*, *nC*) = (2, 1)	0.2907
	(*nP*, *nC*) = (3, 1)	0.2934
NPP	*nP* = 2	0.3066
	nP = 3	0.2827
	nP = 4	0.2583

We conducted system testing using the Intel Core i7-8700 CPU @ 3.2 GHz six core processor with 16GB of shared memory. Our goal was to expose the CS-PSTM to maximal concurrency by running child processes within the workloads on separate processor cores. Since at least one of the cores must be reserved for the operating system and one for the PSTM server, the number of child processes in the workloads had to be up to 4 (we used 1, 2, 3 or 4).

Next, we introduce the application parameters. The parameter *nItems* is the number of items to be transferred over the stack, and is set to 99 (in OPOC) or 120 (otherwise, to make it divisible by the number of child processes). For the applications NP, NC, and NPP, the parameter *nP* is the total number of child processes, and is set to 2, 3 or 4.

For the application NPNC, the parameter *nP* is a number of

producers, whereas *nC* is a number of consumers, and these two parameters are set together as a pair (*nP*, *nC*) to pairs of values (1, 2), (1, 3), (2, 2), (2, 1) or (3, 1), because there were 4 cores available so the sum of *nP* and *nC* must be equal to 4.

The overview of system tests is given in Tab. I. The primary purpose of these tests was to verify the CS-PSTM. Additionally, we used them to observe CS-PSTM performance by measuring the application execution times. The last column in Tab. I contains the execution times averaged over 10 measurements. The performance is as expected. CS-PSTM behaves as a typical concurrent data structure. Generally, as the number of child processes within an application increases, the application execution time (or its makespan) decreases, i.e. the execution speed up increases.

For example, the execution time for the application OPOC in the parallel mode is less than in the sequential mode. Similarly, the execution time for the applications NP and NPP decrease as the value of the parameter *nP* increases. There is a single exception to this trend, which was observed for the application NC and for *nP* = 3 (the execution time for *nP* = 3 is greater than the execution time for *nP* = 2), which may be explained by imperfectness of experimental measurements.

The execution times for the application NC are slightly greater than for the application NP, which is most probably caused by exception handling within the function *pop*. Similarly, the execution times for the application NPNC shows that execution time is smaller (i.e. execution speed is greater) when there are more consumers and a single producer than when there are more producers and a single consumer, which implies that the sequential overhead for the function *pop* is greater than for the function *push*.

After highlighting the above mentioned general trends, we explain execution time results more specifically. We start with results for the applications NP, NC, and NPP. As can be seen from Table I, as the number of child processes (i.e. the parameter *nP*) increases from 2 to 3, and to 4:

- The execution time for the application NP expectedly decreases from 0.176s to 0.1686s, and to 0.1671s.
- The execution time for the application NC (as already explained above) unexpectedly increases from 0.2152s to 0.2189s and then expectedly decreases to 0.2137s.
- The execution time for the application NPP expectedly decreases from 0.3066s to 0.2827s, and to 0.2583s.

Next we explain the results for the application NPNC. We analyze the following three cases:

- Case 1: *nP* fixed to 1 and *nC* increases from 2 to 3. The execution time expectedly decreases from 0.2722s to 0.2695s.
- Case 2: *nC* fixed to 1 and *nP* increases from 2 to 3. The execution time somewhat unexpectedly increases from 0.2907s to 0.2934s, which may be because the total load balancing among cores becomes worst when more faster producers are combined with one slower consumer, or because of imperfectness of experimental measurements.
- Case 3: *nP* = 2 and *nC* = 2. The execution time for this case is not exactly, but somewhere in between, the cases 1 and 2. This result is as expected, because NPNC workloads with more producers (case 2) are heavier than the NPNC workloads with more consumers (case 1), and the case 3 has equal number of producers and consumers, so its NPNC workload is somewhere in between.

At this point it seems appropriate to comment the speed-up of the proposed solution. The application workload execution time decreases as the number of child processes (and thus the number of used cores) increases, but this decrease is really rather small. As can be seen from Table I, the execution time on 2, 3, and 4 cores is almost identical, so it seems that there is no performance gain when using more than 2 cores. The reason for this is that child processes just use the concurrent stack to push and pull items to/from the stack, but they do not perform any processing of data carried by the items, thus there is not enough work for child processes to share. Clearly, the performance gain would become significant, and expectedly proportional to the number of cores, if the child processes would perform more heavy data processing. We might prove this expectation by conducting appropriate experiments in our future work.

V. Conclusion

In this paper, we developed the concurrent stack on Python STM (CS-PSTM), and verified it using unit and system testing. For unit testing we developed nine test cases, whereas for system testing we developed the five application workloads (briefly referred to as OPOC, NP, NC, NPNC, and NPP). The CS-PSTM successfully passed all of the unit and the system tests.

We used the system tests also to estimate the CS-PSTM performance by measuring their execution times. The experimental results indicate that CS-PSTM provides better performance when used by more concurrent processes. Another advantage of CS-PSTM is that it may be easily adapted to be used in distributed settings.

For our future work, we plan to develop a Distributed PSTM (DPSTM) and to adapt CS-PSTM accordingly. We also plan to research on other concurrent data structures in this context.

Acknowledgment

This work was partially supported by the Ministry of Education, Science and Technology Development of Republic of Serbia under Grant III-44009.

References

[1] M. Herlihy and N. Shavit, *The art of multiprocessor programming*, 2nd edition, Morgan Kaufmann, 2008.

[2] D. Lea, *Concurrent programming in Java: design principles and patterns*, 2nd edition, Addison-Wesley Professional, 2000

[3] O. Shacham, N. G. Bronson, A. Aiken, M. Sagiv, M. T. Vechev, and E. Yahav, "Testing atomicity of composed concurrent operations", Proc. of the 2011 ACM international conference on Object Oriented Programming Systems Languages and Applications, pp 51–64, 2011.

[4] M. Herlihy and J.E.B. Moss, “Transactional memory: architectural support for lockfree data structures,” Proc. of the 20th Annual International Symposium on Computer Architecture, ACM, New York, NY, pp. 289–300, 1993.

[5] T. Harris, J.R. Larus, and R. Rajwar, *Transactional Memory*, 2nd edition, Morgan and Claypool, 2010.

[6] A. Discolo, T. Harris, S. Marlow, S.P. Jones, and S. Singh, “Lock Free Data Structures using STM in Haskell”, Proc. of the 8th international conference on Functional and Logic Programming, pp. 65-80, 2006.

[7] M. Popovic, B. Kordic, M. Popovic, and I. Basicevic, “A Solution of Concurrent List on PSTM,” Proc. 5th IcETRAN, Article RTI2.1, pp. 1-6, 2018.

[8] M. Popovic, B. Kordic, M. Popovic, and I. Basicevic, “A Solution of Concurrent Queue on PSTM,” Proc. 25th IEEE Telecommunications Forum, Article RTI2.1, pp. 735-738, 2018.

[9] M. Popovic, and B. Kordic, “PSTM: Python Software Transactional Memory,” Proc. 22nd IEEE Telecommunications Forum, pp. 1106-1109, 2014.

[10] A Spiegelman, G. Golan-Gueta, and I. Keidar, “Transactional Data Structure Libraries”, Proc. of the 37th ACM SIGPLAN Conference on Programming Language Design and Implementation, pp. 682-696, 2016.

[11] T.D. Dickerson, P. Gazzillo, M. Herlihy, and E. Koskinen, “Proust: A Design Space for Highly-Concurrent Transactional Data Structures”, Cornell University Library, arXiv:1702.04866v2 [cs.DC] 26 Jun 2017.

Implementation and evaluation of video conferencing system on public cloud

Vladimir M. Ciric, Oliver M. Vojinovic, Ivan Z. Milentijevic

Abstract— **With constant growth of capacity and bandwidth of computer networks, resource demanding network applications that emerged as concepts a long time ago got their chance to gain a worldwide popularity and widespread usage. It is not the question anymore whether we can implement such applications, but rather how easy they are to implement, and how much their usage costs. In cloud-centric world of today this becomes the matter of the amount of required resources. The aim of this paper is to present the resource requirements of a video conferencing system based on open source BigBlueButton platform, while it is implemented on the Amazon AWS public cloud. The system will be implemented as the part of blended learning system. In order to determine the system limits and to give recommendations for implementation of similar systems, the results of the stress test will be given.**

Index Terms—**Video conferencing, blended learning, cloud applications.**

I. Introduction

The roots of video telephone technology can be traced back to the late 1920s and can be found with the AT&T company Bell Labs and John Logie Baird, who experimented with video phones in 1927 [1]. Videotelephony was developed in parallel with voice telephone systems from the mid-to-late 20th century. A lot of effort was put in the field by Bell Labs during the 1950s and 1960s, which lead to AT&T's Picturephone. However, the market didn't respond well, and the service was withdrawn [2].

A lot of researches and companies continued development throughout the 1980s and 1990s, which led to a number of videoconferencing systems. The systems evolved from proprietary equipment to standard technologies that were available to the general public at reasonable cost. Early attempts have failed mainly due to the cost and the quality of both equipment and resources [2]. However, videotelephony become a practical technology for regular use with the introduction of powerful video codecs in the late 20th century, combined with high-speed Internet service [3].

The constant bandwidth growth resulted with a shift from transmission of a few low-quality images per second to high-definition real-time video conferencing [4]. The price of the video conference systems got significantly lower, resulting with an adoption of such systems in different applications. One of such applications is in education, namely blended learning [5]. In a nutshell, blended learning requires the presence of both a teacher and a student, where the student has control to a certain extent over time, place, and path [6].

In recent years blended learning is gaining a lot of attention due to the growing demand for distance learning and life-long leaning [6]. The popularity emerges from the fact that it combines online digital media with traditional classroom methods, elevating the learning experience. Blended learning, armed with video conferencing, exerts good sides of both traditional and distance learning. Video conferencing becomes a common classroom activity in the online learning stage of blended learning [7].

There are a lot of educational institutions that own blended learning platforms, and even more that don't. The era of cloud computing offers a possibility to institutions which don't own such platforms to utilize public cloud resources for courses delivery. Cloud providers offers different sets of resources in the form of virtual machines, called instances, starting with low price instances with modest resources at the cost of about $0.02 per hour, up to very powerful instances which can cost more than $10 per hour [8]. Observed on a monthly basis, the price may vary from $15-$20 per month, to a few hundreds, or even thousands of dollars. Speaking about the cost difference of up to 500 times, the choice of an appropriate instance for the particular application becomes an important question.

The aim of this paper is to present the resource requirements of a video conferencing system, regarding the number of users, based on open source BigBlueButton platform, while it is implemented on the Amazon AWS public cloud. The system will be implemented as the part of blended learning system. In order to determine the system limits and give recommendations for implementation of similar systems, the results of the stress test will be given. The measured parameters will include CPU, memory and network throughput. The system will be tested on three different AWS instance types, simulating up to 400 users. Recommended user number limits will be given for each of the tested instance types.

This paper is organized as follows. Section 2 presents the proposed implementation of blended learning system, based on video conferencing platform. Section 3 is devoted to the evaluation of the system. The concluding remarks are given in Section 4.

II. Blended learning support system implementation on public AWS cloud

In order to enable students to choose their learning path and create an environment where they can schedule the

Vladimir M. Ćirić is with the University of Niš, Faculty of Electronic Engineering, Aleksandra Medvedeva 14, 18000 Niš, Serbia (e-mail: vladimir.ciric@ elfak.ni.ac.rs).

Oliver M. Vojinović is with the University of Niš, Faculty of Electronic Engineering, Aleksandra Medvedeva 14, 18000 Niš, Serbia (e-mail: oliver.vojinovic@elfak.ni.ac.rs).

Ivan Z. Milentijević is with the University of Niš, Faculty of Electronic Engineering, Aleksandra Medvedeva 14, 18000 Niš, Serbia (e-mail: ivan.milentijevic@elfak.ni.ac.rs).

learning time on their own, blended learning support systems usually include several components [5,7]. These components should make learning material available online, should provide quizzes and tests, and not mandatory, but preferably enable two-way communication between teacher and students [5]. There are different approaches in the implementation of two-way teacher-students communication. E-mails are the oldest and the simplest solution. Nowadays, the most common asynchronous communication tools are discussion forums. Simple, but effective synchronous tool that can be found in such systems is chat. The most powerful two-way communication tool that is gaining more and more popularity is video conference [7]. Fig. 1 shows the common blended learning system components and their relation to teacher and students.

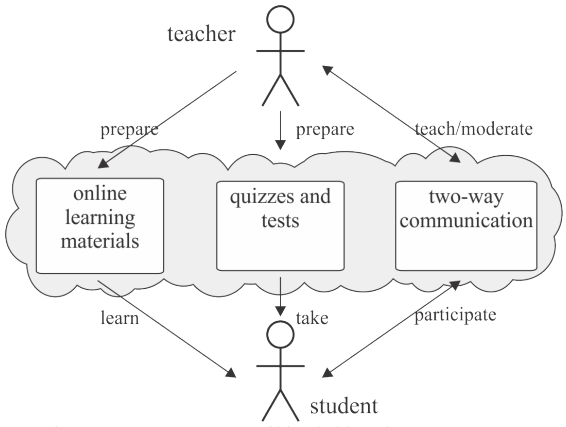


Fig. 1. The common components of blended learning systems

In order to support blended learning paradigm, the video conferencing system, if included, should provide multipoint conference, as well as video recording and streaming, with aim to enable students to schedule their learning time on their own. To provide the components from Fig. 1 two different platforms are required: distance learning platform, and video conferencing platform. Modern distance learning platforms always support both online learning materials delivery and testing of the students.

Both distance learning and video conferencing platforms are common in lecture delivery today, and there are a lot of available options for choosing the systems [9]. Seeking for a low-cost open-source and general solution, we chose the Moodle[1] as the primary learning management system, while for video conferencing we chose BigBlueButton[2] (BBB). The Moodle is chosen as the primary platform due to its ability to integrate BBB through its plug-in system, giving the students impression of accessing one platform only.

The BBB has the following features:

- it does not require any special software on the client side other than standard web-browser;
- good support and integration with Moodle;
- ability of sessions recording and streaming;
- up to 3 simultaneous videos, and 1 audio stream.

Video and audio streams that BBB can simultaneously have are: 1 video and 1 audio stream for the instructor's camera and microphone, 1 video stream for the presentation (PowerPoint, PDF, etc.), and 1 video stream for desktop share. Video stream for the presentation can be additionally used as a whiteboard. In addition, desktop sharing feature of BBB is available to create a live-stream of chosen desktop region to students.

Both Moodle and BBB systems are suitable for either private or public cloud implementation. The Amazon AWS provides a template that contains the software configuration (operating system, application server, and applications), called AMI, required to easily launch the instance with Moodle or BigBlueButton [10]. The chosen building blocks and their relations are shown in Fig. 2.

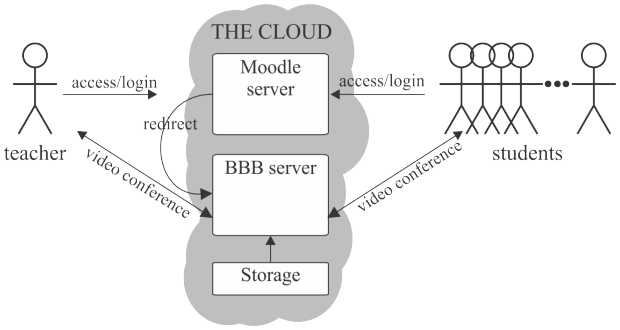


Fig. 2. The blended learning system building blocks

The Moodle can provide both online learning materials distribution, and testing the students' knowledge. Furthermore, it has tools for two-way communication with the students such as forums and chats. The BBB is an open-source project which can be easily included as a video conferencing extension of the Moodle, using available plugins for Moodle. Students are logged in using Moodle, and then, transparently, without additional login redirected to BBB video conferencing (Fig. 2). In order to store the video conference recordings additional storage is required (Fig. 2).

Fig. 3 shows the example of Moodle page prepared by the teacher. The figure shows sections in Moodle page, each devoted to one topic. In our implementation a section contains: learning objectives and learning materials, virtual classroom (given as a link to BBB Moodle plugin), additional learning materials, and online test.

Fig. 4 presents a screenshot of a web browser during a lecture in the virtual BBB classroom, which can be seen by following BBB links (the "b" icons on the Moodle page from Fig. 3). The left column of the screen contains the list of participants and the instructor video camera stream. The central part is devoted to the drawing canvas, while the right part of the screen contains the screen share showing the network simulator.

[1] Moodle is a free and open-source learning management system (LMS), originally developed by Martin Dougiamas, which is distributed under the GNU General Public License. URL: https://moodle.org/

[2] BigBlueButton is an open source video conferencing system, based on GNU/Linux operating system. URL: https://bigbluebutton.org/

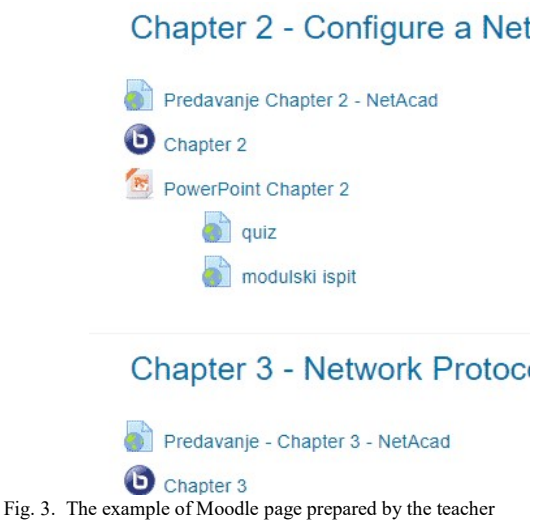


Fig. 3. The example of Moodle page prepared by the teacher

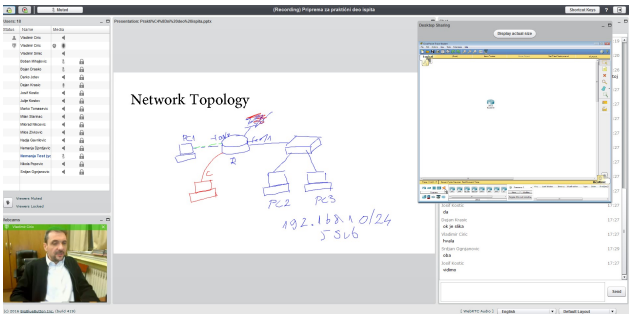


Fig. 4. Screenshot of a web browser showing virtual BBB classroom

III. THE RESULTS OF THE SYSTEM STRESS TEST

The system from Fig. 2 is implemented on Amazon AWS cloud. The resources for learning materials distribution are not critical, thus the Moodle is implemented using *t2.small* instance (1 vCPU, 2 GB RAM). In order to observe the system scaling capabilities and determine limits for different instance types regarding the number of users, we tested video conferencing part of the system using three different instances: *t2.small*, *t2.medium* (2 vCPUs, 4 GB RAM), and *t2.large* (2 vCPUs, 8 GB RAM). Linux CentOS 7 is installed with Moodle v3.2 and Ubuntu 16.04 with BigBlueButton v1.1. Each instance was created with 25 GB of EBS disk reserved for operating system and applications. AWS S3 is used as associated storage (Fig. 2).

In order to perform the stress test, we connected the teacher with one audio stream, one video stream sharing teacher's camera in low resolution (320x240px), and one video stream for the presentation. We used 80 physical PCs to simulate students' connections. Students were viewers only. The connection rate was one new connection per 10 - 20 seconds. We created 5 - 10 connections per PC. During the test we observed the system behavior and user experience, and we measured objective system parameters of BBB server.

For each tested instance we recorded three events: the first time when the system unexpectedly disconnected some of already connected users due to the system load (event "D"); the time needed for establishing a new connection is significantly slowed down (event "S"); the system limit (event "L"). It is interesting to notice that when the limit is reached the system doesn't fail. It rather disconnects some (or more) of already connected users in order to connect the new ones. Moreover, connected users don't experience quality degradation. Table 1 gives the number of users for which events "D", "S", and "L" were recorded.

TABLE I
THE OCCURRENCES OF SIGNIFICANT EVENTS FOR DIFFERENT INSTANCE TYPES DURING THE SYSTEM STRESS TEST

Event \ Inst.type	*t2.small*	*t2.medium*	*t2.large*
"D"	80	200	290
"S"	140	300	300
"L"	415	371	400

The parameters measured during the stress test are given in Fig. 5. We were measuring the following: CPU utilization (Fig. 5a – Fig 5c); memory consumption (Fig. 5d – Fig. 5h); and network throughput (Fig. 5i). The parameters were measured using Linux commands *free*, *uptime*, and *vmstat*.

Figs. 5a – 5c show CPU utilization. Regarding the CPU, we were measuring the time that CPU spends in idle (Fig. 5a), the number of interrupts per second (Fig. 5b), and the number of context switches per second (Fig. 5c). All three parameters give a good insight how the CPU performs with the increase of the number of users. The official BBB documentation recommends the CPU load below 80% for system stability. This is aligned with our findings: the event "D" for *t2.small* instance appeared with 80 users (53% of time CPU spent on idle, Fig 5b), and the event "S" started with 140 users (24% of CPU idle time). We were able to connect more than 400 users (Table 1), but we don't recommend this instance type for more than 120 - 140 users due to the frequent connection losses. This can be noticed in inability for the CPU to increase the number of interrupts and context switches (Figs. 5b and 5c). This instance type was stable for up to 70 - 80 users, when the first disconnections appeared ("D", Table 1). The instance *t2.medium* with its 2 vCPUs performed much better, as expected (Table 1 and Figs. 5a – 5c).

Figs. 5d – 5h show the memory parameters. From Fig. 5e it can be seen that used memory is the same regardless the instance type, which is expected (between 1.1 and 1.6 GB). Instances *t2.*medium and *t2.large* have more available memory (Fig. 5d), which they can spend for kernel disk caches and buffers (Fig. 5h), but this didn't influence the user experiences significantly in our test. Thus, we can conclude that the memory is not a limiting factor.

Due to the completeness of the results, let us briefly mention the storage requirements, disregarding the fact that it is not a limiting factor. For hard disk capacity we chose 25GB for the Linux operating system. The requirements of the storage from Fig. 2 are as follows. The instructor's audio stream with the presentation or drawing canvas requires 110 MB per recorded hour (Mbph) for RAW recorded materials, or 10 MBph for post-processed materials. If the instructor's video camera is added, 130 MBph are required for unprocessed and 50 MBph are required for post-processed materials. When all three streams are used simultaneously, they require 240 MBph and 75 MBph for RAW and post-processed materials, respectively.

In the end, we can conclude that for system stability

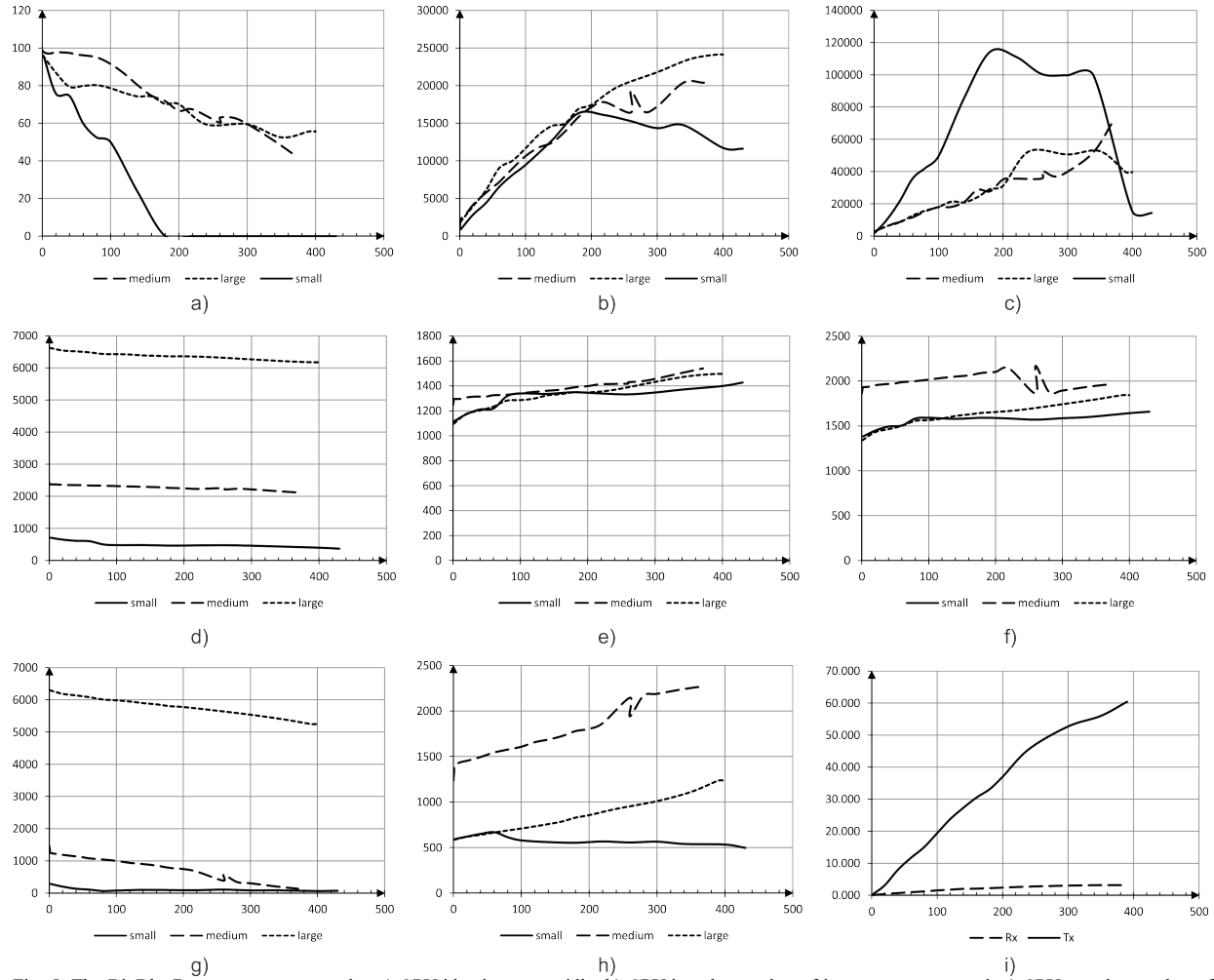


Fig. 5. The BigBlueButton stress test results: a) CPU.id - time spent idle, b) CPU.in - the number of interrupts per second, c) CPU.cs - the number of context switches per second, d) memory - available, e) memory - used, f) memory – active, g) memory – free, h) memory – kernel buffers and caches, i) network throughput

t2.small instance should be used for up to 70 users, while *t2.medium* is sufficient for up to 150 - 200 users. There was no significant gain with *t2.large* instance type.

IV. Conclusion

In this paper the resource requirements of a video conferencing system, based on open source BigBlueButton platform and implemented on the Amazon AWS public cloud, are presented. The system was implemented as the central part of blended learning system. In order to determine the system limits and give the recommendations for implementation of similar systems, the results of the stress test are given. The given parameters include CPU, memory and network throughput. The system was tested on three different AWS instance types, simulating up to 400 simultaneous users. For tested instances the recommended number of users were given.

Acknowledgment

The research was supported in part by the Serbian Ministry of Education, Science and Technological Development (Project TR32012).

References

[1] Logie, Baird John. "Apparatus for transmitting views or images to a distance." U.S. Patent 1,699,270, issued January 15, 1929.

[2] Noll, A. Michael. "Anatomy of a failure: Picturephone revisited." Telecommunications policy 16, no. 4 (1992): 307-316.

[3] Jacobs, Marco, and Jonah Probell. "A brief history of video coding." ARC International (2007): 1-8.

[4] De Serres, Yves, and Lawrence Hegarty. "Value-added services in the converged network." IEEE Communications Magazine 39, no. 9 (2001): 146-154.

[5] Ellis, Robert A., Abelardo Pardo, and Feifei Han. "Quality in blended learning environments–Significant differences in how students approach learning collaborations." Computers & Education 102 (2016): 90-102.

[6] Moskal, Patsy, Charles Dziuban, and Joel Hartman. "Blended learning: A dangerous idea?" The Internet and Higher Education 18 (2013): 15-23.

[7] Köse, Utku. "A blended learning model supported with Web 2.0 technologies." Procedia-Social and Behavioral Sciences 2, no. 2 (2010): 2794-2802.

[8] Weinman, Joe. "Cloud pricing and markets." IEEE Cloud Computing 2, no. 1 (2015): 10-13.

[9] McKenzie, Wendy A., Eloise Perini, Vanessa Rohlf, Samia Toukhsati, Russell Conduit, and Gordon Sanson. "A blended learning lecture delivery model for large and diverse undergraduate cohorts." Computers & Education 64 (2013): 116-126.

[10] Salah, Khaled, Mohammad Hammoud, and Sherali Zeadally. "Teaching cybersecurity using the cloud." IEEE Transactions on Learning Technologies 8, no. 4 (2015): 383-392.

Agile Method and ROS in Automotive Software development processes, practice, and teaching

Momčilo Krunić, Vlado Krunić, Milan Stankić, and Miroslav Popović, *Member, IEEE*

***Abstract* — Software development in Automotive industry experiencing exponential ascent of complexity these days. This is a direct implication of leading trends such as ADAS (Advanced Driver Assistance Systems) and Autonomous vehicles. In order to achieve these goals, new technologies have been introduced in Automotive software industry that have not been traditionally present, such as: Machine learning, data mining, computer vision, object fusion, … This paper describes new approach that has been introduced in Automotive software development processes and practice, such as Agile methodology and Robot Operating System (ROS), which aim is to enable seamless integration of new technologies into the vehicles, since traditional approach didn't provide optimal results. Focus of the paper is implementation of such processes and practices in the course: "Automotive Software Development Processes", at the Faculty of Technical Sciences (FTN), Novi Sad. The main idea behind this was to teach students about processes and practices, by conducting them through lectures and exercises. It has been used Autoware for this purpose. open-source platform for self-driving vehicles, based on ROS, C++ language for implementation, and LeSS (Large-Scale Scrum) framework for agile development. Implemented solutions was developed using TDD (Test Driven Development) methodology and Google test framework (gtest).**

***Index Terms*—Agile Method; ROS; Automotive Software; Processes; TDD; ADAS; Autonomous vehicles; AUTOSAR Adaptive.**

I. Introduction

THE complexity of software development in the automotive industry has grown exponentially recently, guided with new trends such as: HAD (Highly Automated Driving), ADAS (Advanced Driver Assistance Systems), and Car2X. Today, 90% of all innovations in automotive industry are coming from the E/E systems and software [1] being deployed in over hundred ECUs (Electronic Control Units). In order to meet the challenges posed by the market, leading OEMs (Original Equipment Manufacturer) and their immediate suppliers have had to adapt traditional approach to automotive software development processes like V-Model [2], to the most up-to-date trends such as scaled agile frameworks and their most influential implementations: [3] and SAFe [4], while retaining good practices that for years have provided a high level of quality of the finial software product through the implementation of the ASPICE standard [5]. The most disadvantage of the traditional approach using the V-Model, through processes defined by ASPICE, is long period between customers feedbacks. This is the main reason, among others, to introduce Agile method [6] into the world of automotive software development. Fast feedback during software development was also a main driver to include Robot Operating System - ROS [7] as a platform for rapid prototyping in order to reveal design flaws in in early phase of the project. This is crucial for cost optimization.

Very important aspect, but also a challenge, is the implementation of the standard for the functional safety ISO26262 [8] in software developed by agile processes, so that the final product can eventually end up on the public roads. Agile method does not impose strict processes during software development, which might be in contradiction with safety standard ISO26262. The practice has shown that paradigm described in "Clean Code" [9] is in compliance with ISO26262, so it is a strong recommendation to apply theses common sense set of rules during software development in order to achieve reliable and easy to maintain final software product.

One of the main topics covered in this paper is also Agile approach in teaching [10] students about such complex multidimensional problems, such us development of self-driving cars, on the Faculty of Technical Sciences (FTN), University of Novi Sad.

In second section it will be presented general and more traditional approach in the automotive software development processes. Third section provides explanation about modern approach in software development. In fourth section it has been introduced ROS and its utilization in automotive industry. Fifth section explains how Agile method and ROS are being used in practice and teaching. Sixth section concludes the paper.

II. Traditional approach in product development

Development of modern days vehicles is conducted over parallel engineering in multiple domains, such as: Mechanical, Electric/Electronic (E/E), software, etc. It is important to emphasize that parallel development is quite challenging, because of a vast amount of interdependences. For example, one supplier is developing a software for the target hardware platform provided by another supplier, but which is not available yet. Since modern vehicle has about 30,000 parts, this requires distributed engineering between OEM, and supply chain of direct (Tier 1) and indirect suppliers (Tier 2, 3). Distributed simultaneous engineering requires well defined interfaces and seamless integration at system level. Before HAD and ADAS features [11] kicked

Momcilo Krunic - Faculty of Tech. Sciences, University of Novi Sad, Dr Zorana Đindica, 21000 Novi Sad, Serbia (email: momcilo.krunic@rt-rk.com).

Vlado Krunic - Department of Mathematics and Informatics, Faculty of Natural Sciences and Mathematics, University of Banja Luka, Bulevar vojvode Petra Bojovića 1A, 78000 Banja Luka, Bosnia and Herzegovina (email: vlado.krunic@pmf.unibl.org).

Milan Stankić - RT-RK, Institute for Computer Based Systems LLC, Narodnog fronta 23a, 21000 Novi Sad, Serbia (email: milan.stankic@rt-rk.com).

Miroslav Popovic - Faculty of Tech. Sciences, University of Novi Sad, Dr Zorana Đindica, 21000 Novi Sad, Serbia (email: miroslav.popovic@rt-rk.com).

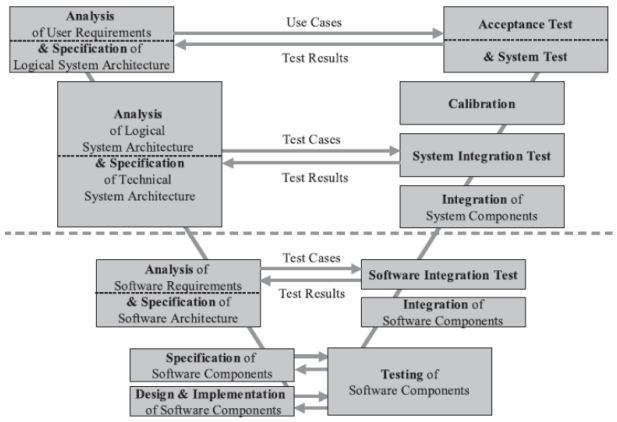


Figure 1. Automotive product development using V-Model.

in, ASPICE [5] and traditional approach has been sufficient to deliver final product with respective quality. Central part of automotive software product development is guided using V-Model [2], which is represented in Figure 1. The V-Model represents coordination of Systems engineering and Software development, where everything starts with User Requirements, which outlines the highest level of abstraction. These requirements are first break down on the System level: Logical and Technical architecture [12], and afterwards on Software level: Architecture, Design, and Implementation. Within each level of abstraction: System and Software, appropriate test cases are specified and linked, in order to verify whether requirement has been fulfilled. This represents one of the crucial points of ASPICE and V-Model, called bidirectional traceability, which is also mandatory by the Functional Safety aspects defined by the standard ISO26262 [13]. Bidirectional

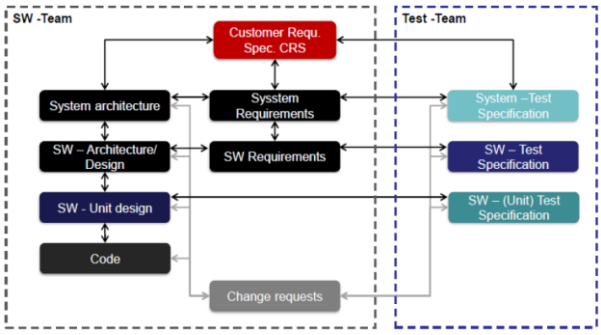


Figure 2. Concept of bidirectional traceability.

traceability enables linking on all levels of abstraction during automotive software development, from User Requirement, to implementation, testing and integration, Figure 2.

It is not hard to conclude that change request on the Customer Requirements level, triggers changes on all levels in the V-Model: System and Software, as well as on both sides: Design and Implementation (left side of V) and Testing and Integration (right side of V). This leads to a very slow process of getting customers feedback. When the changed requirement is finally fulfilled, and if the customer is not satisfied with the end result, this will implicate another change request and so on. The later in the project lifecycle change request occurs, it is harder to implement it, and it is certainly much more expensive. Today, maybe more than ever, due to the fact of exponential rise of complexity and involvement of modern trends such as: HAD, machine learning, Adaptive AUTOSAR, etc., customers' requests are changing more often, so getting a fast feedback from the customers is becoming one of the most important aspects of automotive software development, but also software development in general. Since traditional approach is not optimized for frequent change requests and getting fast feedback, this is where Agile method steps in as an alternative framework which defines certain processes and events to ensure change requests and customers feedbacks every few weeks. On the other hand, ROS [7] role is dual, but also related to fast feedback during development. In early stage of the project ROS is used as a platform for rapid prototyping whilst nor hardware or software are available yet, but algorithms need to be developed and tested. Later ROS is used for debugging and testing, because it provides powerful environment for visualization and simulation.

III. ROS AS PLATFORM FOR RAPID PROTOTYPING

Development of autonomous vehicles is in a domain of robotics. In order to improve and speed up development of such complex software systems, some of the leading OEMs (such as BMW, Daimler, Toyota, Uber, etc.) have adopted the ROS as a rapid prototyping platform, so that in the early phase of the project, while the hardware and software platform (AUTOSAR Adaptive) is not yet available, engineers will be enabled for software development and simulation. ROS was the first choice for many OEMs, because: it has a strong community (still do), it follows the similar design concepts as AUTOSAR Adaptive [14] (Service Oriented Architecture - SOA), it is portable, it is open source, and it is simple. During development, the ROS designers have been guided with following philosophical goals:

- "Hardware agnosticism"
- Peer to peer
- Tools based software design
- Multiple language support (C++/Java/Python/LISP)
- Lightweight: runs only at the edge of your modules
- Open source
- Suitable for large scale research and industry

In order to deal with complexity of robotics, the ROS design concepts are quite simple. First, it has been applied principle: "divide and conquer". All software modules have been divided into the separate processes, called nodes, which communicates over messaging interfaces. So, complexity is handled by composition of processes. There are three main parts in the ROS concept:

1. roscore: ROS Master, Parameter server, and rosout
2. Package: A virtual directory holding one or more executables (nodes)
3. Node: An agent communicating with ROS and other nodes via:
   a. Topics (publish / subscribe) using typed messages
   b. Services: Request / Response paradigm (think of method or operation) via typed messages

It is important to emphasize that ROS Master [15] is used to facilitates communication between nodes as a lookup table. Each node (publishers, subscribers, and services), when it is launched, is first registered on the ROS Master. Communication between nodes is peer-to-peer. The ROS Master is responsible for enabling that communication (Figure 3.).

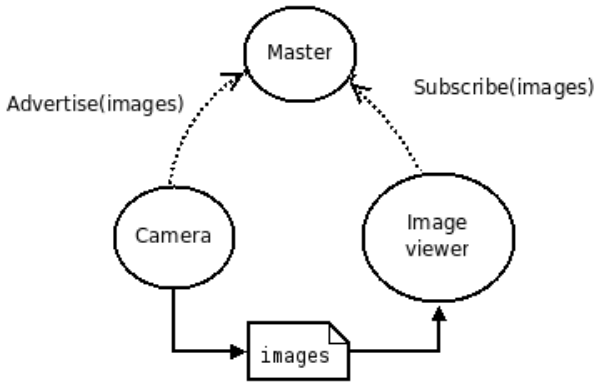


Figure 3. Messaging mechanism in inter-node communication.

For example, in Figure 3., "Camera" node is registered on the ROS Master as a publisher on a topic called "images", and "Image viewer" node is registered as a subscriber on the same topic. Each time "Camera" publishes something on the "images", "image viewer" node is being notified, and appropriate callback function is called for processing new data. This simple mechanism has been introduced to handle asynchronous communication. Synchronous communication in the ROS environment is handled by invocation of services. Beside these simple communication interfaces, ROS supports Tools based software design, by providing various set of tools for: building, running, debugging, simulation, etc. One of the most valuable features of ROS is possibility to record all messages, being sent during the session, in the so-called rosbag file. Afterwards, rosbag file can be played in the rviz [16] tool, which is named for visualization/simulation of recorded messages. In that way entire session can be reproduced and visualized which makes development and debugging much easier.

Simplicity of the entire ROS concept is the main reason of exponential adaptation on various robotics projects of today, but in automotive and teaching as well.

IV. Agile Method in Practice

Agile software development represents process framework for organizing the most modern complex projects, because it follows three simple practices: transparency, inspection, and adaptation. The main reason for publishing Agile Manifesto [17], back in 2001., was inability of traditional approach, like Waterfall [18], to handle complexity in organization of modern software development projects.

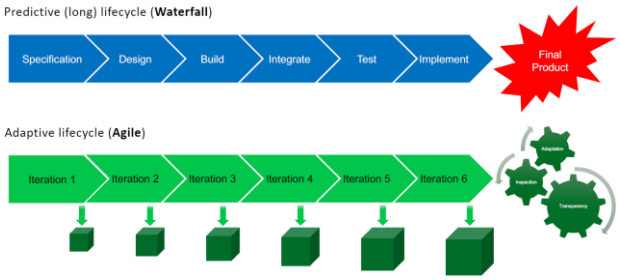


Figure 4. Product development process lifecycle – different approaches.

Figure 4. depicts main differences between two different approaches in software development processes. Traditional approach requires that product requirements, with all details specified, must be defined up front, and only then development can be initiated. This is especially hard to achieve in projects that have long lifecycle, such are in automotive industry, because during that period technology evolves and requirements are being changed. Unlike Waterfall, Agile method does not follows the strict plan, but product is being developed in short iterative cycles which enables a fast Feedback Loop: Build > Measure > Learn cycle. This way it is more likely that final product will satisfy the customers' requirements, because requirements can evolve during the product development lifecycle.

Agile manifesto, with four values and twelve principles [19], provides a philosophical vison how software product should be developed, but does not offers concrete process definitions. There are many frameworks out there that implements Agile manifesto, such as: Scrum, Kanban, Crystal, XP, etc., but the problem is that these frameworks can be applied on a small-scale project. Because of that, several frameworks have been developed for scaling Agile practices. One of the most influential frameworks for scaling Scrum is the Large-Scale Scrum (LeSS) [3], or even bigger Less Huge. Less and Less Huge gain its popularity because of simplicity. Again, as in ROS, applying simple concept on complex projects proved to be a winning ticket.

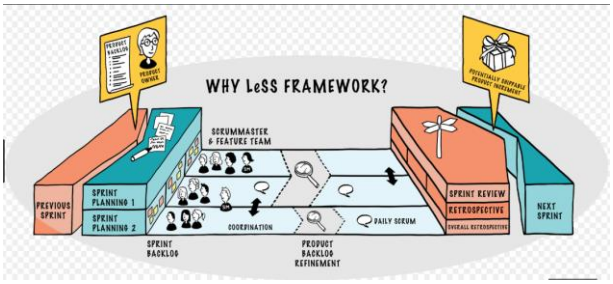


Figure 5. LeSS - framework for scaling Scrum to more than one team **[24]**.

As it can be seen in Figure 5., LeSS framework proposes just a few project roles: Customers, Product Owner (PO), Scrum Masters, and Feature Teams. It is important to emphasize that LeSS and LeSS Huge philosophy is a whole product focus. Therefore, it is important to have one PO and one product backlog.

The PO is responsible for the whole product vision and prioritization of product backlog in order to ensure efficient Return Of Investment (ROI). He/She works together with Feature teams, Scrum Masters and Customers on Product Backlog Refinement (PBR), which is an event that occurs in each Sprint (2 to 4 weeks iterative cycle) where items are broken down to a smaller tasks and owned by Feature teams, which represents cross-component and cross-functional teams that are focused on implementing and delivering a single customer feature. Afterwards, Feature teams organize Sprint planning 1 and 2, where tasks are prioritized in the teams Sprint backlogs and assigned to appropriate team

members. Each day of the Sprint teams conducting Daily Scrums, which is one of the most important events in the Sprint, because it enables transparency, one of the key principles of Agile. At the end of the Sprint, Feature teams, PO, Scrum Masters, and Customers organize an event called Sprint Review Bazaar, where teams demonstrate to the PO and Customers what has been done during the Sprint. This event is important because it enables fast feedback loop from the PO and Customers, therefor implements inspection and adaptation, another two key principles of Agile. Finally, at the end of the Sprint one of the most important event for the team occurs, called Sprint Retrospective, where each team addresses what went good and bad during the Sprint, afterwards making Action Items to improve the processes and itself. One of the key participants of LeSS and LeSS Huge is a Scrum Master. His/Hers role is to teach PO, Customers, and Feature teams to become more Agile, to practice all important aspects of Scrum, and to improve the system.

One of the pioneers in the automotive industry that applied Agile method through the LeSS and LeSS Huge frameworks is the BMW. First, they started LeSS adaptation in 2012. on the project which aim was to create the new BMW i car direct-sales process of the BMW Group. After developing for more than 2 years, Unified Sales Platform (USP) was released in time, with high customer satisfaction and high quality [20]. Because of that outcome, they decided to apply LeSS Huge, starting from 2017., on the one of the most complexed projects of today to bring autonomous driving, levels 3 and 4, on the road until 2021 [21].

In order to enable fast feedback loop and rapid prototyping BMW also decided to use ROS as a platform for simulation and development, since it has similar design concepts as AUTOSAR Adaptive, which will be used in the final product.

V. Agile Method in Teaching

In the past three years, course: "Automotive Software Development Processes (ASDP)", has been presented to students, at the Master level studies, Faculty of Technical Sciences, University of Novi Sad. In order to follow up with the latest trends in the automotive industry, and to prepare students for the future engineering work, a significant reorganization of the ASDP course has been made. First, Agile method was introduced to students by simulating the industrial processes during the teaching, so they can learn about the processes while conducting it. Second, students worked on the ROS and Autoware platform [22], which is a first open source platform for development of self-driving vehicles.

At the very beginning, Course (Product) Backlog was presented, where the professor as Course (Product) Owner (Professor), presented Stories that will be processed/completed during the next two Sprints (2 weeks), duration of lectures. At the beginning of each working day, students, together with the professor, recall what was done the previous day, then introduce what will be done on that day, and finally, if there are some confusions those are clarified. The purpose of these events was to demonstrate standard Daily Scrums. At the end of each Sprint (working week), an event called the Review bazaar was held, where students divided into the Scrum teams presents on tables what they have learned during the Sprint, similarly as Scrum teams in industry, when teams presenting what has been done during the Sprint.

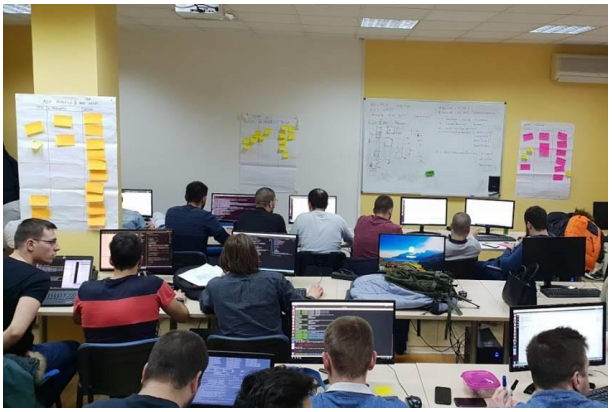


Figure 6. Scrum teams and Sprint boards.

After successfully completing the first two Sprint (lectures), the last two Sprints (Exercises) followed, where students learned about the processes that need to be followed during the development of software in the automotive industry through practical lessons. At the beginning of the Sprint, Course (Product) Backlog Refinement was carried out, where Course (Product) Owner introduced and explained the content of the new Backlog, followed by interactive discussion. Afterwards, Each Scrum team took out, from the Course (Product) Backlog, a project to work on during the Sprint. Scrum teams organized a Sprint planning event and accordingly set tasks to the team's scrum boards and started the realization (Figure 6.).

Figure 7. Scrum Teams demonstrates their projects on the Sprint Review.

Each day of the Sprint teams performed Daily Scrums and update status of tasks on the board. After successfully completing the second Sprint (Exercise), Scrum Teams demonstrated their projects at an event called Sprint Review, where each team on the projector presented their solutions to the rest of the group (Figure 7.).

Figure 8. Overall Retrospective and Action Items.

In order to conclude the Agile process, an Overall (Course) Retrospective event was held (Figure 8.) at the end of the course, where Scrum Teams displayed their impressions, bad and good, followed by a vote and defining the Action Items, whose implementation should improve the next iteration of the course. This was very important event because it provided the Feedback Loop to the Professor and Assistants.

VI. Conclusion

In this paper it has been presented modern trends in automotive industry, such as prototyping self-driving vehicles in the ROS environment, using Agile method for software development processes, as well as their impact on the approach in teaching ASDP.

During the course students were successfully realized following projects:

1. Autoware ROS node for monitoring stopping distance of vehicle based on its velocity and warn if there is a lidar detection in that area.
2. Autoware ROS node for monitoring number of lidar detections in each cell of a 2D grid around the vehicle.
3. Autoware ROS node for detecting objects in 2D grid around the vehicle based on of lidar detections.

Bearing in mind that at the beginning of the course students did not have any previous knowledge about the ROS, and the fact that all students at the end of the course successfully realized difficult projects, implicate that Agile approach in teaching was very effective.

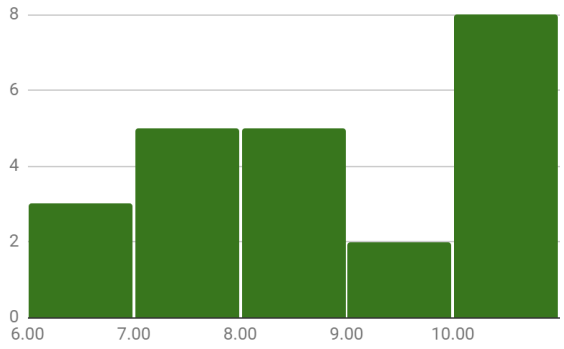


Figure 9. Distribution of grades - Theoretical part of ASDP.

At the end of the second Sprint, after the Review Bazaar, comprehensive exam took a place. It is worthwhile to notice that students did not have much time to learn at home, since an exam was during the lectures, but nevertheless distribution of grades were very positive (Figure 9.).

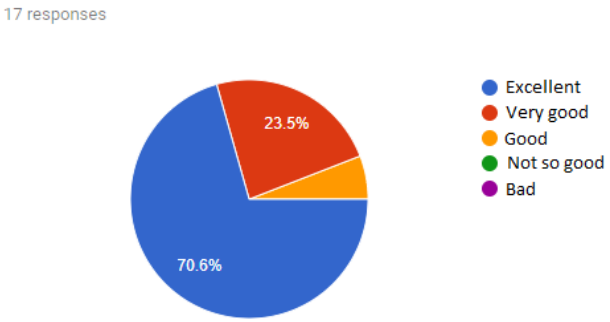


Figure 10. Anonymous survey about the overall organization of the ASDP course.

Finally, at the end of the course anonymous survey (Figure 10.) has been conducted, which showed that Agile method and new platform (ROS) induce very positive opinion among students about overall organization of the course.

Acknowledgment

This work was partially supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia, under grant number: III_044009_2.

References

[1] "News Europe," [Online]. Available: https://www.eenewseurope.com/news/innovation-car-90-comes-electronics-and-software. [Accessed 16 05 2019].

[2] Y. . Han, D. . Lee, B. . Choi, M. G. Hinchey and H. P. In, "Value-Driven V-Model: From Requirements Analysis to Acceptance Testing," *IEICE Transactions on Information and Systems*, vol. , no. 7, pp. 1776-1785, 2016.

[3] "Large-Scale Scrum (LeSS)," , 2014. [Online]. Available: http://less.works. [Accessed 29 4 2019].

[4] D. . Leffingwell, "Scaled Agile Framework," , . [Online]. Available: http://www.scaledagileframework.com/. [Accessed 29 4 2019].

[5] A. . Orecka, S. . Dawid and R. . Dzianach, "Best Practices for Achieving Automotive SPICE Capability Level 3," , 2012. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-30439-2_27. [Accessed 29 4 2019].

[6] D. . Parsons, H. . Ryu and R. . Lal, "The Impact of Methods and Techniques on Outcomes from Agile Software Development Projects," , 2007. [Online]. Available: https://link.springer.com/chapter/10.1007/978-0-387-72804-9_16. [Accessed 29 4 2019].

[7] "Robot Operating System," , . [Online]. Available: https://magazine.engineerjobs.com/2013/robot-operating-system.htm. [Accessed 29 4 2019].

[8] A. . George and J. J. Nelson, "Managing Functional Safety (ISO26262) in Projects," , 2017. [Online]. Available: https://sae.org/publications/technical-papers/content/2017-01-0064. [Accessed 29 4 2019].

[9] R. C. Martin, Clean Code: A Handbook of Agile Software Craftsmanship, ed., vol. , , : Prentice Hall, , p. .

[10] D. A. Dewi and M. . Muniandy, "The agility of agile methodology for teaching and learning activities," , 2014. [Online]. Available: https://igi-global.com/chapter/the-agility-of-agile-methodology-for-teaching-and-learning-activities/145044. [Accessed 29 4 2019].

[11] K. . Talmi and O. . Klenke, "From ADAS to HAD Flexible Software Components and Open Hardware Platforms," *ATZelektronik worldwide,* vol. 12, no. 2, pp. 42-47, 2017.

[12] M. v. d. Beeck, "Development of logical and technical architectures for automotive systems," *Software and Systems Modeling,* vol. 6, no. 2, pp. 205-219, 2007.

[13] P. . Oliveira, A. L. Ferreira, D. . Dias, T. . Pereira, P. . Monteiro and R. J. Machado, "An Analysis of the Commonality and Differences Between ASPICE and ISO26262 in the Context of Software Development," , 2017. [Online]. Available: https://link.springer.com/content/pdf/10.1007/978-3-319-64218-5_17.pdf. [Accessed 29 4 2019].

[14] "AUTOSAR_Release_18_03_EN," , . [Online]. Available: https://www.autosar.org/fileadmin/user_upload/AUTOSAR_Release_18_03_EN.pdf. [Accessed 29 4 2019].

[15] A. . Tiderko, F. . Hoeller and T. . Röhling, "The ROS Multimaster Extension for Simplified Deployment of Multi-Robot Systems," , 2016. [Online]. Available: https://rd.springer.com/chapter/10.1007/978-3-319-26054-9_24. [Accessed 29 4 2019].

[16] H. R. Kam, S. H. Lee, T. . Park and C.-H. . Kim, "RViz: a toolkit for real domain data visualization," *Telecommunication Systems,* vol. 60, no. 2, pp. 337-345, 2015.

[17] K. . Beck, J. . Grenning, R. C. Martin, M. . Beedle, J. . Highsmith, S. . Mellor, A. v. Bennekum, A. . Hunt, K. . Schwaber, A. . Cockburn, R. . Jeffries, J. . Sutherland, W. . Cunningham, J. . Kern, D. . Thomas, M. . Fowler and B. . Marick, "Manifesto for Agile Software Development," , 2001. [Online]. Available: http://agilemanifesto.org. [Accessed 29 4 2019].

[18] K. . Petersen, K. . Petersen, C. . Wohlin, D. . Baca and D. . Baca, "The Waterfall Model in Large-Scale Development," , 2009. [Online]. Available: http://wohlin.eu/profes09.pdf. [Accessed 29 4 2019].

[19] K. . Beck, J. . Grenning, R. C. Martin, M. . Beedle, J. . Highsmith, S. . Mellor, A. v. Bennekum, A. . Hunt, K. . Schwaber, A. . Cockburn, R. . Jeffries, J. . Sutherland, W. . Cunningham, J. . Kern, D. . Thomas, M. . Fowler and B. . Marick, "Principles behind the Agile Manifesto," , 2001. [Online]. Available: http://agilemanifesto.org/principles.html. [Accessed 29 4 2019].

[20] LeSS, "LeSS adoption at a bavarian car manufacturer," [Online]. Available: https://less.works/case-studies/bmw-group.html#Background. [Accessed 29. 04. 2019.].

[21] "LeSS Huge at BMW," [Online]. Available: https://www.infoq.com/presentations/bmw-less-huge. [Accessed 29. 04. 2019.].

[22] S. . Kato, S. . Tokunaga, Y. . Maruyama, S. . Maeda, M. . Hirabayashi, Y. . Kitsukawa, A. . Monrroy, T. . Ando, Y. . Fujii and T. . Azumi, "Autoware on board: enabling autonomous vehicles with embedded systems," , 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8443742. [Accessed 29 4 2019].

[23] M. Biehl, R. I. o. T. S. S. Embedded Control Syst. and M. Torngren, "A Cost-Efficiency Model for Tool Chains," in *Global Software Engineering Workshops (ICGSEW), 2012 IEEE Seventh International Conference on*, Porto Alegre, 2012.

[24] "LeSS Framework," [Online]. Available: https://less.works/less/framework/index.html. [Accessed 16 05 2019].

Churn Prediction in Telco Industry Leveraging Call Center Data

Nenad Petrović

***Abstract*— Identifying the customers that are likely to leave and move from one service provider to another (churners) is of utmost importance in telecommunications (telco) industry due to fact that it is several times more expensive to acquire a new customer than retaining an old one. Therefore, it is mission-critical for Business Intelligence (BI) systems to identify the churning customer timely in order to support the decisioning mechanisms within the customer relationship management (CRM) software and reduce the number of lost subscribers. In this paper, it is explored how the data about customers collected by mobile operator's call center software platform can be leveraged to identify churners together with other features. As an outcome, several experiments based on classification techniques are presented. Performance of two classification approaches is compared – k-NN and neural network. Moreover, for the second approach, CPU and GPU execution time is discussed.**

***Index Terms*—CRM; data mining; deep learning; machine learning; business Intelligence.**

I. Introduction

Due to large competitiveness and high demand, customers in telecommunications (telco) industry are likely to leave and switch to another operator if they find a more suitable offer that will satisfy their current needs for lower price. This process is known as *churn*. Every year, telco companies suffer from huge losses caused by churn, as it is much more expensive (5 to 10 times) to acquire a new customer than to retain the existing one [1]. Therefore, it is of utmost importance in telco industry to recognize the customers that are likely to leave (*churners*) timely and react accordingly in order to retain them [1-4]. Once they are identified, more attractive products and services can be offered to these users in order to prevent them from changing their provider. Moreover, it was shown that even small improvement in churn identification could be highly beneficial [2].

For this reason, one of the main roles of Business Intelligence (BI) [5] in Customer Relationship Management Software used by telco operators is to identify customers that are about to leave operator and leverage that information to adapt the underlying decisioning mechanisms in order to execute the corresponding strategy that could possibly lead to reduced number of lost customers.

In this paper, it is explored how data analysis techniques based on data mining and machine learning algorithms can be leveraged to detect the potential churners. As an outcome, classification approach for churn prediction is proposed, its results presented and evaluated. Moreover, two solutions are compared – the one that is leveraging the transaction data coming from call center software together with customer characteristics (such as number of active/suspended SIMs, spending) and another one that does not involve the data coming from call center. The goal of the experiment is to examine whether the analysis of call center transaction data is beneficial for churn prediction or no. Moreover, the processing time necessary for training and prediction is discussed, considering the results achieved by execution on CPU and GPU. Finally, two classification implementations are compared: deep learning-based and k-nearest neighbours.

Nenad Petrović is with the Faculty of Electronic Engineering, University of Niš, Aleksandra Medvedeva 14, 18000 Niš, Serbia (e-mail: nenad.petrovic@elfak.ni.ac.rs).

II. Background and related work

A. Data analysis techniques

Data analysis techniques include various data mining, machine learning, artificial intelligence and statistical algorithms that are used to extract useful patterns, relations and recognize trends from large quantity of stored data. An overview of data analysis techniques widely used in telecommunications industry customer relationship management is given in Table I [6, 7].

TABLE I
OVERVIEW OF DATA ANALYSIS TECHNIQUES

Technique	Description
Clustering	The task of grouping a set of observed objects in such way that objects within the same group are more similar to each other than to those in other groups.
Classification	Identifying to which of a set of categories, a new observation belongs to, on the basis of a training dataset containing observation whose category membership is already known.
Regression	A set of statistical processes for estimating how the typical value of the dependent variable changes when one of the independent variables varies, while others independent variables are fixed.
Association rule discovery	Association rule discovery has purpose to identify the associations and relationships between the items or events that occur within the input dataset.

When the goal of data analysis mechanism is to make predictions (as churn prediction that is main topic of this paper), the data is split into two datasets: training and test set. Training set is used to fit the parameters of the model using a supervised learning method. Moreover, a label variable is added to the dataset to denote the category where the observation belongs. After that, the constructed model is used to predict responses for new observations coming from another dataset – called test set.

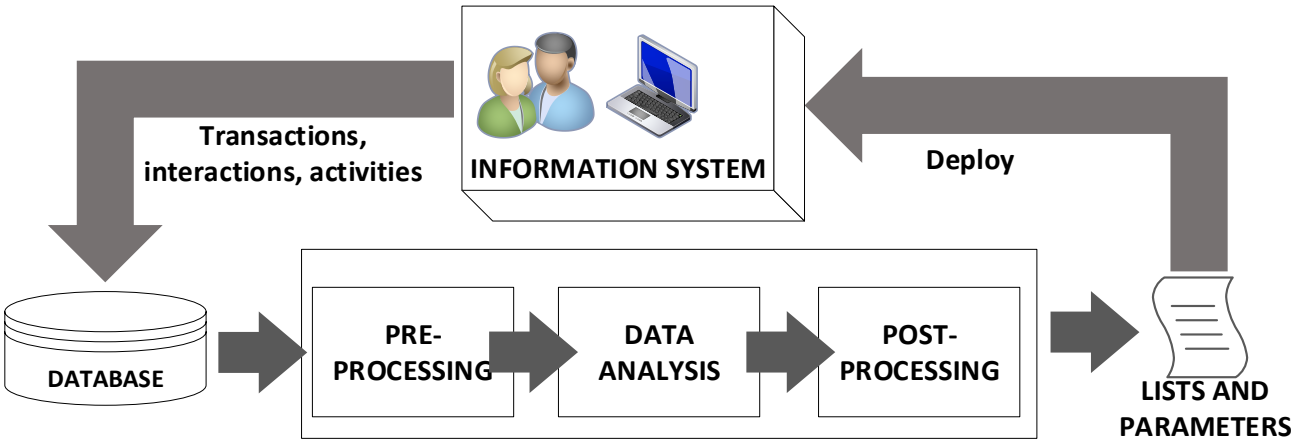


Fig. 1. Adopting BI in enterprise information systems

In [6], clustering was used to derive thresholds for business rules related to product availability, while association rule analysis was leveraged to design product bundles in telco operator's call center platform for different groups of customers. In [3] and [8], several classification algorithms were used for churn prediction based on telco customer features. Moreover, in [2], an approach for context-aware churn prediction in telco industry relying on ensemble of several algorithms was presented. In this paper, the decision is to rely on classification methods due to effectiveness of the approach, satisfactory performance and simplicity at the same time [3, 7, 8].

B. Business Intelligence (BI)

Also referred to as business analytics, it is a term used to describe a range of different applications and technologies used to extract and analyze large amounts of data to aid the process of decision making. BI includes both data analysis and querying tools [5]. Typical process of adopting BI within enterprise information systems is shown in Fig. 1.

First, the step of pre-processing is performed to transform the collected data about customer transactions, interactions and activities to a form suitable for corresponding data analysis algorithms. Queries are executed in order to filter data and select the desired attributes. Optionally, data integration can be done if it comes from heterogeneous information systems. Once the data is prepared, the algorithms for data analysis can be executed. Furthermore, the obtained results are interpreted and stored in a form that is deployable within the information systems environment, such as customer lists and budget parameters so they can be immediately used to drive the decisioning process [5].

C. Deep learning

Deep learning is a family of machine learning methods based on artificial neural networks. The learning process itself can be either supervised, semi-supervised or unsupervised. Deep learning has been applied in various areas, such as computer vision, speech recognition, natural language processing, audio recognition and bioinformatics.

Artificial neural network (ANN) is a group of nodes interconnected through weighted links. A node (also referred to as *neuron* or *perceptron*) is a computational unit containing one or more weighted input connections, a transfer function that combines the inputs, and an output connection. It receives the signal, processes it and then forwards it to other nodes connected to it.

In neural networks, three types of layers are identified: 1) input layer, that corresponds to the input variables 2) hidden layers – the nodes between input and output layers, more than one layer like this can be present within the network architecture 3) output layer – a layer of nodes that produce the output variables.

A *deep neural network* (DNN) is an artificial neural network (ANN) with multiple layers between the input and output layers [9]. A single-layer neural network can be only used to represent linearly separable functions, which is the case for very simple problems. However, most problems are not linearly separable. In [10], it was stated that two hidden layers are sufficient for creating classification regions of any desired shape.

In this paper, neural networks are used for classification approach to customer churn prediction based on supervised learning.

III. CALL CENTER CASE STUDY

Call center operators access a web application that supports the process of negotiation with customers in cases of tariff plan change, activation of discounts or additional services (such as extra GBs for mobile data traffic, unlimited SMS messages), as illustrated in Fig. 2. Whether it is possible to activate some service for the selected customer is determined by the mechanisms of decision logic. For decisioning, several factors are taken into account. First, there is a matrix of compatible products within the catalog which determines the services that go together with a given tariff plan. Moreover, the lists of customers are considered. Certain types of tariffs and services might be reserved only for specific customer categories (big enterprise, small company or churner). On the other side, the ability to activate some service is also determined by customer's budget, which is based on average monthly bill value. Budget calculation formulas contain thresholds and coefficients that are customer type-dependent. Customer lists are updated regularly to take into account the new interactions with customers, transactions, activities and bill changes. Once call center operator finds the right deal for the customer, it is added to the customer basket. Finally, if user accepts the deal offered by the operator, it is submitted for activation.

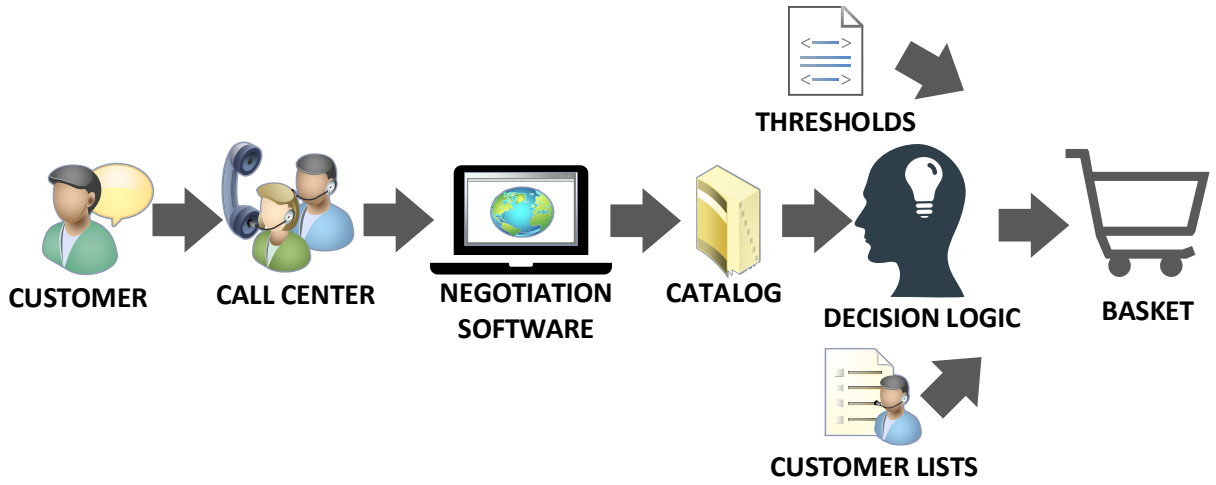


Fig. 2. Call center software platform illustration

Separating potential churners from other customers is of utmost importance for this application, as they are treated differently. The budget thresholds are often more tolerant in this case and some specific services might be offered in order to retain the customers with high churn risk The whole business process of churning customer prediction and their treatment is shown in Fig. 3, from two different perspectives – data analyst and call center operator. The task of data analyst is to construct the lists of customers that are likely to leave applying data analysis techniques that are deployed to the call center application. After that, while call center operators negotiates with customers, it is checked whether they belong to the list of churners. If it is true, another perspective in application is opened, showing special services that are not available for other users.

IV. Churn Prediction Experiments and Results

In this section, an approach of churn identification based on call center transaction data using classification algorithm is presented. Moreover, the obtained results are evaluated, discussed and compared to the case when call center transaction data is not taken into account. The data was prepared using SAS Enterprise Guide[1], while the predictions themselves were done using two open-source libraries: 1) TensorFlow[2], a GPU-supported library for Python programming language in case of deep learning-based approach. Deep learning refers to a class of machine learning algorithms relying on artificial neural networks, that consist of a cascade of multiple layers of nonlinear processing units where each successive layer uses the output

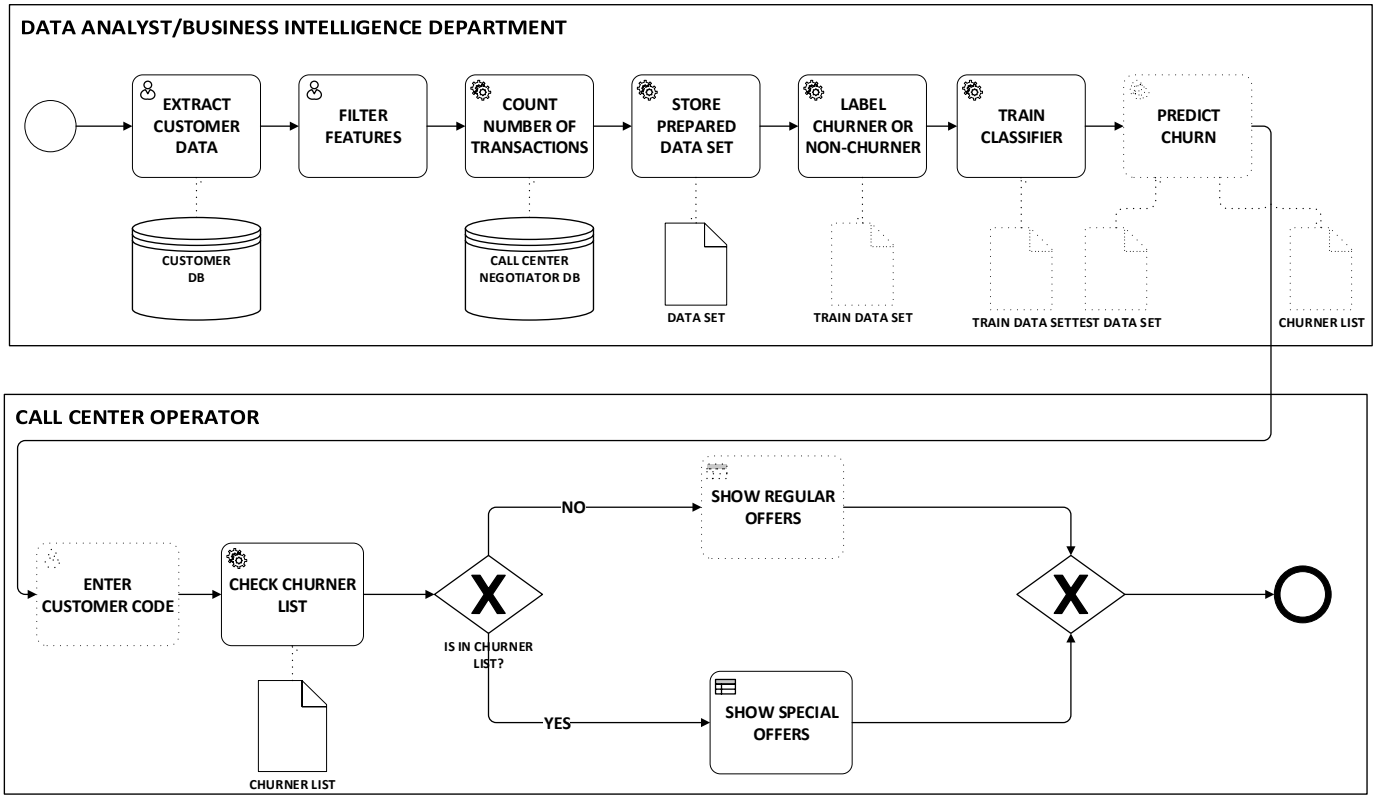


Fig. 3. Business process of churning customer prediction and their treatment in BPMN notation

[1] http://support.sas.com/software/products/enterprise-guide/index.html
[2] https://www.tensorflow.org/

from the previous layer as input. In this paper, deep learning is used for supervised learning. 2) Java-ML[3] library for Java programming language in case of k-nearest neighbours method [11]. In k-NN classification, an object is assigned to the class most common among its k nearest neighbours (k – positive integer, typically small) [12].

The performance evaluation was done on a server equipped with AMD Ryzen 7 1700X octa-core CPU running at 3.80GHz, 64GB DDR4 RAM and NVIDIA Quadro P2000 GPU with 4GB VRAM. For GPU-supported execution, a CUDA-compatible hardware is needed, which is the case with the described configuration.

In pre-processing phase, the data was prepared by executing queries against the negotiator software database. For 200 anonymized customers, the features shown in Table II were extracted for period from 1st January 2016 to 31st December 2018. Five features were taken into account: 1) number of service activations (transactions) via call center software platform 2) average spending taking into account other services provided by the operator, such as DSL and PBX 3) average spending per SIM 4) number of active SIMs 5) number of suspended SIMs. Additional column (label) with value 0 for non-churners and 1 for churners was added, considered as a dependent (outcome) variable, while other features were treated as independent In training set, this label is used for supervised learning, while in test set it is used to check whether the prediction is correct or not.

TABLE II
FEATURES FOR CLASSIFICATION-BASED CHURN PREDICTION

Number of transactions	Avg. spending	Avg. spending per SIM	Active SIMs	Suspended SIMs

After preparation, data was divided into disjoint sets – train and test set, in different ratios. In Table III, an overview of the achieved results in case of different size of training and test set is given for case when call center transactions are taken into account (third column) and not (fourth column) for TensorFlow implementation based on deep learning.

TABLE III
HOW CALL CENTER TRANSACTION DATA AFFECTS THE CHURN PREDICTION PERFORMANCE IN CASE OF DEEP LEARNING

Training set size	Test set size	Correct/Test size [%] (with call center data)	Correct/Test size [%] (w/o call center data)
50	150	90.67	82.67
75	125	91.20	84.00
100	100	94.00	88.00
150	50	96.00	92.00

According to the results, it can be noticed that churn prediction performs slightly better in case when number of call center transactions for each customer is also included. This outcome can be explained by the fact that number of executed transactions via call center platform encapsulates an important aspect of customer behavior.

Moreover, in Table IV, the results of processing time compared in case of GPU and CPU execution are given for both the training and test phases. In this case, a neural network with 2 hidden layers was used (3 nodes per layer). Due to insufficient amount of customer data available, synthetic data was generated using a custom random value generator written in Java for purposes of processing time evaluation.

TABLE IV
COMPARISON OF CPU AND GPU EXECUTION TIME

Set size	CPU train	CPU pred.	GPU train	GPU pred.
100	4.73	3.29	5.13	3.14
1000	5.13	5.05	5.26	3.89
10000	25.5	11.23	24.74	9.42
100000	101.31	22.43	50.23	12.68
500000	161.17	29.14	79.21	15.37

According to the obtained results, it can be noticed that the execution of both training and prediction is in favor of GPU, as the set size increases. However, in case of the smallest dataset (100 observations), the execution of training phase was faster on CPU, while they are comparable in case of medium-sized datasets (1000 and 10000). In case of last two largest datasets, the advantage of GPU becomes more significant and is a bit more beneficial in training phase. When it comes to the execution speed of churn prediction, a similar trend was noticed for larger datasets, while the time needed for prediction was comparable even for the smallest dataset. This can be explained by the fact that GPU calculation parallelism is better exploited by TensorFlow when it comes to deep neural network learning for larger amount of data.

Furthermore, in Table V, the performance of churn prediction for different neural network configurations (number of hidden layers and nodes per layer) is shown. It can be noticed that the number of correct predictions is similar for configurations with more than two nodes per layer. When it comes to the number of layers, it can be noticed that the performance increases up to 3, while the further increase of layer number does not seem beneficial.

TABLE V
CHURN PREDICTION IN CASE OF DIFFERENT NEURAL NETWORK CONFIGURATIONS

Number of hidden layers	Nodes per layer	Correct/Test size [%] (with call center data)
1	1	62
1	2	78
1	3	92
1	4	92
1	5	94
1	6	92
2	3	96
3	3	98
4	5	92
2	5	96
2	10	96

On the other side, in Table VI, an overview of the results obtained in case of k-NN approach is given when call center

[3] http://java-ml.sourceforge.net/

transaction data was included. It can be noticed that it gives slightly larger number of correct predictions compared to deep learning implementation in case of smaller training sets, while deep learning performs better in case of larger training sets.

TABLE VI
CHURN PREDICTION IN CASE OF K-NN ALGORITHM LEVERAGING CALL CENTER TRANSACTION DATA

Training set size	Test set size	Correct/Test size [%] (with call center data)
50	150	92.00
75	125	92.80
100	100	93.00
150	50	94.00

In Fig. 4, it is shown how the percentage of correct predictions depends on k parameter in k-nearest neighbours algorithm in case of training set that consists of 150 observations and test set of 50 observations. It can be noticed that it increases until value 4. After that, it starts decreasing, with significant drop at k=6. For this reason, it was decided to use k=4 in the experiments.

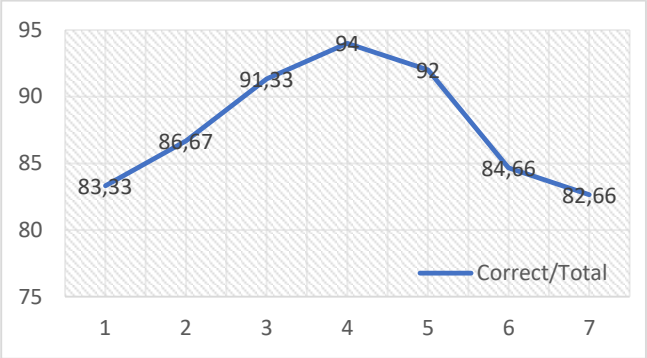


Fig. 4. Illustration how ratio of correct/total [%] predictions depends on k parameter value in k-NN algorithm

Moreover, association rule discovery algorithm was executed against the call center transaction data using SAS Enterprise Miner[4] in order to explore churning customer behavior when it comes to activation of digital mobile services, using the code similar to the one presented in [6]. However, due to relatively small number of transactions involving churn customers (except several outliers), it was not possible to discover rules with significant support and confidence that could illustrate behavior and leveraged in combination with other techniques such as clustering or classification. The value for minimum confidence was 75, while minimum support was 20. Further decrease of these thresholds did not lead to discovery of meaningful patterns.

A number of call center-aided transactions per customer for 50 churning (red) and 50 non-churning (blue) customers extracted from the dataset is illustrated in Fig. 5. The similar trend was also noticed for overall average spending and average spending per SIM, showing much lower values for churning than in case of non-churning users. Therefore, these features contribute to churn prediction due to their discriminative nature noticed in this case.

[4] https://www.sas.com/en_us/software/enterprise-miner.html

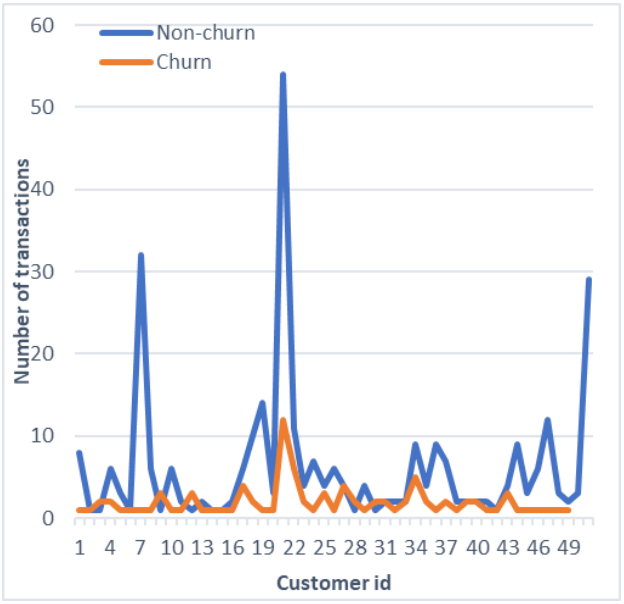


Fig. 5. Number of call center-aided transactions per customer for 50 churning (red) and 50 non-churning (blue) customers

V. CONCLUSION AND FUTURE WORK

In this paper, a method for churn prediction in telco industry leveraging the customer transaction data produced by call center software platform together with other customer features was presented. It can be concluded that addition of call center transaction-related features leads to slightly better churn prediction. However, when it comes to churn prediction, it is already known that even small improvement can save telco company from significant revenue loss. Moreover, it was shown that utilization of GPU-enabled machine learning libraries can lead to faster execution on adequate hardware in case of large customer databases, which is the case of large-scale and international telco operators. On the other side, leveraging GPU for churn prediction based on small customer bases might not be beneficial at all. Furthermore, an approach to exploit association rule discovery for churn prediction was not successful, due to insufficient number of transactions available for churning customers. Finally, it can be concluded that k-NN classification performs better in case of smaller training sets, while deep learning approach is more beneficial for larger training sets. It is planned in future to experiment with different customer features and ensemble of various data analysis techniques. Also, the performance of the developed methods will be evaluated on larger non-synthetic datasets and performance illustrated with more parameters included.

ACKNOWLEDGEMENT

The access to proprietary software tools, licenses, customer data and partial financial support were provided by Tech Rain S.p.A., Via Bisceglie, 76, 20152 Milano, Italy.

REFERENCES

[1] N. Hashmi et al., "Customer Churn Prediction in Telecommunication A Decade Review and Classification", IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 5, No 2, September 2013, pp. 271-282, 2013.

[2] R. Bai et al., "Context aware Telco Churn Prediction Powered by Temporal Feature Engineering", PerCtowd'18 – International Workshop on Context-Awareness for Multi-Device Pervasive Computing, pp. 161-166, 2018.

[3] G. M. Apurva Sree et al, “Churn prediction in Telecom using classification algorithms”, International Journal of Emerging Technology and Innovative Engineering vol. 5 (2), February 2019, pp. 19-28, 2019.
[4] M. Bagri et. al, “Churn Analysis in Telecommunication Industry”, 2018 International Conference on Automation and Computational Engineering (ICACE), pp. 126-132, 2018.
[5] C. Anderson, “Business Intelligence”, Data Science in Practice, Studies in Big Data 46, pp. 97-118, Springer, 2019.
[6] N. Petrovic, “Adopting Data Mining Techniques in Telecommunications Industry: Call Center Case Study”, IEEESTEC – 11th Student Projects Conference, pp. 11-14, 2018.
[7] V. Mahajan, R. Misra, R. Mahajan, “Review of Data Mining Techniques for Churn Prediction in Telecom”, Journal of Information and Organizational Sciences vol. 39, no. 2 (2015), pp. 183-197, 2015.
[8] M. Makhtar et al., “Churn Classification Model for Local Telecommunication Company Based on Rough Set Theory”, Journal of Fundamental and Applied Sciences, 2017, 9(6S), pp. 854-868, 2017.
[9] Y. Bengio, “Learning Deep Architectures for AI”, Foundations and Trends in Machine Learning 2(1), pp. 1-127, 2009.
[10] R. Lippmann, “An introduction to computing with neural nets”, IEEE ASSP Magazine, vol.4(2), pp. 4-22, 1987.
[11] T. Abeel, Y. V. de Peer, Y. Saeys, “Java-ML: A Machine Learning Library”, Journal of Machine Learning Research, vol. 10, pp. 931-934, 2009.
[12] T. Cover and P. Hart, “Nearest neighbor pattern classification”, IEEE Transactions on Information Theory, 13(1), pp. 21–27, 1967.

An implementation of the ARINC 653 APEX API services

Anja Veselinović, Branislav M. Todorović, Miloš Pilipović

***Abstract* — ARINC 653 is a specification used for integrating avionics system on a modern aircraft. The APEX (Application/Executive) interface, between the application software and the Core Software, defines a set of facilities which the system will provide for application software to control the scheduling, communication and status information of its internal processing elements. The interface and the behavior of the API (Application Program Interface) services are specified by ARINC 653. The ARINC 653 APEX API provides services to the applications. The aim of this paper is to present principles of ARINC 653 services implementation.**

***Keywords* — ARINC 653, APEX interface, API services, Integrated Modular Avionics, Partitioning, Partition Management, Partition Communication, Health Monitoring**

I. Introduction

Nowadays, the avionics industry is moving towards the use of a new architecture for aircraft systems called Integrated Modular Avionics (IMA). The IMA architecture consists of a distributed system, where many aircraft applications can be executed in the same hardware module, sharing computing resources, communications and input and output devices. The ARINC 653 [1] specification is one of the most important blocks from the IMA definition, where the partitioning concept emerges as a way to ensure protection and functional separation between applications, usually for fault containment and ease of verification, validation and certification. ARINC 653 is used extensively on new civil and military aircrafts produced by Airbus, Boeing and others [2].

The ARINC 653 defines a general purpose APEX interface Application Program Interface (API) between the Core Software (CSW) of an Avionics Computer Resource (ACR) and the application software. The software specifications of the APEX interface are High-Order Language (HOL) independent, allowing systems using different languages to follow this interface. ARINC 653 time and space separation architectures enable applications with different levels of safety criticality to share the common compute platform, thereby optimizing the use of computer resources and allowing the Human–Machine Interface (HMI), flight controls, and aircraft systems to safely share the common compute resource. ARINC 653 was designed for safety certification, which decreased aircraft programs' certification risk.

This paper is organized as follows. In Section II we present basic concepts of the execution environment of the system and interaction between the individual elements within the system. ARINC 653 defines Basic [1] and Extended Services [3] for the API (specified in section III). The data type names, service request names, parameter names and order of parameters are defined by this standard. ARINC 653 was developed for use in avionics systems, but it is equally suitable for any system that requires API services, and partitioning mechanisms to enable multiple hosted applications (of the same or differing safety levels) to operate on platform hardware. The Ada and C language are recommended and supported by airlines for application software development, but other languages can also be used. C language is used in example of Health Monitoring services implementation, given in section IV.

II. IMA System Architecture

The first level of decomposition of an integrated module IMA consists of two major categories:

- Partitioned application software
- Core module

A core module can contain one or more individual processors, each consisting of one or more processor cores. The core module architecture has influence on the O/S implementation, but not on the APEX interface used by the application software of each partition. Application software should be portable between core modules and between individual processor of core module without modifying its interface with the O/S.

In order to isolate multiple partitions in a shared resource environment, the hardware should provide the O/S with the ability to restrict memory spaces processing time and access to I/O for each individual partition.

III. The ARINC 653 APEX Services

The ARINC 653 APEX [4] services are API calls belonging in six categories:

This work was partially funded by the Ministry of Education, Science and Technological Development of the Republic of Serbia under Grant TR-32030.

Anja Veselinović, RT-RK Institute for Computer Based Systems, Narodnog fronta 23A, 21000 Novi Sad, Serbia (e-mail: anja.veselinovic@rt-rk.com)

Branislav M. Todorović, RT-RK Institute for Computer Based Systems, Narodnog fronta 23A, 21000 Novi Sad, Serbia (e-mail: branislav.todorovic@rt-rk.com).

Miloš Pilipović RT-RK Institute for Computer Based Systems, Narodnog fronta 23A, 21000 Novi Sad, Serbia (e-mail: milos.pilipovic@rt-rk.com).

A. Partition Management

Central to the ARINC 653 philosophy is the concept of partitioning. Partitions are scheduled on a fixed, cyclic basic. Partitions are activated by being assigned to one or more partition time windows within major time frame. Each partition time window is defined by its offset from the start of the major time frame and expected duration. Configuration of all partitions throughout the whole system is expected to be under the control of the system integrator and maintained with configuration tables.

1) Partition Control

The O/S starts the application partitions when the O/S enters operational state. The application is responsible for invoking the appropriate APEX calls to transition the partition from one operational mode to another.

2) Partition Scheduling

Scheduling of partitions is deterministic over time. The main characteristics of the module scheduling model are:

- From the application developer perspective the scheduling unit is a partition
- Partitions have no priority
- The module scheduling algorithm is predetermined, repetitive with a fixed periodicity (major time frame).

3) Partition Operating Modes

Operating mode represents the current execution state of the partition. Partition operating modes (shown in Figure 1) are:

- IDLE – In this mode the partition is not eligible for application execution.
- NORMAL – The partition's initialization is complete and process scheduler is active.
- COLD_START – In this mode the partition's initialization phase is in progress.
- WARM_START – Application is executing its respective initialization code. This mode is similar to the COLD_START, but the hardware context in which the partition starts may be different.

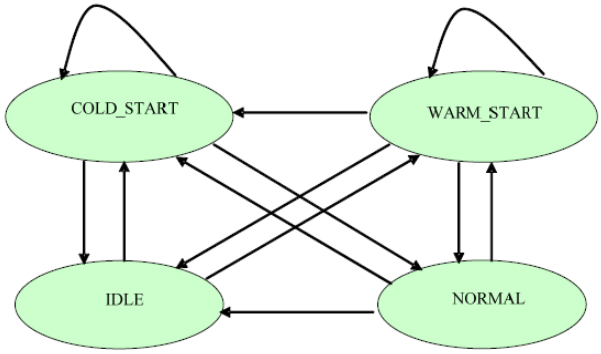


Fig. 1: Partition Operating Modes

B. Process Management

A process is a programming unit contained within a partition which executes concurrently with other processes of the same partition. Process may be designed for periodic or aperiodic execution. The processes are not directly visible outside of the partition. The partition should be responsible for the behavior of its defined processes.

C. Time Management

Time management is an important characteristic of an O/S used in real-time systems. The O/S provides time slicing for module scheduling, deadline, periodicity and delays for process scheduling and time-outs for intrapartition and interpartition communication.

As long as a process performs its entire processing without using its whole time capacity, the deadline is met. If a process requires more processing than the time capacity, the deadline is missed. When the deadline is missed, a health monitoring error will be raised (described in sections III and IV).

D. Memory Management

Partitions and their associated memory spaces are defined during system configuration and initialization. There are no memory allocation services in the APEX interface.

E. Interpartition Communication

A major part of ARINC 653 standard is the definition of the communication between APEX partitions, presented in Figure 2. All interpartition communication is conducted via messages. A message is defined as a contiguous block of data of finite length. A message is sent from a single source to one or more destinations. From the application viewpoint, a message is a collection of data which is sent or received to/from a specific port.

A channel is basic mechanism for linking partitions by messages. It defines a logical link between one source and one or more destinations. Partitions have access to channels in defined access points called ports. A port provides the required resources that allow a specific partition to send or receive messages in a specific channel.

At the application level, messages are atomic entries (either the whole message is received or nothing). The core module is responsible for encapsulating and transporting messages.

Messages may be communicated across different communication levels:

- Within core modules – allows messages to be passed between partitions supported by the same core module
- Between the core modules – allows messages to be passed between multiple core modules via a communication bus
- Between core modules and a non-ARINC 653 component – allows messages to be passed between core modules and devices that do not host an ARINC 653 O/S (traditional LRUs, sensors, etc.) via various communication buses.

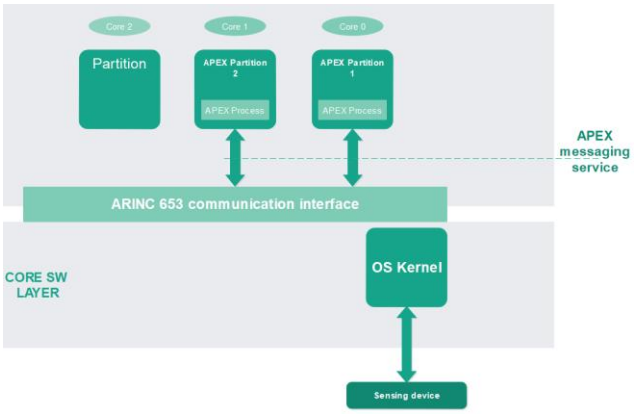


Figure 2: APEX Interpartition Communication

Communications between partitions, or between partitions and external entities, use the same services and are independent of the underlying transport mechanism. The core module is responsible for encapsulating and transporting messages.

F. Intrapartition Communication

Three communication mechanisms are available for intrapartition communication. The first mechanism allows inter-process communication and synchronization via partition-defined buffers and blackboards. The second allows inter-process synchronization via partition-defined counting semaphores and events. The third allows inter-process mutual exclusion via partition-defined mutexes.

1) Buffer Services

A buffer is a communication object used by processes of a same partition to send or receive messages. The maximum number of messages supported and maximum message size is defined at buffer creation.
A buffer is created during the partition initialization phase.

2) Blackboard Services

A blackboard is a communication object used by processes of a same partition to send or receive messages. A blackboard does not use message queues. Each new occurrence of a message overwrites any other.

3) Semaphore Services

A counting semaphore is a synchronization object used to provide access to partition resources.

4) Event Services

An event is a synchronization object used to notify the occurrence of a condition to processes that may be waiting for condition to occur. An event must be created during the partition's initialization phase before it can be used.

5) Mutex Services

A mutex is a synchronization object commonly used to control access to partition resources. Only one process at a time can own a specific mutex.

6) Health Monitor

In general, Health Monitor (HM) functions are responsible for responding to and reporting hardware, application and O/S software errors and failures [5]. ARINC 653 supports HM by providing HM configuration tables and an application level error handler process. Figure 3 illustrates the ARINC 653 HM decision logic. HM services are described in section IV.

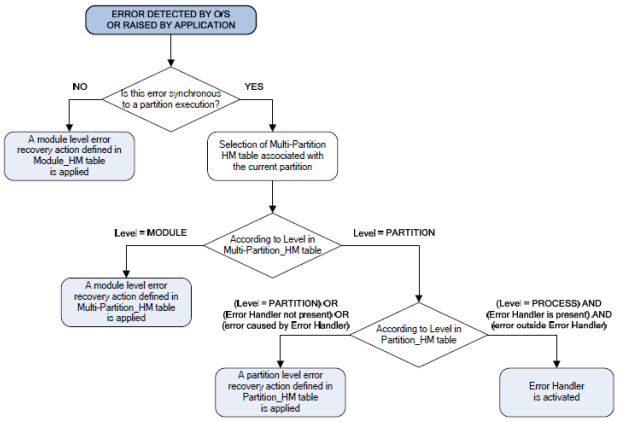


Figure 3: HM Decision Logic

IV. HEALTH MONITORING SERVICES IMPLEMENTATION

The HM is invoked by an application calling the RAISE_APPLICATION_ERROR service or by the O/S or hardware detecting an error. The recovery action is dependent on the error level. The recovery actions for process level errors are defined by the application programmer in a special error handler process.

A. REPORT_APPLICATION_MESSAGE

This service request allows the current partition to transmit a message to the HM function. Code implementation is presented in Figure 3.

```
extern void REPORT_APPLICATION_MESSAGE (
/*in */ MESSAGE_ADDR_TYPE MESSAGE_ADDR,
/*in */ MESSAGE_SIZE_TYPE LENGTH,
/*out*/ RETURN_CODE_TYPE *RETURN_CODE );
```

Figure 3: REPORT_APPLICATION_MESSAGE

B. CREATE_ERROR_HANDLER

This service request creates an error handler process for the current partition. The process has no identifier (ID) and cannot be accessed by the other processes of the partition. It is a special aperiodic process with the highest priority, no deadline, and entry point defined by this service is invoked by the O/S when a detected error is evaluated as a process level error. Figure 4 presents code implementation of this service.

The error handler is written by application programmer. It can stop and restart the failed process with the STOP and START services, restart (COLD_START or WARM_START) or shutdown the partition by the SET_PARTITION_MODE (IDLE) service. The error handler cannot be interrupted or blocked.

```
extern void CREATE_ERROR_HANDLER (
/*in */ SYSTEM_ADDRESS_TYPE ENTRY_POINT,
/*in */ STACK_SIZE_TYPE STACK_SIZE,
/*out*/ RETURN_CODE_TYPE *RETURN_CODE );
```

Figure 4: CREATE_ERROR_HANDLER

C. GET_ERROR_STATUS

This service must be used by the error handler process to determine the error code, the identifier of the faulty process, the address at which the error occurs, and the message associated with the fault. If more than one process is in error, the service may be called in a loop in the error handler until there are no more processes reported as being in error (Code implementation is shown in Figure 5).

```
extern void GET_ERROR_STATUS (
/*out*/ ERROR_STATUS_TYPE *ERROR_STATUS,
/*out*/ RETURN_CODE_TYPE *RETURN_CODE );
```

Figure 5: GET_ERROR_STATUS

D. RAISE_APPLICATION_ERROR

The RAISE_APPLICATION_ERROR (shown in Figure 6) service request allows a running process to invoke the error handler process for the specific error code APPLICATION_ERROR. The error handler process can read the message passed using the GET_ERROR_STATUS. The error handler of the partition is then started (if created) to take the recovery action for the process which raises the error code.

The ERROR_CODE parameter has been retained for backwards compatibility. It can only be set to APPLICATION_ERROR.

```
extern void RAISE_APPLICATION_ERROR (
/*in */ ERROR_CODE_TYPE ERROR_CODE,
/*in */ MESSAGE_ADDR_TYPE MESSAGE_ADDR,
/*in */ ERROR_MESSAGE_SIZE_TYPE LENGTH,
/*out*/ RETURN_CODE_TYPE *RETURN_CODE );
```

Figure 6: RAISE_APPLICATION_ERROR

E. CONFIGURE_ERROR_HANDLER

This (implementation presented in Figure 7) service request allows the current partition to configure the behavior of how process scheduling of other processor cores will be managed for the partition when the error handler process is executing. This service configures whether processes on the other processor cores run concurrently or do not run concurrently (i.e., pause) while the error handler process is running.

```
extern void CONFIGURE_ERROR_HANDLER (
/*in */ ERROR_HANDLER_CONCURRENCY_CONTROL_TYPE CONCURRENCY_CONTROL,
/*in */ PROCESSOR_CORE_ID_TYPE PROCESSOR_CORE_ID,
/*out*/ RETURN_CODE_TYPE *RETURN_CODE );
```

Figure 7: CONFIGURE_ERROR_HANDLER

Health monitoring error code, status and control types are presented in Figure 8.

```
typedef
enum {
DEADLINE_MISSED = 0,
APPLICATION_ERROR = 1,
NUMERIC_ERROR = 2,
ILLEGAL_REQUEST = 3,
STACK_OVERFLOW = 4,
MEMORY_VIOLATION = 5,
HARDWARE_FAULT = 6,
POWER_FAIL = 7
} ERROR_CODE_TYPE;

typedef
struct {
ERROR_CODE_TYPE ERROR_CODE;
ERROR_MESSAGE_SIZE_TYPE LENGTH;
PROCESS_ID_TYPE FAILED_PROCESS_ID;
SYSTEM_ADDRESS_TYPE FAILED_ADDRESS;
ERROR_MESSAGE_TYPE MESSAGE;
} ERROR_STATUS_TYPE;

typedef
enum {
PROCESSES_PAUSE = 0,
PROCESSES_SCHEDULED = 1
} ERROR_HANDLER_CONCURRENCY_CONTROL_TYPE;
```

Figure 8: HM types

V. CONCLUSION

The following objectives are achieved with the ARINC 653 APEX interface:

1. Portability: The ARINC 653 APEX interface facilitates portability of the software. ARINC 653 APEX interface specification is language and hardware independent.
2. Reusability: The ARINC 653 APEX interface allows the production of reusable application code for IMA systems. This interface will reduce the amount of customizing required when a component is reused.
3. Modularity: The ARINC 653 APEX interface provides the benefits of modularity when developing application software.
4. Integration of Software of Multiple Criticalities: The ARINC 653 APEX interface supports the ability to collocate application software of different levels of criticality.

The set of ARINC 653 APEX API provide handy and easy interface across the application and the underlying operating system and enhances developer productivity, reduce complexity and effort involved in certification.

REFERENCES

[1] Airlines Electronic Engineering Committee, “Avionics Application Software Standard Interface Set, Part 0”, ARINC SPECIFICATION 653, SAE-ITC 16701 Melford Blvd., Suite 120, Bowie, Maryland 20715 USA, 2015.

[2] Paul Parkinson, Larry Kinnan, “Safety-Critical Software Development for Integrated Modular Avionics”, Wind River Systems, 2007

[3] Airlines Electronic Engineering Committee, “Avionics Application Software Standard Interface Set, Part 2”, ARINC SPECIFICATION 653, SAE-ITC 16701 Melford Blvd., Suite 120, Bowie, Maryland 20715 USA, 2015

[4] Airlines Electronic Engineering Committee, “Avionics Application Software Standard Interface Set, Part 1”, ARINC SPECIFICATION 653, SAE-ITC 16701 Melford Blvd., Suite 120, Bowie, Maryland 20715 USA, 2015

[5] Slawomir Samolej, “ARINC Specification 653 Based real-Time Software Engineering”, e-Informatica Software Engineering Journal, Volume 3, Issue 1, 2009

RAID 0 on paired magnetic disk arrays

Nikola Davidović, Borislav Đorđević, *Member, IEEE*, Valentina Timčenko, *Member, IEEE*,
Slobodan Obradović, Bojan Škorić

***Abstract* – The connection of multiple secondary memory devices in RAID 0 aims to improve three performance parameters of the secondary memory system: greater storage capacity, the increase of the read data access speed and increase of the write data access speed. This paper analyses the system for storing the data on the paired arrays of magnetic disks in RAID 0 formation, with the number of queue entries for overlapped I/O, where queue depth parameter has the value of 4. The paper presents a range of the testing results and analysis for RAID 0 series for defined workload characteristics. The tests were done in the Microsoft Windows Server 2008 R2 Standard operating system, using 2, 3, 4 and 6 paired magnetic disks and controlled by Dell PERC 6/i hardware RAID controller. For the needs of obtaining the measurement results, we have used the ATTO Disk Benchmark. We have analyzed the obtained results and compared to the expected behaviours.**

Key words - HDD; secondary memory; magnetic disk; performances; RAID 0; ATTO Disk Benchmark; Windows.

I. Introduction

The tendency of different user needs for the increase of the exchanged amount of data has implicated the necessity for providing larger storage space, as well as for the development of faster access and easier management of data. Such trends have led to an inevitable increase in the risks related to the loss or/and compromise of data security. When it comes to the performance of the secondary memory, it is usually measured through three components: available capacity, data access speed and reliability while storing and keeping safe the data. Further improvements in the performance of secondary memory can be achieved either by the use of the parallel disks or by the application of some up to date technology.

The storage space is actually not a performance parameter but still can have an impact on the overall system performances. Thus some recent technological improvements have indicated need of considering storage space as one of the important aspects when designing the system storage characteristics. Despite the emergence of new technology in the production of secondary memory devices - the semiconductor memory storage SSD (Solid State Disk) [1], the magnetic drives technology HDD still plays a significant role as secondary memory, primarily thanks to its high capacity and cost per MB [2]. One of the biggest disadvantages of using magnetic disks is the achievable read and write data access speeds.

The RAID (Redundant Array of Inexpensive Disks) represents a solution where the data is stored on the disk architecture that relies on the combination of several physical disks into one logical unit. The goal is to enhance performances and increase storage capacity [3]. The performance enhancement can be achieved by the increase of the speed of read/write data operations, from and to the disk and by the introduction of specific safety improvements (fault tolerance). This methodology strongly relies on the enforcement of data redundancy. The performances and reliability improvement depend on the applied RAID configuration type. RAID technology is defined through seven (7) different levels of data storage on multiple disks, which in such a way organized present a logical storage space. Seven RAID levels, though having their descriptive names in practice, are usually referred to with their numbers [4][5]. Using the standard seven levels of RAID technology, it is possible to realize the so-called "nested" or hybrid RAID. RAID levels can be realized in two ways: hardware and software RAID.

The support for specific RAID levels can be also provided by operating systems, such as with some Windows versions, or even provided on the file system level such as when using Oracle/Solaris ZFS [6][7]. In addition, the software RAID can be also found as a stand-alone application [8].

II. Redundant Array of Inexpensive Disks – type 0

When compared to all other RAID levels, RAID 0 offers the highest degree of storage space used for storing data. In addition, this RAID level provides the best read and write performance, but it does not offer redundancy.

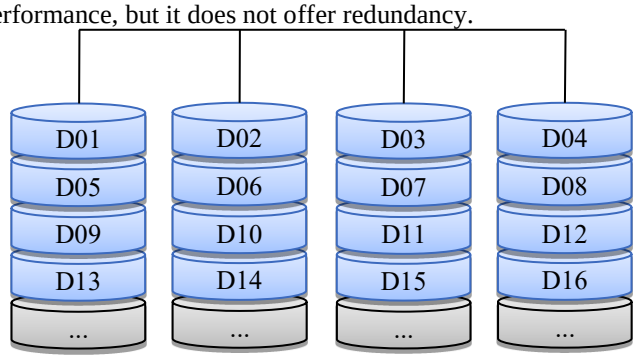


Fig. 1. RAID 0 – apply striping to four secondary memory devices.

Nikola Davidović – University of East Sarajevo, Faculty of Electrical Engineering, Vuka Karadzica 30, 71123 East Sarajevo, RS, Bosnia and Herzegovina, (nikola.davidovic@etf.ues.rs.ba)

Borislav Đorđević – Institute Mihailo Pupin, Volgina 15, 11000 Belgrade, Serbia, (borislav.djordjevic@pupin.rs)

Slobodan Obradović - Information Tehnology School, Belgrade, Serbia, (slobodan.obradovic@its.edu.rs)

Valentina Timčenko - Institute Mihailo Pupin, School of Electrical Engineering, Belgrade, Serbia, (valentina.timcenko@pupin.rs)

Bojan Škorić – VISER, Belgrade, Serbia, (bojan.skoric@viser.edu.rs)

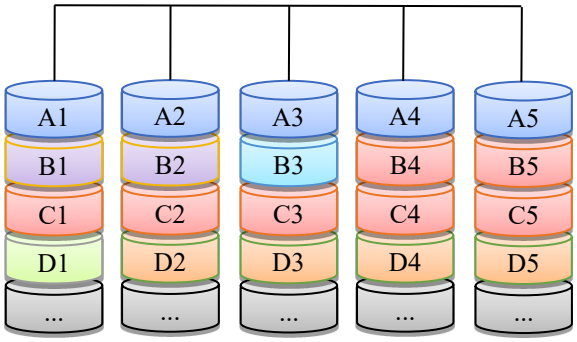


Fig. 2. RAID 0 – parallelity and competitiveness – each color is another complete piece of information

Despite the disadvantages it has, and thanks to its unmatched performance, RAID 0 is used in systems where data access speeds and storage space size play the key role. Figures 1 and 2 provide an overview of the stripping procedure and methodology applied in RAID 0. Figure 1 explains the principles of the basic stripping technique, while Figure 2 provides more detailed example of stripping several data over five different disks that are structured in RAID 0 architecture. Each exemplar data is provided in different colour, thus we can see that for instance data A is stripped on all the five disks, while data B, being smaller in size, is stripped over two disks from the array.

Ideally, we can consider that RAID 0 with N secondary memory devices (disks), when compared to the case of having just one secondary memory device, brings N times better access time for sequential and random data reading and writing. It is described with formulas 1 and 2 [9].

$$T_{seq\_rw}^{RAID-0} \approx T_{seq\_rw\_one\_seq\_m\_d}/N \quad (1)$$

$$T_{random\_rw}^{RAID-0} \approx T_{random\_one\_seq\_m\_d}/N \quad (2)$$

The practical application of the “striping by blocks”, which is typical characteristic of the RAID 0 level, works quite similar to the ideal model but still shows some inconsistencies. The Figure 1 shows procedure of storing data by dividing it and storing on 4 secondary memory devices.

The data writing or reading from the secondary memory device may have smaller or larger size than the defined block (which is essentially the data carrier). Regardless of the size of the data in the block, the secondary memory device accesses the entire block. Therefore, block size estimation is of great importance when designing RAID system. The estimation of the block size has an impact on memory space utilization and plays a significant role in maximizing the RAID performances.

When estimating the size of a block, it is necessary to take into the account the parallelism and competitiveness (Figure 2). In case when the block size is determined so that the data unit exactly occupies the defined disk memory unit (on the RAID full stripe), we can expect the increase of the speed of data access N times the number of secondary memory devices. When configured in this way, RAID 0 supports parallelism and high sequential performance. However, if the goal is to increase the competitiveness and the random access speed performance, then the size of the SU (Strip Unit) needs to be adjusted exactly to the size of the data unit.

III. Test Configuration

A. Hardware configuration

The hardware configuration is shown in Table I. The tests are carried on by the Microsoft Windows Server 2008 R2 Standard operating system. No server components or functions are added to the basic installation of the operating system, except for the necessary drivers - the RAID controller and other specific hardware drivers.

TABLE I
HARDWARE CONFIGURATION

Hardware	Specification
Server	Dell PowerEdge™ T610
RAM	8 GB, 4 x 2 GB DDR3-SDRAM
CPU Model	Intel Xeon E5530 @ 2,40 GHz (4 cores)
BIOS	Dell Inc. v.2.2.10 (9.11.2010.)
Video adapter	Matrox G200
PCIe x4 storage slot	Dell PERC 6/i
Disk	Hitachi Deskstar 250GB SATA2 x 6
PCIe x4 slot	Dell SAS 5/i
Disk	Hitachi Ultra Star 300GB SAS
OS	Microsoft Windows Server 2008 R2 Standard

The hardware RAID controller, Dell PERC 6/i, which is used for test procedures, supports devices with the second generation SATA/SAS interface (3Gb/s), while 2 SAS channels allow up to 32 connected devices. It has 256MB of its own DDR2 cache for quick storage, which can be optionally supported by a battery. It supports operation with RAID levels 0, 1, 5, 6, 10, 50 and 60. The ATTO Disk Benchmark is used for the needs of testing the impact of the RAID 0 storage size on the overall system performances.

B. ATTO Disk Benchmark

ATTO Disk Benchmark is a freeware software which helps the measurement of the storage system performance [10]. ATTO identifies the performance levels of the hard drives, solid state drives, RAID arrays, as well as the of the host connection to the attached storage. One of advantages of this benchmark is the ability to control the proces of writes and reads, while the drawback is the inability to test the random data access speed. The ATTO Disk Benchmark is compatible with Microsoft Windows and supports the File Allocation Table (FAT) and New Tehnology File System (NTFS).

Some of the setting options over which ATTO Disk benchmark can affect system performance or can isolate certain situations in practical work, are:

- *Total lenght* – this parameter specifies the test file lenght, which is the total size of data file that is created on the test drive. After finishing the testing procedure, this file is deleted.

- *Force write access* – this option allows to bypass the drive write cache. Otherwise, if this option is not selected, the drive write caching is determined by the drive settings.
- *Direct I/O* – use of the system buffering. File I/O on the test drive is performed with no system buffering or caching.
- *I/O Comparison* – compares the input and output data to detect errors. This option allows the comparison of the data from the test file with the data written on a per block basis.
- *Overlapped I/O* – this option performs queued I/O testing. The factor that specifies the I/O overlapping is the queue depth. Queue depth specifies the number of queue entries for overlapped I/O, i.e. the maximum number of read/write commands that can be executed during one time interval.
- *Neither* – do not perform overlapped I/O or I/O comparisons. The transfer requestes are sent one by one.

IV. TEST RESULTS

Certain restrictions were set in order to get as highest speed as possible for reading and writing during the test procedures.

The first limitation that is set, is to use only a specific part of the magnetic disk for testing, since magnetic discs do not have the same data transfer speed at the beginning and at the end of the disk. It is configured to use only the first 10 GB of each magnetic disks of the RAID 0 array. In this way, in the case for 2, 3, 4 and 6 magnetic disks, an entry space of 20, 30, 40 and 60 GB is obtained, respectively. Since these 10 GB make up less than 5% of the disk space, the limitations of the data rate at the beginning (in the middle of the disk) and at the end (disk circumference) of the magnetic disk has been avoided during testing.

Data caching feature could give wrong results so that they would not show the real performance of the magnetic disk itself but the performance of the cache. In addition, a significant effect may also be achieved by caching at the level of a single magnetic disk, controller or operating system itself. Because of this, when configuring each array, we have used the option to bypass the cache of the disks, as well as to generate the caching on the controller itself. In ATTO disk benchmark this is enabled by the Force Write Access and Direct I/O options, and represents the second limitation.

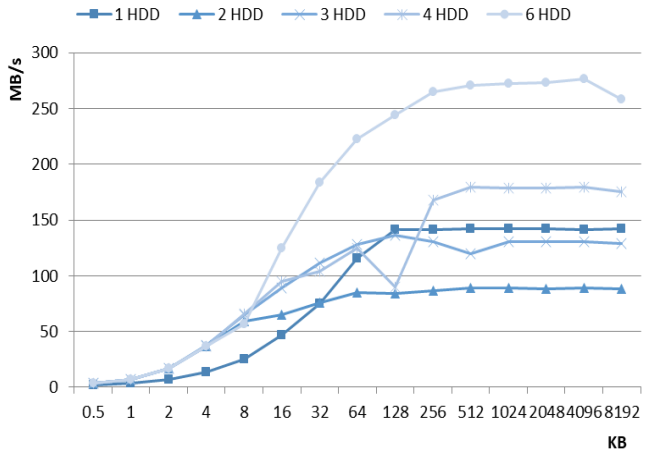
The third limitation is the number of multiple transfer requests which define the maximum number of read/write commands that can be executed in one time interval. The queue depth factor specifies the number of queue entries for overlapped I/O. In this way, the ability to test competitiveness is not eliminated.

In the following test we have used 1GB size NTFS partition, as the test file space was limited to 512 MB. The larger file was selected in order to get the better average values for large transfers, which can also be assumed as the sequential data access test. For the allocation unit we have used the standard size of 4kB. When testing the RAID 0 string, the three block sizes were used:

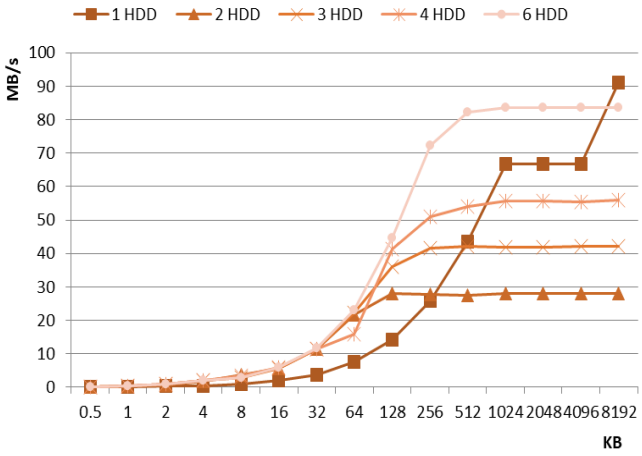
- 8kB, the smallest block that the controller supports;
- 64kB, default value;
- 1MB, the largest block that the controller supports.

TABLE II
TEST RESULTS

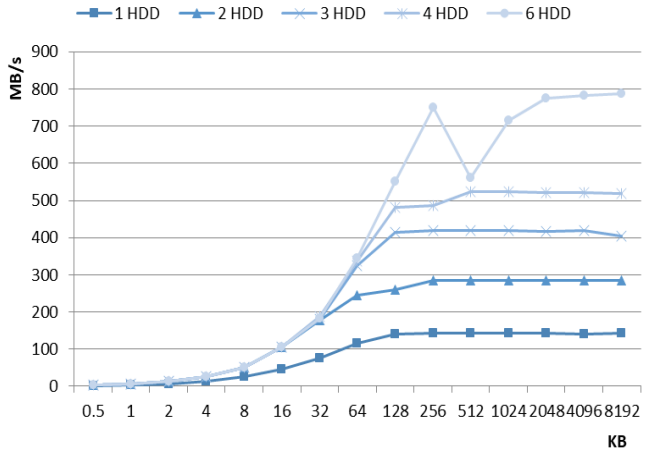
	SU	MB/s	0.5 KB	1 KB	2 KB	4 KB	8 KB	16 KB	32 KB	64 KB	128 KB	256 KB	512 KB	1024 KB	2048 KB	4096 KB	8192 KB
READ	8 KB	1 HDD	1670	3471	6925	13245	25356	46923	74642	115992	141214	141669	142056	141841	141841	141096	142217
		2 HDD	3371	7296	16811	37145	59362	65209	76204	85111	84200	86659	88712	88885	88592	88738	87867
		3 HDD	3278	7260	16646	37328	65048	89219	111890	127875	136593	130745	119974	130944	130625	130308	128900
		4 HDD	3278	6759	17024	36499	65935	95209	103898	125128	90187	167508	179249	178362	178659	179555	175735
		6 HDD	3363	7207	17235	37145	56917	124792	183312	222508	243944	264628	270852	272062	272985	276737	258732
	64 KB	1 HDD	1670	3471	6925	13245	25356	46923	74642	115992	141214	141669	142056	141841	141841	141096	142217
		2 HDD	3303	6622	13050	26360	51447	105441	178028	245227	260087	284837	284346	284058	284560	284560	284058
		3 HDD	3215	6793	12987	26360	51697	106594	180215	323459	414686	419900	418607	418612	417798	419430	404422
		4 HDD	3404	6509	13278	25472	51200	104407	186434	337692	480534	485204	523259	522502	521233	521233	518715
		6 HDD	3041	6557	13312	25661	51200	106230	183369	344329	551708	749682	559848	715827	776198	781850	787585
	1 MB	1 HDD	1670	3471	6925	13245	25356	46923	74642	115992	141214	141669	142056	141841	141841	141096	142217
		2 HDD	3242	6368	12892	24794	46180	80511	129134	187122	255086	283026	284346	284058	282563	284058	284058
		3 HDD	3011	6509	13019	25036	45732	79533	128817	184957	260087	406237	412940	389036	416197	416987	415374
		4 HDD	3223	6098	13147	25098	46293	80511	128187	188446	263969	418941	516925	514984	528936	528936	527637
		6 HDD	3176	6446	13050	24914	46293	79921	128187	188803	262014	409793	511966	697234	766958	778073	778073
WRITE	8 KB	1 HDD	61	121	245	487	974	1918	3819	7447	14185	25826	43509	66858	66692	66774	91147
		2 HDD	196	441	660	1900	3792	5738	11358	21593	27947	27710	27623	28093	28035	28167	28137
		3 HDD	194	414	966	2022	2904	5748	11397	22253	35959	41487	42077	41975	41975	42074	42173
		4 HDD	187	398	894	1918	3515	5789	11437	15791	41217	51000	54050	55634	55749	55461	56099
		6 HDD	195	421	829	1969	2978	5840	11619	22954	44582	72415	82176	83624	83755	83755	83624
	64 KB	1 HDD	61	121	245	487	974	1918	3819	7447	14185	25826	43509	66858	66692	66774	91147
		2 HDD	184	375	765	1606	3585	6742	14340	21881	38607	92794	152465	180764	180764	182920	186090
		3 HDD	179	371	757	1575	3323	7602	14928	22215	42625	80166	167375	267766	275789	273913	275318
		4 HDD	176	336	740	1551	3218	6313	13347	25954	43115	81537	160781	290200	354760	371965	371965
		6 HDD	171	169	750	1525	3162	6986	15312	23116	44506	84836	156052	302746	460833	536870	545600
	1 MB	1 HDD	61	121	245	487	974	1918	3819	7447	14185	25826	43509	66858	66692	66774	91147
		2 HDD	181	358	718	1437	2899	5789	11457	22329	77710	107216	137100	155389	133549	282068	283060
		3 HDD	180	363	731	1432	2810	5748	11260	22140	78800	103614	158145	226050	305619	238080	412997
		4 HDD	181	366	723	1468	2925	5820	11457	22637	82091	105916	152684	243478	268883	267543	512525
		6 HDD	184	367	735	1468	2920	5563	11338	22520	85296	105278	149669	233422	387166	512525	517465



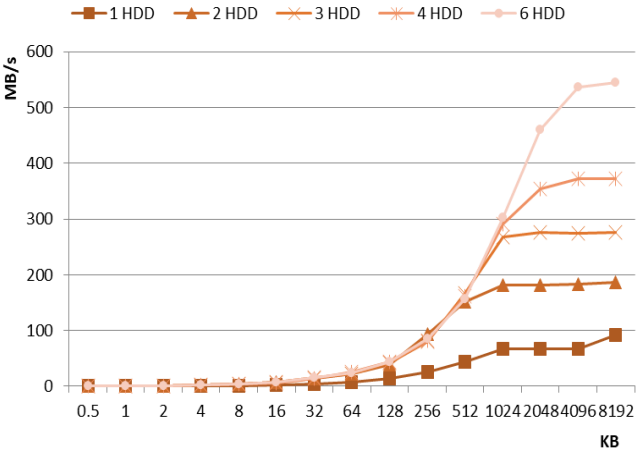
a1)



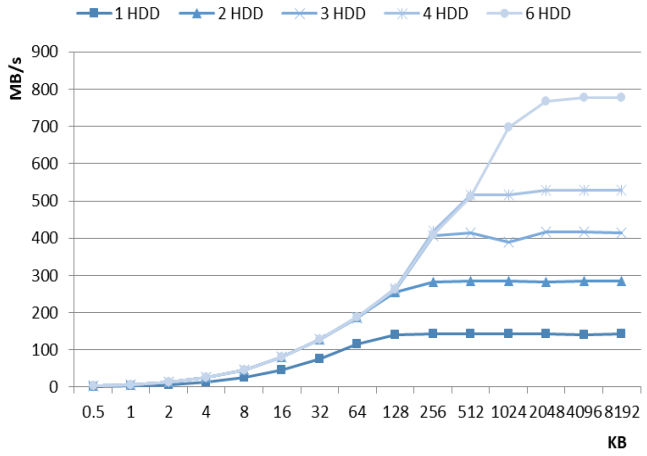
b1)



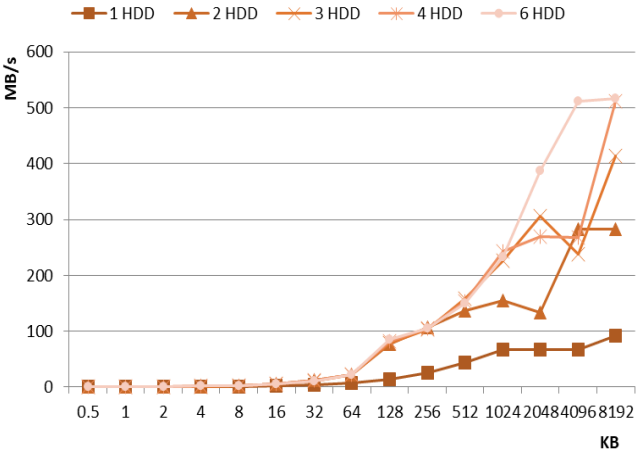
a2)



b2)



a3)



b3)

Fig. 3. Reading speed a) and writing speed b) data for single magnetic disk and paired string of 2, 3, 4 and 6 magnetic disk drivers in RAID 0 at the blok size 1) KB, 2) 64 KB and 3) 1 MB

The values of SUs in the Table II are given in the first column. In the second column of the Table II we have presented the number of magnetic disks on which the tests were performed for different SUs. The test procedure starts with the evaluation of 512 bytes and ends with 8192 KB, each step dobles the data size from the previous iterration.

Figure 3.a1 shows the read speed from a single magnetic disk and for the paired array of magnetic disks RAID 0. We have used the 8 KB block size and different amounts of data, while the measurement results are shown in the Table II. The RAID 0 with two magnetic disks has provided the expected results for small amounts of data, and in accordance with formulas (1) and (2). For the small amounts of data the other RAID 0 configurations (with 3, 4 and 6 magnetic disks) have also performed with better results than the single disk, but with the queue depth factor limitation. This is one of the reasons for slighther improvements of these configurations over the RAID 0 with 2 disks.

The figure 3.a1 and Table II show better results for RAID 0 with 3, 4 and 6 discs, when compared to RAID 0 with 2 disks, which appears with the raise of the data transfer size above 8KB. The queue depth factor effect, the stripe unit size and the data size had the mayor impact to the obtained results.

It is notable from Table II and figures 3.a1, 3.a2 and 3.a3 that RAID 0 with 2 paired disks has the best results (ratio MB/s and number of disk) for small data reads, regardless the size of the SU. In addition, it is noticeable that when increasing the SU size, there is a slight data read speed decrease, while for the larger amounts of data there is a significant improvement of the data read speed. RAID effects are noticeable only for the cases where the size of the RAID block is 64KB or 1MB (figures 3.a2, 3.a3, 3.b2 and 3.b3 and Table II). The SU size impact is evident, as for the case of the larger amount of data the obtained results are much better than in the case of the block size of 8KB.

From figure 3.b1 and Table II we can have similar conclusions as in the case shown on the figure 3.a1, but with more degradation effects due to the fact that all the configured RAID 0 systems have poorer writing performances than the single magnetic disk for the larger amount of data. This behavior is the result of the queue depth factor effect, the size of the SU and the data size. In booth cases shown on the figures 3.a1 and 3.b1, the depth factor effect can not compensate for the losses caused by a bad selection of the SU size.

When comparing the obtained values (Table II) for the data size of 64KB, the improvements relative to a single magnetic disk during reading, for the maximum amount of data transferred, are 200%, 284%, 364% and 554%. For writing operation we have gained 204%, 302%, 407 % and 599% for 2, 3, 4, and 6 paired magnetic disks in RAID 0, respectively. Similar but somewhat minor improvements are obtained for 1 MB block size. When comparing the results for the block size of 64KB and 1MB (Table II), it is noticeable that the increase to the 1MB block size does not significantly improve the performances when compared to the improvements gained with 64kB block size increase. The performance gain is approximately equal when comparing read and write speeds for blocks of the size 64KB and 1MB. Although the improvements for a block size of 64KB and 1MB are better when compared to the single magnetic disk, as well as with RAID 0 with a block size of 8 KB, it is noticeable that they are smaller than in the ideal case (formula 1).

In all the performed tests, it is noticeable that the writing performance results are far worse than for reading. One of the reasons for these results is that during the measurement, the force write access option was applied. It obviates the use of caching on a disk or controller when writing data, while reading the effects of caching are expressed.

V. CONCLUSION

In this paper we have analysed the RAID 0 data storage system on paired magnetic disk arrays. The testing results have shown that the RAID 0 performance depends on four factors: queue depth, amount of data, SU size and the number of disks. Inadequate block size selection for a RAID 0 string can significantly degrade the performance of the system. In the case of the adequate configuration, RAID 0 shows a direct gain in performance with an increase of the number of disks (formulas 1 and 2). The performance gain is approximately the same for both data reading and writing operations. By analysing the obtained results, it is possible to conclude that when using RAID 0 series it is possible to achieve better performances for reading and writing operations for the semiconductor disks, but there is also the issue of increased risk of data loss.

ACKNOWLEDGEMENT

The work presented in this paper has partially been funded by the Ministry of Education, Science and Technological Development of the Republic of Serbia: V. Timcenko by grants TR-32037, TR -32025, and B. Djordjevic by grant III-43002.

LITERATURE

[1] R. Micheloni, A. Marelli, K. Eshghi, Inside Solid State Drives (SSDs), Springer, Heidelberg, 2013

[2] [DD Modeling] Ruemmler, Chris, and John Wilkes. "An introduction to disk drive modeling." Computer 27.3 (1994): 17-28.

[3] McDonald, James Arthur, et al. "Disk array system for processing and tracking the completion of I/O requests." U.S. Patent No. 6,098,114. 1 Aug. 2000.

[4] Patterson, David; Gibson, Garth A.; Katz, Randy, “A Case for Redundant Arrays of Inexpensive Disks (RAID)”, www.eecs. berkeley.edu/Pubs/TechRpts/1987/CSD-87-391.pdf, december 2018.

[5] William Stallings, “Computer Organization and Architecture ”, ISBN 978-86-7991-361-6

[6] Oracle Solaris ZFS, https://docs.oracle.com/cd/E18752_01/html/819-5461/gavwn.html, february 2019.

[7] Microsoft storage, https://docs.microsoft.com/en-us/windows-server/storage/storage-spaces/storage-spaces-fault-tolerance, decembar 2018.

[8] Software RAID, https://www.softraid.com/pages/features/software_raid_benefits.html, december 2018.

[9] V. Timcenko, B. Djordjevic, “The comprehensive performance analysis of striped disk array organizations - RAID-0, ” invited paper, in proc. of Proceedings of the 2013 International Conference on Information Systems and Design of Communication, Lisbon, Portugal, 2013.

[10] ATTO benchmark, https://www.atto.com/disk-benchmark/, february 2018.

Ažuriranje Android operativnog sistema upotrebom Push VoD tehnologije

Miloš Ivanković, dr Ilija Bašičević, MSc Goran Stupar

***Apstrakt*—U ovom radu prezentovano je jedno rešenje ažuriranja korisničkih Set-Top Box uređaja zasnovanih na Android operativnom sistemu u slučaju ograničenog pristupa ili potpune nemogućnosti pristupa internet mreži. Kao specifična okolnost koja se može javiti u slučaju distribucije digitalnog televizijskog signala, ovom prilikom je rešena prenosom podataka neophodnih za ažuriranje uređaja putem prenosnog toka kojim se distribuira digitalni televizijski signal. Ovaj rad predstavlja funkcionalno rešenje primenljivo i u drugim sličnim implementacijama obrade prenosnog toka digitalnog televizijskog signala.**

***Ključne reči*—Digitalna televizija (DTV), Android OS, Push VoD (video sadržaj na zahtev), Set-Top Box (STB)**

I. Uvod

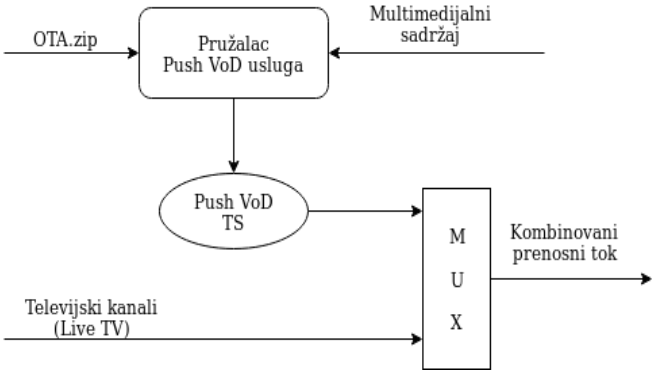
Mnogi moderni STB uređaji zasnovani su na Android platformi. Periodično ažuriranje platforme ključno je za ispravno funkcionisanje uređaja. Novim izmenama sistema ispravljaju se greške u radu, unapređuje sistem zaštite, unapređuje stabilnost uređaja u celini a ovde su uključene i izmene u domenu grafičke korisničke sprege. Naročito su od velikog značaja izmene sistema zaštite koje vremenom postaju sve podložnije napadima.

Posebno je od značaja mogućnost ažuriranja udaljenih uređaja odnosno uređaja koji se već nalaze u upotrebi. Sa razvojem i rastom popularnosti pametnih prenosivih uređaja poput mobilnih telefona i tableta raste potreba za redovnom nadogradnjom postojećih uređaja. Naročito popularna tehnologija u ovom domenu jeste OTA (eng. Over-the-air). Osnovna ideja jeste distribucija specijalizovanih OTA paketa podataka koje u sebi sadrže izmene koje treba primeniti. Standardni način za pribavljanje ovakvih paketa jeste posredstvom internet konekcije kojom se korisniku podaci dostavljaju. Međutim u slučaju digitalne televizije može se javiti specifična situacija u kojoj u određenoj regiji postoji dovoljno dobra razvijena televezijska ali ne i internet infrastruktura. U tom slučaju krajnji korisnici STB uređaja mogu uživati u konzumiranju A/V sadržaja međutim imaju ograničen pristup internetu ili pak uopšte ne postoji mogućnost priključenja na internet. Tada je rešenje moguće naći u integraciji OTA paketa sa standardnim televizijskim signalom. Ovakav pristup zahteva prilagođavanje svih slojeva programskog koda koji se bave obradom digitalnog televizijskog signala, kao i izmene strukture samog digitalnog signala opisanih u narednim poglavljima.

Push Vod predstavlja mogućnost dobavljanja multimedijalnog sadržaja na izričit korisnički zahtev. Pored toga korisnik se dostavljaju i dodatne informacije vezane za sadržaj pa je tako korisnik u mogućnosti da sazna nešto više o sadržaju što je od posebnog značaja ukoliko se od korisnika zahteva dodatna naplata ovakve usluge.

II. Arhitektura rešenja

Posao stvaranja prenosnog toka poveren je specijalizovanim kompanijama koje se bave isporukom televizijskog sadržaja. Prenosni tok u sebi sadrži kako sadržaj koji se emituje u realnom vremenu, tako i sadržaj koji je predmet usluge Video na zahtev (eng. Video On Demand). Ranije pomenuti OTA paketi se dostavljaju ovakvim kompanijama koje ih potom uključuju u transportni tok koji korisnik može preuzeti. OTA paketi se u transportni tok uključuju zajedno sa sadržajima namenjenim preuzimanju na zahtev. Na narednom dijagramu dat je sažet pregled ovakvog rešenja.



Sl. 1. Šematski prikaz formiranja prenosnog toka koji se isporučuje korisnicima

Trasportni tok nastao kombinovanjem multimedijalnog sadržaja namenjenog preuzimanju na zahtev i OTA podataka na dijagramu je označen kao Push Vod TS. Dalje se on kombinuje sa živim televizijsim signalom odnosno televizijskim sadržajem namenjenim emitovanju u realnom vremenu. Ova dva toka se potom kombinuju upotrebom namenske specijalizovane jedinice nazvane multiplekser. Kao

Miloš Ivanković – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail: ivankovicm02@gmail.com).

Dr Ilija Bašičević – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail: ilibas@uns.ac.rs).

Msc Goran Stupar – Istraživačko-razvojni Institut RT-RK, Novi Sad, Srbija (e-mail: goran.stupar@rt-rk.com).

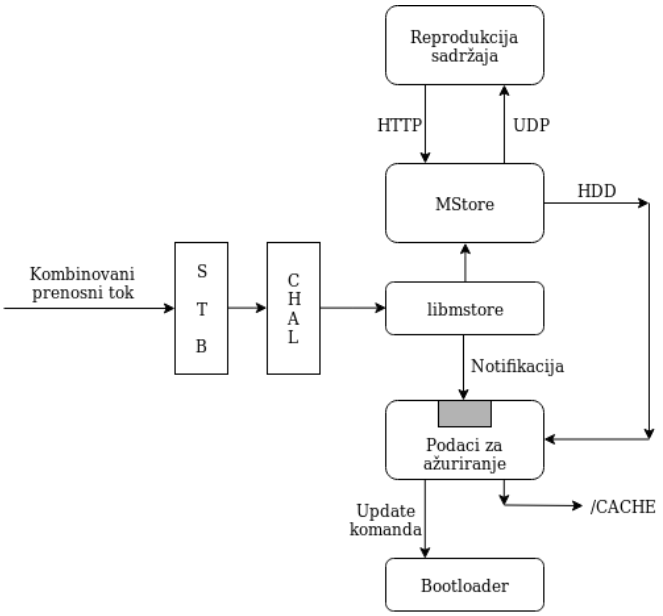
izlaz ovog modula (bez obzira da li je hardverski ili softverski implementiran) dobijamo kombinovan prenosni tok koji se potom isporučuje krajnjem korisniku.

Međutim ažuriranje nije potrebno obavljati svaki put kada korisnički uređaj u prenosnom toku prepozna OTA sadržaj nego se korisniku dodatno signalizira trenutak kada ažuriranje treba izvršiti. Kako se unutar prenosnog toka zajedno sa audio i video sadržajem za reprodukciju prenose i dodatni podaci vezani za televizijski sadržaj u vidu specijalizovanih tabela podataka, to znači da je dozvolu za ažuriranje moguće dostaviti upotrebom standardom predviđenih polja već postojećih tabela. DVB standard jasno propisuje koja polja mogu činiti jednu tabelu kao i redosled pojavljivanja svih polja u tabeli. Pri tome su pojedina polja označena kao opciona te se samim tim ne moraju koristiti prilikom isporuke prenosnog toka. Tako na izlazu dobijamo prenosno tok podataka koji se od do sada korišćenog razlikuje u tome što sadrži dodatne podatke u vidu OTA paketa kao i dodatne signalizacione informacije postavljenje u opciona polja tabela dodatnih podataka. Svi oni sačinjavaju prenosni tok koji se isporučuje korisniku.

STB uređaj poseduje namenski prilagodni programski sloj čiji je zadatak prihvat, obrada i dalja distribucija digitalnog prenosnog toka. Sastoji se od tri celine i to: sloja zaduženog za komunikaciju sa fizičkom arhitekturom uređaja, sloja za obradu digitalnog prenosnog toka i na kraju sloja zaduženog za komunikaciju sa Android platformom. Najniži sloj zadužen za komunikaciju sa fizičkim komponentama zadužen je za prihvat Push VoD sadržaja ali i ostalog digitalnog sadržaja (filtriranje metapodataka, audio i video komponenata prenosnog toka). Ovo je moguće zbog toga što se prenosni tok sadrži od niza paketa jedinstveno označenih nizom identifikacionih brojnih vrednosti. Tako je moguće iz prenosnog toka dobaviti samo pakete od interesa. Prikupljanje delova OTA paketa kao i formiranje jedinstvene datoteke povereno je korisničkoj biblioteci libmstore. Načelno ova korisnička biblioteka ima za zadatak isporuku audio i video sadržaja dobavljenog na zahtev koji se potom isporučuje korisničkoj aplikaciji. Ovaj sadržaj se potom može reprodukovati na zahtev korisnika. Proširenjem ovog korisničkog servisa omogućen je prihvat OTA podataka. Jedan od osnovnih zadataka pomenute korisničke biblioteke jeste i slanje obaveštenja višim programskim slojevima o uspešnom prispeću Push VOD sadržaja u trenutku kada je sav digitalni sadržaj preuzet iz prenosnog toka. Ovaj mehanizam je proširen time što sada postoji mogućnost slanja obaveštenja o prihvatu OTA paketa.

Ranije je pomenuto da se korisničkom STB uređaju posredstvom tabela sa metapodacima signalizira trenutak kada je uređaj potrebno ažurirati. Programski sloj zadužen za komunikaciju sa elementima fizičke arhitekture je zadužen za upravljanje komponentama zaduženim za filtriranje metapodataka prenosnog toka. Među ovim podacima nalaze se brojne standardne DVB tabele (PAT, PMT, SDT, TOT, TDT…). Jedna od ovih tabela zadužena mrežnu identifikaciju namenski je proširena signalizacionim podacima u okviru Linkage descriptor-a na osnovu kojih se odobrava ili ne odobrava ažuriranje. Nakon pristizanja tabela se na osnovu njihovih jedinstenih identifikacionih brojnih oznaka iz prenosnog toka izdvaja NIT tabela. Njen sadržaj se potom iščitava uz pomoć programskog sloja zaduženog za obradu prenosnog toka. Kako je NIT tabela namenski proširena signalizacionim podacima, tako je bilo potrebno namenski proširiti programske module namenjene iščitavanju sadržaja tabela i na taj način omogućio pristup i odredio sadržaj signalizacionih podataka. Pristizanje ove tabele je periodično, što je predviđeno DVB standardom. Kada je ažuriranje dozvoljeno tada se unapred registrovanom metodom javlja ostatku sistema da je ažuriranje neophodno. Nakon toga je dobavljeni OTA paket je potrebno prebaciti u odgovajući keš direktorijum bootloader platformskog alata. Finalni korak predstavlja prosleđivanje komande za ažuriranje bootloader komponenti nakon čega se pokreće proces ažuriranja sistema koji za ulazni parameter ima prethodno pomenuti OTA paket.

Na narednom dijagramu dat je sažet pregled toka obrade na klijentskom STB uređaju.



Sl. 2. Šematski prikaz koraka neophodnih za pokretanje ažuriranja sistema

III. NAMENSKA PROŠIRENJA POSTOJEĆEG REŠENJA

Kao što je ranije pomenuto, do određenih metapodataka dolazimo filtriranjem ulaznog prenosnog toka podataka. Na ovaj način možemo izdvojiti i NIT tabele. Jednom preuzeta, ona se potom prosleđuje programskom modulu namenjenom njenom iščitavanju. Ovaj modul je bilo neophodno proširiti kako bi bile uzete u obzir izmene nastale dodavanjem ranije pomenutih signalizacionih podataka.

Potrebno je bilo implementirati mehanizam kojim bi se drugim programskim modulima moglo javiti da je ažuriranje uređaja dozvoljeno. U ovu svrhu implentiran je niz povratnih poziva koji se aktiviraju svaki put kada signalizacioni podaci pristignu.

mstore predstavlja korisničku biblioteku čija je osnovna funkcionalnost prosleđivanje Push Vod sadržaja Mstore korisničkoj aplikaciji koja ih potom skladišti u trajnoj memoriji uređaja na unapred određenoj lokaciji. Ciljna aplikacija koja je namenjena reproduckiji video sadržaja u ovom slučaju nema pravo pristupa OTA podacima. Kompletna funkcionalnost vezana za rad sa OTA i signalizacionim podacima dodata je ovoj korisničkoj biblioteci.

Prethodno su opisani programski slojevi koji za cilj imaju reagovanje na određene događaje u sistemu kao i slanje odgovarajućih obaveštenja višim programskim slojevima. Stoga je bilo neophodno razviti namenski programski modul koji će predstavljati krajnju putanju ovih obaveštenja. Na prethodnom dijagramu je ovaj blok predstavljen kao blok namenjen skladištenju podataka za ažuriranje i koji ima dvojaku ulogu. Sa jedne strane se nalazi u komunikaciji sa MStore aplikacijom od koje pristižu obaveštenja o tome da je OTA paket pristigao, uspešno sastavljen i sačuvan u unapred definisanoj lokaciji trajne memorije. Komunikacija se odvija putem UDP mrežnog protokola. Po pristizanju obaveštenja, OTA podaci se kopiraju u /CACHE direktorijum bootloader alata. Sa druge strane se opet posredstvom UDP protokola dostavlja obaveštenje o pristizanju dozvole za pokretanje ažuriranja. Ovo za rezultat ima prosleđivanje komande za pokretanje ažuriranja bootaloder alatu. Nakon toga se unapred određenim procedurama operativni sistem ažurira a korisnik je u mogućnosti da nakon automatskog ponovnog pokretanja uređaja koristi novu verziju operativnog sistema.

IV. Zaključak

Moderne implementacije DVB standarda praćene savremenom mrežnom infrastrukturom omogućavaju ažuriranje operativnog sistema korisničkog uređaja posredstvom mrežne infrastrukture. Odstupanje od ovakvog pristupa sa sobom nosi brojne poteškoće s obzirom da podrazumeva rešenja koja nisu deo uobičajene prakse. Izmene obuhvataju veći broj modula raspoređenih u različitim programskim slojevima sistema za distribuciju i obradu digitalnog televizijskog signala, razvijenih u različitim tehnologijama. Implementacija takvog rešenje zahteva saradnju nekoliko ne samo timova nego i kompanija što dodatno otežava implementaciju. Upravo takav jedan problem predmet je uspešne implementacije opisane u ovom radu.

Zahvalnica

Ovaj rad je delimično finansiran od strane Ministarstva za prosvetu, nauku i tehnološki razvoj Republike Srbije, na projektu broj: TR32030.

Literatura

[1] M.Z. Bjelica, N. Teslić, V. Mihić, “Softver u televiziji i obradi slike 1”, FTN, radna verzija, 2016.
[2] Benoit, H. “Digital Television – Satellite, Cable, Terrestrial, IPTV, Mobile TV in the DVB Framework”, Focal Press, 2008.
[3] ETSI EN 302 769 v1.3.1, https://www.etsi.org/deliver/etsi_en/302700_302799/302769/01.03.01_60/en_302769v010301p.pdf, 2015.
[4] Pap dr Ištvan, Lukić dr Nemanja, “Projektovanje i arhitekture softverskih sistema – Sistemi zasnovani na Androidu”.
[5] Android Developers documentation, https://developer.android.com/
[6] Watkinson J. “The MPEG Handbook”, 2001.
[7] Morris S, Smith Chaigneau A. “Interactive TV Standards: A Guide to MHP, OCAP, and JavaTV”, 2005.
[8] Fischer, W. “Digital Video and Audio Broadcasting Technology – A practical Engineering Guide”, Springer-Verlag, 2010.
[9] Richardson, I. E. G. “H.264 and MPEG-4 Video Compression”, Wiley, 2004.
[10] I. Bašičević, M. Popović, V. Kovačević, “Osnovi računarskih mreža 1”, FTN, 2017.

Abstract

In this paper we present one solution for updating user owned Set-Top Box device based on the Android operating system in case of limited access or complete inability to acces the Internet. As a specific circumstance that occurs in the case of the distribution of a digital television signal, on this occasion it is solved by the transmission of the data necessary for updating the device through the tranpost stream used to distribute the contents of the digital television signal. This paper presents a functional solution applicable also in other similar implementations of the processing of the transport stream of the digital television signal.

Android operating system updating with the use of Push VoD technology

Miloš Ivanković, dr Ilija Bašičević, MSc Goran Stupar

Ažuriranje Android baziranog digitalnog TV prijemnika u slučaju onemogućene internet konekcije

Nataša Bogdanović, Ilija Bašičević, Goran Stupar

[1]

***Apstrakt*—Ovaj rad predstaviće jedno od rešenja ažuriranja korisničkih Set-Top Box uređaja zasnovanih na Android operativnom sistemu u uslovima kada je internet konekcija ograničena ili onemogućena. Nedostatak kvalitetne internet konekcije predstavlja jedan od mogućih problema koji se javljaju prilikom distribucije digitalnog televizijskog signala. Rešenje koje je predstavljeno u ovom radu koristi specijalne kontrolne strimove koji sadrže signalne tabele kako bi se kontrolisalo vreme ažuriranja.**

***Ključne reči*—Digitalna televizija (DTV), Android OS, OTA (eng. Over-the-air), Set-Top Box (STB)**

I. Uvod

U današnje vreme, televizija je svakako jedna od tehnologija koje su u velikoj meri pod uticajem tehnološkog razvoja. Kada se govori o tehnološkom razvoju bilo koje tehnologije, to podrazumeva pre svega prolazak određene tehnologije kroz niz procesa promena i unapređenja. Razvoj televizije može da se posmatra kroz mnoge različite aspekte kao što su, na primer, razvoj načina emitovanja, promene i unapređenja u tehnikama emitovanja slike, razvoj prenosnog puta i mnogi drugi. Razvoj televizije je svakako usko povezan sa razvojem televizijskih prijemnika i usmeren je najviše na prilagođavanje zahtevima tržišta koji se sve brže menjaju.

Promena koja je najviše doprinela razvoju i ekspanziji televizije je promena prenosa signala. Pod promenom prenosa signala podrazumeva se prelazak sa analognog na digitalni signal, što uvodi novi pojam u televizijskoj industriji,a to je pojam digitalne televizije. Pojava i razvoj digitalne televizije donosi mnoge prednosti, a samim tim otvara i nove pravce i mogućnosti za dalji razvoj, ali uz to donosi i mnoge probleme na tržištu. Smena analogne televizije digitalnom predstavljala je najveći problem za korisnike čiji prijemnik nije mogao da obradi digitalni signal. Da bi se ovaj problem rešio, televizijska industrija je predstavila novi proizvod, STB (set-top-box) uređaje. Pojava STB uređaja na tržištu omogućila je korisnicima prednosti koje digitalna televizija pruža.

Potreba da se u što većoj meri korisnicima obezbede televizijski sadržaji visokog kvaliteta predstavlja veliki izazov za televizijsku industriju. Prednost digitalnog prenosa podataka se ogleda pre svega u tome što obezbeđuje sliku i zvuk boljeg kvaliteta. Kod digitalnog prenosa više ne postoji problem ometanja interferencijom sa drugim signalima, čak i kod prenosa slike i zvuka na velikom rastojanju. Sadržaji koji se isporučuju putem digitalnog prenosa isti su kao na izvoru emitovanja. Postoji više različitih digitalnih standarda koji se koriste u različitim delovima sveta. Standardi koriste MPEG-2 tehnologiju za audio i video kodiranje i multipleksiranje kako bi se omogućio odgovarajući protok podataka potrebnih za podršku televizije visoke rezolucije (engl. High Definition Television, HDTV) ili televiziju standardne rezolucije (engl. Standard Definition Television, SDTV).

Prijemnik digitalnog televizijskog signala je uređaj koji je zadužen za konverziju digitalnog televizijskog signala da bi isporučeni sadržaj mogao biti prikazan na televizorima. Prijemnik može biti ugrađen ili odvojeni uređaj. Prijemnici koji su odvojeni uređaji nazivaju se STB uređaji. Uloga STB uređaja je pre svega prijem signala, ali razvojem televizijske industrije ovo postaje uređaj koji ima sve više osobina personalnog računara i multimedijalnog uređaja. Najvažnije funkcionalnosti STB uređaja su obrada emitovanog signala, zaštita signala od neovlašćenog pristupa, obrada audio i video podataka, kao i mogućnost interakcije sa korisnikom.

II. Definicija problema

Jedan od problema koji svakako prate razvoj u televizijskoj industiji je i ažuriranje udaljenih uređaja. Pojam udaljeni uređaji se najviše odnosi na uređaje koji su već u korisničkoj upotrebi. Sve brži razvoj u ovoj oblasti donosi potrebu za unapređenjem već postojećih uređaja, kako bi korisnici mogli da dobiju najnovije promene i unapređenja u što kraćem roku. Kao jedno od najvažnijih rešenja ovog problema predstavlja se OTA (eng. Over-the-air) tehnologija. Ova tehnologija podrazumeva distribuciju specijalizovanih OTA paketa podataka. OTA paketi sadrže unapređenja koja treba da se primene na već postojeće uređaje. Najrasprostranjeniji način korišćenja OTA paketa je dobavljanje istih putem internet konekcije. U najvećem broju slučajeva ovakav pristup je moguće i najjednostavnije rešenje, ali postoje i situacije u kojima je potrebno unaprediti određeni uređaj bez prisustva

[1]

Nataša Bogdanović – Naučno-istraživački institut RT-RK, Narodnog Fronta 23a, 21000 Novi Sad, Srbija (e-mail: natasa.bogdanovic@rt-rk.com).

Goran Stupar – Naučno-istraživački institut RT-RK, Narodnog Fronta 23a, 21000 Novi Sad, Srbija (e-mail: goran.stupar@rt-rk.com).

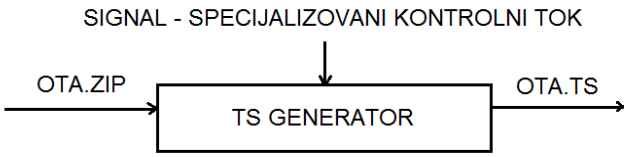
Ilija Bašičević – Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail: ilibas@uns.ac.rs).

internet konekcije. Ovakve situacije mogu da zahvate veliki broj korisnika, pa je veoma važno imati rešenje i u ovakvim okolnostima.

Jedno od rešenja predstavlja integraciju OTA paketa sa standardnim televizijskim signalom. Integracija OTA paketa sa standardnim televizijskim signalom dovodi do problema kontrolisanja vremena ažuriranja uređaja.

III. ARHITEKTURA REŠENJA

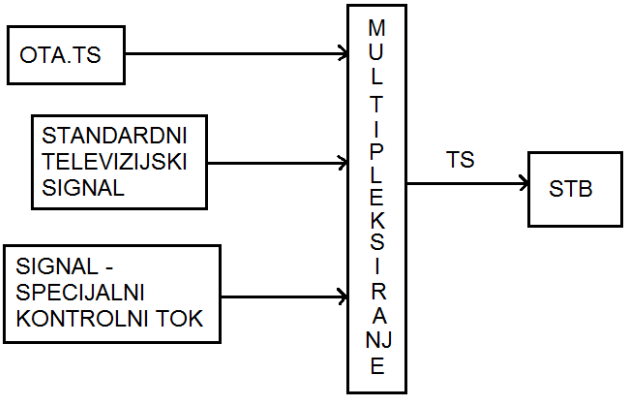
Osnovna ideja rešenja datog problema je upotreba specijalnih kontrolnih tokova koji sadrže signalne tabele. Postoje dva načina da se ovakvi kontrolni tokovi uključe u standardni televizijski signal. Prvi način, koji se koristi samo za testiranje, podrazumeva uključivanje kontrolnog toka prilikom prilagođavanja OTA paketa za integraciju sa standardnim televizijskim signalom. Ovakvo rešenje prikazano je na Slici 1.



Slika 1

Da bi se OTA paket mogao integrisati sa standardnim televizijskim signalom, potrebno je izvršiti konverziju paketa u prenosni tok. Ovo rešenje uključuje specijalizovani kontrolni tok prilikom konverzije. TS generator generise DSM-CC sekcije. DSM-CC (Digital Storage Media – Command and Control) je novi ISO/IEC standard za isporuku multimedijalnih usluga širokog obima. DSM-CC tabele u sebi sadrže komande koje se šalju prijemniku. Ovaj protokol koristi klijent-server model komunikacije i koristi se za kontrolu prijema signala. Protokol između ostalog sadrži uobičajene funkcije ubrzanog premotavanja, pauziranja i sl. Kao što je već navedeno, ovo rešenje se koristi samo za testiranje, pa je samim tim drugo rešenje mnogo značajnije jer ima širu upotrebu.

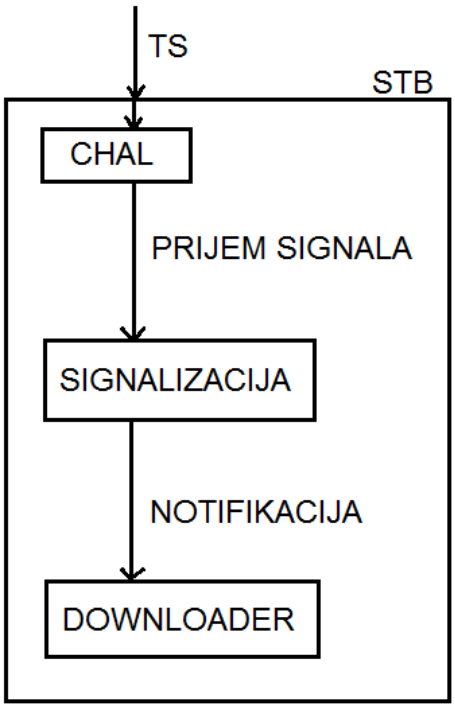
Drugo rešenje navedenog problema podrazumeva uključivanje specijalizovanog kontrolnog toka za signalizaciju prilikom multipleksiranja. Multipleksiranje je proces kombinovanja MPEG-2 elementarnih tokova u jedinstveni informacioni tok u skladu sa definisanim formatom. Multipleksiranje se koristi za integraciju OTA.TS sa standardnim televizijskim signalom i u ovom slučaju sa specijalnim kontrolnim tokom za signalizaciju. Kao rezultat multipleksiranja dobija se jedinstveni prenosni tok koji će da prima prijemnik na STB. Ovaj process je prikazan na Slici 2.



Slika 2

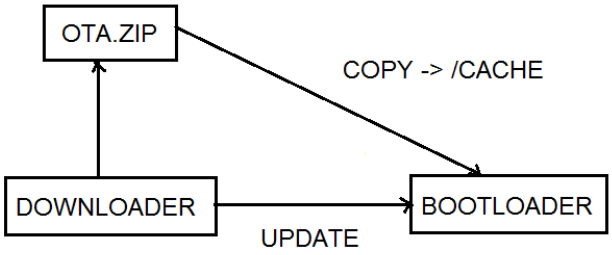
Kada prenosni tok generisan na ovakav način stigne do STB prijemnika, potrebno je obezbediti odgovarajuće rešenje za primenu dobijenih podataka. Kako ažuriranje nije potrebno uraditi uvek kada uređaj prepozna OTA paket, specijalni kontrolni tokovi koji su uključeni u prenosni tok će signalizirati korisniku kada je potrebno izvršiti ažuriranje uređaja. Korisnički STB urežaj u sebi sadrži programski sloj koji prima i vrši dalju obradu i distribuciju primljenog prenosnog toka. Najniži programski deo – CHAL koristi se za komunikaciju sa fizičkom arhitekturom STB uređaja i služi za prijem digitalnog sadržaja u vidu prenosnog toka. Kako je prenosni tok sastavljen od niza paketa koji su identifikovani jedinstvenim brojnim vrednostima, moguće je pristupiti konkretnim paketima iz prenosnog toka koji su potrebni. Nakon prijema paketa dalje se obavlja obrada i distribucija određenih podataka.

U ovom konkretnom slučaju, kada posmatramo OTA paket koji je primljen putem prenosnog toka, treba dobaviti informaciju o tome da li je ažuriranje potrebno. Za to su zaduženi specijalni kontrolni tokovi koji su takođe poslati putem prenosnog toka. Oni sadrže signalne tabele u kojima se nalaze informacije o tome da li je ažuriranje potrebno i kada. Ukoliko stigne informacija da je ažuriranje potrebno, uređaj će primiti signal i poslati notifikaciju posebnom programskom sloju prilagođenom za dobavljanje podataka – DOWNLOADER. Slika 3 prikazuje ovaj proces, od prijema podataka na korisnički STB uređaj do slanja notifikacije za ažuriranje. Slika 3 predstavlja nastavak na process prikazan na Slici 2.



Slika 3

Downloader služi za preuzimanje OTA paketa za ažuriranje koji je poslat preko prenosnog toka. Koristi DSM-CC protokol koji se sastoji od data carousel i object carousel dela. Data carousel se koristi za slanje podataka preko prenosnog toka, a object carousel je njegovo proširenje koje se koristi za slanje fascikli sa fajlovima i slično. Dowloader je zadužen da iz primljenog prenosnog toka koji je TS (engl. Transport stream) izdvoji OTA paket koji sadrži podatke za ažuriranje. Da bi se OTA paket mogao primeniti na korisnički uređaj, downloader će izdvojiti OTA.ZIP iz prenosnog toka. Kada je ovaj proces završen, OTA podaci će se kopirati u /cache particiju bootloader alata. Bootloader alat će primiti notifikaciju za pokretanje procesa ažuriranja. Nakon ažuriranja i automatskog ponovnog pokretanja uređaja korisnik može da koristi novu verziju operativnog sistema. Ovaj proces je prikazan na Slici 4.



Slika 4

IV. IDEJE ZA BUDUĆI RAZVOJ

Opisano rešenje još uvek nije u potpunosti implementirano i pojedini delovi u ovom radu su samo idejno i teorijski predstavljeni, pa se kao prvi sledeći korak svakako mora posmatrati implementacija. Nakon završene implementacije naredni korak je testiranje i analiza, kao i poboljšanja na osnovu dobijenih rezultata testiranja. U prethodnom odeljku opisana su dva načina za integraciju signala u prenosni tok. Kao što je već navedeno, prvo rešenje će svoju primenu naći prilikom testiranja downloadera.

Dalji razvoj mogao bi doneti i širu primenu opisanog rešenja. Ovakav pristup koji ne zavisi od internet konekcije može da se iskoristi ponovo prilikom rešavanja nekog drugog problema.

V. ZAKLJUČAK

U ovom radu je opisano jedno od rešenja problema ažuriranja korisničkih STB uređaja koji su već u upotrebi kada je onemogućena internet konekcija. Na ovaj način korisnicima je obezbeđeno da prate trendove usavršavanja proizvoda televizijske industrije koje koriste. Opisano rešenje će učiniti ažuriranje uređaja mnogo jednostavnijim i doneće velikom broju korisnika nove mogućnosti.

ZAHVALNICA

Ovaj rad je delimično finansiran od strane Ministarstva prosvete, nauke i tehnološkog razvoja Republike Srbije, projekat TR32030.

LITERATURA

[1] M.Z. Bjelica, N. Teslić, V. Mihić, "Softver u televiziji i obradi slike 1", FTN, radna verzija, 2016.
[2] Benoit, H. "Digital Television – Satellite, Cable, Terrestrial, IPTV, Mobile TV in the DVB Framework", Focal Press, 2008.
[3] ETSI EN 302 769 v1.3.1, https://www.etsi.org/deliver/etsi_en/302700_302799/302769/01.03.01_60/en_302769v010301p.pdf, 2015.
[4] Pap dr Ištvan, Lukić dr Nemanja, "Projektovanje i arhitekture softverskih sistema – Sistemi zasnovani na Androidu".
[5] Android Developers documentation, https://developer.android.com/
[6] Watkinson J. "The MPEG Handbook", 2001.
[7] Morris S, Smith Chaigneau A. "Interactive TV Standards: A Guide to MHP, OCAP, and JavaTV", 2005.
[8] Fischer, W. "Digital Video and Audio Broadcasting Technology – A practical Engineering Guide", Springer-Verlag, 2010.
[9] Richardson, I. E. G. "H.264 and MPEG-4 Video Compression", Wiley, 2004.
[10] I. Bašičević, M. Popović, V. Kovačević, "Osnovi računarskih mreža 1", FTN, 2017.

ABSTRACT

This paper presents a solution for updating Set-Top Boxes based on the Android operating system in circumstances where the Internet connection is limited or disabled. The lack of a quality internet connection is one of the possible problems that occur when distributing a digital television signal. The solution presented in this paper uses special control streams that contain signal tables to control the update time.

Nataša Bogdanović, Ilija Bašičević, Goran Stupar

Realizacija upravljačke korisničke sprege za kontrolu softvera za snimanje i uređivanje zvuka

Milan Vuletić, Sergej Furtula, Jelena Kovačević, *Member, IEEE*

***Apstrakt*—Ovim radom je predstavljen jedan od načina unapređenja ispitnog okruženja za ispitivanje uređaja namenjinih za dekodovanje, obradu i reprodukciju Audio sadržaja pomoću programskog alata Audacity. Ova ispitivanja se vrše u dve faze, prva faza na nivou procesora a druga na nivou proizvoda. Ispitno okruženje za ispitivanje obično koristi APx (Audio Precision) merni uređaj, a unapređenje zapravo pretstavlja integraciju Audacity programskog alata u ova okruženja. Takođe, za potrebe dodatne automatizacije ispitnih slučajeva, realizovano je automatsko izvršavanje komandi za reprodukovanje, snimanje, generisanje i analiziranje audio zapisa.**

Rezultat ovog unapređenja se ogleda u smanjenju greške tokom ispitivanja i analize, izazvane „ljudskim faktorom“ kao i smanjenju vremena potrebnog za izvršavanje celokupne procedure.

***Ključne reči*—Automatizacija, programski alati, ispitivanje.**

I. Uvod

Sa porastom zahteva u području audio i video tehnologije, javlja se potreba za sve efikasnijim i komplikovanijim uređajima za audio i video obradu. Razvoj audio tehnologije direktno utiče na razvoj hardverskih jedinica, a još više na razvoj softverske podrške. Kako bi se na tržištu pojavio u što kraćem roku, nov audio uređaj, potrebno je obaviti i odgovarajuća ispitivanja tog uređaja sa ciljem zadovoljenja različitih sertifikata i željenog kvaliteta od strane samog proizvođača. U cilju bržeg i kvalitetnijeg ispitivanja audio uređaja potrebno je omogućiti automatizovano radno okruženje čime se postiže i smanjenje faktora “ljudske greške“. Ispitivanje audio uređaja se obično izvršava kroz dve faze, ispitivanja na nivou procesora koji direktno vrši audio obradu i na nivou gotovog proizvoda.

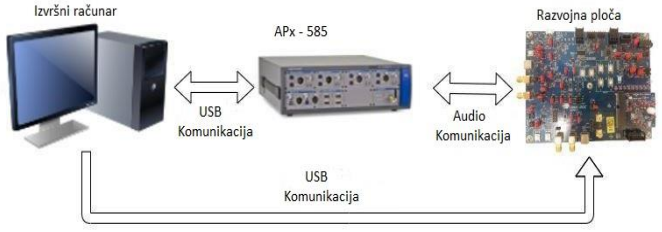
Razvoj softvera za dekodovanje, obradu i reprodukciju audio sadržaja obično se vrši na razvojnoj ploči sa odgovarajućim procesorom, odnosno arhitekturom prilagođenom za obradu audio signala (DSP). U prvoj fazi ispitivanja ispitni slučajevi pokrivaju scenarije koji za cilj imaju da verifikuju funkcionalnost softverske podrške na procesorkom nivou. U drugoj fazi se verifikuju signali na sistemskom nivou i ispituje se odziv uređaja kao proizvoda. Ispitivanje ispravnosti na nivou procesora i na nivou proizvoda razlikuje se u načinu verifikovanja (od načina dekodovanja, obrade pa sve do reprodukcije audio sadržaja), te su za te potrebe razvijene posebne ispitne procedure od samih IP providera, koji razvijaju i pružaju audio tehnologije (Dolby, DTS).

Razvoj proizvoda podrazumeva softversku podršku i QA (“Quality Assurance“) podršku. Neke procene govore da se za ispitivanje i pronalaženje grešaka utroši i do 50% ukupnog vremena potrebnog za razvoj [1]. Da bi uređaj u što kraćem roku ušao u fazu proizvodnje veoma je bitno da u fazi razvoja softverske podrške verifikacija bude usklađena sa razvojem. Postizanjem što većeg nivoa automatizacije, kako ispitivanja tako i analize rezultata, omogućava se brži razvoj softvera i brži razvoj uređaja [2].

U ovom radu je analizirano jedno rešenje automatizacije ispitnih okruženja za audio i video uređaje. U njemu je predstavljen jedan od načina unapređenja kroz automatizaciju ispitnih procedura sa ciljem da se omogući brži razvoj i sistemska integracija softvera za audio obradu u audio i video uređajima.

II. Opis ispitnog okruženja

Ispitno okruženje je razvijeno unutar Istraživačkog-razvojnog Instituta „RT-RK“ pod nazivom UTS (Unified Test Setup), koje sadrži program za pokretanje ispitnih slučajeva, neophodne sistemske biblioteke koje se koriste prilikom ispitne procedure i potrebne alate za obradu rezultata. Na Sl. 1 je prikazana osnovna postavka UTS-a:



Sl. 1 – Prikaz ispitnog okruženja

Kako se vidi na Sl. 1, ispitno okruženje se sastoji od izvršnog računara, mernog uređaja “APx-585 multichannel audio analyzer“ proizvođača “Audio Precision“ (u daljem tekstu audio analizator) i razvojne ploče na kojoj se izvršava softver koji se verifikuje. Na izvršnom računaru se pokreće Audio Precision ispitno okruženje audio analizatora. Audio analizator ima mogućnost reprodukucije i snimanja audio zapisa. Razvojna ploča simulira prisustvo audio uređaja sa USB konekcijom i na njoj se izvršava softverska podrška koja je predmet ispitivanja.

Milan Vuletić – Istraživačko-razvojni Institut RT-RK, Novi Sad, Srbija (e-mail: Milan.Vuletic@rt-rk.com)

Sergej Furtula – Istraživačko-razvojni Institut RT-RK, Novi Sad, Srbija (e-mail: Sergej.Furtula@rt-rk.com)

Jelena Kovačević – Fakultet Tahničkih Nauka, Odsek za računarsku tehniku i računarske komunikacije, Novi Sad, Srbija (e-mail: Jelena.Kovacevic@rt-rk.com)

Ispitni slučajevi u okviru UTS-a se pokreću kroz Audio Precision softvreske podrške za audio analizator [3]. U okviru UTS-a analiza rezultata je mogla biti izvršena vizuelno ili preslušavanjem audio zapisa. Prilikom izvršavanja ispitnih procedura u okviru ispitnog okruženja, dolazilo je do sledećih problema:

- uticaj subjektivnog aspekta u odlučivanju da li je ispitni slučaj uspešan (PASS), neuspešan (FAIL) ili nedefinisan (UNRESOLVED),
- ručno pokretanje koje kao posledicu ima duže vreme izvršavanja same ispitne procedure.

A. Opis unapređenog ispitnog okruženja

Unapređenje ispitnog okruženja se ogleda u korišćenju programskog alata Audacity, odnosno iskorišćenju njegovih funkcionalnosti kako bi se povećao nivo automatizacije, a samim tim smanjila interakcija QA inženjera.

Programski alat Audacity je pre svega besplatan, veoma je korišćen za uređivanje audio zapisa i jednostavan za korišćenje [4]. Da bi se ovaj programski alat uspešno koristio u procesu ispitivanja, neophodno je nadograditi njegov aplikativni deo kako bi se koristile komande za reprodukovanje, snimanje i generisanje audio zapisa. Od verzije 2.2.2 programskog alata Audacity, korisnicima je omogućeno da pristupe aplikativnom delu samog alata i da putem komandne linije proslede parametar za reprodukovanje, snimanje, generisanje, grafičku analizu u frekventom domenu i za reprodukovanje višekanalnih audio zapisa. Pored svih pogodnosti koje Audacity omogućava, Audacity takođe ima podršku i za ostale operativne sisteme (Windows 7, Windows 10, Linux i Mac OS X). Takođe, Audacity ima mogućnost i grafičkog prikaza što je pogodno za eventualno ručno pokretanje pojedinačnih slučajeva ili za analizu rezultata. Smanjenjem faktora ljudske greške je smanjena mogućnost reprodukovanja pogrešno prosleđenog audio zapisa, kao i smanjen subjektivni osećaj da li je ispitni slučaj uspešan, neuspešan ili pak nedefinisan.

Za realizaciju komunikacije Audacity-a sa softverskom podrškom audio analizatora a time i generalno sa UTS-om, neophodno je unutar Audacity-a omogućiti modul pod nazivom "mod-script-pipe". Mod-script-pipe je komunikacioni kanal koji služi za pristupanje aplikativnom delu Audacity softverskog paketa preko koga se takođe prosleđuju parametri kroz komandnu liniju. Kako bi se omogućio dodatni modul "mod-script-pipe" u Audacity-u, neophodno je obezbediti biblioteke ovog modula, odnosno kopirati ih u direktorijum u kome se nalazi instalacija Audacity progrmaskog alata. Početna verzija ovih biblioteka se može naći u izvornom "C" kodu na zvaničnom sajtu Audacity [5]. Za nadogradnju programskog alata Audacity dodatnim bibliotekama, odnosno modulima, korišćen je programski jezik Python. Jedan od ciljeva rada je da se kao rezultat dobije biblioteka u obliku Python modula sa metodama potrebnim za nadogradnju koja se kasnije može koristiti i za druga ispitivanja. Python je programski jezik opšte namene koji je veoma pogodan za automatizaciju i lak za učenje i razumevanje [6]. Za potrebe razvoja biblioteke za automatizaciju, korišćeni su već postojeći ispitni slučajevi koji su bili napisani u Python-u. Potrebno je napomenuti da mod-script-pipe takođe podržava Python.

III. REALIZACIJA

U okviru dodatne biblioteke, razvijeni su moduli za podešavanje alata, reprodukovanje, snimanje i analizu audio zapisa [7]. Zasebnim modulom za izvršavanje komandi je vršeno ispitivanje u okviru izvršnog računara i audio uređaja sa USB konekcijom. Na izvršnom računaru je pokretan Audacity i reprodukovani audio zapisi na audio uređajima.

A. Konfiguracija uređaja

Audio uređaji sa USB konekcijom, nad kojima su se izvršavale ispitne procedure, prikazani su u Tabeli 1**.** U tabeli su takođe prikazane različite konfiguracije uređaja:

Topology	Speed	UAC	Mode	Channel	Render SR	Render BD	Capture SR	Capture BD
Headset	High	2.0	Sync	Stereo / Mono	92 kHz	24 - bit	48 kHz	24 - bit
Headphones	Full	1.0	Sync	Stereo	44.1 kHz	16 - bit		
Microphone	Full	1.0	Sync	Stereo	48 kHz	24 - bit		

Tabela 1 – Prikaz konfiguracije za perifeje prilikom testiranja

Biblioteka za automatizaciju je verifikovana kroz primenu u postojećim sistemskim procedurama za ispitivanja audio zapisa na "Cirrus Logic CDB46L60-A1" razvojnoj ploči (audio uređaji sa USB-C konekcijom).

Platformski softver (engl. firmware) potrebno je konfigurisati kroz parametre iz Tabele 1. Ti parametri imaju sledeće značenje:

- "Topology" označava tip uređaja za testiranje (mikrofon, slušalice ili slušalice sa mikrofonom),

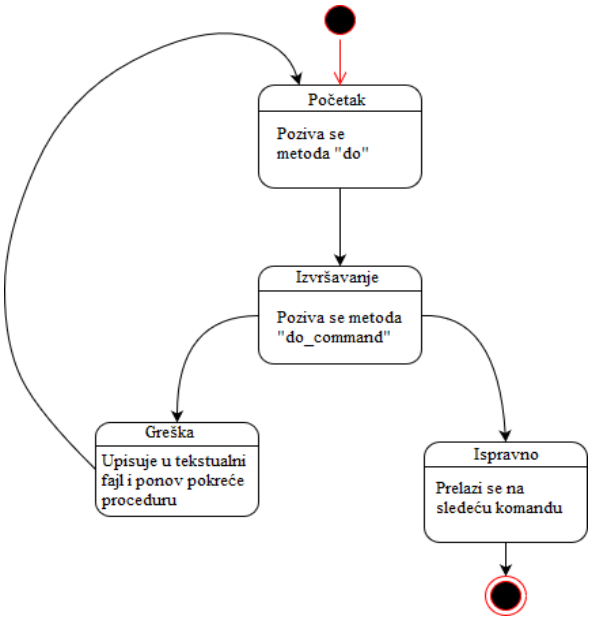
- “Speed“ označava brzinu prenosa padataka USB komunikacije,
- “UAC“ označava USB audio klasu,
- “Mode“ označava tip prenosa podataka (sinhroni ili asinhroni),
- “Channel“ označava odabir kanala (mono ili stereo),
- “Render SR“ označava učestanost odabiranja izvornog audio signala (8 kHz - 384 kHz),
- “Render BD“ označava bitsku dubinu podataka izvornog audio signala (16, 24, 32),
- “Capture SR“ označava učestanost odabiranja izlaznog audio signala,
- “Capture BD“ označava bitsku dubinu izlaznog audio signala.

Uređaj nad kojim se vrši ispitivanje, potrebno je konfigurisati za pojedinačne ispitne slučajeve i tako konfigurisan uređaj se može ispitivati na više operativnih sistema (Windows 7, Windows 10, Mac OS X, Linux).

B. Ispitivanje ispravnosti testova

Da bi se dokazala ispravnost testova, ispitivanje je vršeno na audio uređajima sa USB konekcijom, kao što su slušalice sa USB konekcijom, mikrofon sa USB konekcijom i slušalice sa mikrofonom sa USB konekcijom.

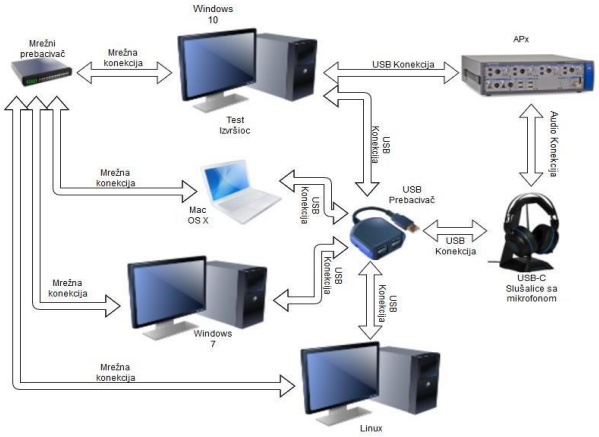
Pokretanje ispitne procedure i softverske podrške audio analizatora je realizovano slanjem naredbi (komandnih linija) u okviru “windows cmd“ konzole. Pomoću komandne linije se takođe prosleđuju parametri za pozivanje UTS-a, softverska podrška audio analizatora i parametri za određenu komandu u zavisnosti od tipa ispitnog slučaja (play, record i generate). Nakon pokrenute ispitne procedure, proverava se svaka prosleđena komanda. Na Sl. 2, se nalazi grafički prikaz procedure ispitivanja ispravnosti komande:



Sl. 2 - Grafički prikaz procedure ispitivanja ispravnosti komande

Kako se vidi na Sl. 2, ispitivanje ispravnosti komandi se zasniva na četiri stanja. Prvo stanje “Početak“ prikupi prosleđenu komandu i prelazi u sledeće stanje “Izvršavanje“. Stanje “Izvršavanja“ komandu pošalje u pomoćnu memoriju, gde se upoređuje prikupljena komanda sa nazivom komande u aplikativnom delu Audacity i vraća se povratna informacija da li je komanda ispravna ili neispravna. Iz stanja “Izvršavanja“ se prelazi u jedno od dva moguća stanja koja zavise od povratne informacije stanja “Izvršavanja“. Stanje “Greška“ neispravnu komandu upisuje u tekstualni dokument i prekida ispitnu proceduru, a korisniku se ispisuju pravila u windows konzoli kako treba da izgleda zapis komande. Stanje “Ispravno“ izvršava komandu.

Pre izvršavanja ispitnih slučajeva QA inženjer povezuje audio analizator sa izvršnim računarom i sa razvojnom pločom koja u ispitnom okruženju simulira audio uređaje sa USB konekcijom. Razvojna ploča omogućava ispitivanje ispitnih slučajeva na nivuo prosecora. Za ispitivanje na nivuo proizvoda se povezuje gotov proizvod sa audio analizatorom. Na Sl. 3 je grafički prikazano povezivanje audio uređaja sa audio analizatorom i izvršnim računarom prilikom ispitivanja na nivou audio uređaja. Ispitno okruženje je povezano preko mrežne konekcije sa izvršnim računarom i ovakva konfiguracija omugućava da se ispitivanje izvršava sa više različitih računara koji ne moraju biti fizički na istoj lokaciji i mogu da imaju različite operativne sisteme:



Sl. 3 – Grafički prikaz ispitnog okruženja za ispitivanje na nivou proizvoda

Kako se vidi na Sl. 3, ispitivanje na nivou proizvoda sa audio analizatorom je moguće izvršavati na više operativnih sistema. Komunikacija sa ostalim operativnim sistemima se odvija putem mrežne komunikacije. Audio analizator se povezuje sa izvršnim računarom i sa ispitnim proizvodom (slušalice sa mikrofonom sa USB konekcijom) putem USB komunikacije.

Nakon uspešnog povezivanja, pomoću komandnih linija se pokreće softverska podrška audio analizatora i UTS-a, unutar kog se poziva biblioteka za automatizaciju komandi u programskom alatu Audacity. U zavisnosti od potrebe ispitivanja, QA inženjer povezuje i dodatnu periferiju, u ovom slučaju USB slušalice sa mikrofonom, USB slušalice ili samo USB mikrofon.

Ispitni slučajevi koje QA inženjer treba da izvrši se mogu podeliti u tri grupe a to su:

- ispitni slučajevi sa USB slušalicama,
- ispitni slučajevi sa USB mikrofonom,
- ispitni slučajevi sa USB slušalicama i mikrofonom.

Za prva dva slučaja sa strane QA inženjera pokretanje ispitnih slučajeva se vrši direktno pomoću komandne linije koristeći samo aplikativni deo razvijene biblioteke. Za ove grupe ispitnih slučajeva QA inženjer bira između samo snimanja ili samo reprodukovanja. Inicijalno QA inženjer treba da proveri ispravnost ispitnog okruženja, tako što pokreće grupu ispitnih slučajeva namenjenih za validaciju ispitnog okruženja. Nakon dobijenog rezultata QA inženjer treba da podnese izveštaj da li je ispitni slučaj uspešan, neuspešan ili nedefinisan.

Za treću grupu ispitnih slučajeva, kod ovog načina ispitivanja potrebno je omogućiti u programskom alatu Audacity opciju da istovremeno može snimati i reprodukovati audio zapis. U ovoj grupi ispitivanja audio analizator generiše audio zapis, pošalje izvršnom računaru gde se čuva audio zapis i zatim vrati do audio analizatora koji reprodukuje audio zapis.

Ukoliko se desi da je ispitni slučaj neuspešan, potrebno je još jednom ili čak više puta pokrenuti problematičan slučaj. Ukoliko se i dalje dobija da je slučaj neuspešan, QA inženjer ima mogućnost da sam unapredi problematičan slučaj. Ako se i sa tim promenama dobija nesupešan rezultat, obaveštava se inženjer koji razvija platformski softver (engl. firmware). Kod ispitnih slučajeva sa nedifinisanim rezultatima ista je procedura provere kao za neuspešne ispitne slučajeve.

Nakon završene procedure ispitivanja, softverska podrška audio analizatora pravi word dokument u kom upisuje podatak o nazivu ispitnog slučaja i dodaje sliku sa karakteristikom u frekvetnom domenu za taj slučaj. Završetkom ispitne procedure UTS automatski prebacuje word dokument u html dokument. Nakon završenih svih ispitnih slučajeva poziva se zaseban manji program za isčitavanje podataka iz html dokumenta i upisivanje tih podataka u poseban excel dokument koji QA inženjer može da koristi za proveru rezultata. Softverska podrška audio analizatora prilikom ispitne procedure ubacuje granice za svaki ispitni slučaj.

Greške koje mogu prouzrokovati neuspešnost ispitnih slučajeva mogu biti:

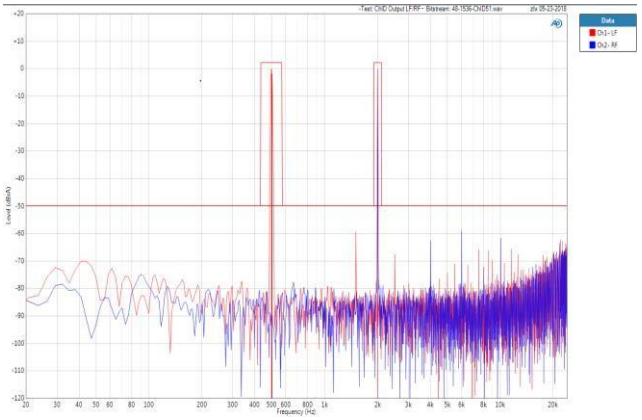
- ne reagovanje razvojne ploče na zadani takt (ovo se uglavnom dešava prilikom zadavanja više opcija)
- ili pogrešno rukovanje softverskim alatima.

Kao jedan način sigurnosti implementirano je da se izvršavanje UTS-a i alata Audacity odvija u pozadini, dok se kroz softversku podršku audio analizatora prikazuje ispravnost ispitnih slučajeva. Poželjno je prilikom ovakvih ispitivanja imati zaseban izvršni računar sa neophodnom hardverskom opremom (razvojnim pločama i mernim uređajima), kao i dodatnim periferijama.

C. Rezultati ispitivanja

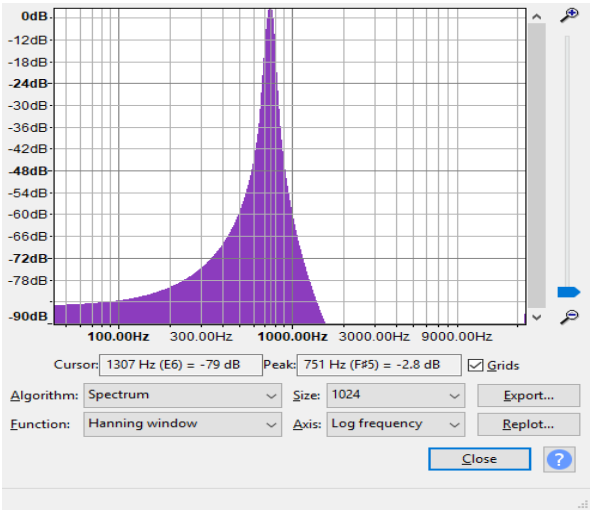
Automatizacijom samog ispitnog okruženja i integracijom programskog alata Audacity u UTS-u dobija se značajna ušteda vremena ispitivanja. Nakon upoređivanja ispitivanja pre i nakon automatizacije, dobija se veoma značajna ušteda od čak 40% prilikom celokupne ispitne procedure.

Na Sl. 4 i Sl. 5 je prikazan rezultat uspešno izvršenog ispitnog slučaja u okviru softverske podrške audio analizatora i programskog alata Audacity u obliku prenosne karakteristike u frekvetnom domenu:



Sl. 4 – Prikaz rezultata uspešno izvršenog ispitnog slučaja u okviru softverske podrške audio analizatora u frekventnom domenu

Kako se vidi na Sl. 4, u okviru softverske podrške audio analizatora su postavljene granice za ispitni slučaj. Ove granice olakšavaju vizuelnu proveru rezultata ispitnog slučaja. Ovaj rezultat se odnosni na ispitni slučaj gde se kao ispitni signal koristi dvokanalni audio zapis.



Sl. 5 – Prikaz uspečno izvršenog ispitnog slučaja u okviru programskom alatu Audacity u frekventnom domenu

Na Sl. 55 se vidi rezultat uspešno izvršenog ispitnog slučajeva koji kao ispitni signal koristi jednokanalni audio zapis.

Za potrebe ispitivanja na samom proizvodu, audio uređaji sa USB konekcijom, ispitivani su po 20 ispitnih slučajeva za svaku od ispitinih grupa. Ispitni slučajevi su verifikovani sa rezultatom uspešnog ispitivanja.

IV. ZAKLJUČAK

U okviru ovog rada implementirana je integracija programskog alata Audacity unutar UTS-a (Unified Test Setup) i Audio Precision ispitnog okruženja. Komande za programski alat Audacity su realizovane u okviru biblioteke koja može da se koristi i u ostalim ispitnim procedurama. Opisan je postupak povezivanja audio uređaja i verifikacija. Automatizovana celokupna ispitna procedura koja olakšava način ispitivanja na nivou procesora i na nivou gotovog proizvoda. Za potrebe verifikacije i ispitnih slučajeva, uspešno je implementirana integracija Audacity unutar UTS-a i Audio Precision ispitnog okruženja.

Mogućnosti za dalja istraživanja i poboljšanja mogu biti u pravcu proširenja aplikativnog dela za specifične ispitne procedure, automatizovanog grafičkog upoređivanja ostvarenih rezultata i omogućavanje ispitivanja ispitnih slučajeva na drugim mernim uređajima, što dovodi do jeftinijih hardverskih jedinica.

ZAHVALNICA

Ovaj rad je delimično finansiran od strane Ministarstva za nauku i tehnologiju Republike Srbije, na projektu tehnološkog razvoja broj: TR32029.

LITERATURA

[1] G.Myers, Art of software testing, ISBN: 0471043281, 1979
[2] "Experiences of Test Automation: Case Studies of Software Test Automation", Dorothy Graham, Mark Fewster, Addison-Wesley Professional, 2012.
[3] Audio Precision, "APx58x B Series Audio Analyzers", 2019. [Online]. Available: https://www.ap.com/analyzers-accessories/apx58x/, [posećen 16.04.2019.]
[4] Audacity, "About", 2019. [Online]. Available: https://www.audacityteam.org/about/, [posećen 16.04.2019.]
[5] Git Hub, "Mod-script-pipe", 2019. [Online]. Available: https://github.com/audacity/audacity/tree/master/lib-src/mod-script-pipe, [posećen 16.04.2019.]
[6] Python, "About", 2019. [Online]. Available: https://www.python.org/about/apps/, [posećen 16.04.2019.]
[7] Albert Sweigart, *Automate The Boring Stuff With Python*, No Starch Press, San Francisko, Kalifornija. 2015

ABSTRACT

This paper presents one of the ways to improve the test environment for testing devices intended for decoding, processing and reproducing Audio content using the Audacity software tool. These tests are performed in two phases, at the processor level and at the product level. The test environment usually uses the APx (Audio Precision) measuring device. The improvement actually represents the integration of the Audacity program tool in these environments. Also, for the purposes of additional automation of test cases, automatic execution of commands for reproducing, recording, generating and analyzing audio records was performed.
The result of this enhancement is to reduce the error during testing and analysis, caused by the "human factor" as well as reducing the time needed to complete the entire procedure.

Implementation of the software user interface for sound recording and editing tool

Milan Vuletić, Sergej Furtula, Jelena Kovačević, Member, IEEE

Povezivanje Android Media API-ja za DASH izvore podataka

Nikola Ječmenica, Marija Jovanović, Dušan Zivkov, Đorđe Glišić

***Apstrakt*— U ovom radu je predstavljena realizacija aplikacije koja ima za cilj da reprodukuje multimedijalni sadržaj koristeći Android API niskog nivoa, pre svega MediaCodec i MediaExtractor. Ona takođe predstavlja teset aplikaciju za MediaExtractor koji kao izvor podataka koristi MediaDataSource interfejs. Aplikacija je pisana u Java programskom jeziku i predstavlja simulaciju DASH protokola prilagođenu Android TV platformi. Aplikacija ima jednostavan GUI (Graphical User Interface).**

***Ključne reči*—MediaCodec, MediaExtractor, AudioTrack, MPEG-DASH, MPD, Android**

I. UVOD

Zahvaljujući velikom broju raznovrsnih API-ja (Application programming interface), developerima je dosta olakšan rad i omogućeno stvaranje projekata bilo kog tipa.

Trenutno najviše korišćen operativni sistem za pametne telefone je Android OS, koji je svoju primenu pronašao i u drugim uređajima kao što su televizori, kamere, igračke itd.[1].Druga po redu najrasprostranjenija upotreba mu je kao operativni sistem pametnih televizora koji postaju sve zastupljeniji. Još uvek nije napisan dovoljan broj aplikacija koje bi se na njima pokretale i nemamo toliki spektar mogućnosti kao što je to slučaj sa mobilnim uređajima. Najpopularnije i najviše korišćene su one za reprodukciju multimedijskih sadžaja.

MPEG-DASH je jedan od protokola za reprodukciju multimedijalnog sadržaja preko interneta. Omogućava visokokvalitetan protok multimedijalnog sadržaja koji se isporučuje preko HTTP servera[9]. Smanjeno kašnjenje prilikom pokretanja, izbegava baferovanje, prilagođavanje situaciji klijentskog propusnog opsega su neke od prednosti u odnosu na starije tehnike kao što su HLS(HTTP Live Streaming)[2].

Jedna od najvećih produkcijskih kuća Netflix, i jedne od najvećih kompanija kao kao što je Google nedavno su prešle na DASH standard, što je znatno doprinelo ovom procentu

Aplikcaija ima za cilj i testiranje MediaExtractor-a, kome smo kao izvor podataka stavili MediaDataSource interfejs. Njega smo impelmentirali tako da čita DASH komponente iz lokalnog direktorijuma, u tom smislu se radi o simulaciji DASH protokola.

U poglavlju II biće detaljan opis Android Medija API-ja nižeg nivoa koji su ključni za implementaciju ove aplikacije.

U poglavlju III biće opisan MPEG-DASH protokol. Biće prikazan način rada, prednosti u odnosu na starije protokole.

U poglavlju IV biće data implementacija aplikacije, način simulacije DASH protokola[6], kao i način korišćenja aplikacije.

Rad je završen poglavljem V u kome su izvedeni zaključci uz predloge daljeg unapređenja aplikacije.

Nikola Ječmenica, Naučno istraživački institut RT-RK, Bulevar Milutina Milankovića 19b, 11000 Beograd, Srbija (telefon: 381-21-480-1143, email: nikola.jecmenica@rt-rk.com)

Marija Jovanović, Naučno istraživački institut RT-RK, Bulevar Milutina Milankovića 19b, 11000 Beograd, Srbija (telefon: 381-21-480-1143, email: marija.jovanovic@rt-rk.com)

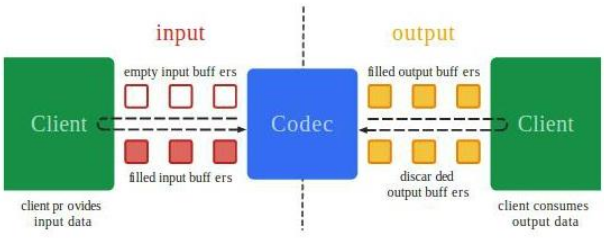
Dušan Živkov, Naučno istraživački institut RT-RK, Bulevar Narodnog fronta 23a Novi Sad, Srbija (telefon: 381-21-480-1297, email: dusan.zivkov@rt-rk.com)

Đorđe Glišić, Naučno istraživački institut RT-RK, Bulevar Milutina Milankovića 19b, 11000 Beograd, Srbija (telefon: 381-21-480-1278, email: djordje.glisic@rt-rk.com)

II. ANDROID MEDIJA API

Što se tiče Android medija API-ja višeg nivoa tu je MediaPlayer koji se koristi za ubacivanje video snimaka u aplikaciju. Ovaj API je veoma jednostavan za korišćenje. Jednostavnost korišćenja ovog API-ja ima svoje mane. Naime, ako neko ima potrebu da doda neke funkcionalnosti koje su potrebne za određenu aplikaciju nije u mogućnosti to da uradi. Ova ograničenja su prevaziđena dodavanjem API-ja nižeg nivoa MediaCodec, MediaExtractor, AudioTrack itd.

MediaCodec se koristi za dekodovanje kompresovanih audio i video formata i kodovanje u RAW format[4].



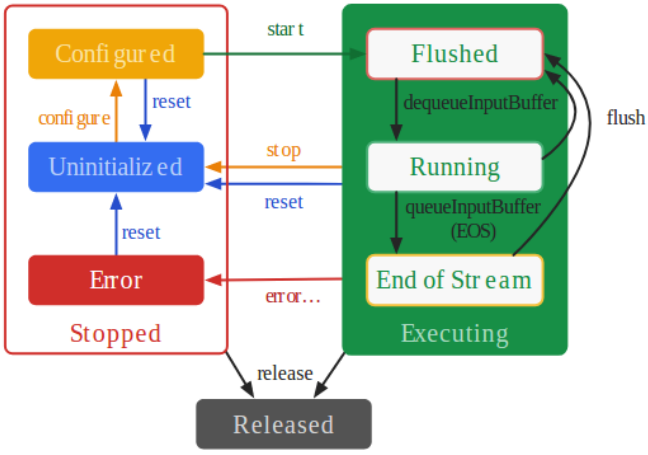
Slika1. Ilustracija rada MediaCodec-a

MediaCodec dobija ulazne podatke, nad njima radi obradu i daje izlazne podatke. Prenos podataka se vrši pomoću ulaznih i izlaznih bafera, pri čemu se obrada nad podacima vrši asinhrono. Prvo klijent zatraži prazan ulazni bafer, napuni ga podacima, zatim pošalje kodeku. Kodek obradi dobijene podatke, a zatim napuni prazan izlazni bafer. Na kraju klijent dobije napunjen bafer, iskoristi podatke zatim ga vrati nazad kodeku. Sve ove operacije se izvršavaju u paraleli. Na slici 1. prikazana je ilustracija rada MediaCodec-a.

Tokom rada MediaCodec se nalazi u jednom od tri stanja: Stopped, Executing i Released. Stopped je stanje u kom je MediaCodec zaustavljen i sastoji se od tri podstanja: Uninitialized, Configured i Error. Executinig je stanje izvršavanja i naizmenično prolazi kroz tri podstanja: Flushed, Running, End-of-stream[4].

Kada se napravi instanca klase MediaCodec ona se nalazi u stanju Stopped, podstanje Uninitialized. Sledeća korak je konfiguracija koja se radi pomoću odgovarajuće metode i zatim MediaCodec prelazi u podstanje Configured. Odavde se, odgovarajućom metodom, prelazi u Executing stanje, Flushed podstanje. Nakon ovog, dobija indeks praznog ulaznog bafera, po završetku punjenja ovog bafera prelazi u Running podstanje. U ovom podstanju MediaCodec više ne prihvata ulazne bafere već samo generiše izlazne. Pomoću odgovarajuće metode u svakom trenutku dok je MediaCodec u Executing stanju moze da se vrati u Flushed podstanje.

Kada se završi sa korišćenjem, neophodno ga je osloboditi čime prelazi u Released stanje. Prelazna stanja su prikazana su na slici 2.



Slika2. Prelazna stanja

MediaExtractor olakšava vađenje podataka koji prosli kroz demukser, uglavnom kodiranih, multimedijalnih podataka iz izvora podataka. Prvo se pravi nova instanca MediaExtractor-a, zatim se postavljaju podaci koji treba da se kodiraju[5]. Prolazi se kroz sve trake u jednoj petlji, i uzimaju se trake odgovarajućeg formata. Postoji metoda pomoću koje se dobija format trake i na osnovu toga se odlučuje da li će traka biti obrađena ili preskočena.

Pretpostavimo da je MediaCodec spreman da primi podatke odnosno da mu je bafer podešen, odgovarajućom metodom dekodovani uzorak iz MediaExtractor-a stavljamo u ulazni bafer počevši od zadatog pomeraja. Kada je završeno sa korišćenjem potrebno je osloboditi resurse[5].

AudioTrack kontroliše i pušta audio resurse za Java aplikaciju. Može da radi u dva režima, statički ili striming.

U režimu strimovanja pomoću blokirajuće funkcije upisuju se tokovi podataka u AudioTrack. Ovaj režim je koristan kada želimo da pustimo audio podatke koji su preveliki da stanu u memoriju. Statički režim je pogodan kada su podaci kratki zvukovi koji staju u memoriju[5].

Po kreiranju objekta AudioTrack-a radi se inicijalizacija njegovih bafera. Veličina ovih bafera određuje koliko dugo će AudioTrack moći da pušta pre nego što ostane bez podataka.

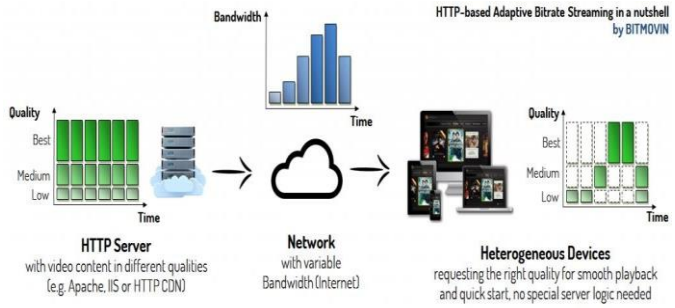
III. MPEG-DASH

MPEG-DASH (Dynamic Adaptive Streaming over HTTP) predstavlja adaptivnu tehniku striminga bitova koja omogućava visokokvalitetan protok multimedijalnog sadržaja preko interneta koji se isporučuje sa HTTP server[8].

MPEG-DASH je dostupan na Android platform pomoću ExoPlayer-a. YouTube i Netflix takođe koriste ovaj protokol.

Za razliku od HLS, HDS i Smooth Streaming-a, DASH je kodek agnostik, što znači da može da koristi sadržaj koji je kodiran bilo kojim formatom kodiranja, kao što su H.265, H.264, VP9 itd[7].

Osnovna ideja ovog protokola je sledeća – sadržaj razdvaja u niz malih segmenata. Segmenti su obezbeđeni na veb serveru i mogu se preuzeti preko HTTP protokola koji su u skladu sa GET zahtevima, kao što je prikazano na slici 3., gde HTTP služi tri različita kvaliteta niski, srednji i visoki iseckana u segmente jednake dužine[9].



Slika3. Način rada MPEG-DASH-a

Klijent automatski bira segment sa najvećom rezolucijom koja se može preuzeti u vremenu za reprodukciju bez izazivanja smetnji i ponovnog učitavanja sadržaja. Stoga se MPEG-DASH klijent može neprimetno prilagoditi promenama mrežnih uslova i obezbediti visokokvalitetnu reprodukciju bez kočenja i ponovnog baferovanja.

Da bi opisali vremenske i strukturne odnose između segmenata MPEG-DASH uvodi MPD (Media Presentation Description) strukturu. MPD je XML datoteka koja predstavlja različite kvalitete multimedijalnog sadržaja i pojedinačnih segmenata svakog kvaliteta. Ova struktura obezbeđuje vezivanje segmenata na rezoluciju, početno vreme trajanja, trajanje segmenata[2].

MPD je hjerarhijski model podataka. Svaki MPD može da sadrži jedan ili više perioda. Svaki od tih perioda sadrži medijske komponente kao što su video komponente npr. različiti uglovi gledanja, audio komponente sa različitim jezicima itd. Ove komponente imaju određene karakteristike kao što su bitrate, framerate, audio kanali itd., koje se ne menjaju tokom jednog perioda[7].

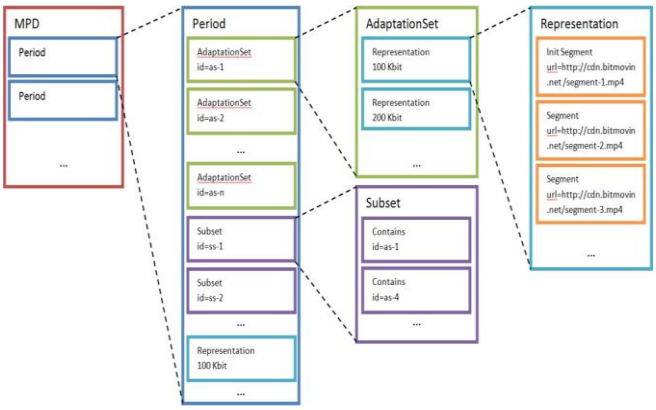
Tipične medijske komponente kao što su video, audio ili

titlovi organizovani su u adaptacioni set. Svaki period sadrži jedan ili više adaptacionih setova koji omogućavaju grupisanje različitih multimedjialnih komponenti koji su logički povezane. Na primer komponente sa istim kodekom, jezikom, rezolucijom, formatom audio kanala itd. mogu biti u istom adaptacionom setu. Ovaj mehanizam omogućuje korisniku da eliminiše niz multimedijalnih komponenti koje ne ispunjavaju određene zahteve.

Adaptacioni set se sastoji od skupa reprezentacija koje sadrže izmenjive verzije odgovarajućeg sadržaja. Iako jedna reprezentacija omogućava strimovanje u jednoj rezoluciji, više reprezentacija omogućava korisniku prilagođavanje medija na trenutne uslove mreže.

Reprezentacija je podeljena u segmente koji omogućavaju prebacivanje između pojedinačnih reprezentacija tokom reprodukcije. Ovi segmenti su opisani URL-om, a u određenim slučajevima dodatnim opsegom bajtova ako su segmenti stavljeni u veće datoteke. Segmenti u reprezentacijama obično imaju istu dužinu u vremenu i uređuju se po vremenskoj liniji što omogućava glatko prebacivanje reprezentacija tokom reprodukcije. Za razliku od drugih sistema, MPEG-DASH ne ograničava dužinu trajanja segmenta ili daje savete o optimalnoj dužini. Ovo se može izabrati od datog scenarija, duzi segmenti omogućavaju bolju kompresiju, a kraći segmenti se koriste za uslove velike promenljive propusnosti kao što su mobilne mreže, jer omogućavaju brže i efikasnije prebacivanje između pojedinačnih bitova.

Segmenti se dalje mogu podeliti na manje podsegmente koji predstavljaju skup manjih pristupnih jedinica u datom segmentu. Na slici 4. prikazana je MPD struktura.



Slika4. Struktura MPD-a

Tokom reprodukcije sadržaja, proizvoljno prebacivanje između reprezentacija nije moguće u bilo kojoj tački u toku i potrebno je razmotriti određena ograničenja. Dakle segmenti se ne smeju preklapati, takodje nisu dozvoljene zavisnosti izmedju segmenata. Da bi omogućio prebacivanje između reprezentacija, MPEG-DASH je predstavio SAP (Stream Access Point) na kojima je to moguće.

IV. REALIZACIJA APLIKACIJE

Aplikacija je u potpunosti napisana u Java programskom jeziku. Isključivo se oslanja na Android medija API niskog nivoa. Aplikacija treba da obezbedi gladak prelaz sa jedne rezolucije na drugu u toku reprodukcije.

Nakon pokretanja aplikacije korisniku se prikazuje ekran sa četiri dugmeta, gde može da izabere da li želi običan strim u jednoj rezoluciji, da li želi promenu dve različite rezolucije, promenu tri različite rezolucije. Na četvrto dugme bira video koji želi da reprodukuje. Na slici 5. prikazan je ekran aplikacije nakon pokretanja.

Slika5. Početni ekran aplikacije

Postavljanje izvora podataka za MediaExtractor vrši funkcija *setDataSource*, koja kao argument prima fajl deskriptor ili putanju do odgovarajućeg fajla koji predstavlja izvor podataka.

Kako ovo treba da bude simulacija DASH-a, nećemo koristiti nijedan od navedenih argumenata već ćemo implementirati posebnu klasu za to. Ovde koristimo video fajlove koji predstavljaju kratke segmente jednog videa. Uvek se počinje sa inicijalnim segmentom, koji ne sadrži video ili audio za reprodukciju, već informacije potrebne za reprodukciju tih segmenata kao celine.

Za ovo smo implementirali apstraktnu klasu *MediaDataSource* koja se upravo koristi kada aplikacija ima posebne zahteve za dobijanje medija podataka. Metode ovog intefejsa se pozivaju iz različitih niti. Sinhronizacija već postoji između poziva metoda.

Funkcija *setDataSource* prima *MediaDataSource* kao argument, na taj način postavljamo izvor podataka za ovu aplikaciju.

Prvo pročitamo inicijalni segment i sačuvam ga kao jedan bajt. Zatim u petlji čitamo sledeće segmente u zavisnosti koliko ih ima i smeštamo u bafer bajtova. Funkcijom *System.arraycopy* spajamo inicijalni segment i ostale segmente u jedan bafer i time hranimo MediaExtractor.

Svi segmenti koji su potrebni aplikaciji nalaze se na Android uređaju, u direktorijumu /Movies/. Fajl, za čiji konstruktor treba putanja do odgovarajućeg segmenta, prosledimo funkciji *readAllBytes* koja vraća bajtove.

Za audio je potpuno ista implementacija samo se stavi putanja do audio segmenata.

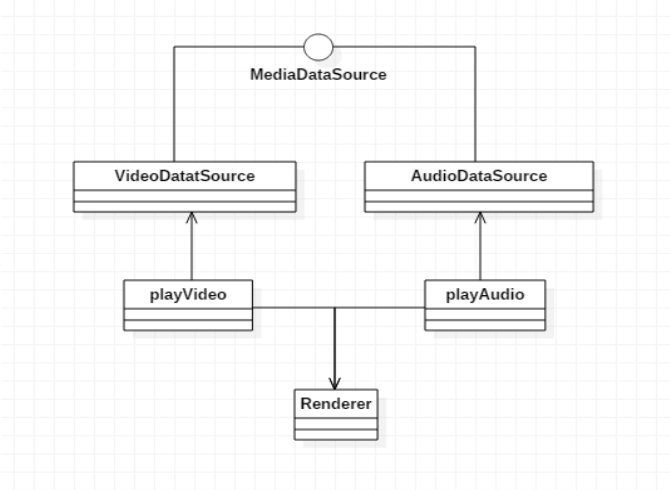
Kako je potrebno reprodukovati audio i video, moramo da koristim dva MediaCodec-a. Jedan za dekodovanje videa, drugi za dekodovanje audia. Po istoj logici je potrebno koristiti i dva MediaExtractor-a.

Na slici 6. prikazan je UML dijagram klasa koje su korišćene.

Kako nakon promene rezolucije potrebno je krenuti ponovo od inicijalnog segmenta za novu rezoluciju, na primer

reprodukuje se: init1, seg1, seg2, seg3 i od trećeg segmenta kreće nova rezolucija, tada je potrebno reprodukovati init2, seg4, seg5, seg6, itd.

Ovde je nastao problem, jer kad sa dva različita init-a hranimo MediaExtractor dolazi do greške.. Zbog ovoga je potrebno da koristimo novi MediaExtractor, odnosno potrebno je da koristimo onoliko MediaExtractor-aa koliko ima različitih rezolucija.



Slika 6. UML dijagram klasa

MediaCodec može da prima podatke od više MediaExtractor-a, samo pre prelaska na novi MediaExtractor, mora se pozvati metoda *release*..

Što se tiče same reprodukcije, prvo smo implementirali interfejs *SurfaceHolder.Callback* koji prati sve promene na površini na kojoj se vrši reprodukcija. Zatim deklarišemo *SurfaceView*, *Surface*, *SurfaceHolder*.

U funkciji *onCreate*, koja se poziva jednom na početku, funkcijom *setContentView* postavljamo odgovarajući XML fajl. Zatim prethodno deklarisani *SurfaceView* povežemo sa *SurfaceView*-om koji smo pstavili u xml fajl.

Kako je za konfiguraciju MediaCodec-a potreban *Surface*, njega dobijamo iz prethodno inicijalizovanog *SurfaceView*-a. Takođe iz *SurfaceView*-a dobijamo *SurfaceHolder*, kome prosleđeujemo implementirani *Callback*, čime prati promene na pozadini odgovarajućeg xml fajla. Na slici 6. prikazana je reprodukcija videa.

Slika6. Reprodukcija videa u aplikaciji

V. ZAKLJUČAK

Android OS TV uređaji su sve zastupljeniji na stranom i našem tržistu, a upravo aplikacije potrebne za takve uređaje su aplikacije koje rade sa multimedijalnim sadržajem. Može se zaključiti da će aplikacije koje reprodukuju multimedijalni sadržaj biti sve potrebnije u budućnosti.

Ova aplikacija je realizovana korišćenjem Java programskog jezika u AndroidStudio okruženju koje omogućava testiranje aplikacija na raznim uređajima, koji podržavaju Android OS, u vidu emulatora.

Deo aplikacije koji predstavlja MediaExtractor sa MediaDataSource iterfejsom kao izvorom podataka može se korsititi za rad sa DASH-om u uslovima kada nema mreže. Ovaj API mogu da koriste aplikacije poput ExoPlayer-a, kako bi promenili izvor podataka.

Dalji razvoj ove aplikacije može da bude opcija da korisnik može sam da bira rezoluciju u kojoj želi da reprodukuje sadržaj.

LITERATURA

[1] Android Developer, https://developer.android.com/

[2] Bitmovin, https://bitmovin.com/dynamic-adaptive-streaming-http-mpeg-dash/

[3] Ištvan Pap, Nemanja Lukić, "Projektovanje i arhitektura softverskih sistema: Sistemi zasnovani na Androidu", FTN Izdavaštvo, Novi Sad, 2015

[4] MediaCodec, https://developer.android.com/reference/android/media/MediaCodec.html

[5] MediaExtractor, https://developer.android.com/reference/android/media/MediaExtractor.html

[6] AudioTrack, https://developer.android.com/reference/android/media/AudioTrack.html

[7] Kasarapu Ramani, "Media Presentation Description over MPEG-DASH", LAMBERT, Academic Publishing, 2018

[8] B. Lazarević, "Development application for recording MPEG-DASH data streams and protection of recorded data on devices with Android operating system", Telecomunications Forum Telfor (TELFOR), Belgrade 2017

[9] N. Mitić, "Realisation of server for adaptive video and audio stream playback", Telecomunication Forum Telfor (TELFOR), Belgrade 2018

ABSTRACT

This paper presents an application that amis to reproduce media content that uses Android APIs, primarly MediaCodec and MediaExtractor. The application is written in Java programming language and represents simulation of the DASH tuning protocol for the Android TV platform. The application has a simple GUI (Graphicla User Interface).

Connecting the Android Media API to DASH data source

Nikola Ječmenica, Marija Jovanović, Dušan Zivkov, Đorđe Glišić

Generalizacija prikaza funkcija i stanja uređaja u IoT sistemima

Lana Salai, Igor Stefanović, Roman Pavlović, Ištvan Pap, Miloš Milanović

***Apstrakt*—U ovom radu opisan je koncept rešenja prikaza funkcionalnosti IP (eng. *Internet protocol)* uređaja krajnjim korisnicima sistema za kućnu automatizaciju na pojednostavljen način. Rešenje podrazumeva generički prikaz fizičkih osobina uređaja, pri čemu se informacije o uređajima dobavljaju korišćenjem WISE (eng. *WiFi Sensors*) protokola. Podelom funkcionalnosti uređaja po prioritetima, moguće je iscrtati kontrolne ekrane za funkcionalno različite uređaje na intuitivan način. To značajno olakšava održavanje i proširivanje sistema pametnih kuća novim uređajima.**

***Ključne reči*—pametna kuća; protokol; WISE; IP; uređaji; mobilna aplikacija;**

I. Uvod

Stalni razvoj tehničkih rešenja, zajedno sa konceptom uređaja povezanih na Internet, doživljava procvat na polju kućne automatizacije. Iz dana u dan raste dostupnost električnih uređaja koji nude direktno povezivanje na Internet. Međutim, koriste razne komunikacione protokole koji su uglavnom namenjeni samo za određeni tip uređaja ili za uređaje jednog proizvođača. Iz tog razloga, trenutno ne postoji jedno univerzalno rešenje programske podrške za upravljanje uređajima različitih funkcionalnosti. S obzirom da sistem pametne kuće sačinjava više različitih tipova namenskih uređaja, uloga glavne komponente ovog sistema, centralnog kontrolera, je da obezbedi uniforman pristup perifernim uređajima različitih protokola.

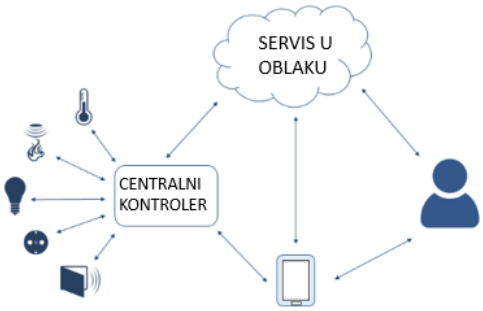
Namenski uređaji niske potrošnje se uglavnom oslanjaju na *ZigBee* [1], *Z-Wave* [2] i *Bluetooth Low Energy (BLE)* [3] protokol. Ovi protokoli nude mogućnost povezivanja uređaja u mrežu *mesh* topologije [4], pri čemu je komunikacija ograničena na lokalnu mrežu. Nasuprot ovim uređajima, uređaji koji se oslanjaju na IP protokol mogu biti dostupni kako u lokalnoj tako i u globalnoj mreži.

Imajući u vidu da su uređaji integrisani unutar sistema pametnih kuća ograničeni sa stanovišta računarskih resursa, protokol koji oni koriste mora da ispunjava uslove veoma ograničenog okruženja (mala memorija, niska potrošnja i procesorske mogućnosti). Kao rešenje predložen je komunikacioni protokol WISE [5], isključivo namenjen realizaciji namenskih uređaja za upotrebu u sistemima kućne automatizacije. On spaja prednosti postojećih, standardizovanih komunikacionih protokola sa prednostima IP protokola, nudeći unificiranu osnovu za realizaciju najrazličitijih uređaja.

II. IoT sistem

Sistem kućne automatizacije čini više distribuiranih komponenti među kojima je glavna komponenta centralni kontroler. On predstavlja posrednika u komunikaciji između klijentskih aplikacija i perifernih uređaja. Uloga centralnog kontrolera je da komande primljene od strane korisnika prosledi perifernim uređajima, kao i da informacije od značaja koje pristižu sa perifernih uređaja prosledi korisniku. Kako bi komunikacija sa uređajima različitog tipa bila moguća, neophodno je da centralni kontroler implementira komunikacione protokole koje koriste ti uređaji.

Zadatak klijentske aplikacije je da pruži korisniku uvid u trenutno stanje svih uređaja i da omogući udaljenu kontrolu tih uređaja. Posredstvom aplikacije, korisnik ima mogućnost da se poveže sa centralnim kontrolerom i u lokalnoj i u globalnoj mreži. U slučaju kada su klijentska aplikacija i centralni kontroler u lokalnoj mreži, komunikacija između njih je bežična i bez posrednika. Međutim, ukoliko se aplikacija nalazi izvan lokalne mreže, tada se javlja potreba za drugim načinom pristupa centralnom kontroleru. Iz tog razloga uvodi se servis u oblaku (eng. *cloud*) koji predstavlja poslužioca (eng. *server*) koji je konstantno povezan sa centralnim kontrolerom i dostupan je na globalnoj mreži.



Sl. 1. Sistem pametne kuće

Lana Salai – OBLO Living LLC, Narodnog fronta 23a, 21000 Novi Sad (e-mail: lana.salai@rt-rk.com).
Igor Stefanović – OBLO Living LLC, Narodnog fronta 23a, 21000 Novi Sad (e-mail: igor.stefanovic@obloliving.com).
Roman Pavlović – OBLO Living LLC, Narodnog fronta 23a, 21000 Novi Sad (e-mail: roman.pavlovic@obloliving.com).
Ištvan Pap – OBLO Living LLC, Narodnog fronta 23a, 21000 Novi Sad (e-mail: istvan.papp@obloliving.com).
Miloš Milanović – OBLO Living LLC, Narodnog fronta 23a, 21000 Novi Sad (e-mail: milos.milanovic@obloliving.com).

III. WISE PROTOKOL

WISE protokol rešava problem uniformne realizacije IP uređaja različitih funkcionalnosti za upotrebu u sistemu pametnih kuća. Definiše i standardizuje uključivanje uređaja u sistem, njihovu konfiguraciju kao i model razmene poruka, odnosno pribavljanje podataka i kontrolu samih uređaja. Protokol definiše različite tipove funkcionalnosti podržanih od strane namenskih uređaja. Ukoliko uređaj podržava funkcionalnosti definisane ovim protokolom, one su prepoznate od strane centralnog kontrolera sistema. Time je omogućen pristup fizičkim osobinama uređaja, dobavljanje informacija od značaja i eventualna promena stanja od strane krajnjih korisnika sistema pametne kuće, posredstvom klijentskih aplikacija. Protokolom je definisano sledeće:

- Režimi rada:
  a) Režim u kojem se centralni kontroler nalazi u istoj mreži kao i IP uređaj. U ovom režimu uređaj pretražuje mrežu kako bi dobio potrebne informacije o centralnom kontroleru upotrebom SSDP (eng. *Simple Service Discovery Protocol*) protokola [6]. Prednosti su brzina razmene informacija i mogućnost te razmene bez pristupa Internetu
  b) Režim u kojem se centralni kontroler izvršava kao deo Internet oblaka. Uređaj se povezuje sa centralnim kontrolerom sa unapred definisanom adresom putem HTTP (eng. *Hypertext Transfer Protocol*) protokola [7]. Dobija se distribuirana mreža uređaja direktno povezanih na Internet, bez postojanja centralnog kontrolera kao fizičke komponente
- Model razmene poruka:
  Sisteme pametnih kuća karakteriše potreba za asinhronom komunikacijom i velikom propusnom moći. Takođe je od velikog značaja odnos zaglavlja i sadržaja poruke koja se razmenjuje među komponentama, pogotovo kod poruka kratkog sadržaja. Kako bi protokol ispunio ove veoma stroge zahteve, kao osnova za njegovu implementaciju prihvaćen je javno dostupan protokol aplikativnog sloja ISO/OSI steka [4] – MQTT (eng. *Message Queuing Telemetry Transport*) protokol [8]. On predstavlja jedan od najzastupljenijih protokola unutar IoT (eng. *Internet of Things*) zbog svoje jednostavnosti, asinhronosti, kompatibilnosti i fleksibilnosti. Baziran je na principu pretplate i objave na određene teme od značaja, pri čemu je posrednik u komunikaciji komponenta nazvana broker.
  WISE protokol definiše formate tema i četiri tipa poruka [9] koje učestvuju u razmeni:
  1. Zahtev
  2. Odgovor
  3. Obaveštenje
  4. Status
- Priključivanje uređaja na Internet i priključivanje uređaja internoj mreži centralnog kontrolera
- Upravljanje i održavanje mreže:
  Uvode se pojmovi „spavajućih“ i „aktivnih“ uređaja, pri čemu prvi tip predstavlja uređaje koji objavljuju svoja stanja centralnom kontroleru, pri čemu kontroler ta stanja može da promeni samo u trenutku kada se spavajući uređaji „probude“. Ovi uređaji nemaju konstantno uspostavljenu vezu sa kontrolerom. Drugi tip predstavlja uređaje koje je moguće kontrolisati u svakom momentu zbog toga što je njihova veza sa kontrolerom konstantna (ukoliko se ne uzme u obzir da usled drugih uticaja može doći do prekida veze). Protokolom su definisana ponašanja uređaja, u pogledu održavanja mreže, u oba slučaja.
- Podržane funkcionalnosti IP uređaja:
  Funkcionalnost IP uređaja (protokolom definisana kao „*Service*“) predstavlja skup više povezanih fizičkih osobina uređaja (definisanih kao „*Property*“). Povezujemo ih u grupe, čime se odvajaju različiti skupovi funkcionalnosti. Ti skupovi su protokolom definisani kao „*Group*“, što sve zajedno čini pojam uređaja.

FUNKCIONALNOSTI UREĐAJA PODRŽANE WISE PROTOKOLOM

Device	Informacije o uređaju
Battery	Informacije o bateriji uređaja
Switch	Sprega ka promeni stanja uređaja
Light	Sprega ka promeni stanja svetla na uređaju
Temperature	Informacija o temperaturi koju je izmerio uređaj
LevelControl	Sprega ka promeni nivoa fizičke osobine uređaja
Dim	Sprega ka promeni nivoa osvetljenja
Color	Sprega ka promeni boja svetla uređaja
ColorTemperature	Sprega ka promeni temperature boje svetla
Alarm	Informacija o stanjima uređaja čije se promene mogu objaviti
Humidity	Informacija o trenutnoj vlažnosti vazduha koju je izmerio uređaj
Gas	Informacija o tome da li je uređaj detektovao neki od potencijalno opasnih gasova
PowerMetering	Informacija o potrošnji energije, itd.
Shutter	Sprega ka podešavanju roletni na prozoru

Thermostat	Sprega ka podešavanju termostata
Diagnostic	Osnovne informacije o dijagnostici uređaja
MotionDetection	Informacija o tome da li je uređaj detektovao pokret u svojoj blizini
Contact	Informacija o tome da li ima kontakta na magnetu uređaja
Flood	Informacija o tome da li je uređaj detektovao poplavu
Smoke	Informacija o tome da li je uređaj detektovao dim

IV. PRIKAZ UREĐAJA I MOGUĆNOSTI

Ideja rada je da se na generički način korisniku prikaže bilo koji IP uređaj dodat u sistem, što je predstavljeno na primeru korišćenja WISE protokola i njegovih funkcionalnosti. Njime je moguće prikazati nebrojeno mnogo IP uređaja koji implementiraju ovaj protokol, a broj funkcionalnosti je lako proširiv ukoliko nastane potreba za opisom novih fizičkih osobina uređaja.

Centralni kontroler, kao posrednik u komunikaciji klijentske aplikacije i krajnjeg uređaja služi da apstrahuje obe strane komunikacije. Na strani kontroler-uređaj, on dobavlja skupove funkcionalnosti IP uređaja i mapira ih na interno definisane funkcionalnosti. Sa druge strane, na korisničkoj aplikaciji moguće je dobaviti sve funkcionalnosti uređaja, kako bi bile prikazane krajnjem korisniku, posredstvom drugog protokola komunikacije, implementiranog i na strani aplikacije i na strani centralnog kontrolera.

U slučaju da je IP uređaj već poznat centralnom kontroleru, odnosno kontroler prepoznaje njegov model i implementira čitav skup servisa koje uređaj poseduje, iscrtavanje informacija o uređaju na krajnjim aplikacijama vrši se statički, koristeći unapred poznate izglede kontrolnih ekrana specifične za svaki od uređaja. Na strani aplikacije, za svaki servis, unapred je određen izgled komponente koja se prikazuje.

Međutim, u slučaju da centralni kontroler ne prepoznaje uređaj po njegovom imenu, modelu i ostalim karakterističnim fizičkim osobinama, on od uređaja zahteva sve potrebne informacije slanjem komandi definisanih WISE protokolom. Tako dobija skup funkcionalnosti koje prenosi do krajnje aplikacije. U samoj aplikaciji, moguće je iscrtati kontrolni ekran na osnovu dobijenih funkcionalnosti. Samo generičko prikazivanje se fokusira na to da su funkcionalnosti grupisane po prioritetima kako kontrolni ekrani ne bi bili pretrpani nekim, korisniku irelevantnim podacima sa uređaja.

Postojeća ideja podele funkcionalnosti po prioritetima se bazira na tome da u prvu grupu prioriteta spadaju sve senzorske funkcionalnosti koje korisniku pružaju informaciju o trenutnom stanju koje IP uređaj očitava, naravno ukoliko su iste podržane na uređaju.

Ukoliko ne postoji ovaj tip funkcionalnosti, proverava se druga prioritetna grupa koju čine jednostavne kontrolne funkcionalnosti kao što su *Switch*, *Dim*, itd.

U treću grupu funkcionalnosti spadaju komplikovanije funkcionalnosti koje sadrže veći skup fizičkih osobina IP uređaja, pa je samim tim i kontrolni ekran za njih složeniji i zahteva više mesta za prikaz.

Kontrolni ekrani služe da prikažu korisniku najznačajnije informacije, kao što su očitana stanja sa senzora, odnosno komponente za kontrolisanje uređaja. Oni se prvi prikazuju korisniku pri dodavanju uređaja u sistem.

V. TESTIRANJE I REALIZACIJA

U okviru postojećeg sistema kućne automatizacije izvršeno je testiranje inicijalne implementacije predložene ideje. Mobilna aplikacija u kojoj su prikazani kontrolni ekrani IP uređaja pisana je za iOS platformu u *Objective-C* programskom jeziku.

Platforma na kojoj se izvršava softver centralnog kontrolera naziva se DOC400 (*Desktop OBLO Controller 400*) [10]. Bazirana je na Qualcomm WD114 sistemu na čipu [11] sa 64MB RAM memorije.

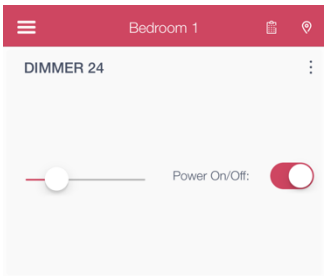
Centralni kontroler šalje informacije klijentskim aplikacijama u JSON (eng. *JavaScript Object Notation*) [12] formatu. Svaka funkcionalnost je struktuirana po određenom standardu. Može sadržati jednu ili više fizičkih osobina od kojih svaka poseduje svoje atribute. Ti atributi su ključni za tačno određen prikaz te fizičke osobine. Atributi sadrže skup vrednosti koje fizička osobina može imati, kao i trenutno podešenu odnosno očitanu vrednost iz tog skupa.

S druge strane, klijentske aplikacije imaju predefinisane klase koje se koriste kako bi se svaka funkcionalnost predstavila na korisničkom interfejsu. Polja tih klasa su tačno predviđena za parsiranje vrednosti iz JSON objekta u primitivne tipove podataka koji se zatim čuvaju u instancama tih klasa. U zavisnosti od tipa funkcionalnosti, iste fizičke osobine se mogu različito mapirati. Na primer, za *MotionDetection* funkcionalnost, koja ima fizičku osobinu *stanje* (skup vrednosti: *true/false*), vrednosti se mapiraju na DA/NE, dok se za *Contact* funkcionalnost, čiji je skup vrednosti isti, vrednosti mapiraju na OTVORENO/ZATVORENO. Za senzorske funkcionalnosti, dodatno se definišu vizuelne komponente koje jasno prikazuju njihovu namenu. Uređaji koji zahtevaju interakciju sa korisnikom sadrže funkcionalnosti kojima je moguće upravljati i one su grafički predstavljene pomoću komponenti kao što su prekidačke (eng. *switch*), odnosno klizne (eng. *slider*) kontrole.

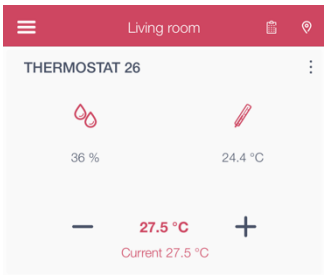
U aplikaciji je definisana klasa koja predstavlja generički uređaj i ona sadrži skup svih funkcionalnosti koje WISE protokol podržava. One su raspoređene po prioritetu i određen je maksimalan broj funkcionalnosti koje je moguće prikazati na jednom kontrolnom ekranu. Nakon što se lista funkcionalnosti dobavi od strane centralnog kontrolera, prolazi se kroz listu podržanih WISE funkcionalnosti i ukoliko se funkcionalnost nalazi u listi dobijenih funkcionalnosti, ona se iscrtava na ekranu.

Uzimajući u obzir da trenutno ne postoje fizički uređaji koji implementiraju WISE protokol, testiranje je obavljeno na simuliranim WISE uređajima koristeći *JMeter* alat [13]. S obzirom da su uređaji virtuelni, njihovi servisi ne sadrže realne karakteristike i služe samo za prikaz informacija korisniku.

Sl. 2. Kontrolni ekran za senzorske funkcionalnosti



Sl. 3. Kontrolni ekran za jednostavniji tip funkcionalnosti



Sl. 4. Kontrolni ekran koji predstavlja i komplikovan tip funkcionalnosti (Postavljanje temperature na termostatu)

Namena *JMeter* alata je simulacija klijenata različitih protokola, što su ovom slučaju IP uređaji. Alat podržava različite protokole i pisan je u Java programskom jeziku.

S obzirom da se oslanja na paralelno izvršavanje, pri čemu svaka nit predstavlja jednog klijenta, iskorišćen je u situaciji gde je jedan klijent zapravo jedan IP uređaj. Svaki od IP uređaja prolazi proces priključivanja internoj mreži centralnog kontrolera. Nakon uspešnog povezivanja, za svaki uređaj simuliran je prijem komandi za dobavljanje informacija o uređaju gde se između ostalog dobija i lista funkcionalnosti koje uređaj podržava.

Od 30 pokrenutih virtuelnih IP uređaja, svih 30 su uspešno prošli proces priključivanja. Prosečno vreme za odgovor na zahtev dobavljanja informacija o uređaju (od strane centralnog kontrolera ka uređaju) je 52ms. Najkraće vreme je 13ms, a najduže vreme izvršavanja 915ms.

Od strane aplikacije je takođe poslat zahtev ka centralnom kontroleru za dobavljanje informacija o uređaju. U odgovoru tog zahteva nalazi se lista funkcionalnosti uređaja. Prolaskom kroz listu, ispitani su prioriteti i funkcionalnosti su povezane sa odgovarajućom grafičkom komponentom koja ih predstavlja.

VI. Zaključak

Protokolom, prikazanim u ovom radu, definisan je proces uparivanja IP uređaja sa centralnim kontrolerom pametne kuće, bez obzira na njegovu lokaciju, kao i ponašanje uređaja i centralnog kontrolera. U internoj mreži centralnom kontroleru je omogućena potpuna interakcija sa IP uređajem kroz protokolom definisanu spregu. S obzirom da uređaji mogu imati različite funkcionalnosti, protokolom je pokriven dovoljan broj funkcionalnosti kako bi se interakcija sa uređajima olakšala, a samim tim i kako bi se nebrojeno mnogo različitih uređaja moglo povezati sa centralnim kontrolerom.

WISE protokol se bazira na MQTT protokolu. Uvođenjem tipova poruka, definisane su oznake unutar MQTT tema za svaki od tipova, a načinom na koji su definisane teme se otvara mogućnost razmene poruka na nivou jedne od funkcionalnosti uređaja. Time je omogućeno pretplaćivanje samo na željenu funkcionalnost sa ciljem dobijanja promena osobina iste, odnosno slanje zahteva radi menjanja neke osobine funkcionalnosti.

Na krajnjoj aplikaciji definisan je prikaz za svaku od funkcionalnosti koja je podržana protokolom. Dobavljanjem niza funkcionalnosti sa uređaja i proverom da li je određena funkcionalnost implementirana protokolom i kom prioritetu pripada, informacije dobijene od nje i komponente za vršenje akcija nad njom, ukoliko te akcije postoje, se prikazuju korisniku.

Prednost ovog rešenja je što centralni kontroler može prepoznati uređaje i u potpunosti upravljati njima čak i ukoliko oni nisu prethodno integrisani u sistem. Pritom, protokol komunikacije opisan ovim radom se bazira na protokolu koji je veoma dobro prihvaćen od strane IoT zajednice po pitanju efikasnosti, što ne dovodi do značajnog usporavanja komunikacije u sistemu.

Zahvalnica

Ovaj rad je delimično finansiran od strane Ministarstva za prosvetu, nauku i tehnološki razvoj Republike Srbije, na projektu broj: TR32014.

Literatura

[1] ZigBee Alliance, *ZigBee specification*, ZigBee document 053474r13, 2006.
[2] C. Paetz, *Z-Wave Basics: Remote Control in Smart Homes*, 2013.
[3] *Bluetooth 4.0: Low Energy*, 2010.
[4] I. Bašičević, M. Popović, V. Kovačević, *Osnovi računarskih mreža 1*, FTN, 2013.
[5] I. Stefanović, *Protokol za komunikaciju IP uređaja sa centralnim kontrolerom pametne kuće*, Univerzitet u Novom Sadu, Fakultet tehničkih nauka, 2018.
[6] A. Donoho, *UPnP Device Architecture 2.0*, Open Connectivity Foundation: Beaverton, OR, 2015
[7] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, T. Barners-Lee, *Hypertext Transfer Protocol – HTTP/1.1*, IETF, 1999
[8] A. Banks, R. Gupta, *MQTT Version 3.1.1*, OASIS Standard, 2014.

[9] M. Tucić, *Protokol za razmenu poruka u sistemima pametnih kuća*, Univerzitet u Novom Sadu, Fakultet tehničkih nauka, 2016.
[10] N. Lukač, *Sistemska programska podrška za DOC400 kontroler zasnovana na OpenWRT sistemu*, Univerzitet u Novom Sadu, Fakultet tehničkih nauka, 2016.
[11] Qualcomm Atheros, Inc, *IoT Client Overview: Control and Automation*, 2015.
[12] *The JavaScript Object Notation (JSON) Data Interchange Format*, IETF, 2017
[13] D. Nevedrov, *Using JMeter to Performance Test Web Services*, Published on dev2dev (http://dev2dev.bea.com/), 2006

ABSTRACT

This paper describes the concept of the solution for displaying the functionality of IP (stands for Internet Protocol) devices to end users of the home automation system in a simplified way. The solution implies a generic representation of the physical properties of the device, whereby device information is supplied using the WISE (stands for WiFi Sensors) protocol. By dividing the functionality of the devices by priority, it is possible to draw the control screens for functionally different devices in an intuitive way. This significantly facilitates the maintenance and expansion of smart home systems for new devices.

Generic representation of functionalities and states of devices in IoT systems

Lana Salai, Igor Stefanovic, Roman Pavlovic, Istvan Papp, Milos Milanovic

Podacima-vođena arhitektura za prilagodljive energetske mreže zasnovana na IoT uređajima

Nenad Petrović, Đorđe Kocić

***Apstrakt*—Povećanje potrošnje električne energije poslednjih godina donosi nove izazove i iziskuje promene. Postojeći sistemi i infrastruktura bi trebalo da postanu fleksibilniji, sa mogućnošću da rade u veoma dinamičnim uslovima, pri čemu treba da reaguju adekvatno na promene koje nastaju u okruženju. Iz tog razloga, u zadnje vreme se teži evoluciji energetske mreže prema takozvanoj pametnoj mreži. U ovom radu, istražuju se mogućnosti primene modernih informacionih i telekomunikacionih tehnologija u pametnim elektroenergetskim mrežama, oslanjajući se na pristupačne IoT uređaje. Kao rezultat istraživanja, predložena je podacima-vođena arhitektura za prilagodljive energetske mreže koja se oslanja na Internet of Things (IoT) uređaje i dat je opis implementacije najbitnijih komponenti sa pojedinim aspektima evaluacije.**

***Ključne reči*— IoT, edge computing, inženjering modelovanja, mašinsko učenje, pametne mreže**

I. UVOD

Električna energija je ključni faktor koji je doveo do munjevitog tehnološkog razvoja ljudskog društva u prošlom veku [1]. Danas, potreba za električnom energijom postaje sve veća, dok je dostupnost neobnovljivih izvora energije sve manja, što vrši pritisak na postojeće distributivne sisteme koji su potencijalno usko grlo [1, 2]. Preopterećenje sistema može izazvati ozbiljne probleme i dramatično uticati na kvalitet prenosa električne energije. Sa druge strane, uprkos povećavanju potrošnje, korisnici očekuju jeftinije cene električne energije. Stoga postoji potreba za unapređenjem postojećih sistema za distribuciju električne energije i za povećanje njihove fleksibilnosti [2]. Mogućnost adekvatnog prilagođavanja promenama i rad u uslovima dinamičnog opterećenja je glavni zahtev modernih sistema i servisa [3].

Iz tog razloga, poslednjih godina je dosta rada uloženo u transformisanje postojećih distributivnih sistema, pomoću informaciono-komunikacionih tehnologija u tzv. *pametne mreže*. Pametna mreža se definiše kao elektroenergetska mreža sledeće generacije, koja je implementirana u vidu dvosmernog sajber-fizičkog sistema sa ugrađenom računarskom inteligencijom koja iskorišćava prikupljene podatke da kontroliše energetsku mrežu sa ciljem stvaranja čiste, bezbedne, pouzdane, otporne i efikasne energetske mreže koja isporučuje električnu energiju krajnjim korisnicima [1, 2, 4, 5]. Implementacija pametnih mreža postaje od strateškog značaja za mnoge zemlje. Jedna od glavnih karakteristika ovakve mreže je mogućnost detekcije događaja bilo gde u mreži i usvajanja odgovarajuće strategije kao odgovor na novonastale uslove rada i sve to skoro u realnom vremenu [1, 2, 4, 5].

U ovom radu istražuje se mogućnost korišćenja najsavremenijih informacionih i telekomunikacionih tehnologija za primenu u unapređivanju energetskih distributivnih mreža oslanjajući se na *Internet of Things* (IoT) uređaje i prikupljene podatke. Kao rezultat istraživanja, predložena je arhitektura, a pojedini aspekti implementacije i evaluacije su detaljnije prikazani.

Nenad Petrović - Elektronski fakultet, Univerzitet u Nišu, Aleksandra Medvedeva 14, 18000 Niš, Srbija (e-mail: nenad.petrovic@elfak.ni.ac.rs).

Đorđe Kocić - Elektronski fakultet, Univerzitet u Nišu, Aleksandra Medvedeva 14, 18000 Niš, Srbija (e-mail: seriousdjoka@gmail.com).

II. OSNOVE I SRODNA ISTRAŽIVANJA

A. Arhitektura pametne mreže

U literaturi postoje mnogi opisi strukturalnih komponenti i modela arhitektura pametnih mreža [4-6]. U Tabeli I, dat je sumarni pregled elemenata koji se najčešće sreću i njihove uloge objašnjene. U ovom radu akcenat je na informacionim i komunikacionim tehnologijama, a ne na komponentama za proizvodnju i prenos električne energije.

TABELA I
KOMPONENTE PAMETNE MREŽE I NJIHOVE ULOGE

Komponenta	Uloga
Energetski podsistem	Proizvodnja, prenos i distribucija električne energije
Sloj komunikacije	Žičane i bežične komunikacione mreže koje omogućavaju prenos informacija između komponenata pametne mreže
Merni uređaji	Uređaji koji vrše merenja električnih i drugih veličina od značaja
Podsistem računarske inteligencije	Izvođenje određenih zaključaka na osnovu prikupljenih podataka i donošenje odluka za izvršenje odgovrajućih akcija
Aplikacije i servisi	Različiti tipovi softvera koje koriste operateri i korisnici pametne mreže, a obuhvata funkcije vizuelizacije, nadzora i kontrole

B. Internet of Things (IoT)

Pojam *Internet of Things* (IoT) se odnosi na sistem raznolikih, međusobno povezanih uređaja koji se koriste u svakodnevnom životu koji ima za cilj automatizaciju određenog domena (od kućnih uređaja, preko proizvodnje do

vojnih primena) [2-6]. Ovi uređaji opremljeni su različitim tipovima senzora, kamerama i modulima za pozicioniranje koji generišu različite vrste podataka. Mogu se takođe opremiti i aktuatorima da bi se njima mogla vršiti i kontrola okruženja. U većini slučajeva njihova procesorska moć je ograničena i moraju da komuniciraju sa drugim uređajima (serverima) da bi postigli svoj cilj. Što se tiče komunikacije koriste se različite tehnologije, kratkog dometa (kao što je Bluetooth) i dugog dometa (Wi-Fi, 4G). Trenutno su na pomolu i 5G mobilne mreže koje omogućavaju veću brzinu prenosa, veći kapacitet prenosa i veću pouzdanost, što će biti jako korisno za primenu sa IoT uređajima [7].

Može se primetiti da IoT ima veliki potencijal za primenu u pametnim mrežama [2, 6]. Pametne mreže zahtevaju da merni i aktuatorski uređaji budu distribuirani širom stambenih i industrijskih oblasti da bi prikupljali neophodne podatke i pružali adekvatne odgovore na novonastale promene. IoT sistemi savršeno odgovaraju toj svrsi, imajući u vidu njihove male dimenzije, priuštivost i činjenicu da je internet konekcija danas široko dostupna.

C. Tehnike analize podataka

Kod pametnih mreža, postoji neophodnost da se analizira velika količina podataka prikupljenih IoT i mernim uređajima i da se iz tih podataka izvuče znanje sa ciljem da se detektuju šabloni ili pojava određenih događaja.

U tu svrhu se koriste različite tehnike mašinskog učenja i *data mining-a*. U većini postojećih radova, fokus je na detekciji anomalija u radu i prognozi opterećenja [4]. U ovim slučajevima najčešće se koriste algoritmi za klastering, klasifikaciju, regresiju i pronalaženja pravila asocijacije, izvršenih nad različitim merenim podacima prikupljenim sa pametnih merača, od električnih parametara do vremenskih prilika [4, 5, 6, 8, 9].

Detekcija anomalija je od najveće važnosti jer pruža mogućnost ranog otkrivanja kvarova i otkaza, tako da pametna mreža može brzo da reaguje i odgovori na promene [4, 5]. U radu [5], korišćen je alogritam klasterovanja zajedno sa otkrivanjem pravila asocijacije sa ciljem da se otkriju problematični događaji, kao što je preopterećenje. Sa druge strane, da bi se optimizovao raspored potražnje energije, potreban je precizan dijagram potrošnje korisnika. Ovde prognoza potrošnje ima bitnu ulogu. U [8] i [9], za prognozu potrošnje koristi se regresija bazirana na pomoćnim vektorskim mašinama.

D. Edge computing

Prikupljanje podataka sa IoT uređaja i njihovih senzora je od velike važnosti za monitoring i odlučivanje u pametnim mrežama. Nažalost, klasičan pristup obradi podataka u oblaku ne daje dobre rezultate, jer bi prenošenje velikog broja podataka koje stvaraju IoT uređaji (senzori) u oblak izazvalo velika kašnjenja u sistemu. Sa druge strane, kako pametna mreža raste, sve veći broj IoT uređaja koji se priključuju izaziva još veće kašnjenje.

S obzirom da pametna mreža mora da odgovori na promene u okruženju u približno realnom vremenu, takva kašnjenja su nedopustiva [10, 11]. Iz ovog razloga, *edge computing* se razmatra kao moguće rešenje. Ideja *edge computing-a* je da se obrada i procesiranje podataka prenesu bliže izvorima stvaranja podataka - *edge* serverima, da bi se omogućilo brže vreme reagovanja [10]. Na primer, u [11] je pokazano da se performanse sistema za nadzor pametne mreže mogu poboljšati čak 10 puta pomeranjem obrade podataka bliže mestima gde se podaci generišu.

E. Semantičke tehnologije

Uloga semantičkih tehnnologija je da opiše značenje podataka i veze između njih odvojeno od samog sadržaja i aplikacionog koda. Na taj način se omogućava da i ljudi i mašine mogu da interpretiraju podatke, razmenjuju i izvode zaključke na osnovu njih. U oblasti semantičkih tehnologija, ontologije se koriste za opis koncepata i relacije među njima u okviru nekog domena. Ovi semantički opisi često čuvaju u formi tripleta (*subjekat, predikat, objekat*) korišćenjem RDF[1] standarda. Za izvršavanje upita nad RDF semantičkim bazama tripleta se koristi SPARQL[2]. Izvršavanjem upita nad semantičkom bazom pribavljaju se rezultati koje je dalje moguće koristiti za mehanizme zaključivanja, sa ciljem da se izvede novo znanje na osnovu postojećih činjenica.

U IoT-baziranim sistemima, semantičke tehnologije se mogu iskoristiti na mnogo različitih načina. Jedna od čestih primena jeste ostvarivanje interoperabilnosti heterogenih IoT uređaja [12]. Osim toga, mogu se koristiti u kompleksnijim scenarijima, poput automatskog generisanja koda sa ciljem koordinacije robota [13, 14]. U [15] je prikazan semantički radni okvir za anotaciju rezultata algoritama računarskog vida u IoT-baziranom sistemu za video-nadzor sa ciljem da se omogući zaključivanje o događajima i reaguje na njih. U ovom radu želimo primeniti kombinaciju navedenih pristupa, pri čemu se umesto procesa koordinacije posmatra adaptacija kao proces.

F. Domenski-specifični jezici

Domenski specifični jezici su specijalizovani za rešavanje problema iz određenog domena, a njihove notacije mogu biti tekstualne ili grafičke (u okviru alata za modelovanje). Uglavnom, u IoT-baziranim sistemima se ovakvi jezici koriste da bi smanjilo kognitivno opterećenje korisnika koje proizilazi iz velike raznorodnosti uređaja i kompleksnosti infrastrukture. U ovom slučaju, kodovi napisani u domenski specifičnim jezicima se prevode na specifične komande pojedinih uređaja.

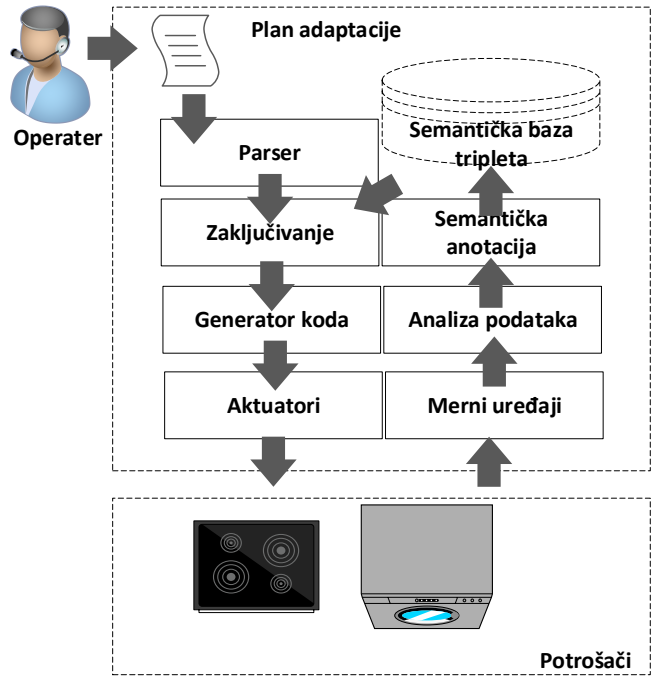
Recimo, u [16] je prikazan EDL jezik za definiciju eksperimenata korišćenjem robotskih IoT uređaja. U kontekstu pametnih električnih mreža, vizuelni alati koji koriste domenski-specifične notacije bi se mogli iskoristiti sa ciljem da se operaterima olakšaju operacije upravljanja. Na ovaj način je omogućena implementacija složenih scenarija bez potrebe da se upušta u implementacione detalje pojedinačnih uređaja. U ovom radu se, konkretno, razmatra modelovanje plana adaptacije kao odgovor pametne mreže na promene iz spoljašnje sredine. Usko povezano postojeće rešenje je Cloud Application Modelling and Execution Language (CAMEL) [17], koji omogućava specificiranje različitih aspekata aplikacija u oblaku, a između ostalog pruža mogućnost modelovanja adaptivnog ponašanja.

[1] https://www.w3.org/RDF/
[2] https://www.w3.org/TR/sparql11-query/

III. PREGLED IMPLEMENTACIJE

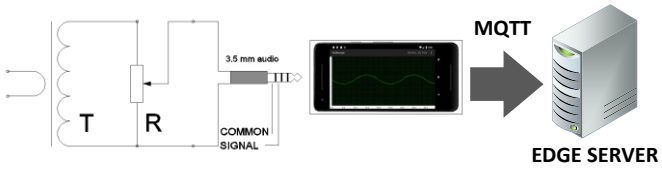
U ovoj sekciji, predložena je arhitektura pametne električne mreže koja se oslanja na IoT uređaje i prethodno navedene tehnologije, a pojedini detalji implementacije detaljnije izloženi.

Arhitektura sistema i pirncip rada: Merenje električnih veličina obavlja se od strane pametnih uređaja za merenje. Ovu ulogu mogu imati mikrokontroleri, single-board računari (poput Raspberry Pi) ili pametni telefoni, kao što je to u našem slučaju. Prednost korišćenja pametnih telefona je u velikom broju dostupnih senzora i ulaza (poput 3.5 mm audio ulaza) Osim toga, njihove punjive baterije i podrška za bežične mobine mreže pružaju mogućnost korišćenja čak i u manje pristupačnim predelima. Zatim, prikupljeni podaci se analiziraju korišćenjem tehnika *data mining-a* i mašinskog učenja. Rezultati dobijeni analizom se semantički anotiraju, tako da se semantičkim zaključivanjem može doći do zaključaka o događajima nastalih u okruženju. Skup akcija koji se izvršava kada su određeni uslovi u okruženju ispunjeni (*plan adaptacije*) se definiše korišćenjem vizuelnog alata za modelovanje koji se oslanja na domenski-specifičnu notaciju. Konačno, na osnovu zadatog plana se generišu komande za specifične uređaje sa ciljem da se odgovori na promene nastale u okolini. Ove komande aktuatori izvršavaju nad potrošačima, ali se sam mehanizam izvršenja komandi i hardverska implementacija aktuatora ne razmatraju u ovom radu. Ilustracija opisanog radnog okvira i principa je data na Sl. 1.



Sl. 1. Podacima-vođen radni okvir za prilagodljive IoT-zasnovane pametne mreže

Pametni merači bazirani na Android platformi: U [18] je prezentovan metod za prikupljanje informacija o električnim merenjima koristeći pristupačne uređaje sa Android platformom. Signali napona i struje se prikupljaju pomoću naponskih i strujnih transformatora sa elektroenergetske mreže. Oba signala se pretvaraju u naponske signale, koje se onda dodatno skaliraju na nivo audio signala posredstvom promenljivih otpornika. Nakon toga signal ide direktno na 3.5 mm audio priključak pametnog uređaja. Pošto većina uređaja baziranih na Androidu podržava stereo ulaz za mikrofon, to u ovom slučaju stvara savršenu dvokanalnu platformu za merenje energetskih signala. Audio sistem uređaja (zvučna kartica) vrši analogno/digitalnu A/D konverziju, tako da se sada podaci mogu obrađivati standardnim metodama obrade digitalnih signala, kao što su algoritmi bazirani na FFT-u. Takođe je korisno snimati i druge podatke koje generiše pametni uređaj kao što su temperatura, vreme i lokacija. Prikupljeni podaci se šalju na *edge* server preko MQTT[3] (Message Queuing Telemetry Transport), lagan, *publish-subscribe* ISO standardizovan protokol za razmenu poruka koji radi povrh TCP/IP protokola. Poruke se šalju kao JSON-kodirani string. U ovom formatu, primer poruke koja sadrži izmerenu vrednost frekvencije električnog signala bi bio: `{"device":"Smarthpone", "sensorType":"frequency", "value":49.9}`. Merni sistem zasnovan na Android pametnim uređajima je ilustrovan na Sl. 2.

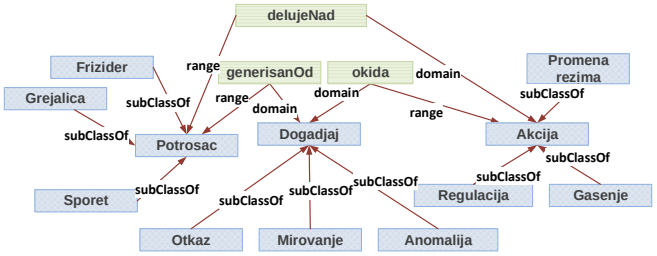


Sl. 2. Merni sistem baziran na Android uređajima

Mehanizmi analize podataka: Za implementaciju mehanizama analize podataka u ovom radu se oslanjamo na TensorFlow za programski jezik Python, biblioteku otvorenog koda za mašinsko učenje. Osim toga, pruža mogućnost izvršenja na GPU, što je pogodno kada je kritično vreme obrade u realnom vremenu. Dva mehanizma su implementirana: 1) detekcija anomalija korišćenjem klasifikacije 2) predviđanje potrošnje zasnovano na regresiji. Klasifikacija je proces identifikacije kom skupu pripada posmatrani uzorak, na bazi trening skupa koji se sastoji od uzoraka čija je pripadnost unapred poznata. Sa druge strane, regresija omogućava da se vrši predviđanje kako se menja vrednost neke zavisne promenljive, kada neka od nezavisnih promenljivih varira. U prvom slučaju, koriste se merenja frekvencije i napona, a treba ih svrstati u jednu od dve kategorije: 1) ok – normalan rad potrošačkog uređaja 2) anomaly – anomalija pri radu. U drugom slučaju koristi se prosečna dnevna temperatura kao nezavisna promenljiva, dok je potrošnja zavisna promenljiva.

Semantički radni okvir: Domenska ontologija (prikazana na Sl. 3) koristi se za semantičku anotaciju rezultata dobijenih nakon analize podataka. Razmatraju se različite vrste događaja koji se mogu detektovati na potrošačkim uređajima: otkazi, anomalije napona, režim mirovanja i slično. Osim toga, za svaki od razmatranih događaja se definiše skup mogućih akcija koje se mogu preduzeti kao reakcija na taj događaj, poput gašenja uređaja, regulacija napona, prebacivanje u režim niske potrošnje.

[3] http://mqtt.org/



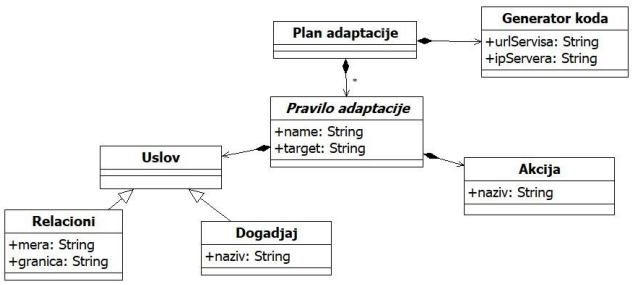
Sl. 3. Domenska ontologija za upravljanje u pametnim mrežama

U okviru Listinga I je dat primer SPARQL upita koji se izvršava kada se proverava uslov da li je detektovana anomalija napona na bar jednom potrošaču.

LISTING I
PRIMER SPARQL UPITA IZVRŠENOG PRILIKOM DETEKCIJE ANOMALIJA

```
PREFIX sgc: <http://www.example.com/resources/sgc/>
SELECT ?Dogadjaj
WHERE {
  GRAPH <http://www.example.com/test> {
  ?Dogadjaj sgc:generisanOd ?Potrosac.
          FILTER(regex(STR(?Dogadjaj), "anomalija-napon"))
  }
}
```

Vizuelni alat za modelovanje plana adaptacije: Razvijen je korišćenjem Node-RED kao osnove. Node-RED je intuitivni i proširivi radni okvir za povezivanje IoT uređaja, aplikacionih programskih interfejsa i servisa na inovativne načine. Okruženju za modelovanje se pristupa korišćenjem web pretraživača, a pruža jednostavan korisnički interfejs sa proširivom paletom elemenata za modelovanje (čvorova). Domenski-specifična notacija korišćena u okviru alata je data u obliku metamodela na Sl. 4. Metamodel predstavlja model programskog jezika koji definiše strukturu i ograničenja za familiju modela. U ovom radu, koristi se za definisanje adaptacionog plana koji se sastoji od niza pravila adaptacije. Svako pravilo adaptacije ima uslov, cilj i akciju koju treba izvršiti nad ciljem kada uslov bude ispunjen. Ovi uslovi mogu biti konkretni događaji (otkazi, anomalije itd.), ali i relacioni izrazi koji se odnose na granice određenih merenih veličina. Osim toga, svaki plan adaptacije mora posedovati i jedan element generatora koda, koji sadrži kao parametre IP servera na kome se generator nalazi i URL za poziv servisa. Na ovaj način je omogućeno da se odgovarajući generator koda odabere zadavanjem parametra, bez modifikacije samog alata za modelovanje, što olakšava proširivost i prilagođavanje za različite platforme pametnih mreža u budućnosti. Okruženje za modelovanje u okviru Node-RED je prikazano na Sl. 5.



Sl. 4. Metamodel plana adaptacije

Zaključivanje i generisanje komandi: Izvršava se po algoritmu datom u Listingu II, inspirisanom mehanizmom predstavljenom u [19]. Svaki uslov se prevodi u SPARQL upit koji se izvršava nad semantičkom RDF bazom znanja. Ukoliko upit vrati rezultat *true*, onda se generiše odgovarajući kod i dodaje u skriptu komandi koje aktuatori treba da izvrše nad potrošačima. Konačno, generisane komande se šalju ciljanim uređajima takođe kao MQTT poruke.

LISTING II
ALGORITAM GENERISANJA KOMANDI

Ulaz: merenja, plan adaptacije
Izlaz: skripta komandi
Steps:
1. Pribaviti sva pravila adaptacije iz plana adaptacije
2. Analizirati senzorske podatke;
3. Izvršiti semantičku anotaciju rezultata iz Koraka 2;
4. For each AdaptationRule
5. uslov:=izvršiUpit(AdaptationRule.condition);
6. If (uslov is true)
7. then generisati komandu za ciljani uređaj;
8. end for each
9. end

IV. EVALUACIJA I REZULTATI

U ovoj sekciji predstavljeni su rezultati ostvareni u eksperimentima koristeći prethodno opisani radni okvir, iz dve različite perspektive. Prvo se razmatraju performanse mehanizama za detekciju anomalija i predviđanje potrošnje. Nakon toga, analizira se vreme neophodno za procesiranje u određenim koracima prilikom generisanja koda, sa ciljem da se utvrdi sposobnost sistema da blagovremeno reaguje na promene. Eksperimenti su izvršavani na serveru sa procesorom AMD Ryzen 7 1700X octa-core CPU 3.80GHz, 64GB DDR4 RAM i NVIDIA Quadro P2000 GPU sa 4GB VRAM.

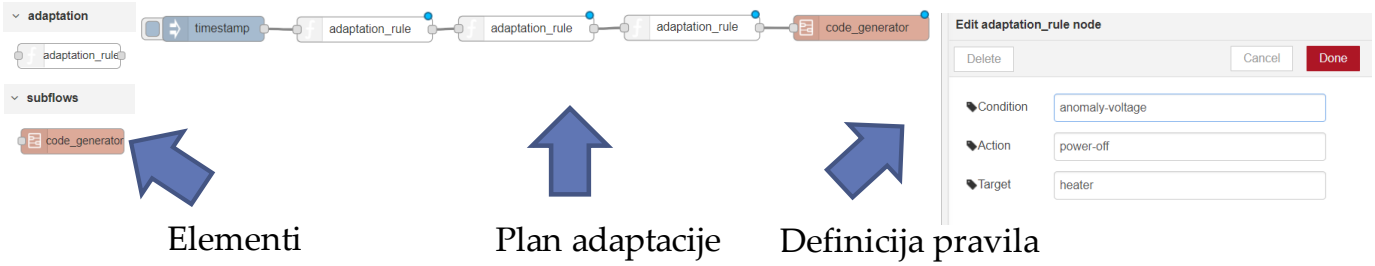
U Tabeli II I Tabeli III, prikazani su rezultati dobijeni prilikom korišćenja predstavljenog pristupa za detekciju anomalija i predviđanje potrošnje, implementirani korišćenjem TensorFlow u Python programskom jeziku. Za različite eksperimente, prikupljeni podaci su podeljeni u dva disjunktna skupa u različitim odnosima – trening i test skup.

TABELA II
REZULTATI DETEKCIJE ANOMALIJE UPOTREBOM KLASIFIKACIJE

Veličina trening skupa	Veličina test skupa	Tačno klasifikovani/ veličina testa [%]
50	100	84.93
75	100	89.58
100	100	92.17

TABLE III
REZULTATI PERDVIĐANJA POTROŠNJE KORIŠĆENJEM REGRESIJE

Veličina trening skupa	Veličina test skupa	Relativna greška [%]
50	100	18,31
100	100	16,29
150	100	15,21



Sl. 5. Okruženje za modelovanje plana adaptacije u okviru Node-RED

Kao što se može videti iz tabela, oba mehanizma analize podataka daju zadovoljavajuće rezultate u različitim odnosima veličine trening i test skupa, s tim da daju bolje rezultate za veće trening skupove, do određene granice. Međutim, nakon daljeg povećanja veličine trening skupova postoji tendencija overfitting-a, pri čemu se ne dolazi do boljih rezultata.

Sa druge strane, u Tabeli IV je dat pregled ostvarenih vremena procesiranja u koracima analize podataka i generisanja koda za različit broj pravila adaptacije u okviru plana adaptacije prilikom detekcije anomalija. Svako pravilo se odnosi na različit potrošački uređaj, dok su podaci merenja prikupljeni tokom 600 sekundi. Što se tiče analize podataka, prikazani rezultati predstavljaju sumu vremena za trening i test (u razmeri: 70% trening skup, 30% test skup), a razdvojeno su razmatrana vremena procesiranja dobijena primenom CPU i GPU. Sva merenja vremena su data u sekundama kao prosek deset uzastopnih merenja u identičnim uslovima.

Prema dobijenim rezultatima, vreme koje je potrebno za analizu podataka povećava se sa brojem razmatranih potrošačkih uređaja iz razloga što više uređaja generiše veću količinu podataka za obradu. Osim toga, i generisanje koda traje duže iz razloga što je potrebno izvršiti veći broj SPARQL upita u slučaju većeg broja pravila adaptacije, s obzirom da svako pravilo sadrži uslov koji se prevodi u SPARQL upit, a i veći je broj parametara koje je neophodno pribaviti i umetnuti u šablon komandi. Može se zapaziti da je u nekim slučajevima vreme generisanja koda za isti broj pravila različito, a to proizilazi iz činjenice da se komande generišu samo onda kada je uslov pravila adaptacije ispunjen, a ne uvek. Što se tiče poređenja performansi izvršenja na CPU i GPU, može se uočiti da se detekcija anomalija izvršava do oko 2.4 puta brže na GPU nego na CPU, pri čemu se ubrzanje povećava sa porastom količine podataka.

TABELA IV
PREGLED VREMENA PROCESIRANJA U SLUČAJU DETEKCIJE ANOMALIJA

Broj pravila	Analiza podataka (CPU) [s]	Analiza podataka (GPU) [s]	Generisanje koda[s]
1	4.31	3.11	1.67
2	7.28	3.67	2.43
2	7.54	3.95	1.91
3	9.82	4.07	3.08

V. ZAKLJUČAK I BUDUĆA ISTRAŽIVANJA

U ovom radu, razmotrene su tehnologije koje omogućavaju adaptaciju u pametnim električnim mrežama koje se oslanjaju na sveprisutne IoT uređaje. Vizuelni alat za modelovanje plana adaptacije daje mogućnost realizacije kompleksnih scenarija bez potrebe poznavanja implementacionih detalja samih uređaja. Sa druge strane, implementirani mehanizmi za detekciju anomalija napona i prognozu potrošnje pokazuju zadovoljavajuće rezultate, pri čemu vreme izvršenja dobijeno na GPU u slučaju detekcije anomalija deluje obećavajuće, čak i za primene u skoro realnom vremenu. Naravno, planiramo da isprobamo i različite implementacione varijante i testiramo njihove performanse. Osim toga, cilj nam je da realizujemo kompletne merne sisteme koristeći još jeftinije uređaje, poput Raspberry Pi i Arduino.

Međutim, u budućnosti ćemo se fokusirati više na implementaciju aktuatorskih mehanizama, koji nisu predstavljeni u ovom radu. Još jedan bitan aspekat koji nije obuhvaćen ovim radom, a od velike je važnosti za IoT-zasnovane pametne mreže, jeste bezbednost, s obzirom da IoT uređaji komuniciraju uglavnom bežičnim kanalima koji su česta meta napada. U ovom slučaju, napadi mogu imati posledice katastrofalnih razmera, poput nestanka električne energije, ali i oštećenja infrastrukture i ugrožavanje živih bića [20]. Prema tome, još jedna od tema naših budućih istraživanja biće upravo bezbednost.

ZAHVALNICA

This work has received funding from the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Grant Agreement No 645220, project RAWFIE (Road-, Air- and Water- based Future Internet Experimentation). Ovaj rad je delimično finansiran od strane Ministarstva prosvete, nauke i tehnološkog razvoja Republike Srbije u okviru projekta III44006.

LITERATURA

[1] A. H. Bagdadee, L. Zhang, Smart Grid: A Brief Assessment of the Smart Grid Technologies for Modern Power System, Journal of Engineering Technology, vol. 8, no. 1, January 2019, pp. 122-142, 2019.

[2] M. Jaradat et al., "The Internet of Energy: Smart Sensor Networks and Big Data Management for Smart Grid", Procedia Computer Science 56 (2015) , pp. 592–597, 2015.

[3] R. Kazhamiakin, S. Benbernou, L. Baresi, P. Plebani, M. Uhlig, O. Barais, "Adaptation of Service-Based Systems", Service Research Challenges and Solutions for the Future Internet, Lecture Notes in Computer Science, vol 6500. Springer, Berlin, Heidelberg, pp. 117156, 2010.
[4] B. Rossi, S. Chren, "Smart Grids Data Analysis: A Systematic Mapping Study" [submitted for publication], pp. 1-26, 2018. Available on: https://arxiv.org/pdf/1808.00156.pdf
[5] B. Rossi, S. Chren, B. Buhnova, T. Pitner, "Anomaly Detection in Smart Grid Data: An Experience Report", IEEE International Conference on Systems, Man, and Cybernetics (SMC 2016), pp. 1-6, 2016.
[6] H. Shahinzadeh, J. Moradi, G. B. Gharehpetian, H. Nafisi, M. Abedi, "IoT Architecture for Smart Grids", 2019 International Conference on Protection and Automation of Power System (IPAPS), pp. 22-30, 2019.
[7] W. Ejaz et al., "Internet of Things (IoT) in 5G wireless communications", IEEE Access 4, pp. 10310–10314, 2016.
[8] M. Božić, M. Stojanović, Z. Stajić, "Short-Term Electric Load Forecasting Using Least Square Support Vector Machines", Facta Universitatis, Series: Automatic Control and Robotics vol. 9, no. 1, pp. 141-150, 2010.
[9] A. B. M. S. Ali and S. Azad, "Demand Forecasting in Smart Grid" Green Energy and Technology, pp. 135–150, 2013. https://doi.org/10.1007/978-1-4471-5210-1_6
[10] W.Z. Khan et al., Edge computing: A survey, Future Generation Computer Systems (2019), pp. 1-44, 2019.
[11] Y. Huang et al., "An Edge Computing Framework for Real-Time Monitoring in Smart Grid", 2018 IEEE International Conference on Industrial Internet (ICII), pp. 99-108, 2018.
[12] R. Agarwal et al., "Unified IoT ontology to enable interoperability and federation of testbeds", 2016 IEEE 3rd World Forum on Internet of Things (WF-IoT), pp. 70-75.
[13] V. Nejkovic, N. Petrovic, N. Milosevic, M. Tosic, "The SCOR Ontologies Framework for Robotics Testbed", 2018 26th Telecommunication Forum (TELFOR), Belgrade, pp. 1-4, 2018. https://doi.org/10.1109/telfor.2018.8611841
[14] N. Petrovic, V. Nejkovic, N. Milosevic, M. Tosic, "A Semantic Framework for Design-Time RIoT Device Mission Coordination", 2018 26th Telecommunication Forum (TELFOR), Belgrade, pp. 1-4, 2018. https://doi.org/10.1109/telfor.2018.8611845
[15] N. Petrovic, "Surveillance System Based on Semantic Video and Audio Annotation Leveraging the Computing Power within the Edge, XIV International SAUM 2018, pp. 281-284, 2018.
[16] K. Kolomvatsos, M. Tsiroukis, S. Hadjiefthymiades, "An Experiment Description Language for Supporting Mobile IoT Applications", FIRE Book, European Commission, River Publishers, 2016, pp. 461-486.
[17] A. Rossini et al., "The cloud application modelling and execution language (CAMEL)," Open Access Repositorium der Universität Ulm, pp. 1-39, 2017.
[18] Đ. Kocić, N. Petrović, "Application of Android Based Devices in Analog Electric Signal Measurement", YuInfo 2019, Kopaonik, Serbia, pp. 1-5, 2019.
[19] N. Petrovic, M. Tosic, V. Nejkovic, N. Milosevic, "Formalizing Device Coordination in IoT Systems: The SCOR Case Study", YuInfo 2019, pp. 1-6, 2019.
[20] V. Delgado-Gomes, J. F. Martins, C. Lima, C., P. N. Borza, "Smart grid security issues", 2015 9th International Conference on Compatibility and Power Electronics, pp. 534-538, 2015.

ABSTRACT

The dramatic increase of demand for electric power in recent years has introduced new challenges and requirements. The existing infrastructure and systems should become more flexible, able to operate in highly dynamic environments and react adequately to changes. For that reason, the existing electric grid evolves towards Smart Grid architecture. In this paper, we explore how state-of-art information and communication technologies can be exploited to enable adaptability in Smart Grid relying on affordable IoT devices. As an outcome, we propose a data-driven architecture of adaptable Smart Grid based on IoT devices. Moreover, some implementation and evaluation aspects are also presented.

Data-driven architecture for adaptable Smart Grid based on IoT devices

Nenad Petrović, Đorđe Kocić

Automatizacija radnog okruženja za ispitivanje složenih audio sistema

Filip Uzunović, Branko Đorđević, Nenad Pekez, Jelena Kovačević

***Apstrakt*—U ovom radu je predstavljeno automatizovano radno okruženje za ispitivanje složenih audio sistema, čijom upotrebom se smanjuje vreme neophodno za ispitivanje do 35 %. Kao precizan i pouzdan uređaj za efikasno ispitivanje ovih sistema koji zamenjuje skupa rešenja na tržištu, korišćen je RT-AG (RT – Audio Graber). RT-Executor koji omogućava automatsko izvršavanje ispitnih slučajeva, predložen je kao programska podrška ispitivanju. U radu je takođe navedeno i na koji način je unapređena sama procedura za ispitivanje.**

***Ključne reči*—Audio sistemi, Ispitivanje, Verifikacija, Automatizacija**

I. Uvod

Audio tehnologije su doživele ogroman napredak od početne dvokanalne (stereo) prezentacije, preko više kanalnih prezentacija, pa sve do savremenih 3D objektnih prezentacija.

Kanalne prezentacije zapravo predstavljaju tradicionalan način prenosa i reprodukovanja zvuka. Svaki kanal je dizajniran tako da se reprodukuje na tačno određenom mestu u sistemu u odnosu na slušaoca. 2D zvučna scena kombinacija je velikog broja pojedinačnih zvučnih elemenata, dijaloga snimljenih na setu, muzike snimljene u studiju itd. Ovi elementi distribuirani su na nekoliko audio zapisa u skladu sa standardizovanom konfiguracijom zvučnika [1]. Krajnji rezultat ovog procesa jeste skup kanala, gde svaki od njih predstavlja sadržaj jednog od zvučnika. Međutim takav signal pravilno će se reprodukovati na samo jednoj izlaznoj konfiguraciji zvučnika odnosno zvučnoj sceni. Naravno postoji mogućnost reprodukcije takvog signala i na drugim zvučnim scenama za koje nije predviđen, ali to zahteva još neke dodatne korake u obradi a sam doživljaj zvuka neće biti prostorni ukoliko zvučnici nisu dobro postavljeni. Danas su audio sistemi bazirani na kanalima i dalje veoma rasprostranjeni i vrlo stabilni, iako postoji sve veća težnja ka sistemima za reprodukciju 3D zvuka.

Za opisivanje prostiranja 3D zvuka koristi se određen broj 3D audio objekata [2]. Za razliku od kanala čije pozicije su tačno definisane, pozicija objekata u prostoru može biti:

- potpuno nezavisna od lokacije prisutnih zvučnika u sistemu
- mogu varirati tokom vremena za modelovanje pokretnih objekata kao što je avion koji se kreće iznad glave slušaoca

Jedno od rešenja za univerzalno i efikasno kodovanje i pružanje kvalitetnog prostornog zvučnog sadržaja jeste 3D objektno orijentisano renderovanje. Osim što je u stanju da obezbedi visoki kvalitet prostornog zvuka, ima mogućnost i da objedini mnoštvo 3D audio formata od nižih stereo, 5.1, do 22.1. Na ovaj način se postiže i kompatibilnost novih formata sa prethodnim uređajima.

Kako su nove generacije audio tehnologija donele određene promene, dovele su i do povećanja složenosti sistema. Samim tim proces verifikacije i ispitivanja ovih rešenja postaje složeniji i predstavlja sve veći izazov za inženjere. Podatak koji najbolje oslikava složenost ovog procesa jeste porast broja ispitnih slučajeva od nekoliko stotina do nekoliko hiljada ispitnih slučajeva. Metode manualnog ispitivanja postaju neefikasne u primeni, kako sa stanovišta utrošenog vremena, tako i pouzdanosti vezane za ocenu ispravnosti testiranog sistema. Ručno ispitivanje je zavisno od ljudske interpretacije, odnosno podložno greškama QA inženjera prilikom ocene ispravnosti sistema, usled nedovoljnog poznavanja specifikacije sistema ili iskustva. Sa druge strane, primena metoda ručnog testiranja oduzima značajan deo vremena potrebnog za razvoj sistema. Usled težnje proizvođača da obezbede konkurentnost na tržištu, razvoj sistema u ovoj oblasti je u velikoj meri uslovljen potrebom za skraćenjem vremena pojavljivanja proizvoda na tržištu [3]. Jedan od načina da se to postigne jeste skraćenje procesa ispitivanja. Sa druge strane, nedovoljno ispitan proizvod dovodi do smanjenja kvaliteta, što negativno utiče na konkurentnost. Da bi se obezbedilo kraće vreme proizvodnje i pružio bolji kvalitet, potrebno je više pažnje posvetiti analizi kvaliteta sistema i unaprediti proces ispitivanja, čime bi se poboljšao njegov kvalitet. Time se javlja potreba za automatizacijom ovog procesa, što predstavlja glavnu temu rada.

U ovom radu predstavljeno je radno okruženje za automatizaciju ispitivanja složenih audio sistema. Prvo je naveden opis hardverskog sistema i tipične ispitne procedure, koja je do sada primenjivana, kao i njeni nedostaci koje ponuđeno radno okruženje otklanja. Potom je u narednim poglavljima predstavljena sama realizacija predloženog rešenja. Opisuje se RT-Audio Graber (RTAG), uređaj razvijen u okviru RT-RK instituta, koji služi za reprodukciju i snimanje audio signala. Zatim je objašnjeno kako je tipična

Filip Uzunović – Istraživačko-razvojni Institut RT-RK, Novi Sad, Srbija (e-mail: filip.uzunovic@rt-rk.com)
Branko Đorđević – Istraživačko-razvojni Institut RT-RK, Novi Sad, Srbija (e-mail: branko.djordjevic@rt-rk.com)
Nenad Pekez – Istraživačko-razvojni Institut RT-RK, Novi Sad, Srbija (e-mail: nenad.pekez@rt-rk.com)
Jelena Kovačević – Istraživačko-razvojni Institut RT-RK, Novi Sad, Srbija (e-mail: jelena.kovacevic@rt-rk.com)

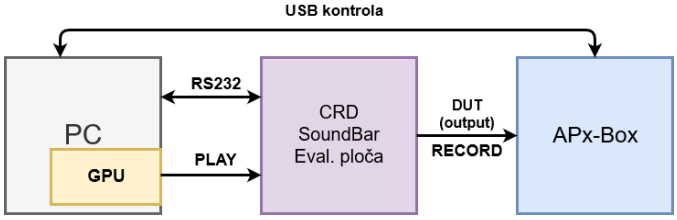
ispitna procedura unapređena i navedeno je na koji način je ova procedura automatizovana. Ističe se kako upotreba RT-Executora, koji osim za automatizaciju izvršavanja ispitnih slučajeva, pomoću svojih ugrađenih algoritama omogućava i analizu snimljenih izlaza. Na kraju su prikazani rezultati primene predloženog radnog okruženja.

II. TIPIČNA ISPITNA PROCEDURA

Kompanije koje obezbeđuju algoritme (eng. IP provider) u isto vreme obezbeđuju i pakete za ispitivanje tih algoritama. Tipična struktura paketa za ispitivanje se sastoji od dokumentacije, test materijala i alata koji se koriste.

Kompanija, kao što je npr. Dolby, obezbeđuje QA inženjeru upitnik čijim popunjavanjem se dobija spisak ispitnih slučajeva koje je neophodno izvršiti kako bi se verifikovala ispravnost ispitivanog uređaja. QA inženjer specificira o kom se konkretnom uređaju radi, koliko izlaznih kanala podržava taj uređaj, koji audio izlazi postoje, koje izlazne konfiguracije zvučnika podržava taj uređaj itd. Popunjavanjem upitnika generiše se spisak ispitnih slučajeva, ispitnih signala za svaki slučaj kao i konfiguracione poruke koje je neophodno poslati uređaju prilikom samog izvršavanja. Na osnovu ovih informacija QA inženjer piše Python skriptu koja služi kao šablon po kom se kreira Audio Precision (APx) projekat u kom se izvršavaju testovi.

Sledeći korak je postavljanje i povezivanje hardverskog sistema, čija struktura je predstavljena na Sl. 1.



Sl. 1Prikaz tipičnog hardverskog sistema

APx je višekanalni audio analizator, koji podržava 8 ili 16 simultanih analognih ulaza i izlaza, napravljen za dizajniranje i ispitivanje korisničkih uređaja [4]. Prikaz APx uređaja dat je na Sl. 2

Sl. 2 Audio Precision APx585

U ispitnoj proceduri nalazi se u ulozi mernog instrumenta, koji reprodukuje ili snima audio signale i ima svoju softversku podršku. PC vrši upravljanje grafičkom karticom i izvršavanje testova po koracima:

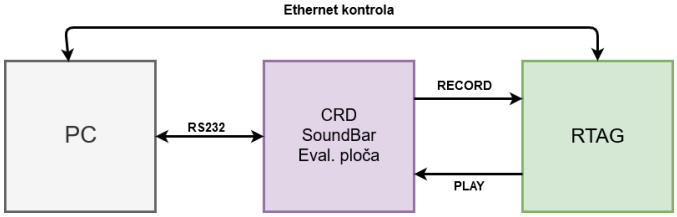
- podešavanje samog uređaja slanjem konfiguracionih poruka
- reprodukovanje ispitnog signala preko grafičke kartice
- merenje i analiza izlaza snimljenih APx mernim instrumentom sa ploče/uređaja

Svaki od navedenih koraka izvršava se zasebno za svaki ispitni slučaj, a samo izvršavanje je ručno. Izvršavanje jednog testa traje oko 5 min a broj ispitnih slučajeva može da ide od nekoliko stotina do nekoliko hiljada. Slanje konfiguracionih poruka je usko vremenski ograničeno i ukoliko se ne ispoštuje tačno definisan vremenski interval neće biti uspešno. Sve ovo zahteva potpunu pažnju i konstantnu prisutnost QA inženjera, uz činjenicu da oduzima značajan broj radnih sati. Takođe sam APx, zbog svoje kompleksnosti, važi za veoma skup uređaj. Samim tim, evidentna je potreba i javlja se prostor za unapređenjem ove procedure i njenu automatizaciju čija realizacija će biti predstavljena u narednom poglavlju.

III. REALIZACIJA REŠENJA

A. *Hardverski sistem*

Hardverski sistem, prikazan na Sl. 3 unapređen je zamenom APx-a sa audio uređajem RTAG, koji je razvijen u razvojno istraživačkom institutu RT-RK. RTAG predstavlja eksternu zvučnu karticu koja obezbeđuje ulaze i izlaze audio signala, do i sa računara, pod kontrolom kompjuterskog programa/aplikacije [5]. Podržava kako reprodukciju tako i snimanje audio signala do 8 kanala na frekvencijama do 192kHz/32b. Konekcija sa računarom ostvaruje se putem Ethernet-a. Primarno RTAG je napravljen da se koristi u kombinaciji sa DSP evaluacionim pločama, ali je moguće koristiti ga u paru sa krajnjim korisničkim audio uređajima kao što su Blu-Ray, Sound Bar, AVR itd.



Sl. 3 Prikaz unapređenog hardverskog sistema

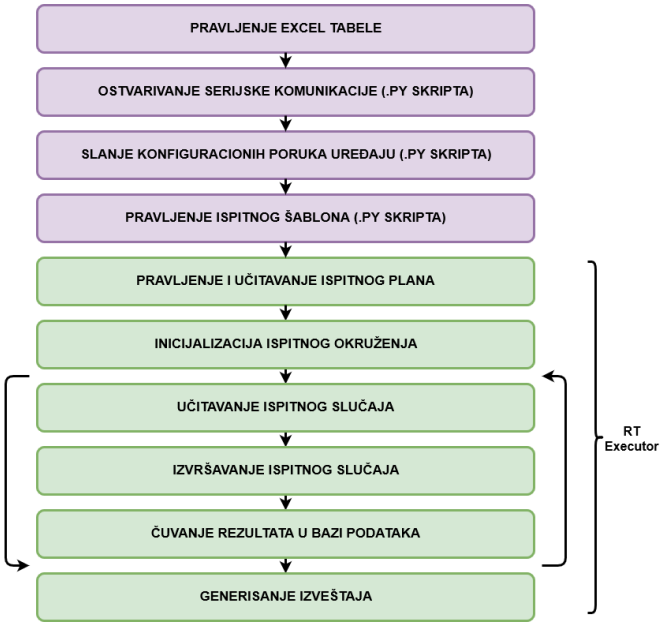
B. *Unapređena ispitna procedura*

Redosled kreiranja i izvršavanja automatskih test slučaja predstavljen je na Sl. 4. Za programsku podršku ispitivanju odabran je RT-Executor, program razvijen u istraživačko razvojnom institutu RT-RK, čime je postignuta automatizacija procesa samog izvršavanja ispitnih slučajeva. RT-Executor je samostalna aplikacija, osposobljena da kontroliše veliki broj softverskih i hardverskih modula neophodnih za automatsko testiranje [6]. Podržava ispitivanje po principu “crne kutije”.

Na osnovu specifikacije ispitivanog uređaja pišu se testovi kojima će se ispitati da li uređaj uspešno izvršava sve

predviđene operacije. Ispitni slučajevi se prave na osnovu šablona, napisanog u programskom jeziku Python, i redom izvršavaju sledeće korake za svaki ispitni slučaj:

- otvaranje serijskog porta
- prebacivanje MCU (Microcontroller Unit) u UART mode (Universal Asynchronous Reciever-Transmitter)
- slanje odgovarajućih konfiguracionih poruka MCU
- reprodukciju odgovarajućeg testnog signala
- upisivanje konfiguracionih poruka iz MCU na DSP (Digital Signal Procesor)
- snimanje izlaza



Sl. 4 Prikaz unapređene ispitne procedure

Slanje konfiguracionih poruka obavlja se putem serijske komunikacije, a poruke za svaki ispitni slučaj mogu biti različite. Poruke se dobijaju pomoću Python skripte koja parsira svaki ispitni slučaj posebno. Prethodno isparsirane poruke za promenu režima rada i individualne poruke, prvenstveno se šalju MCU koji ih dalje upisuje na DSP. Ove korake takođe izvršava zasebna Python skripta koja se poziva unutar testa za taj konkretan ispitivan slučaj.

Kada su napravljeni ispitni slučajevi koji pokrivaju kompletnu funkcionalnost uređaja i kada se proveri da su ispravno napisani od njih se pravi projekat koji se pokreće u RT-Executor-u.

RT-Executor smešta rezultate ispitnih slučajeva u bazu podataka iz koje generiše izveštaj u vidu HTML (Hyper Text Markup Language) stranice. Rezultat izvršavanja može biti jedan od tri statusa: PASS, FAIL, UNRESOLVED koji jasno govore da li je ispitni slučaj uspešno izvršen (PASS), neuspešno izvršen (FAIL) ili je njegov slučaj nerazrešen (UNRESOLVED). Ukoliko je isitni slučaj iz određenih razloga neuspešno izvršen ili je status nerazrešen, QA inženjer kroz sam izveštaj može da vidi i opis greške koja postoji. Da bi proglasili da je uređaj uspešno ispitan, svi ispitni slučajevi moraju biti "PASS". U slučaju da nakon provere i ponavljanja ispitnih slučajeva rezultati ponovo nisu pozitivni, QA inženjer prijavljuje grešku.

C. *Analiza snimljenih izlaza*

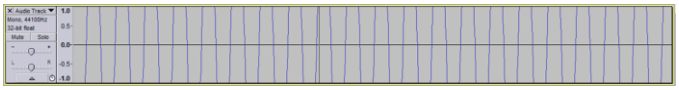
RT-Executor pomoću svojih ugrađenih alata omogućava jednostavnu analizu audio izlaza sa ispitivanog uređaja, po svakom snimljenom kanalu. Jedan od takvih alata je Audio Quality Meter (AQM).

AQM je softverski dodatak za RT-Executor koji u sebi ima audio algoritme za prepoznavanje različitih audio artifakta kao što su isecanje, prekidi, zamrzavanje i izostanak audio signala. Na kraju svakog ispitnog slučaja vrši se pozivanje određenog AQM audio algoritma koji nam javlja da li je prisutan neki od audio artifakta. Ukoliko ne postoje, RT-Executor proglašava ishod ispitnog slučaja kao PASS, ili u suprotnom će ishod biti FAIL.

Algoritam odsustva zvuka analizira audio signal za nultu vremensku aktivnost. Ako audio signal nedostaje duže od navedenog vremena, detektuje se odsustvo. Postoje dva tipa algoritama odsustva: provera odsutnosti u vremenskim i frekvencijskim domenima. Algoritam takođe podržava dodavanje nivoa buke u dB (podrazumevano = -60 dB) koju treba zanemariti tokom analize. Primer odsustva zvuka u audio signalu prikazan je na Sl. 5

Sl. 5 Primer odsustva zvuka u audio signalu

Isecanje predstavlja izobličenje talasnog oblika koji se javlja kada je pojačalo preopterećeno i pokušava da isporuči izlazni napon ili struju van svojih opsega. Algoritam za audio isecanje u RT-Executoru vrši analizu u vremenskim i frekvencijskim domenima i detektuje periode kada je signal isečen zbog vrlo visokog nivoa amplitude. Primer takvog signala dat je na Sl. 6



Sl. 6 Primer isecanja u audio signal

IV. REZULTATI

Predloženo radno okruženje i ispitna procedura primenjena je prilikom ispitivanja Sound Bar uređaja, za odabranu konfiguraciju zvučnika 5.1.2, na audio tehnologiji Dolby Atmos HT 1.6.2.

PRIMENJIVANA METODA ISPITIVANJA	BROJ ISPITNIH SLUČAJEVA	VREME POTREBNO ZA ISPITIVANJE JEDNOG SLUČAJA	BROJ PONAVLJANIH TESTOVA	UKUPNO VREME POTREBNO ZA ISPITIVANJE UREDAJA
STARA METODOLOGIJA	138	5 min	15	~ 13h
NOVA METODOLOGIJA	138	3 min	3	~ 8,5 h

Sl. 7 Rezultati ispitivanja Sound Bar uređaja i konfiguraciju zvučnika 5.1.2

U tabeli sa Sl. 7 jasno se vidi kako ponuđeno rešenje utiče na pouzdanost samog ispitivanja. Na broj ispitnih slučajeva koji se moraju ponavljati najviše je uticao faktor ljudske greške koji je u novoj proceduri, upotrebom RT-Executora, uspešno eliminisan. Broj ovakvih ispitnih slučajeva sveden je na minimum i smanjen čak 5 puta. Ukupno vreme neophodno za ispitivanje uređaja smanjeno je za više od 30%. Ušteda vremena još je značajnija ukoliko se uzme u obzir podatak da je ukupan broj ispitnih slučajeva neophodnih za verifikaciju Dolby tehnologija porastao sa 4226 za Dolby Atmos HT 1.5 na 6539 za Dolby Atmos HT 1.6.2.

V. Zaključak

Ponuđeno radno okruženje zajedno sa celokupnim sistemom za automatsko ispitivanje eliminiše sve nedostatke ručnog ispitivanja. Automatizacijom ovog procesa dobija se značajna ušteda na vremenu. Čak i do 35% manje vremena je potrebno kako bi se verifikovala ispravnost jednog uređaja. Takođe, kako se testovi automatski izvršavaju, QA inženjer ne mora konstantno nadgledati proces, što mu ostavlja prostor da obavlja još jedan posao istovremeno čime se dodatno dobija na vremenu. Sve ovo smanjuje vreme potrebno za pojavljivanje samog proizvoda na tržištu. Upotreba RTAG-a, čija cena je manja čak i do 30 puta u odnosu na prethodno korišćeni APx, donosi značajnu finansijsku uštedu.

VI. Zahvalnica

Ovaj rad je delimično finansiran od strane Ministarstva za nauku i tehnologiju Republike Srbije, na projektu tehnološkog razvoja broj TR32029.

Literatura

[1] Dolby Laboratories Inc: Surround Sound Past, Present, and Future, 1999.

[2] B.Claypool, W. Van Baelen, B. Van Daele, "Auro 11.1 versus object-based sound in 3D", 2015.

[3] N. Paunović, J. Kovačević, I. Rešetar, "Jedno rešenje metodologije ispitivanja složenih sistema profesionalne elektronike", Conference for Electronics, Telecommunications, Computers, Automatic Control and Nuclear Engineering (ETRAN), Zlatibor, Srbija, Jun, 2011.

[4] Audio Precision, "APx58x B Series Audio Analyzers", 2019. [Online]. Available: https://www.ap.com/analyzers-accessories/apx58x/, [posećen 03.04.2019.]

[5] N. Pekez, J. Kovačević, M. Beljulji, "Tehničko rešenje softverske arhitekture digitalnog audio grebera i plejera RT-AG", Conference for Electronics, Telecommunications, Computers, Automatic Control and Nuclear Engineering (ETRAN), Zlatibor, Srbija, Jun, 2017.

[6] V. Peković, N. Teslić, I. Rešetar, T. Tekcan, „Test Management and Test Execution System for Automated Verification of Digital Television Systems", 14th International Symposium on Consumer Electronics, Braunschweig, Germany, Jun, 2010.

Abstract

This paper presents an automated framework for testing complex audio systems, whose use reduces the time required for testing up to 35%. RT-Audio Graber was used as an accurate and reliable device for efficient testing of these systems, which replaces the expensive solution on the market. RT-Executor, which allows the automatic execution of test cases, is proposed as a software support for testing. The paper also outlines the way in which the test procedure itself has been improved.

Automated framework for testing of complex audio systems

Filip Uzunović, Branko Đorđević, Nenad Pekez, Jelena Kovačević

Realizacija aplikacija za konkurentno snimanje i reprodukciju multimedijalnog sadržaja na uređajima sa Android operativnim sistemom

Marko Milovanović, Marčeta Zoran, Milan Ačanski, Nikola Vranić, Istraživačko-razvojni institut RT-RK, Novi Sad, Srbija

***Apstrakt* — Android je najpopularniji operativni sistem i platforma za mobilne telefone i tablet računare. Milioni mobilnih uređaja širom sveta danas koriste Android kao primarnu softversku platformu. Iako je prvenstveno dizajniran za uređaje sa ekranom osetljivim na dodir, Android postaje deo širokog spektra platformi, pa tako i digitalnih prijemnika. Opremljeni bogatim multimedijalnim opcijama i podrškom za instalaciju besplatnih Android aplikacija, digitalni prijemnici su odlično rešenje za nadogradnju običnog televizora na Smart TV. Snimanje multimedijalnog sadržaja, kao jedna od opcija Androidovog multimedijalnog podsistema, podržano je za više kontejnerskih formata, ali Android trenutno ne podržava snimanje datoteka u MPEG-TS formatu, najpopularnijem formatu u oblasti digitalne televizije. U ovom radu opisano je proširenje multimedijalnog podsistema za snimanje TS tokova, koje je iskorišćeno za potrebe realizacije aplikacija, čija je glavna funkcionalnost mogućnost istovremenog snimanja i reprodukcije multimedijalnog sadržaja.**

***Ključne reči* — set-top box, Media Muxer, Media Codec, MPEG-TS, stagefright**

I. UVOD

Nakon pojave televizije u boji sredinom pedesetih godina 20-og veka, prva značajna evolucija u oblasti televizije bila je pojava digitalne televizije. Tranzicija sa analogne na digitalnu televiziju započeta je početkom 21-og veka i danas je digitalna televizija gotovo u potpunosti preuzela primat.

Uporedo sa ovim globalnim trendom, razvijali su se i digitalni TV prijemnici (eng. set-top box, STB), čija je prvobitna uloga bila prijem i konverzija digitalnog u analogni signal, čime je omogućeno emitovanje digitalne televizije na starijim, analognim TV uređajima. Njihov razvoj bio je direktno povezan sa napretkom tehnologije. Vremenom su dodate nove funkcionalnosti. Elektronski programski vodič, roditeljska zaštita, izbor omiljenih kanala i snimanje postali su standardni deo implementacije svakog STB uređaja.

Nakon više od 10 godina na tržištu, Android operativni sistem nalazi primenu na velikom broju platformi, pa tako i na STB uređajima. Jedan od razloga za njegovo prihvatanje je poznato okruženje i brojne funkcionalnosti koje krajnji korisnici već poseduju na svojim mobilnim telefonima – pristup internetu, izvršavanje aplikacija najšireg skupa funkcionalnosti i mnoge druge. [1]

Snimanje multimedijalnog sadržaja jedna je od standardnih funkcionalnosti digitalnog TV prijemnika. Najkorišćeniji format za prenos digitalnih TV podataka u oblasti digitalne televizije naziva se MPEG-TS (skraćeno TS). Android trenutno ne podržava snimanje TS tokova, pa je za potrebe realizacije aplikacija iskorišćen jedan predlog proširenja mltimedijalnog podsistema.Glavna zamisao realizovanih aplikacija je da simuliraju rad STB uređaja. Aplikacije imaju mogućnost istovremenog snimanja i reprodukcije multimedijalnog sadržaja u TS formatu. Snimanje prosleđenih audio i video tokova u odredišnu datoteku sa strane aplikacije podražava snimanje sadržaja jednog ili više TV servisa sa strane STB uređaja, dok reprodukcija sa strane aplikacije imitira emitovanje sadržaja jednog TV servisa sa strane STB uređaja. Realizovane su dve aplikacije: native, napisana u C/C++ programskim jezicima i Java Android aplikacija, napisana u Java programskog jeziku. Obe aplikacije poseduju isti skup funkcionalnosti, ali se razlikuju u sprezi koju pružaju korisniku.

Na samom početku dat je opis radnog okruženja. Par reči biće posvećeno digitalnoj televiziji i formatima za prenos digitalnih TV podataka. U četvrtom poglavlju opisana je arhitektura Androidovog sistema. Peto poglavlje daje opis proširenja sistema za snimanje, kao i realizovanih aplikacija, dok je šesto poglavlje posvećeno testiranju funkcionalnosti navedenih aplikacija.

II. RADNO OKRUŽENJE

Kako se Android razvijao, javila se potreba i mogućnost korišćenja Androida na širokom spektru platformi. Google je

Marko Milovanović, Istraživačko razvojni institut RT-RK, Narodnog Fronta 23a, Novi Sad, Srbija (telefon: 381-21-480-1234, email: marko.milovanovic@rt-rk.com)

Nikola Vranić, Istraživačko razvojni institut RT-RK, Narodnog Fronta 23a, Novi Sad, Srbija (telefon: 381-21-480-1238, email: nikola.vranic@rt-rk.com)

Zoran Marčeta, Istraživačko razvojni institut RT-RK, Narodnog Fronta 23a, Novi Sad, Srbija (telefon: 381-21-480-1297, email: zoran.marceta@rt-rk.com)

Milan Ačanski, Istraživačko razvojni institut RT-RK, Narodnog Fronta 23a, Novi Sad, Srbija (telefon: 381-21-480- 1256, email: milan.acanski@rt-rk.com)

razvio porodicu prenosnih uređaja baziranih na Androidu, Google Nexus, u koju su uključeni mobilni telefoni (Nexus 6P, Nexus 5X), tablet uređaji (Nexus 9), kao i plejeri za reprodukciju multimedijalnog sadržaja (Nexus Player). [2]

Uređaj na kom je testirano dato proširenje sistema, kao i realizovane aplikacije, je navedeni multimedijalni plejer, Nexus Player.

Na samom početku je bilo neophodno skinuti Android projekat (eng. Android Open Source Project, AOSP). Svi neophodni sistemski rukovaoci, kao i verzije projekta za ovaj uređaj nalaze se na sajtu. Aktuelna verzija Android operativnog sistema dostupna za Nexus Player je Android Oreo.

Nakon skidanja koda i drajvera, i primene zakrpe nad kodom u cilju proširenja sistema za snimanje, bilo je neophodno izvršiti prevođenje projekta. Uspostavljane okruženja za prevođenje postiže se upotrebom lunch komande, koja je zadužena za odabir određene konfiguracije okruženja za generisanje programske slike, dok se komandom make započinje proces prevođenja.

Java aplikacija je u velikoj meri realizovana u razvojnom okruženju Android Studio, ali proces prevođenja kao i pokretanje aplikacije nije bilo moguće ostvariti u njemu. Razlog za to je upotreba AOSP biblioteka, neophodnih za izradu aplikacija. Budući da Android Studio koristi drugačiji sistem za prevođenje u odnosu na Android projekat, projekat je prebačen u stablo AOSP-a uz par izmena kako bi se izvršio proces prevođenja.

III. Digitalizacija televizije

Digitalizacija televizije je fenomen koji prati nekoliko povezanih trendova:

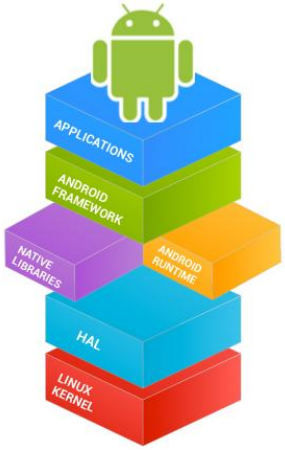
- konstantno povećavanje dimenzije ekrana
- smanjivanje veličine uređaja
- konstantno povećanje rezolucije slike
- integracija multimedijalnih i internet tehnologija
- nove digitalne usluge i aplikacije na TV ekranu

Razvoj televizije i razvoj različitih oblasti se, u svetlu gore navedenih oblasti, nadopunjuju.

Digitalna televizija obuhvata prenos televizijskog signala od emitera do gledaoca digitalnim putem, korišćenjem digitalnih modulacionih tehnika, u obliku digitalnog prenosnog toka u okviru kojeg se vremenski multipleksirano prenose paketi vezani za audio i video sadržaje, metapodatke i dodatne sadržaje koji se odnose na više servisa. Format za prenos digitalnih TV podataka, kontejnerski format, oblikovan je kao bitski tok sa tačno definisanom paketskom organizacijom. Postoji više standarda za kontejnerske formate kojima se prenose multimedijalne informacije, a najčešće korišćen standard za kontejnerske formate koji se koriste u digitalnoj televiziji je MPEG-2 Verzija 1. Ovim standardom definisana su dva tipa kontejnerskog formata: MPEG-PS format, koji se koristi kod prenosnih medijuma koji nisu podložni pojavi grešaka, i MPEG-TS format, koji se koristi kod prenosnih medijuma podložnim pojavi grešaka pri prenosu emisionim putem. [3]

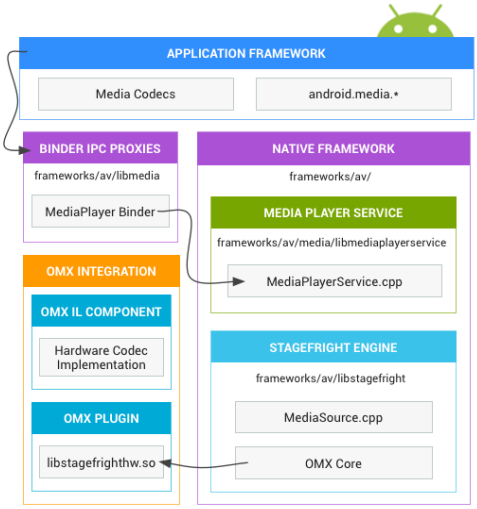
IV. Arhitektura Androida

Android operativni sistem sadrži specifičnu programsku podršku koja je podeljena u nekoliko suštinski različitih slojeva (Sl. 1), a prednosti su brojne: jasno definisanje uloge svakog sloja omogućuje visok stepen modularnosti i enkapsulacije, izolacija pojednih slojeva olakšava paralelizaciju razvoja, moguća je verifikacija svakog sloja ponaosob, što olakšava pronalaženje grešaka i povećava kvalitet programske podrške i dr. [4]



Sl. 1. Arhitektura Android operativnog sistema

Arhitektura multimedijalnog podsistema Androida je po ugledu na celokupnu arhitekturu takođe slojevita, sa puno nivoa indirekcije (Sl. 2).



Sl. 2. Multimedijalni podsistem

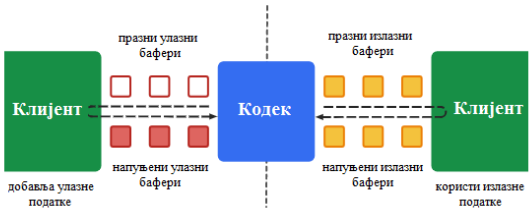
Aplikacije komuniciraju sa nižim delovima sistema preko sloja programskih okvira, koji nudi interfejse za pristupanje i manipulaciju nad multimedijalnim sadržajem. Ispod ovog sloja, stvari se komplikuju jer delovi sloja programskih okvira komuniciraju sa native komponentama preko JNI interfejsa.

Zahtevi iz Java programske podrške se prosleđuju MediaPlayer servisu putem Inter Process Comunication (IPC) poziva. U okviru MediaPlayer servisa pravi se komponenta za reprodukciju audio i video sadržaja odgovarajućeg tipa, a osnovna komponenta je StagefrightPlayer. [5] On koristi

StagefrightEngine, u kom su smeštene sve funkcionalnosti vezane za reprodukciju i snimanje audio i video datoteka. Deo StagefrightEngine-a je Stagefright biblioteka, čije su najbitnije komponente MediaCodec, MediaExtractor i MediaMuxer. MediaCodec je klasa koja apstrahuje enkoder ili dekoder, MediaExtractor je klasa koja služi za dohvatanje traka iz prosleđenog toka i čitanje podataka iz njih, a MediaMuxer je klasa namenjena za snimanje tokova u fajl. [6]

Ove klase su deo android.media API-ja, pa je moguće praviti multimedijalne aplikacije koje ne koriste celokupan definisani stek, već samo deo od Stagefright biblioteke na dole. Prednost ovakvog pristupa je brže izvršavanje aplikacija ali, zbog kompleksnosti nižih slojeva i njihove slabe dokumentovanosti, potrebno je uložiti dodatno programersko vreme.

Na slici 3 je na uprošćen način prikazan rad kodeka. Kodek procesira ulazne podatke kako bi generisao izlazne podatke. Podaci se obrađuju asinhrono, koristeći ulazne i izlazne bafere. Klijent šalje zahtev za dobijanje praznog ulaznog bafera od kodeka, zatim ih puni podacima, dobijenim najčešće od MediaExtractor klase, i šalje ih kodeku na dekodovanje. Kodek nakon dekodovanja primeljenih podataka puni jedan od izlaznih bafera. Klijent šalje zahtev za dobijanje izlaznog bafera, koristi njegov sadržaj i vraća ih natrag kodeku. Dekodovani podaci se mogu prikazati korišćenjem odgovarajućih apstrakcija za prikaz audio/video sadržaja. Osim prikaza podaci se mogu proslediti MediaMuxer klasi koja će vršiti njihov upis u fajl. Na taj način vršimo snimanje željenog multimedijalnog sadržaja. [7]



Sl. 3. Dekodovanje podataka pomoću kodeka

V. OPIS REŠENJA

A. *Proširenje sistema za snimanje*

Formati koje MediaMuxer trenutno podržava su MPEG4, Webm i 3GP. Dakle, Android ne podržava snimanje sadržaja u najkorišćenijem formatu za prenos digitalnih podataka – MPEG Transport Stream.

Glavni deo komponente MediaMuxer je writer, tj. instanca bilo koje klase koja nasleđuje MediaWriter apstraktnu klasu. U korišćenoj verziji Android-a, mukser je imao mogućnost kreiranja dve vrste MediaWriter-a: MPEG4Writer i WebmWriter, ali istraživanjem ostatka biblioteke uočena je i klasa MPEG2TSWriter. Najveći deo izmena koda vezan je za ovu klasu, ali određene izmene je trebalo odraditi i u klasi MediaMuxer. U MPEG2TSWriter klasi nalazi se sva funkcionalnost specifična za snimanje sadržaja u TS formatu. Proširenjem je omogućeno snimanje TS fajla sa jednim programom i po jednom audio i video trakom, po ugledu na podržano snimanje MP4 formata. Moguće je snimanje samo sa kodecima Advanced Audio Coding (AAC) za audio trake i Advanced Video Coding (AVC) za video trake, jer Android trenutno podržava samo reprodukciju TS-a sa navedenim kodecima. [8]

Izmene u kodu realizovane su u vidu zakrpe (eng. patch) koju je trebalo primeniti nad kodom Android projekta pre same izrade aplikacija.

B. *Native aplikacija, MediaRecorder*

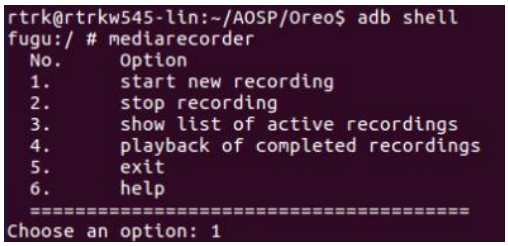
Glavni delovi native aplikacije su klase DeltaPlayer i MediaRecorder. Za potrebe ovog rada napisana je biblioteka DeltaPlayer i ona je namenjena isključivo reprodukciji multimedijalnog sadržaja. Reprodukcija sadržaja u TS formatu je podržana u Androidu, ali samo uz korišćenje AVC kodeka za video i AAC kodeka za audio. DeltaPlayer preveden je kao deljena biblioteka libdeltaplayer.

Osnovna funkcionalnost MediaRecorder klase je snimanje audio i video tokova izvorišnog fajla u odredišni fajl pomoću muksera, odnosno objekta klase MediaMuxer. Nakon inicijalizacije date klase, poziva se njena start() metoda, u kojoj se pokreće mukser i startuje nit nad threadLoop() metodom, koja je zadužena za glavnu obradu, snimanje odgovarajućih tokova u fajl (Sl. 4).

```
void* MediaRecorder::threadLoop(void* arg) {
    while (true) {
        AMediaExtractor_getSampleTrackIndex();
        AMediaExtractor_getSampleTime();
        AMediaExtractor_getSampleFlags();

        buffer = malloc(bufferSize);

        //! Fill input buffer with coded data
        if (AMediaExtractor_readSampleData(buffer) == -1) {
            break;
        }

        //! Setup buffer for muxing
        buffer = ...

        //! Mux data
        AMediaMuxer_writeSampleData(buffer);

        free(buffer);

        //! Read next data
        AMediaExtractor_advance();
    }

    return NULL;
}
```

Sl. 4. Telo threadLoop() metode

Ulazna tačka aplikacije je funkcija main(). Pregled svih funkcionalnosti prikazuje se korisniku putem komandne linije, preko koje se i obavlja potrebna komunikacija (Sl. 5).



Sl. 5. Izgled native aplikacije

Aplikacija ima sledeće funkcionalnosti:

- pokretanje novog snimanja
- zaustavljanje snimanja
- pregled snimaka u toku
- reprodukcija završenih snimaka

Aplikacija je konkurenta, jer je omogućeno snimanje više tokova istovremeno. Pokretanje novog snimanja podrazumeva kreiranje novog objekta klase MediaRecorder, pri čemu putanje do izvorišnog i odredišnog fajla zadaje korisnik. Prilikom zaustavljanja snimanja korisnik unosi redni broj koji predstavlja identifikator na osnovu kojeg se dohvata odgovarajući objekat klase MediaRecorder, a zatim poziva stop() metoda nad datim objektom. Pregled snimaka u toku daje korisniku uvid u informacije vezane za aktivne snimke, odnosno snimke u toku. Kod reprodukcije završenih snimaka, najpre se prikazuju informacije o snimcima. Korisnik zatim unosi redni broj snimka, na osnovu kojeg se iz baze dohvata putanja do fajla, odnosno snimka koji treba prikazati. Nakon toga instancira se komponenta za reprodukciju audio i video sadržaja, odnosno objekat klase DeltaPlayer, čijem konstuktoru se prosleđuje putanja do fajla koji je potrebno prikazati. Nakon inicijalizacije, komponenta za reprodukciju je spremna za pokretanje. Funkcionalnosti vezane za rad sa snimcima su:

- reprodukcija snimljenog sadržaja
- pauziranje reprodukcije snimljenog sadržaja
- nastavak reprodukcije snimljenog sadržaja
- zaustavljanje reprodukcije snimljenog sadržaja

Izvorni fajl u kom se nalazi klasa MediaRecorder, kao i funkcija main(), prevedeni su u izvršni fajl mediarecorder. Zajedno sa bibliotekom libdeltaplayer, izvršni fajl je prebačen na NexusPlayer pomoću adb alata. Za pokretanje native aplikacije najpre treba pristupiti terminalu uređaja, a zatim ukucati ime programa koji je potrebno pokrenuti.

C. *Java Android aplikacija, MediaRecorderApp*

Sve funkcionalnosti native aplikacije, implementirane su i grafičko-korisničkoj sprezi. Od postojeće native aplikacije napravljena je deljena biblioteka, uz određene izmene. Sama aplikacija sastoji se iz dve aktivnosti: MainActivity i PlaybackActivity. U MainActivity klasi obezbeđene su sve funkcionalnosti vezane za snimanje multimedijalnog sadržaja, dok je PlaybackActivity klasa zadužena za reprodukciju snimljenog sadržaja. Da bi aplikacija komunicirala sa native bibliotekama, bilo je neophodno uspostaviti vezu posredstvom JNI interfejsa jer nije moguće direktno pozivanje C koda u Java kodu. Zato sa jedne strane imamo Java klasu sa deklaracijama potrebnih metoda koje se koriste u navedenim aktivnostima (obratiti pažnju na ključnu reč native), a sa druge strane korespodentne C funkcije. Prilikom statičke inicijalizacije Java klase vrši se učitavanje biblioteke libmediarecorderjni, u kojoj se nalaze implementacije svih nabrojanih metoda, napisane na C programskom jeziku. Za generisanje zaglavlja u kom se nalaze deklaracije C funkcija koristi se alat javah.

Nakon prevođenja unutar stabla AOSP-a, kreirani .apk fajl se prebacuje na Nexus Player pomoću adb alata. Izgled početnog ekrana aplikacije prikazan je na Sl. 6.

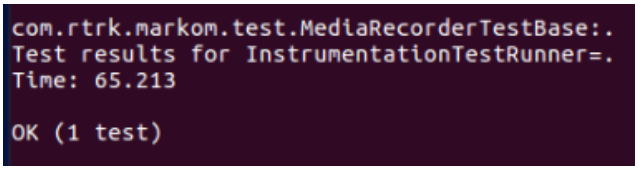
Sl. 6. Izgled početnog ekrana aplikacije

VI. TESTIRANJE

Nakon realizacije aplikacija bilo je neophodno verifikovati ispravnost njihovih funkcionalnosti. Napravljena je test aplikacija MediaRecorderTest, koja ne predstavlja ništa drugo nego aplikaciju sa testovima, realizovanu pomoću test klasa.

Sami testovi realiozovani su kao test metode klase MediaRecorderTestBase. Ona nasleđuje ActivityInstrumentationTestCase2 klasu koja je deo android-test biblioteke.

MediaRecorderTestBase klasa sadrži jednu test metodu, testPlayVideo(), u okviru koje se najpre vrši snimanje audio i video tokova iz izvorišnog u odredišni fajl sa predefinisanih putanja, a nakon snimanja testiraju se sve funkcionalnosti vezane sa rad sa snimcima. Prevođenje se obavlja na identičan način kao i prevođenje MediaRecorderApp aplikacije. Za pokretanje test aplikacije, najpre treba pristupiti terminalu uređaja, a zatim uneti odgovarajući komandu kojom se pokreću testovi. Ono što je neophodno navesti je ime test paketa u kom se nalaze napisani testovi i naziv *runner* klase koja ih pokreće. Klasa koja je korišćena za pokretanje testova je android.test.InstrumentationTestRunner i ona je navedena unutar manifesta date aplikacije. InstrumentationTestRunner najpre instancira objekat klase MediaRecorderTestBase, a zatim pokreće njene test metode, odnosno testPlayVideo() metodu. Po završetku testova ispisuje se naziv klase u okviru koje se nalaze test metode, vreme njihovog izvršavanja i broj testova sa konačnim rezultatom (Sl. 7).



Sl. 7. Rezultati testova

VII. ZAKLJUČAK

Nakon proširenja multimedijalnog podsistema za snimanje TS tokova, realizovane su dve varijante komponente za reprodukciju audio i video sadržaja, jedne u obliku native aplikacije, MediaRecorder i druge u obliku Java Android aplikacije, MediaRecorderApp. Obe aplikacije poseduju iste funkcionalnosti, ali se razlikuju u načinu komunikacije sa korisnikom, odnosno sprezi koji pružaju. Glavna odlika ovih komponenti je simultano snimanje i reprodukcija multimedijalnog sadržaja, čime se simulira rad digitalnih TV prijemnika u jednom od režima njihovog rada.

Jedna od mogućnosti u pogledu unapređenja datih aplikacija bila bi istovremena reprodukcija više snimaka, čime bi se simulirao rad STB uređaja u režimu emitovanja više TV servisa, zatim unapređenje interfejsa, kao i uvođenje dodatnih funkcionalnosti karakterističnih za STB uređaje.

ZAHVALNICA

Ovaj rad je delimično finansiran od strane Ministarstva za prosvetu, nauku i tehnološki razvoj Republike Srbije, na projektu broj: TR32029.

LITERATURA

[1] „Android (operating system)“, https://en.wikipedia.org/wiki/Android_(operating_system), pristupano april 2019.

[2] „Google Nexus“, https://en.wikipedia.org/wiki/Google_Nexus

[3] M. Bjelica, N. Teslić, V. Mihić, „Softver u digitalnoj televiziji 1: Osnove digitalne televizije i video kodovanja“, FTN Izdavaštvo, Novi Sad, 2017.

[4] I. Pan, N. Lukić, „Projektovanje i arhitekture softverskih sistema: Sistem zasnovani na Androidu“, FTN Izdavaštvo, Novi Sad, 2015

[5] „Android's Stagefright Media Player Architecture”, https://quandarypeak.com/2013/08/androids-stagefright-media-player-architecture/ , pristupano april 2019.

[6] „Media“, https://source.android.com/devices/media/, pristupano april 2019.

[7] „MediaCodec“, https://developer.android.com/reference/android/media/MediaCodec, pristupano april 2019.

[8] D. Petrović, M. Zeković, N. Vranić, „Jedno rešenje proširenja sistema za snimanje multimedijalnog sadržaja na Android baziranim uređajima“, 25th Telecommunications forum TELFOR, Beograd, 2017

ABSTRACT

Android is the most popular operating system and platform for mobile phones and tablets. Millions of mobile devices around the world today use Android as the primary software platform. Although primarily designed for touch screen devices, Android is becoming part of a broad spectrum of platforms, including digital receivers. Equipped with rich multimedia options and support for the installation of free Android applications, digital receivers are a great solution for upgrading a standard TV to Smart TV. Media recording, one of the functionalities of the Android multimedia subsystem, is supported for multiple container formats, but Android currently does not support recording in the MPEG-TS format, the most popular format in the field of digital television. This paper describes the extension of the multimedia subsystem for recording TS streams, which is used in applications, whose main functionality is concurrent media recording and playback.

Applications for concurrent media recording and playback on Android devices

Marko Milovanović, Marčeta Zoran, Milan Ačanski, Nikola Vranić

Razvoj mobilne aplikacije zasnovan na testovima korišćenjem XCTest okruženja

Dražen Drašković

***Apstrakt* – U ovom radu opisana je implementacija mobilne aplikacije kviz znanja za iOS platformu sa akcentom na testiranju takve aplikacije. Aplikacija je projektovana kao sistem sa *Model View Controler* (MVC) arhitekturom. Tokom razvoja aplikacije korišćena je metodologija planiranja i izvršavanja testova u toku implementacije, a kao alat za pisanje testova korišćeno je XCTest okruženje. Razvoj aplikacija zasnovan na testovima je vrlo važna tehnika kod softverskih inženjera, pošto podiže kvalitet softvera i smanjuje nedostatke u softveru. U istraživanju su korišćene strategije bele kutije. Realizovani su jedinični testovi za testiranje logike aplikacije i jedinični testovi za testiranje korisničkog interfejsa. Dobijena pokrivenost koda realizovanim jediničnim testovima je 68.85%, a pokrivenost testovima korisničkog interfejsa je 75.02%.**

***Ključne reči* – testiranje softvera, razvoj mobilne aplikacije, *Test Driven Development*, jedinično testiranje, testiranje korisničkog interfejsa, testiranje performansi.**

I. Uvod

Razvojem mobilnih tehnologija povećala se i potreba za razvojem specifičnih alata i okruženja. Najpopularnije platforme za mobilne uređaje današnjice su Android i iOS. Prema podacima iz prvog kvartala 2019. godine ukupan broj aplikacija na portalu Google Play bio je oko 2.1 miliona, a ukupan broj aplikacija na portalu App Store oko 1.8 miliona mobilnih aplikacija [1]. Android i iOS imaju primat i u broju korisnika, sa oko 85% i 13% tržišta respektivno [2].

Aplikacije zastarevaju zbog promena tehnologija, a usled nedostatka vremena i resursa da se aplikacija piše iz početka, prebacivanje aplikacije sa jednu na drugu tehnologiju često je podložno velikom broju grešaka u programskom kodu. Pored toga jako je bitna pouzdanost aplikacije, koja se postiže unapred definisanim procesom testiranja. Da bi se zaštitili od kasnijih grešaka, softverski inženjeri su počeli da pišu testove u sklopu programskog koda aplikacije (eng. *Test Driven Development*, skraćeno: TDD*)* ili neposredno pre pisanja programskog koda funkcionalnosti aplikacije. TDD ciklus obuhvata učitavanje testnih podataka, pokretanje delova programskog koda aplikacije sa učitanim podacima i potvrđivanje rezultata testa. Potrebno je da se nakon izvršenog testa dobije očekivani rezultat ili se rade izmene na programskom kodu i proces refaktorizacije.

Dražen Drašković – Elektrotehnički fakultet, Univerzitet u Beogradu, Bulevar kralja Aleksandra 73, 11020 Beograd, Srbija (e-mail: drazen.draskovic@etf.bg.ac.rs).

Cilj ovog rada bio je da se prikažu prednosti pisanja i izvršavanja testova tokom ciklusa razvoja aplikacije za mobilnu platformu. Postoji dosta alata i biblioteka za testiranje iOS aplikacija tokom samog ciklusa razvoja, a jedna od njih je XCTest. To je biblioteka koja služi za razvoj aplikacije korišćenjem testova bele kutije. Jedinični testovi predstavljaju formu testiranja, gde se testovi pišu za manje celine softvera koji se razvija. Sa XCTest je moguće pisati jedinične testove za testiranje logike aplikacije i testove za testiranje korisničkog interfejsa. U razvoju aplikacije osim XCTest biblioteke, korišćeni su programski jezici Python i Swift i MySQL relaciona baza podataka za čuvanje podataka. Kao mobilna aplikacija, koja prati realizovan proces testiranja, implementirana je igra kviz znanja sa nekoliko funkcionalnosti.

U drugom poglavlju ovog rada prikazan je pregled postojećih mobilnih aplikacija sa sličnom tematikom i kratak opis korišćenih tehnologija. U poglavlju III opisane su realizovane funkcionalnosti mobilne aplikacije. Poglavlje IV prikazuje ključne implementacione detalje, a poglavlje V proces testiranja, od dizajniranja test plana i test primera, do izvršavanja testova i analize dobijenih rezultata. Na kraju rada dat je zaključak.

II. Pregled postojećih rešenja i korišćenih tehnologija

U ovom poglavlju dat je pregled postojećih najpopularnijih mobilnih aplikacija sa kvizovima i njihova uporedna analiza. Sve aplikacije koje su obuhvaćene ovom analizom su besplatne za korišćenje.

A. Pregled postojećih kviz mobilnih aplikacija

QuizUp [3] je kviz sa velikim izborom kategorija pitanja. Na raspolaganju su teme iz oblasti filma, muzike, sporta, kulture, nauke i drugih. Korisnici mogu pristupiti ovoj aplikaciji korišćenjem svog naloga na nekoj drugoj društvenoj mreži poput Facebook-a ili Twitter-a. Na osnovu toga ova kviz aplikacija omogućava korisniku i da izazove svoje prijatelje koje ima na društvenoj mreži.

DK Quiz [4] je aplikacija koja pokriva preko 100 tema i podtema iz oblasti istorije, nauke ili umetnosti. Može se igrati samostalno ili protiv prijatelja sa društvenih mreža. Takođe, aplikacija pravi rang listu, pa se mogu uporediti rezultati sa drugim učesnicima kviza.

Quiz Your Friends [5] je aplikacija u kojoj izazivate u kvizu

prijatelja. Pozivnica za učešće u kvizu se može poslati prijateljima, može se igrati protiv nekog nepoznatog korisnika, videti rezultat ostalih korisnika, ali i napraviti kviz sa svojim pitanjima.

Logos Quiz [6] je kviz za prepoznavanje logoa svetskih kompanija. Ova aplikacija omogućava igranje protiv pravog protivnika, ili protiv kompjuterskog inteligentnog agenta. Ova kviz aplikacija ima nivoe, pa kako se prelaze lakši nivoi, kviz postepeno postaje sve teži.

Trivia Crack [7] je kviz u kome se na slučajan način na početku kviza dobija jedna od šest mogućih kategorija pitanja. Svaka igra se sastoji od pet pitanja. Algoritam igre je takav da ukoliko se što brže odgovori, igrač dobija više poena. Može se videti kako protivnik napreduje po tome kako se ikonica kreće na traci iznad prostora sa pitanjima. Pitanja su različite težine i shodno tome donose različit broj bodova. Ova igra se može igrati sa prijateljem koji ima aktivan nalog na društvenoj mreži Facebook, tako što ga lično pozovete, ili protiv slučajno odabrane osobe, koja se trenutno nalazi unutar kviza i želi da otpočne novu igru.

U tabeli I dat je uporedni prikaz mobilnih aplikacija sa kvizovima znanja, kao i funkcionalnosti koje su podržane.

TABELA I
UPOREDNI PRIKAZ MOBILNIH APLIKACIJA SA KVIZOVIMA

Funkcionalnost	QuizUp	DK Quiz	Quiz Your Friends	Logos Quiz	TriviaCrack
Izbor kategorija pitanja	Da	Ne	Ne	Ne	Da
Logovanje preko društvenih mreža	Da	Da	Da	Da	Da
Pozivanje prijatelja	Da	Ne	Ne	Ne	Da
Izazivanje prijatelja na igru	Da	Da	Da	Da	Da
Upoređivanje rezultata sa drugima	Ne	Da	Da	Da	Ne
Postavljanje kviza na društvenu mrežu	Ne	Ne	Da	Ne	Ne
Nivoi u kvizu	Ne	Da	Ne	Da	Da
Prikaz trenutno aktivnih prijatelja u igri	Da	Ne	Ne	Ne	Da
Slanje kviza prijateljima	Ne	Da	Da	Ne	Ne
Merenje brzine odgovora	Ne	Ne	Ne	Ne	Da
Prosečan broj pitanja u kvizu po jednoj partiji	20	20	10	34	5
Unos pitanja od strane korisnika	Da	Ne	Ne	Da	Ne
Igranje bez konekcije na internet	Da	Da	Da	Da	Da

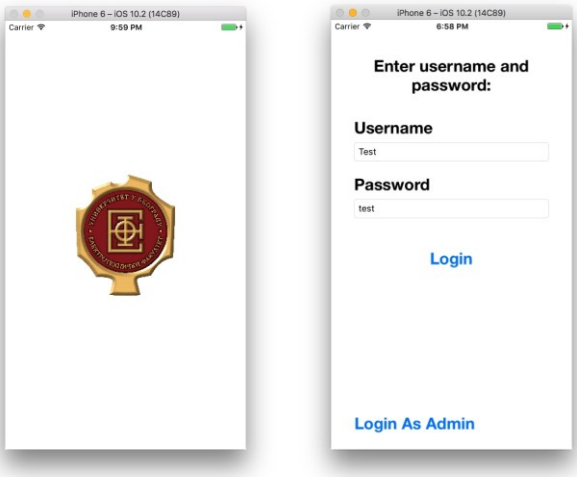
B. Pregled korišćenih tehnologija

Swift je objektno orijentisani programski jezik, koji služi za razvoj Mac i iOS aplikacija. Dizajniran je da bude kompatabilan sa Objective C jezikom. Neke od važnih karakteristika Swift jezika su: brži je od Objective C i Python programskog jezika, poseduje closure, generike i tipsko zaključivanje, ne zahteva stvaranje zasebnih interfejsa i implementacionih fajlova za prilagođene klase i strukture i lako obrađuje velike nizove. Takođe, definicija klase i strukture su u potpisu fajla, a spoljni interfejs te klase i strukture je automatski dostupan. Koristi Objective C runtime, omogućavajući Objective C, C++ i Swift kod u okviru jednog programa, tako da aplikacije koje su već napravljene korišćenjem Objective C jezika mogu da se menjaju pomoću Swift jezika.

Python je programski jezik opšte namene i pripada grupi jezika visokog nivoa. Prednosti Python jezika su: dobra čitljivost koda, sintaksa koja omogućava programerima da izraze koncepte u manje linija koda nego što je moguće u jezicima kao što su C++ ili Java, zatim dinamičko određivanje tipa podataka i automatsko upravljanje memorijom. Ovaj jezik ima veliku i sveobuhvatnu standardnu biblioteku. Takođe, ovaj jezik podržava više programskih paradigmi: objektno orijentisano programiranje, imperativno programiranje, funkcionalno programiranje i proceduralno programiranje. Jedan od nedostataka ovog programskog jezika je u brzini. On je sporiji nego kompajlerski jezici, kao što je na primer C.

III. OPIS REALIZOVANIH FUNKCIONALNOSTI

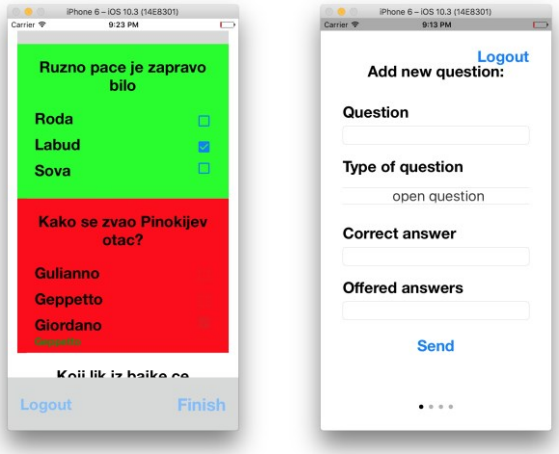
Aplikacija ima intuitivan korisnički interfejs koji je u skladu sa drugim popularnim iOS aplikacijama koje se svakodnevno koriste. Prilikom pokretanja aplikacije pojavljuje se početni ekran i mogućnost da korisnik odabere da li želi da se uloguje na sistem ili da se registruje ukoliko je nov korisnik sistema, kao što je prikazano na Sl. 1. Za obe funkcionalnosti urađena je validacija podataka.



Sl. 1. Uvodni ekran i ekran za logovanje postojećeg korisnika sistema.

Realizovanoj aplikaciji mogu pristupati dva tipa korisnika: standardni korisnici i administratori. Nakon uspešnog logovanja korisnika, on dobija mogućnost da odabere trenutno dostupan kviz i da odgovara na pitanja. Kviz se završava nakon što istekne vreme kviza ili nakon što korisnik završi kviz pre isteka predviđenog vremena. Ukoliko se administrator uloguje, on dobija korisnički meni sa više privilegija od standardnog korisnika. Trenutno podržane

funkcionalnosti administratora u ovoj mobilnoj aplikaciji su: kreiranje kviza, dodavanje pitanje u kviz, kreiranje novog korisnika i pregled rezultata postojećih korisnika. Ako administrator želi da doda novo pitanje, potrebno je da: unese tekst pitanja, odabere tip pitanja, tačan odgovor i u slučaju da je tip pitanja izbora jednog od više ponuđenih odgovora, potrebno je da unese i ponuđene odgovore. Ako ne ispoštuje potrebna pravila za unos pitanja, pitanje neće biti unešeno u bazu. Na Sl. 2 (levo) prikazan je završeni kviz od strane standardnog korisnika, a na Sl. 2 (desno) je prikazan unos pitanja od strane administratora sistema.



Sl. 2. Prikaz završenog kviza i provera odgovora od strane sistema i prikaz unosa pitanja od strane administratora

IV. Implementacioni detalji

Realizovani softverski sistem sastoji se iz aplikacije, baze podataka, API interfejsa (eng. *Application Programming Interface*) i testova. Serverski deo aplikacije pisan je u skladu sa RESTful API i napravljen je tako da može da na zahtev dobijen od aplikacije smesti podatak u bazu. Kroz API je moguće dodati novog korisnika u bazu, proveriti da li su korisnički kredencijali validni, upisati rezultat, odnosno odgovor korisnika u bazu, dodati novo pitanje i druge funkcionalnosti koje su opisane u poglavlju III. Tokom pisanja koda, vodilo se računa da kod bude čitak i da jedna funkcija ima jednu funkcionalnost. Takođe, funkcije su logički povezane u klase. Aplikacija je logički podeljena na deo za administratora i deo za standardnog korisnika.

Kontroleri prikaza su temelj arhitekture mobilne aplikacije. Svaki kontroler prikaza upravlja delom aplikacije korisničkog interfejsa, i obavlja interakciju između tog interfejsa i podataka koji su mu potrebni da ih prikaže. Kontroleri prikaza olakšavaju prelaz između različitih delova korisničkog interfejsa.

UIViewController klasa definiše metode i svojstva koji upravljaju prikazima i događajima, prelaze iz jednog kontrolera prikaza do drugog. U potklasu UIViewController dodaje se prilagođeni kod koji treba da sprovede ponašanje aplikacije. Postoje dva tipa UIViewContollera:

- Prilagođeni kontroleri prikaza - upravljaju diskretnim delovima aplikacije.
- Kontejnerski kontroleri ekrana - prikupljaju informacije iz drugih kontrolera prikaza (engl. *child view contoller*) i predstavljaju ih na način koji olakšava navigaciju ili predstavlja sadržaj tih kontrolera prikaza.

U aplikaciji su korišćeni prilagođeni kontroleri prikaza i kontejnerski kontroler prikaza. *UITableView* kontroler je korišćen za prikaz pitanja u aplikaciji, a *UIPageView* kontroler za navigaciju između stranica administratorskog dela aplikacije. *UITableView* prikazuje listu stavki u jednoj koloni. *UITableView* je podgrupa *UIScrollView*, koja omogućava korisnicima da se vertikalno kreću kroz tabelu. *UITableView* se sastoji od ćelija koje su izvedene iz *UITableViewCell* i koje mogu biti podrazumevane ili prilagođene. Mogu se kombinovati različite ćelije. Sastoje se od sadržaja, zaglavlja i podnožja. Visina može da se određuje dinamički i statički.

UIPageView kontroler omogućava korisnicima da se kreću prevlačenjem između njegovih stranica. *UIPageView* kontroler se sastoji iz niza kontrolera prikaza koji su međusobno nezavisni i svaki upravlja sopstvenim prikazom. Navigaciju može kontrolisati programski aplikacija ili korisnik koristeći pokrete na ekranu. Kada se kreće od stranice do stranice, kontroler koristi prelaz koji je naveden da animira promene.

Za komunikaciju sa serverom se koristi *NSURLSession*. *NSURLSession* klasa i srodne klase pružaju API za preuzimanje sadržaja preko HTTP-a. Ovaj API nudi podršku za autentifikaciju i daje mogućnost da se u pozadini preuzimaju podaci kada aplikacija nije aktivna ili dok se na glavnoj niti radi nešto drugo.

Kada se koristi *NSURLSession* API, aplikacija stvara niz sesija, od kojih svaki koordinira grupu zadataka povezanih sa prenosom podataka. U okviru svake sesije aplikacija dodaje niz zadataka. Svaka sesija predstavlja zahtev ka određenoj URL adresi. *NSURLSession* je asinhroni. Ako se koristi podrazumevano podešavanje sa delegatom, potrebno je, kada se prenos završi uspešno ili sa greškom, obezbediti blok završetka koji vraća podatke u aplikaciju. *NSURLSession* API podržava tri vrste sesija:

- Podrazumevane sesije rade na glavoj niti i blokiraju izvršavanje svega ostalog. Čuvaju podatke i u keš memoriji i na disku.
- Prolazne sesije ne čuvaju nikakve podatke na disku, već se sve čuva u keš memoriji.
- Pozadinske sesije su slične podrazumevanim sesijama, osim što je poseban proces upravljanja svim prenosima podataka.

NSURLSession klasa podržava tri vrste zadataka: za prenos podataka, za skidanje sa servera, i za postavljanje na server. Zadaci za prenos i prijem podataka preko *NSData* objekta namenjeni su za kratke, često interaktivne zahteve iz aplikacije na server. Zadaci za skidanje podataka i zadaci za slanje podataka preuzimaju, odnosno šalju podatke u obliku datoteke i imaju podršku za preuzimanje/slanje u pozadini.

V. TESTIRANJE APLIKACIJE

Testovi su podeljeni na jedinične testove logike apikacije, testove korisničkog interfejsa (skraćeno: UI testovi) i testove performansi. Prva grupa testova obuhvata proveru osnovnih funkcionalnosti na nivou testiranja tehnikom bele kutije. Druga grupa testova pokazuje da li su svi očekivani elementi prisutni i da li se očekivani tekst nalazi unutar tih elemenata. Rađeni su i testovi performansi da bi se videlo koliko je vremena potrebno da se izvrši neka funkcija.

UI testovi testiraju između ostalog da li je neki element na odgovarajućoj poziciji na ekranu i da li je navigacija kroz aplikaciju do određenog prikaza implementirana kako je zamislila osoba koja je zadužena za UX (eng. *User Expiriance*). Ovo su vrlo korisni testovi posebno kada se radi naknadni proces testiranja nakon nekog refaktorisanja koda ili dodavanja neke nove funkcionalnosti u aplikaciji. Jedinični testovi testiraju da li su kontroleri prikaza povezani pravilno i da li server vraća očekivane vrednosti.

Pravila za pisanje testova koja su definisana su: testovi treba da su brzi, da su izolovani koliko god da je to moguće od ostatka sistema, da su postojani, svaki put kada se izvrše da treba da daju isti rezultat i idealno treba da su napisani pre samog pisanja koda, što je i krajnji cilj ovakvog procesa testiranja prilikom razvoja. U tabelama II i III dat je prikaz nekih značajnih jediničnih testova i UI testova.

TABELA II
PRIKAZ REALIZOVANIH JEDINIČNIH TESTOVA

Funkcija	Stanje u sistemu	Šta testiramo
testLoginFuncTrue()	Forma za logovanje, uneti podaci postojećeg korisnika	Logovanje korisnika koji postoji
testLoginFuncFalse()	Forma za logovanje, uneti podaci korisnika koji ne postoji	Logovanje korisnika koji ne postoji
testRegisterFuncTrue()	Popunjena forma za registraciju (kor. ime ne postoji u sistemu)	Dodavanje korisnika koji ne postoji
testRegisterFuncFalse()	Popunjena forma za registraciju (kor. ime postoji u sistemu)	Dodavanje korisnika koji ne postoji
testPostNumQue()	Upisuje se u bazu broj pitanja	Dodavanje broja pitanja na server
testPostScoreTrue()	Popunjeno kor. ime korisnika i rezultat koji se dodeljuje	Upisivanje rezultata korisniku
testGetScoreFunc()	Popunjeno ime korisnika kome znamo da ima neki rezultat	Dohvatanje rezultata korisnika
testReturnQuestionsTrue()	Popunjeno korisničko ime	Dohvatanje broja pitanja
testReturnsData()	Popunjeno pogrešno korisničko ime	Dohvatanje pitanja za korisnika koji ne postoji
testThatViewLoads()	Dohvata se korisnički prikaz	Da li se prikaz dobro instancirao
testThatTableViewLoads()	Dohvata se korisnički prikaz iz memorije	Da li je prikaz tabela dobro učitan
testThatViewConformsToUITableViewDataSource()	Dohvata se DataSource tabele	Da li je DataSource povezan
testThatViewConformsToUITableViewDelegate()	Dohvata se Delegat tabele	Da li je delegat povezan
testNumOfCells()	Dohvataju se pitanja za korisnika sa servera	Da li je broj ćelija jednak broju pitanja

TABELA III
PRIKAZ REALIZOVANIH UI TESTOVA

Funkcionalnost	Stanje u sistemu	Šta testiramo
testLoginOrRegister()	Aplikacija startovana	Pojavljivanje login forme i forme za registraciju
testRegister()	Uneto korisničko ime koje ne postoji u bazi	Da li je moguće registrovati novog korisnika
testLogin()	Uneti su podaci korisnika koji postoji u bazi	Da li je moguće ulogovati korisnika
testLogout()	Ulogovan korisnik ima link da se izloguje iz sistema	Da li je moguće izlogovati se
testFinalScreen()	Na korisničkoj strani klikne se na link za kraj rada	Da li radi navigacija do finalnog prikaza
testNavigationLeftAtLoginAsAdmin()	Korisnik ulogovan kao administrator	Da li je moguće pristupiti svim administratorskim funkcionalnostima
testLoginAsAdmin()	Korisnik pokušava da se uloguje kao administrator	Da li vidi administratorski navigacioni meni
testAddQuestionTitleFalse()	Prazan tekst pitanja	Da li je moguće poslati u bazu pitanje bez teksta
testAddQuestionCorrectAnswerFalse()	Prazno polje za tačan odgovor	Da li je moguće poslati u bazu pitanje bez odgovora
testAddOpenQuestion()	Upisan tekst pitanja, tačan odgovor i izabran tip “open question”	Da li se pitanje sa odgovorom korektno šalje na server i upisuje u bazu
testAddCheckBoxQuestion()	Upisan tekst pitanja, tačan odgovor i tačno tri ponuđena odgovora kod checkbox tipa	Da li se pitanje tipa checkbox uspešno šalje na server i upisuje u bazu
testAddCheckBoxQuestionFalse()	Upisan tekst pitanja, tačan odgovor i manje od tri ponuđena odgovora	Da li je moguće poslati pitanje bez 3 ponuđena odgovora

testAddDateQuestion()	Upisan tekst pitanja, tačan odgovor i izabran "date question"	Da li se pri pravilno unetim podacima pitanje šalje serveru i upisuje u bazu
testAddUser()	Upisano korisničko ime, lozinka i adresa e-pošte korisnika koji ne postoji	Da li je moguće da administrator u bazu doda korisnika koji ne postoji
testAddUserFails()	Upisano korisničko ime, lozinka i adresa e-pošte korisnika koji postoji	Da li je moguće da administrator u bazu doda korisnika koji postoji
testGetScore()	Upisano korisničko ime korisnika koji postoji	Pokušava se dohvatiti rezultat korisnika koji postoji
testGetScoreFails()	Upisano korisničko ime korisnika koji ne postoji	Pokušava se dohvatiti rezultat korisnika koji ne postoji
testAddNumQuestions()	Upisuje se broj pitanja po kvizu	Testira se da li administrator može da unese broj pitanja

Fajl u kome pišemo UI test treba da učita *XCTest* okvir. Klasa koja vrši testiranje izvedena je iz klase *XCTestCase* i ona instancira objekat klase *XCUIApplication*, koji treba da pristupi elementima aplikacije. Pristupanje se vrši tako što objekat identifikuje aplikaciju navedenu kao atribut *target*. Klasa preklapa dve funkcije: *setUp()* i *tearDown()*. Funkcija *setUp()* se izvršava uvek na početku testa i u njenom telu se prvo poziva *setUp()* funkcija natklase. Ukoliko napišemo promenljivu *continueAfterFailure* i postavimo je na vrednost *false*, to znači da se testiranje ne nastavlja, ako ovaj test padne. Nakon ove funkcije, vrši se pokretanje aplikacije i izvršavanje testa. Kada test krene svoje izvršavanje, od prvog elementa nastaje stablo, koje se dalje gradi, počevši od korenog čvora – *Application* objekta, pa dalje ka listovima. Na kraju svakog testa pokreće se funkcija *tearDown()* koja gasi aplikaciju. U listingu 1 prikazan je inicijalni sadržaj testa.

```
class quizAppUITests : XCTestCase {
  let app = XCUIApplication()
  override func setUp() {
    super.setUp()
    continueAfterFailure = false
    XCUIApplication().launch()
  }
  override func tearDown() {
    super.tearDown()
  }
```

Listing 1. Inicijalni sadržaj UI testa.

Za razliku od *JUnit*-a i sličnih okruženja za pisanje jediničnih testova, u *XCUnit*-u svaka funkcija koja je test funkcija mora započeti rezervisanom reči *test*, a u okviru nje se mogu pozivati ostale funkcije. Dugmadima se pristupa prema dodeljenom imenu. Kada se dohvate i smeste u lokalne promenljive može da se proveri da li je njihov atribut *title* korektan. Takođe, može se ispitati postojanje određenog elementa korišćenjem *exists*.

Navigacija se testira tako što se odabere dugme za logovanje ili registrovanje novog korisnika pozivom funkcije *tap()* nad jednim od ta dva dugmeta. Nakon toga treba verifikovati postojanje prikaza ekrana za logovanje ili registraciju. Instanca *let window = app.children (matching: .window).element(boundBy: 0)* vraća prvi iz niza prozora u *iOS* sistemu. Prvi je uvek onaj koji je trenutno aktivan. Konstanta *let element = window. children(matching: .other).element. children(matching: .other).element* se inicijalizuje tako što se pozivaju deca *window* promenljive koja su tipa *.other* i dohvataju se njihovi elementi. Zatim se ponovi radnja i dohvate se svi elementi koji se trenutno prikazuju. Oni se instanciraju preko indeksa, na primer: *XCTAssertTrue (element.children (matching: .textField). element(boundBy: 0).exists).* Na ovaj način je verifikovano da postoji *textField* na ekranu za registraciju. U listingu 2 prikazan je primer testa prvog ekrana u aplikaciji kviz.

```
func loginOrRegistersFunc(){
  let registerButton = app.buttons["Register"]
  let loginButton = app.buttons["Login"]
  XCTAssertTrue(app.staticTexts["Login or
  Register"].exists)
  XCTAssertTrue(registerButton.exists)
  XCTAssertTrue(loginButton.exists)
  XCTAssertEqual("Login", loginButton.title)
  XCTAssertEqual("Register",
                 registerButton.title)
  XCTAssertNotEqual("Register",
                    loginButton.title)
}
func testLoginOrRegister() {
  self.loginOrRegistersFunc()
}
```

Listing 2. Primer testa za dugmad na prvom korisničkom ekranu.

U ovoj mobilnoj aplikaciji očekujemo pitanja iz baze koja se nalazi na serveru, pa iz tog razloga moramo da se povežemo na server i prihvatimo podatke sa servera na klijentskoj strani. Kada se dobiju podaci, proverava se da li su različiti od *nil* pomoću metode *waitForExpectations*. Testirano je da li se dobijaju validni podaci sa servera, odnosno da li pri logovanju za korisnika čiji su kredencijali u redu, dobijen očekivan rezultat, a realizovani su i testovi za logovanje korisnika sa pogrešnom lozinkom i korisnika čije korisničko ime ne postoji. Slično je urađeno i kod registracije novog korisnika.

Realizovan je i jedan broj testova performansi servera i mreže. Ovo testiranje se koristi da se izmeri koliko je vremena potrebno da se izvrši neka funkcija u kodu. U listingu 3 dat je primer testova performansi. Prikazano je kako se testira vreme koje je potrebno da se dohvate podaci o korisniku i njegov rezultat iz baze i koliko je potrebno da se dohvati jedno pitanje iz baze. Izlazni kod ovih testova dat je u listingu 4. Realizovana je i funkcija za dohvatanje svih pitanja iz baze. Test za dohvatanje jednog pitanja prolazi za 0.256 sekundi, a test za dohvatanje svih 80 pitanja za 0.277 sekundi.

```
func testPerformanceScore() {
self.measure{
   ProxyInterface.sharedInstance.scoreFunc
   ("drasko", password: "sifra123",
     completion:{(retVal) -> Void in
})}}

func testPerformanceGetQuestionsOne() {
  UserDefaultsHelper().writeUserDefaults(string: "70")
   self.measure {
      ProxyInterface.sharedInstance.
      getQuestions(completion:{(retVal) -> Void in
          let _ = Questions().parsingQue(getQuestions:
                                         retVal).count
})}}
```

Listing 3. Testovi performansi.

```
Test Case '-[quazzAppTests.TestNetworkingClass
testPerformanceGetQuestionsAll]' passed (0.277
seconds).

Test Case '-[quazzAppTests.TestNetworkingClass
testPerformanceGetQuestionsOne]' passed (0.256
seconds).

Test Case '-[quazzAppTests.TestNetworkingClass
testPerformanceScore]' passed (0.260 seconds).
```

Listing 4. Izlazni kod testiranja performansi.

Pokrivenost koda realizovanim jediničnim testovima je 68.85 %, a pokrivenost koda UI testovima je 75.02 %. Ovo se smatra srednjom ka visokoj pokrivenosti koda [8-10].

VI. Zaključak

U ovom radu prikazan je razvoj jedne mobilne aplikacije i način pisanja testova tokom razvoja. Aplikacija je razvijana sa trojslojnom arhitekturom i ima klijentski deo za korisnike, serverski deo sa logikom aplikacije i sloj podataka.

Veliki broj naučnih radova i istraživanja ističe benefit TDD pristupa. Takođe, mnoge organizacije su radile istraživanja u kojima se pokazuje merljivost softverskih bagova kada se koristi TDD u razvoju softveru i kada se ne koristi taj pristup. TDD se može koristiti kako u tradicionalnim metodologijama razvoja softvera, poput vodopada i V-modela, tako i u agilnim metodologijama [11].

U ovom istraživanju pokazano je da i kod jednostavne mobilne aplikacije sa svega nekoliko funkcionalnosti, pristup pisanja jediničnih testova i testova korisničkog interfejsa može da ubrza implementaciju i stavljanje aplikacije u produkciju, sa dosta visokim stepenom pokrivenosti koda.

Zahvalnica

Ovaj rad je delimično bio finansiran od strane Ministarstva prosvete, nauke i tehnološkog razvoja Republike Srbije i projekta TR 32047.

Literatura

[1] https://www.statista.com/statistics/276623/number-of-apps-available-in-leading-app-stores/, posećen dana 6. aprila 2019.

[2] https://techcrunch.com, posećen dana 6. aprila 2019.

[3] QuizUp, dostupno na: https://www.quizup.com/en, posećen dana 5. oktobra 2017.

[4] DK Quiz, dostupno na: https://itunes.apple.com/us/app/dk-quiz/id556884950, posećen dana 7. oktobra 2017.

[5] Quiz Your Friends, dostupno na: https://itunes.apple.com/us/app/quiz-your-friends-see-who/id919383759, posećen dana 8. oktobra 2017.

[6] Logos Quiz, dostupno na: https://itunes.apple.com/us/app/logos-quiz-guess-the-logos!/id478364212, posećen dana 11. oktobra 2017.

[7] Trivia Crack, dostupno na: https://www.triviacrack.com/, posećen dana 20. oktbora 2017.

[8] Andrei Chirila, "Test Coverage at Google", dostupno na: https://docs.google.com/presentation/d/1god5fDDd1aP6PwhPodOnAZSPpD80lqYDrHhuhyD7Tvg/edit#slide=id.g3f5c82004_8636

[9] Brian Marick, "How to Misuse Code Coverage," dostupno na: http://www.exampler.com/testing-com/writings/coverage.pdf

[10] Ayewah, Pugh, Hovemeyer, Morgenthaler, Penix, "Using Static Analysis to Find Bugs," IEEE Software, vol. 25, no. 5, 2008.

[11] P. Kayongo et al., "Why Do Software Developers Practice Test-Driven Development," International Conference on Advances in Computing and Communication Engineering, Durban, South Africa, 2016.

Abstract

This paper describes the implementation of the mobile Quiz application for the iOS platform with the emphasis on software testing process. The application is designed using Model View Controller (MVC) pattern. During the development of the application, the Test Driven Development (TDD) methodology was applied and the XCTest environment was used. Development of tests during implementation phase is a very important process for software engineers, and it raises the quality of the software and reduces the deficiencies in the software. Unit tests for testing the business logic of the mobile application and unit tests for testing the user interface were realized. The coverage of the code by realized unit tests is 68.85%, and the coverage of code by realized user interface tests is 75.02%.

Test Driven Development of Mobile Application using XCTest framework

Drazen Draskovic

Automatsko generisanje testova za automotive sisteme zasnovane na AUTOSAR modelu

Aleksandar Lukić, Dragan Kukolj, Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Velibor Ilić, Istraživačko-razvojni Institut RT-RK, Novi Sad, Srbija, Milena Milošević, Fakultet tehničkih nauka, Univerzitet u Novom Sadu

***Apstrakt*– U radu je opisan postupak generisanja testova pomoću kojih se proverava da li se *RTE* (*runtime environment*) sprege izvršavaju kako je predviđeno na ugrađenoj (eng. *embedded*) *automotive* platformi za konkretne namenske kontrolne jedinice (eng. *ECU - Electronic Control Unit*). Predstavljen je jedan od načina za automatsko generisanje testova pomoću kojih se proverava rad softverskih komponenti koje pripadaju *AUTOSAR* modelu. Opisani su procesi od izvlačenja informacija potrebnih za pravljenje testova do samog izvršavanja testa. Za potrebe testiranja su implementirana dva moda, *FreedomFromInterference* i *MiddlewareConnections*. Za proveru funkcionalosti korištena je *Autonomus driving* platforma sa dva domaćina (eng. *host*) i VN89000 uređaj na kome je realizovan *framework* za komunikaciju između namenskog sistema i testnog računara uz pomoć *ethernet* veze.**

***Ključne reči–Automotive*; *AUTOSAR*; Communication; Testing; *RTE*;**

I. Uvod

Automotive inženjering je deo inženjeringa za vozila, koji osim razvoja vozila za drumski saobraćaj ima i primenu u vazdušno-kosmičkim i vodenim vozilima. Ova vrsta inženjeringa se bavi mehaničkim, električnim, elektronskim, programskim i bezbednosnim elementima na osnovu kojih se dizajniraju i projektuju motocikli, automobili, kamioni i slična vozila. Kreiranje, razvoj i proizvodnja su glavne funkcije ove grane nauke.

Tokom istorije prevozna sredstva su se menjala kao i način njihove proizvodnje. Razlika od prvog prevoznog vozila do danas je ogromna i promene su išle u skladu sa razvojem tehnologije i potrebama ljudi. Menjana je veličina automobila, izgled, pogonsko sredstvo, snaga, elektronika, razni dodaci i pomagala radi efikasnije kontrole i upotrebe vozila. Sve to je dovelo do toga da su vozila danas znatno kompleksnija, a samim tim i njihova izrada.

Aleksandar Lukić – Fakultet Tehničkih Nauka, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail: aleksandar.lukic@rt-rk.com)

Dragan Kukolj - Fakultet Tehničkih Nauka, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail: dragan.kukolj@rt-rk.uns.ac.rs)

Velibor Ilić - RT-RK.doo, Narodnog Fronta 23a, 21000 Novi Sad, Srbija (e-mail: velibor.ilic@rt-rk.com)

Milena Milošević - Fakultet Tehničkih Nauka, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail: milena.milosevic@rt-rk.uns.ac.rs)

Tako da danas automobili sadrže različite aplikacije kojima je potrebno kontrolisati tok izvršavanja i životni ciklus. Upravljanje njima je komplikovan zadatak i u autoindustriji nema mnogo opcija za ovaj izazov.

Trenutno najbolje i rešenje koje se najčešće koristi proizvela je organizacija *AUTOSAR* (*Automotive Open System Architecture*). Ova organizacija, koju čine partnerska grupa proizvođača automobila i automobilske opreme, se bavi standardizacijom platforme za razvoj softvera za *automotive* industriju.

U samom automobilu se nalaze mnogobrojne elektronske kontrolne jedinice na kojima se izvršavaju softverske komponete koje se ugrađuju da bi se smanjio prostor za grešku u samom radu automobila. Zbog toga je potrebno testirati i proveriti funkcionalnost ovog složenog sistema različitim testovima kako bi se te greške mogle na vreme detektovati i ispraviti.

U ovom radu prikazano je jedno rešenje koje automatsko generiše testove za *AUTOSAR* model kojim se proverava funkcionalnost sistema. Pod automatskim testiranjem podrazumevamo da se sam rad komponente simulira i izvršava automatski, a ne da se deo koji se testira kontroliše ručno, korak po korak i samim tim automatsko testiranje je mnogo brže od ručnog testiranja. Opisane su osnove *AUTOSAR* standarda, izvršno okruženje i testiranje sprega. Dat je pregled operacija kroz koje je potrebno proći da bi generisali i izvršili testove, i prikazana je realizacija tih operacija (raščlanjivanje modela, generisanje potrebnih informacija, izvršavanje), i na kraju se nalazi zaključak.

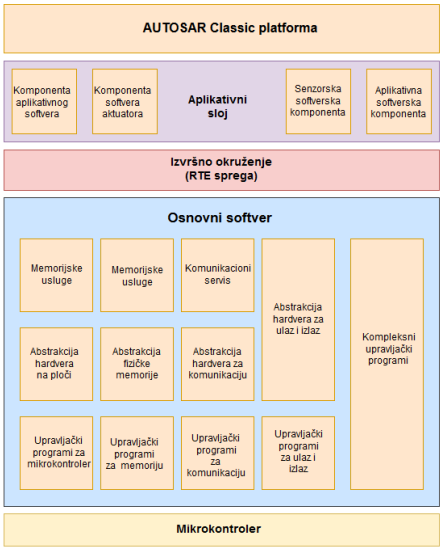
II. Teorijske osnove

U ovom poglavlju je definisano šta znače *AUTOSAR* platforma, izvršno okruženje (eng. *RTE – runtime environment*) i testiranje sprega *(*eng. *Interface testing*). Pomoću ovih osnova razumevanje rada će biti lakše.

A. AUTOSAR platforme

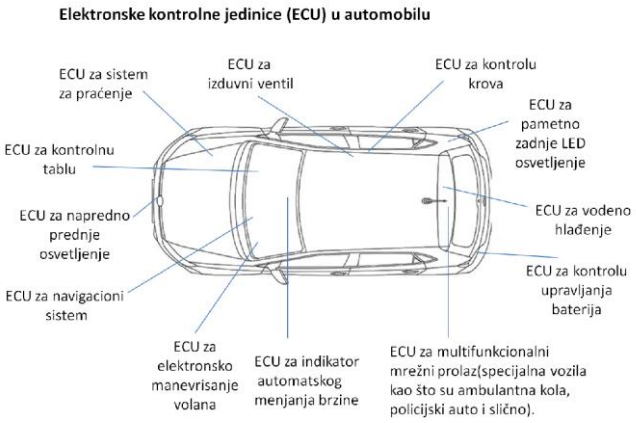
AUTOSAR je grupa osnovana 2003. godine koju danas čine proizvođači automobila, automobilske opreme, proizvođači alata i proizvođači poluprovodnika. *AUTOSAR* grupa je razvila standard koji je jedan od vodećih u automobilskoj industriji za razvoj softvera[1]. Postoje dve softverske platforme *Classic* i

Adaptive. *Classic* je u široj upotrebi i koristi se uglavnom za namenske *ECU* jedinice, dok se *Adaptive* platforma bazira za razvoj aplikacija kojima je potrebnija veća snaga i moć računarske obrade. *AUTOSAR Classic* omogućuje komunikaciju unutar elektronskih kontrolnih jedinica kao i između više njih.



Sl. 1. *AUTOSAR Classic Platform* arhitektura

Arhitektura prikazana na Sl. 1 raspoznaje tri softverska sloja koja se pokreću na mikrokontroleru, a ta tri sloja su aplikativni, *RTE* sprega i osnovni softver. Jedna od osnovnih i najbitnijih delova ove platforme je *VFB* (eng. *Virtual Functional Bus*) [2]. Ova virtuelna magistrala je skup *RTE* sprega koje još nisu raspoređene konkretnoj elektronskoj kontrolnoj jedinici[3]. *VFB* upravlja komunikacijom *ECU* jedinica sa drugim *ECU* jedinicama[4]. Na Sl. 2 se mogu videti neki primeri *ECU* jedinica u automobilu. Ova komunikacija se odvija preko namenskih priključaka (eng. *port*), tako da komunikacione sprege moraju biti povezane sa tim priključcima.

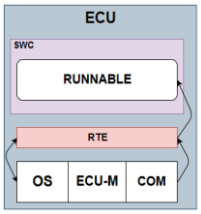


Sl. 2. Primer ECU jedinica u automobilu

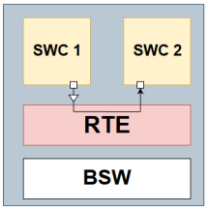
B. *Izvršno okruženje (RTE – runtime environment)*

U *AUTOSAR* arhitekturi *RTE* sprega realizuje komunikaciju između softverskih komponenti (eng. *SWC – Software Component*) i osnovnog softvera (eng. *BSW – Basic Software Component*).

Softverske komponente komuniciraju sa drugim komponentama i/ili modulima osnovnog softvera isključivo preko *RTE* sprega, što omogućava da softverske komponente budu međusobno nezavisne, ali i nezavisne od pojedinačnih *ECU* jedinica. *RTE* sprega predstavlja jednu *ECU* jedinicu i generiše se posebno za svaku konfiguraciju *ECU* jedinice. Svaka particija ili jezgro *ECU* jedinice generiše jednu *RTE* spregu [5]. *RTE* predstavlja apstrakciju operativnog sistema, servisa za komunikaciju, hardverske sprege i planiranje rada softverskih komponenti kao što se može videti na Sl. 3 i 4.



Sl. 3. *RTE* za *runnable* komponentu



Sl. 4. *RTE* kao komunikaciona sprega

Postoje dve vrste komunikacije koje *RTE* realizuje. To su Klijent/Server komunikacija i Pošiljaoc/Prijemnik (eng. *producer/receiver*) komunikacija, a ona se deli na:

- direktno (*Rte_Read*, *Rte_Write*).
- indirektno (*Rte_IRead*, *Rte_IWrite*, *Rte_IwriteRef*).
- redno (*Rte_Receive*, *Rte_Send*).

Pošiljaoc/Prijemnik komunikacija poseduje dva moda: eksplicitni (koristi eskplicitne *RTE API* (eng. *application programming interface*) pozive kako bi primio i poslao delove podataka) i implicitni (*RTE* automatski čita određene setove podataka pre nego što se *runnable* komponenta pozove). Klijent/Server komunikacija takođe poseduje dva moda: sinhroni (rade u istom kontekstu) i asinhroni (rade u različitim kontekstima).

C. *Testiranje sprega (Interface testing)*

Testiranje sprega služi kako bi se izvršila provera da li u sistemu koji proveravamo, njene komponente vrše pravilan prenos podataka između softverskih komponenti, ali i proveru ispravnosti tih podataka kada dođe do izmena nad njima. Ovo testiranje proverava da li interakcija odnosno komunikacija između njih funkcioniše kako treba[6,7]. To nije jedini način da proverimo funkcionalnost softverskih komponenti, postoje i drugi tipovi testova koji proveravaju različite funkcionalnosti. U ovom radu ovo testiranje se izvršava u dva moda koja su delimično modifikovana u ovom rešenju:

- *FreedomFromInterference* (*FFI*) – proveravamo da li su sve veze koje su opisane kroz model prenesene i u sam sistem.
- *MiddlewareConnections* (*MDWConn*) – proveravamo da li se pri svakom pozivu unutar jedne instance statičke planer (eng. *scheduler*) tabele koja se ponavlja, prenos podataka za datu spregu (eng. *interface*) uspešno obavlja.

III. Koncept rešenja

U datom rešenju korišten je jedan sistem za autonomnu vožnju (eng. *Autonomus driving system*) koji je opisan *AUTOSAR* modelom koji je opisan u *.rxml* formatu. Iz sistema za autonomnu vožnju koji testiramo izvučene su potrebne informacije za testiranje željenih softverskih komponenti koje su takođe definisane u modelu. Te komponente smo popunili sa testnim kodom koji predstavlja jedan *framework* koji omogućava kontrolisano izvršavanje testova (*FFI* i *MDWConn*) za svaku *RTE* spregu u sistemu. Kontrolu izvršavanja testnog koda vršimo kroz *framework* na testnom računaru. *Framework* je implementiran kroz *CAPL* (eng. *Communication Access Programming Language*) kod koji se izvršava u kontekstu *Vector CANoe* alata. Ono se oslanja na mrežni interakcioni sloj (eng. *Ethernet Interaction Layer*) koji pruža *CANoe* alat, za slanje i primanje kontrolnih poruka sa namenskih (eng. *embedded*) sistema kao i modul za izveštaje koji nam prikazuje rezultate testova. Glavni zadatak ovog *framework* koda je da kreiranjem odgovarajućih komandnih poruka i tumačenjem odgovara sa testnog sistema kontroliše izvršavanje testova i prikazivanje rezultata testova kroz izveštaj, koji je u *html* formatu[8]. Ovo je sve uspešno odrađeno kroz tri glavne operacije:

- raščlanivanje modela,
- generisanje potrebnih informacija,
- izvršavanje.

IV. Opis realizacije

Aplikacija je izrađena u *Microsoft Visual Studio Community*. Cilj ove aplikacije je da se raščlanjivanjem *AUTOSAR* modela dobiju neophodne informacije kako bi se generisali testovi i neophodni ulazi za *framework* kako bi se testovi izvršili. Samo raščlanjivanje *AUTOSAR* modela odrađeno je pomoću *TinyXML2 XML* [9] raščlanjivača (eng. *parser*) za C++ programski jezik. Ovaj raščlanjivač je izabran zbog toga što je jednostavan za upotrebu i može da obradi veće količine podataka.

A. *Raščlanjivanje modela*

Prvi korak je raščlanivanje *AUTOSAR* modela odnosno *.rxml* datoteke koja sadrži njegov opis. Jedna datoteka predstavlja jedan računar (eng. *host*). Nakon raščlanjivanja dobija se informacija o svim *SWC* jedinicama za dati *host*, o *runnable* komponentama koje su vezane za date *SWC* kao i *RTE* sprege dostupne datoj *runnable* komponenti. Na osnovu ovih informacija pravi se jedna baza podataka koja sadrži veze ovih komponenti koje su potrebne kasnije prilikom generisanja testova odnosno proizvođač/potrošač (eng. *producer*/*consumer*) veza. Na Sl. 5 ovaj korak je predstavljen sa prva dva bloka i dobijene informacije su označene većim sivim blokom.

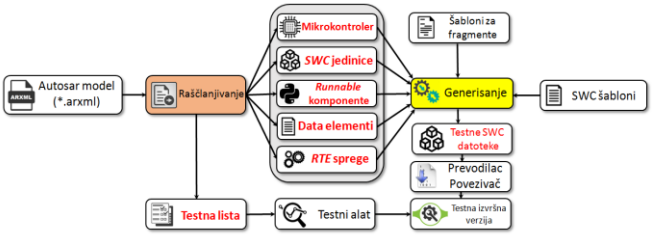
B. *Generisanje potrebnih informacija*

Ulazni podaci za generisanje su baza podataka dobijena parsiranjem, šabloni koji su prethodno generisani od strane *Vector* alata i fragmenti koda koje smo mi kreirali. Ove informacije prikazane su na Sl. 5 kao blokovi koji pokazuju na žuti blok koji obeležava generisanje podataka.

Fragmenti sadrže kod koji će vršiti testiranje *RTE* apstrakcije na sistemu i mogu se generisati sledeće sekcije u *SWC* jedinici:

- ***INCLUDE*** – sekcija u kojoj se definišu promenljive neophodne za kontrolu izvršavanja testova i prenos podataka preko date *RTE* sprege i uključivanje potrebnih zaglavlja.
- ***INIT*** - sekcija koja će biti generisana u inicijalnom delu *runnable* komponente date *SWC* jedinice.
- ***MAIN*** – deo koji će se generisati u cikličnoj *runnable* komponenti date *SWC* jedinice i koji sadrži glavni deo testa za svaku *RTE* spregu.
- ***FUNCTION*** – sekcija za dodatne funkcije koju je neophodno generisati u okviru pojedinačne *SWC* jedinice.

Svaki fragment može da ima neke ili sve od pomenutih sekcija. Napisani su različiti fragmenti u zavisnosti od tipa *RTE* sprege koju testiraju, opisani u sekciji B poglavlja II. Šablone koji su prethodno automatski izgenerisani na osnovu našeg rešenja, popunjavamo ovim sekcijama na odgovarajuća mesta koja su definisana u samom šablonu. Testni kod kojim smo popunili ove šablone nam omogućava da kontrolišemo koji mod je izabran (*FFI* ili *MDWConn*), izvršavanje testa ukoliko je to zahtevano, kontrolu nad željenom *RTE* spregom, rezultate testa i slanje rezultata testa na testni računar.



Sl. 5. Testno okruženje

Sem testnih *SWC* jedinica generišemo i testnu listu (eng. *Test Case List*). Testna lista služi kao početna tačka *framework* kodu

za izvršavanje testova i proveru dobijenih rezultata. Kontrola testova se vrši kroz specijalne linije u testnoj listi koje konfigurišu sam test. One sadrže sve *SWC* jedinice i njima pripadajuće *runnable* komponente. *Framework* raščlanjuje ove informacije i na osnovu njih pravi testne komande koje, kada se protumače na namenskom sistemu, postavljaju početne uslove za izvršavanje testova u datom modu. Glavni deo testne liste su testovi za pojedinačne *RTE* sprege.

One su generisane kao grupa koju čine - proizvođač (eng. *producer*) sa svim svojim relevantnim potrošačima (eng. *consumer*), napravljene na osnovu istog priključka i strukture podata koji idu kroz *RTE* spregu. Bitno je naglasiti da *RTE* sprege koje predstavljaju proizvođače i *RTE* sprege koje predstavljaju potrošače nisu istog tipa (proizvođači: *write, iwrite, iwriteref, client, send*; potrošači: *read, iread, receive, server*), ali imaju zajednički priključak za komunikaciju i strukturu podataka (eng. *data element*) koji se prenose kroz spregu.

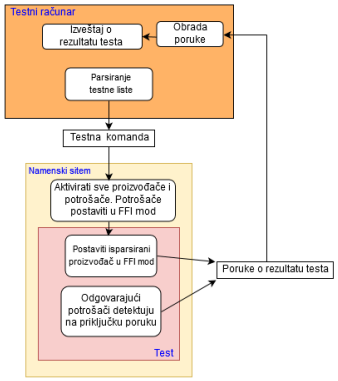
C. *Izvršavanje*

Nakon što su generisane testne *SWC* datoteke i testna lista, spremno je testno okruženje. Da bi se ovo okruženje koristilo za testiranje potrebno je da se izgenerisani kod prevede, uveže (eng. *linking*) i spusti na namenski sistem. Sl. 5 pokazuje tok ovih operacija koji počinje od blokova *Parsing* i *Generate*.

Nakon što se prebaci kod na namenski sistem i ostvari fizička gigabit *ethernet* veza između testnog računara i namenskog sistema, ostaje da se podesi mod izvršavanja testova u konfiguracionoj datoteci i da se pokrene *CANoe* alat sa unapred kreiranom konfiguracijom koja sadrži naš *framework* i da se počne sa izvršavanjem testova[10]. Implementirani *framework* će da izvrši raščlanjivanje testne liste, napraviće komandne poruke i poslati ih na namenski sistem. Nakon što se na namenskom sistemu izvrše testovi aktivirani komandnim porukama, vratiće nazad odgovor i *framework* će da protumači te odgovore i sastaviti izveštaj o uspešnosti izvršavanja testa. Same informacije koje se nalaze u tom izveštaju zavisiće od moda u kojem su se testovi izvršavali.

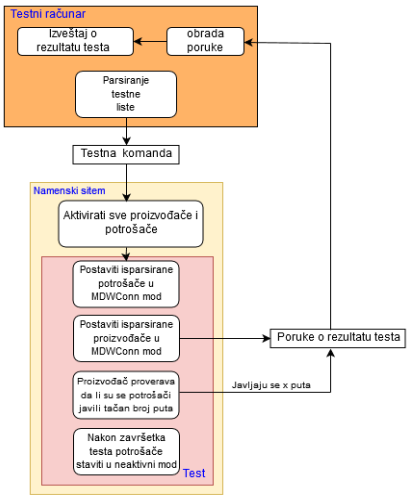
U *FreedomFromInterference* modu želimo da proverimo da su napravljene veze dobro prenesene iz modela koji opisuje realan sistem. U konfiguracionom delu (*startMeasure*) sve *RTE* sprege koje predstavljaju potrošače aktiviramo. To znači da će se u svakoj iteraciji izvršavanja bilo koje *runnable* komponente sve sprege koje su potrošači aktivirati i proveriti ispravnost podataka koji su oni pročitali. Zatim se pojedinačno aktiviraju proizvođači koji su upisani u testnoj listi. Pod aktiviranjem podrazumevamo da se podaci prvo modifikuju korišćenjem algoritma koji je poznat i potrošačima i proizvođačima, a zatim upišu u datu *RTE* spregu. Nakon ovoga *framework* čeka odgovore od potrošačke strane. *Framework* zna uz pomoć testne liste koji potrošač treba da se javi. Sam testni kod na potrošačkoj strani je napisan tako da će se ukoliko se nešto pročita na datoj *RTE* sprezi proveriti da li su pročitani podaci u skladu sa algoritmom promene podataka definisanim za naše testove i zatim se šalje poruka o ispravnosti podataka. *Framework* treba da primi odgovarajući broj poruka, po jednu od svakog definisanog potrošača i samo od njih, sa statusom da su pročitani očekivani podaci.

Ukoliko se to i desi smatra se da je test za datu *RTE* spregu prošao. Na Sl. 6 je prikazan pojednostavljen tok samog testa u ovom modu.



Sl. 6. Tok testa u *FFI* modu

U *MDWConn* modu proveravamo da li se za svaki poziv unutar jedne instance statičke planer tabele, koja se ponavlja, podaci uspešno prenose za datu spregu, kao što je prikazano na Sl. 7. Cilj nam je da proverimo da planer *SWC* jedinice koju testiramo bude ispoštovan, odnosno da su periodi pozivanja proizvođačke *runnable* komponente, potrošačke *runnable* komponente i *middleware runnable* komponente tačni i da će svaki poziv date *RTE* sprege u okviru jednog perioda (eng. *hyperperiod*) biti odrađen kako treba. Informacije o planeru smo prethodno izvukli iz raščlanjenog modela u kome je za svaku *runnable* komponentu definisan period pozivanja.



Sl. 7. Tok testa u *MDWConn* modu

Razlika u odnosu na *FFI* na samom početku je u tome što ne aktiviramo potrošače u konfiguracionom delu već kad započnemo testiranje. Svi potrošači koji se nalaze u testnoj listi se aktiviraju jedan po jedan, a aktiviranje ima isto značenje kao i u *FFI* modu. Nakon toga se aktiviraju proizvođači. Njihovo aktiviranje u *MWDConn* modu podrazumeva da će pozvati datu *RTE* spregu onoliko puta koliko će se data *runnable* komponenta pozvati u datom periodu (eng. *hyperperiod*). Da bi test bio uspešan potrebno je da potrošači odgovore onoliko puta koliko je proizvođač puta bio aktiviran i da sa sobom nosi status da su dobri podaci pročitani. Na kraju jednog prolaza se deaktiviraju potrošači koji su bili testirani u tom prolazu.

Kada prođu svi slučajevi definisani u testnoj listi dobijamo izveštaj i komunikacija se završava.

V. Zaključak

U okviru rada je prikazano jedno rešenje koje pokazuje kako je moguće automatski generisati testove koji proveravaju rad različitih softverskih komponenti koje su deo *AUTOSAR* modela. Odrađen je ceo proces od izvlačenja samih informacija potrebnih za pravljenje testa do samog izvršavanja testa. Kreirani testovi se karakterišu kao integracioni sprežni testovi i napisani su na osnovu *AUTOSAR* specifikacije. Testovi obuhvataju sve moguće slučajeve, jer uzimamo sve sprege iz modela prilikom raščlanjivanja modela. Za potrebene našeg testiranja mi smo implementirali dva moda, *FreedomFromInterference* i *MiddlewareConnections*. Za proveru funkcionalosti ovog rešenja korištena je *Autonomus driving platforma* sa dva domaćina i VN89000 uređaj [11] koji je realizovao *framework* komunikacije između namenskog sistema i testnog računara uz pomoć *ethernet* veze. Rešenje je moguće proširiti tako da je testiranje opširnije, da se proverava više uslova, da se izvršavaju detaljnije provere i analize u samom testu prilikom njegovog izvršavanja ili uz konkretnije modifikacije čak i raščlaniti drugi model.

ZAHVALNICA

Ovaj rad je delimicno finansiran od strane Ministarstva za nauku i tehnologiju Republike Srbije, na projektu tehnoloskog razvoja broj: III44009-6.

LITERATURA

[1] Fürst S., Mossinger J., Bunzel S., Weber T., Kirschke-Biller F., Heitkamper P., Kinkelin G., Nishikawa K., Lange K. (2009) "AUTOSAR - A Worldwide Standard is on the Road", Internationaler Kongress: 14, Elektronik im Kraftfahrzeug 2009, Baden-Baden, Germany

[2] AUTOSAR Classic platform , pristupljeno novembar 2018. https://en.wikipedia.org/wiki/AUTOSAR#Classic_Platform

[3] Hermans T., Denil J., Anthonis J., De Meulenaere P., Ramaekers P., (2011) " Incorporation of AUTOSAR in an Embedded Systems Development Process: a Case Study", EUROMICRO 2011, Oulu, Finland

[4] Schreiner D., Goshka K., (2007) " A Component Model for AUTOSAR Virtual Function Bus", COMPSAC 2007, Beijing, China

[5] Chul Jo H., Piao S., Rae Cho S., Young Jung W., (2008) "RTE Template Structure for AUTOSAR based Embedded Software Platform", IEEE/ASME 2008, Beijing , China

[6] Ćorluka G., Ilić V., Marijan M., Spasojević D., (2018) "Testiranje komunikacije između softverskih komponenti na elektronskim upravljačkim jedinicama", ETRAN 2018, Palic, Serbia

[7] Vujanović V., Popić S., Ilić V., Pap M., (2016) "Analiza rezultata testova prilikom razvoja kompleksnih elektronskih upravljačkih jedinica", ETRAN 2016, Zlatibor, Serbia

[8] Kovačić M., Ilić V., Popić S., (2016) "Data flow in automated testing of complex automotive electronic units", ZINC 2016 , Novi Sad, Serbia

[9] TinyXML, pristupljeno novembar 2018. http://www.grinninglizard.com/tinyxml2

[10] Moon H., Kim G., Kim Y., Shin S., Kim K., Im S. (2009) "Automation Test Method for Automotive Embedded Software Based on AUTOSAR", Fourth International Conference on Software Engineering Advances 2009, Porto, Portugal

[11] VN8900 - Modular FlexRay/CAN FD/LIN/J1708/K-Line Network Interface with up to 8 Channels, pristupljeno novembar 2018. https://www.vector.com/int/en/products/products-a-z/hardware/network-interfaces/vn89xx/

ABSTRACT

In this paper it is described procedure of generating tests wich are checking if *RTE* connections are working as intended on embedded automotive platform for specific *ECU* components. It is presented one way for automatic generation of tests which are testing the functionality of software components that are part of *AUTOSAR* model. Process of retaining information needed for creating those tests and their use are also described in this paper. For testing purposes two modes were implemented, FreedomFromInterference and MiddlewareConnections. To check functionality of tests Autonomus driving platform with two hosts and VN89000 device, whom provided framework for communication between embedded system and test PC via ethernet, were used.

Automatic generation of tests for automotive systems based on AUTOSAR model

Aleksandar Lukić, Dragan Kukolj, Velibor Ilić, Milena Milošević

Unapređenje programskog prevodioca Clang sa podrškom za standard MISRA/AUTOSAR

Đorđe Milićević, Mirko Brkušanin, Milena Vujošević Janičić, Teodora Novković, Petar Jovanović

Poštovanje standarda kodiranja je posebno važno u automobilskoj industriji jer greške u softveru automobila mogu imati fatalne posledice. U radu je predstavljeno unapređenje programskog prevodioca Kleng (eng. *Clang*) dodavanjem mogućnosti provere ispunjenosti 164 pravila iz standarda MISRA/AUTOSAR. Predstavljeni su opis i struktura standarda AUTOSAR, kompajlerske infrastrukture LLVM i programskog prevodioca Kleng, kao i detalji realizacije i način integrisanja skupa podržanih pravila u prednji deo infrastrukture LLVM. U okviru realizacije, napravljene su različite vrste proširenja leksičkog, sintaksičkog i semantičkog analizatora. Evaluacijom je utvrđeno da realizovane dodatne analize usporavaju proces prevođenja za svega 19,15%, što čini ovu realizaciju izuzetno efikasnom.

***Ključne reči*: verifikacija softvera; AUTOSAR; MISRA; statički analizator; Kleng; LLVM; kompajler.**

I. Uvod

Softver je postao značajan deo svih aspekata naših života. Značajan deo procesa razvoja softvera odnosi se na process utvrđivanja njegove ispravnosti i kvaliteta, odnosno, verifikaciju softvera. Posledice neispravnosti softvera su mnogobrojne i protežu se od lakših, koje predstavljaju osećaj neprijatnosti usled nestabilnosti rada određene aplikacije, preko teških, koje predstavljaju pojavu ogromnih materijalnih gubitaka, pa sve do fatalnih posledica koje uključuju ljudske žrtve. Iako nije moguće napraviti program koji bi potpuno automatski, u konačnom vremenu, koristeći konačne resurse, mogao da utvrdi ispravnost proizvoljnog programa potpuno precizno, to nikako ne znači nepostojanje odgovarajućeg kompromisa [1].

Posebno važnu oblast primene verifikacije softvera predstavlja automobilska industrija, što za posledicu ima postojanje velikog broja standarda koji propisuju odgovarajuća pravila kodiranja (eng. *coding rules*). Jedan od njih je standard AUTOSAR (eng. *Automotive Open System Architecture*) C++ 14 kojim je propisano više stotina pravila kodiranja koja se odnose na programski jezik C++ [2].

Đorđe Milićević – Istraživačko-razvojni Institut RT-RK, Narodnog fronta 23a, 21000 Novi Sad, Srbija (e-mail: Djordje.Milicevic@rt-rk.com)
Mirko Brkušanin – Istraživačko-razvojni Institut RT-RK, Narodnog fronta 23a, 21000 Novi Sad, Srbija (e-mail: Mirko.Brkusanin@rt-rk.com)
Milena Vujošević Janičić – Matematički fakultet, Univerzitet u Beogradu, Studentski trg 16, 11000 Beograd (e-mail: milena@matf.bg.ac.rs)
Teodora Novković – Istraživačko-razvojni Institut RT-RK, Narodnog fronta 23a, 21000 Novi Sad, Srbija (e-mail: Teodora.Novkovic@rt-rk.com)
Petar Jovanović – Istraživačko-razvojni Institut RT-RK, Narodnog fronta 23a, 21000 Novi Sad, Srbija (e-mail: Petar.Jovanovic@rt-rk.com)

U radu je predstavljeno unapređenje programskog prevodioca Kleng (eng. *Clang*) [3] realizacijom skupa leksičkih, sintaksičkih i semantičkih pravila, propisanih od strane standarda AUTOSAR C++ 14. U glavi II predstavljeni su razlozi za nastanak i razvoj pomenutog standarda, brojnost i različite vrste klasifikacija propisanih pravila, postojeći statički analizatori koji imaju podršku za navedeni standard, kao i primer koji daje uvid u izazovnost realizacije statičkog analizatora koji podržava dati standard. Glava III opisuje strukturu projekta LLVM (sa fokusom na prednji deo kompajlera), implementacione detalje podržanih pravila, način upotrebe dijagnostike za prijavljivanje upozorenja, kao i način integrisanja razvijenog analizatora u postojeći kompajler. U glavi IV predstavljeni su rezultati evaluacije, dok je u glavi V izveden zaključak i date su smernice za buduća unapređenja.

II. Standardi AUTOSAR i MISRA

Standard AUTOSAR C++ 14 [2] nastao je usled nedostatka odgovarajućih standarda za programski jezik C++ (verzije 11 i 14) koji bi se primenjivali na sigurnosne (eng. *safety-related*) i kritične (eng. *critical*) sisteme. Standard MISRA (eng. *Motor Industry Software Reliability Association*) C++ 2008 [4] odnosio se na programski jezik C++ (verzija 03), koji je u vremenu razvoja gorepomenutog standarda AUTOSAR, bio već trinaest godina star i nije pokrivao napredne konstrukcije jezika koje su nove verzije C++ standarda donosile sa sobom. Štaviše, standard MISRA u potpunosti je zabranjivao korišćenje dinamičke memorije, dok je upotreba standardnih biblioteka bila nedovoljno dobro pokrivena. Glavni sektor primene standarda AUTOSAR predstavlja automobilska industrija (eng. *automotive*), ali se oblast primene može proširiti i na druge vrste ugrađenih (eng. *embedded*) sistema.

A. Klasifikacije pravila

Standard AUTOSAR C++ 14 (verzija 18-03) propisuje 402 pravila od kojih je deo pravila usvojen iz postojećeg standarda MISRA C++ 2008 (148 pravila, što predstavlja 64% MISRA standarda), deo pravila izveden iz postojećih standarda programskog jezika C++ (195 pravila), a preostala pravila zasnovana su na različitim vrstama istraživačkih radova (57 pravila) [2].

Pravila iz standarda AUTOSAR moguće je klasifikovati na nekoliko načina: prema **nivou važnosti** (eng. *obligation level*), **primenljivosti statičke analize** (eng. *enforcement by static analysis*) i **oblasti primene** (eng. *allocated target*) [2].

Klasifikacija prema nivou važnosti obuhvata podelu pravila na **obavezna** (eng. *required*) i **savetodavna** (eng. *advisory*). Da bi se za neki program moglo reći da ispunjava zahteve standarda AUTOSAR, sva pravila iz skupa obaveznih pravila

moraju biti ispunjena. Iako pravila iz skupa savetodavnih pravila nisu obavezna, to nikako ne znači da se ona mogu ignorisati – naprotiv, poželjno je pridržavati ih se u što većoj meri [2].

Klasifikacija prema primenljivosti statičke analize obuhvata podelu pravila na **automatizovana** (eng. *automated*), **delimično automatizovana** (eng. *partially automated*) i **neautomatizovana** (eng. *non-automated*). Automatizovana pravila predstavljaju skup pravila koje je moguće automatizovati u smislu statičke analize. Delimično automatizovana pravila moguće je automatizovati korišćenjem odgovarajućih heuristika, tako da pokriju najčešće scenarije koji vode ka narušavanju ispravnosti programa. Neautomatizovana pravila nije moguće automatizovati u smislu statičke analize, već zahtevaju ručno analiziranje programa ili korišćenje nekih drugih, dodatnih alata [2].

Klasifikacija prema oblasti primene obuhvata podelu pravila na **implementaciona** (eng. *implementation*), **verifikaciona** (eng. *verification*), **infrastrukturna** (eng. *infrastructure*) i pravila koja se odnose na **iskorišćene alate** (eng. *toolchain*). Implementaciona pravila odnose se na implementaciju projekta, odnosno, programski kôd, softverski dizajn i arhitekturu. Verifikaciona pravila odnose se na verifikacione aktivnosti poput pregleda, analize i testiranja koda. Infrastrukturna pravila odnose se na operativni sistem i hardver, dok se poslednji skup pravila odnosi na iskorišćene alate poput pretprocesora, kompajlera, linkera i biblioteka [2].

B. *Primer neodlučivog pravila*

U skupu pravila koje standard propisuje, postoje pravila koja su po svojoj prirodi neodlučiva (eng. *undecidable*) u opštem slučaju. Svojstvo neodlučivosti sugeriše kompleksnost algoritama koji se koriste pri implementaciji takvih pravila i korišćenje odgovarajućih aproksimacija, koje za posledicu imaju generisanje takozvanih lažnih uzbuna (eng. *false alarm*). U nastavku teksta će na jednom od pravila, koje ne dozvoljava postojanje mrtvog koda (eng. *dead code*), biti prikazana teškoća rezonovanja o tome da li je neki deo koda mrtav, odnosno, da li će neki deo koda biti izvršen u fazi izvršavanja programa. Uzmimo za primer *while* petlju u kojoj se u zavisnosti od uslova određuje dalji tok izvršavanja programa. Ukoliko uslov petlje ne može biti evaluiran u vremenu prevođenja (jer zavisi od okruženja u kojem se dati program izvršava), iz toga sledi da nije moguće sa sigurnošću utvrditi tok kojim će se izvršavanje programa nastaviti, a samim tim ni rezonovati o tome da li je neki deo programa mrtav ili ne. Algoritam, kojim je spomenuto pravilo implementirano, ne bi trebalo da proglasi određeni deo koda mrtvim (ukoliko postoje tendencije da u nekom slučaju taj deo koda bude izvršen), ali takođe broj lažnih uzbuna mora biti sveden na minimum. Iz toga jasno proističe zaključak da su neodlučiva pravila veoma izazovna za implementaciju i testiranje, kao i da se pri njihovoj implementaji moraju koristiti odgovarajuće heuristike i aproksimacije.

III. STATIČKA ANALIZA KODA

Statička analiza koda obuhvata sve tehnike analize karakteristika koda bez njegovog izvršavanja, a koje imaju za cilj pronalaženje defekata u kodu. Statička analiza može biti neautomatizovana, u obliku pregleda koda (eng. *code review*) ili automatizovana. U okviru automatizovane statičke analize vrše se različite provere kvaliteta koda, počevši od izračunavanja mnogobrojnih softverskih metrika do provera ispravnosti koda u kontekstu nepostojanja grešaka u fazi izvršavanja (eng. *runtime error*), kao što su deljenje nulom, pisanje van granica niza (eng. *buffer overflow*), dereferenciranje neispravnog pokazivača, postojanje mrtvog koda i slično. Dok je izračunavanje metrika koda veoma jednostavno, za potrebe statičkih provera koje se odnose na osobine softvera koje se ispoljavaju u fazi izvršavanja, razvijene su kompleksne tehnike apstraktne interpretacije [5] (primeri alata su **Astree** [6], **Coverty** [7], **Polyspace Bug Finder** [8]), proveravanja modela [9] (primeri alata su **CBMC** [10], **LLBMC** [11], **LAV** [12]) i simboličkog izvršavanja [13] (primeri alata su **KLEE** [14] i **PEX** [15]). Osnovne razlike ovih tehnika su u preciznosti i efikasnosti analize koju obavljaju. Pomenute tehnike mogu se koristiti i u okviru provere ispunjenosti kodiranja po standardu AUTOSAR C++ 14.

A. *Statički analizatori za standard C++ 14*

Neki od postojećih statičkih analizatora koda, koji imaju podršku za AUTOSAR C++ 14 standard, su: **Helix QAC** (kompanije Perforce) [16], **Klocwork** (kompanije Rogue Wave) [17] i statički analizator kompanije LDRA [18]. Prema dostupnim informacijama kompanije Rogue Wave, statički analizator Klocwork podržava 181 pravilo iz standarda AUTOSAR C++ 14 (verzija 18-03) [17].

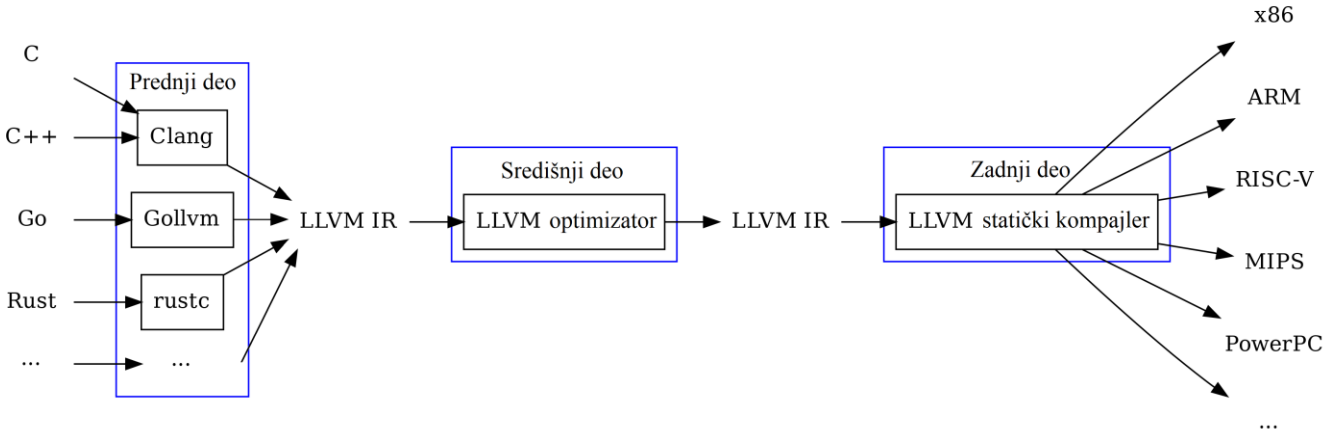
IV. REALIZACIJA

Projekat LLVM [19] sastoji se iz više biblioteka i alata koji zajedno čine veliku kompajlersku infrastrukturu. Osnovna filozofija LLVM-a je da je „svaki deo neka biblioteka" i veliki deo koda je ponovno upotrebljiv. Projekat je započet 2000. godine na univerzitetu u Ilinoisu, kao istraživački rad sa ciljem proučavanja tehnika kompajliranja i kompajlerskih optimizacija [20]. Danas je LLVM prerastao u sveobuhvatni naziv za više projekata koji zajedno čine potpun kompajler: prednji deo (eng. *frontend*), središnji deo (eng. *middleend*), zadnji deo (eng. *backend*), optimizatore, asemblere, linkere, libc++ i druge komponente. Projekat je napisan u programskom jeziku C++, koristeći prednosti objektno-orijentisane paradigme, generičkog programiranja (upotrebom šablona), a takođe sadrži i svoje implementacije raznih struktura podataka koje se javljaju u standardnim bibliotekama programskog jezika C/C++, kao i novu klasu za efikasnije upravljanje niskama karaktera (eng. *string*).

Termin LLVM se nekada može koristiti i sa drugim značenjem. Tako na primer možemo govoriti o kompajleru zasnovanom na LLVM-u, koji može biti u potpunosti ili delimično izgrađen od LLVM strukture. Možemo koristiti prednji i zadnji deo iz LLVM-a, a povezivati (eng. *linking*) sa GCC-om i GNU-ovim bibliotekama.

A. *Struktura LLVM projekta*

Osnovni delovi projekta LLVM su prednji deo, središnji deo i zadnji deo (Sl. 1.) [19]:



Sl. 1 Podela LLVM-a na predni deo, IR (srednji deo) i zadnji deo.

Prednji deo prevodi izvorni kôd (eng. *source code*) nekog od podržanih programskih jezika u međureprezentaciju LLVM IR (eng. *intermediate representation*). Ovaj deo podrazumeva leksički, sintaksički, semantički analizator i generator LLVM međureprezentacije. Kleng je podprojekat LLVM-a koji predstavlja prednji deo za jezike: C, C++ i Objective-C. Pored spomenutog on nudi još i dodatne alate i biblioteke za statičku analizu koda.

IR je programski jezik niskog nivoa, blizak asembleru. IR je jezik u formi SSA (eng. *static single assignment*) koju odlikuje beskonačan broj registara. U **središnjem delu** procesa kompilacije nalazi se veliki broj optimizacionih prolaza. Optimizaciju možemo posmatrati kao prevođenje početnog koda u semantički ekvivalentan, ali efikasniji kôd.

Zadnji deo služi da generiše izvršni kôd za specifičnu arhitekturu. Dobijeni IR će se prevesti u odgovarajuće mašinske instrukcije. Takođe će obaviti i neke dodatne transformacije kao što je registarska alokacija, transformacije petlji i razne druge optimizacije koje su specifične za platformu za koju se prevodi.

Svaku od ovih komponenata moguće je pokrenuti pojedinačno sa ulaznim i izlaznim datotekama odgovarajućeg formata. Ukoliko nam uvid u razne međurezultate nije potreban, a samo želimo da dobijemo izvršnu datoteku, onda za to možemo upotrebiti Kleng. Pored biblioteka za prednji deo on se može koristiti i kao alat koji upravlja celokupnim procesom kompajliranja. Jednostavno će izlaz jedne faze proslediti narednoj i tako ne moramo biti upoznati sa pojedinačnim alatima i njihovim načinom korišćenja. Uz pomoć ovog interfejsa Kleng možemo koristiti kao bilo koji drugi kompajler. Takođe je kompatibilan sa flegovima kompajlera GCC, pa LLVM možemo lako koristiti kao zamenu [19].

Kada se govori o Klengu potrebno je razlikovati rukovaoca (eng. *driver*) koji upravlja procesom kompilacije od prednjeg dela koji se bavi obradom izvornog koda i generisanja IR-a.

B. Implementacija statičke analize

Kako se standardi AUTOSAR i MISRA bave propisivanjem pravila za C++, nama će od značaja biti samo Kleng, odnosno, samo one biblioteke iz prednjeg dela koje direktno zavise od programskog jezika C/C++, a to su: *Lex* (pretprocesiranje i leksička analiza), *Parse* (sintaksička analiza) i *Sema* (semantička analiza). Celokupna implementacija do sada ugrađenih pravila nalazi se u ove tri biblioteke. Projekat je urađen kao nadogradnja LLVM verzije 7.0.1.

Rezultat rada nakon leksičke, sintaksičke i semantičke analize je ispravno apstraktno sintaksičko stablo (eng. AST, *abstract syntax tree*). Ova struktura, iako ne čuva sve detalje sintakse, jasno oslikava strukturu koda i veliki deo detalja, što je čini pogodnijom za dalje ispitivanje i obrade od obične tekstualne reprezentacije. Svaki čvor stabla odgovara jednom konstruktu jezika. Na primer, možemo imati čvor koji odgovara operatoru sabiranja ili množenja. Potomci tog čvora su operandi koji mogu biti konstante, promenljive ili neki drugi složeni izrazi.

Implementacija svakog pravila je enkapsulirana u zasebnu funkciju ili klasu. Izuzeci postoje za vrlo jednostavna i slična pravila kao što su, na primer, zabrana upotrebe pojedinih zaglavlja. Postoje odvojena pravila koja kažu da nije dozvoljno koristiti mogućnosti biblioteka: *csignal*, *cstdio*, *ctime* i *clocale*. Sve četiri provere se razlikuju samo u poređenju sa odgovarajućom niskom karaktera i vrlo ih je prirodno spojiti.

Način prijavljivanja prekršaja nekog pravila vrši se pomoću Klengovih mehanizama dijagnostike. Dovoljno je da znamo lokaciju u kodu gde se javlja element koji ne poštuje standard, kao i naziv upozorenja koji odgovara pravilu koje prijavljujemo. Za svako pravilo postoji zaseban fleg, kao što je *–Wvolatile-keyword-used* (slika 2), koji kontroliše da li će se ono ispitati i prijaviti odgovarajuće upozorenje korisniku. Ukoliko želimo da uključimo sva upozorenja tada, prilikom prevođenja, koristimo opciju *-Wautosar-cxx14*. Takođe su dostupni i flegovi za manje grupe pravila koje odgovaraju podeli iz standarda AUTOSAR (grupa pravila za klase, izraze, deklaracije i sl.). Jednostavnim upitom nad klasom za dijagnostiku možemo proveriti da li je neki od flegova prisutan i tako kontrolisati izvršavanje naših dodatnih provera.

```
test.cpp:9:3: warning: volatile keyword shall not be used
 volatile int i;
 ^
```

Sl. 2 Primer ispisa Klenga za prekršaj pravila.

Trenutna implementacija sadrži 164 pravila, od toga je 50 implementirano kao prosti funkcijski pozivi, 70 kao zasebne klase koje se oslanjaju na neku od pomoćnih struktura iz Klenga, 31 je u potpunosti podržano od strane Klenga kao i 13 koja ne pokrivaju sve slučajeve koje AUTOSAR zahteva i potrebno ih je dopuniti.

Pravila implementirana kao funkcije. Prilikom obrade izvornog koda kroz biblioteke *Lex, Parse* i *Sema* vrše se provere da li je kôd ispravno i smisleno napisan. Ukoliko su korišćene naredbe ili elementi jezika koji ne postoje ili nisu lepo upareni kompajler će nam prijaviti grešku ili upozorenje. Neka od prostojih pravila standarda AUTOSAR i MISRA predstavljaju zabranu upotrebe pojedinih elemenata, struktura ili funkcija iz programskog jezika C++. Prema tome, prirodno je implementirati ih baš na mestu gde se oni obrađuju. Ugrađivanjem dodatnih provera i ograničenja možemo prepoznati situacije koje su nam od značaja. Kako bi što više zadržali nezavisnost originalnog Kleng koda od našeg, sve dodatne provere se vrše u zasebnim funkcijama koje samo pozivamo na odgovarajućim mestima. Tako, na primer, funkcija koja sadrži logiku za pravilo koje ograničava upotrebu ključne reči *typedef* će se pozvati na mestu u parseru neposredno nakon obrade te ključne reči.

Pravila implementirana kao klase. Ukolko je pravilo zahtevnije onda će i logika koja ga opisuje biti složenija pa se ne može efikasno implementirati pomoću obične funkcije. Tu su nam od velike pomoći rekurzivni AST posetioci (eng. *Recursive AST Visitors*). Ova klasa nam nudi mogućnost obilaska celog AST stabla pretragom u dubinu. Za svaki tip čvora u stablu, koji nam je od značaja, možemo definisati odgovarajući metod posete. Potrebno je naslediti klasu *RecursiveASTVisitor* i predefinisati željeni metod. Na primer, ukoliko nas interesuju upotrebe binarnih operatora onda ćemo definisati metod *VisitBinaryOperator(BinaryOperator *)*. Svi metodi posete kao argument dobijaju pokazivač na čvor koji se trenutno obilazi odakle možemo dobiti sve informacije koje nas interesuju. Tako će, na primer, pravilo koje ispituje da li se koriste ispravni operatori nad promeljivama tipa *bool* da poseti čvorove u stablu koji opisuju binarne i unarne operatore. Metod posete će se sastojati od provere da li se koristi neki nedozvoljen operator zajedno sa proverom da li su operandi tipa *bool*.

Pored metoda posete dostupni su nam i metodi obilaska. Njih koristimo kada želimo da obiđemo stablo u dubinu od nekog konkretnog čvora, a možemo ih i predefinisti ako nam je potreban neki drugi redosled obilaska. Najčešće ih koristimo kada želimo da se naše pravilo ispita za celokupni kôd. Potrebo je pozvati već definisani metod obilaska *TraverseDecl(Decl *)* sa korenim čvorom AST-a kao argumentom.

Drugi korisni mehanizam koji koristimo je klasa *PPCallbacks*. Ona predstavlja interfejs za praćenje rada pretprocesora. Za neka pravila koja se odnose na pretprocesorske direktive ne možemo na mestu njihove obrade u biblioteci *Lex* znati da li su ispoštovana ili ne. Jedno od takvih pravila se odnosi na proveru neiskorićenih zaglavlja. To možemo utvrditi tek kada imamo uvid u potpun AST ulaznog koda i znamo da li se neka funkcija, struktura ili bilo koji drugi element deklarisan u nekom od zaglavlja zapravo koristi. Tada ćemo nasleđivanjem klase *PPCallback* i predefinisanjem željenih metoda za obradu „include" direktiva, makroa i drugih elementa sakupiti sve podatke koji će nam biti od koristi. Tek u nekoj kasnijoj fazi kompilacije kada nam budu dostupne preostale informacije o ulaznom kodu zajedno sa podacima iz pretprocesora možemo doneti odluku o traženim pravilima.

Integracija sa Klengom. Sav kôd koji je dodat se nalazi u novim i odvojenim datotekama. Jedine izmene u postojećem kodu su u .cpp datotekama i predstavljaju samo umetnute pozive funkcija zajedno sa uključivanjem neophodnih zaglavlja na odgovarajućem mestima u bibliotekama *Lex*, *Parse* i *Sema*. Te funkcije ili vrše jednostavnu proveru pojedinih pravila ili instanciraju rekurzivne posetioce za obilazak AST-a. Kao argumente, te funkcije uzimaju podatke koji su im dostupni ali ih nikada ne menjaju i time ni na koji način ne utiču na postojeću funkcionalnost Klenga niti je ometaju.

V. Evaluacija

Pored samog broja implementiranih pravila bitna nam je i efikasnost tih provera. Svaka nova provera troši dodatno procesorsko vreme. Pokretanje našeg koda, koji podržava 164 pravila, troši 32,5% više vremena za izgradnju svih testova grupa *SingleSource* i *MultiSource,* dostupnim u okviru LLVM test-suite-a, kada su sve provere uključne. Glavni razlog ovog usporenja je veliki broj prijava pojedinih pravila usled testiranja na kodu koji nije pisan po standardu AUTOSAR ili MISRA. Primer takvog pravila je zabrana upotrebe celobrojnih tipova koji nisu fiksne veličine. Dakle, umesto: short, int, long i dr. potrebno je koristiti tipove int8_t, int16_t, int32_t, int64_t ili njihove ekvivalente bez znaka. Kako je upotreba ovih tipova vrlo česta javlja se i veliki broj upozorenja za ovo pravilo i to čak 18.85% od svih upozorenja. Sledeća mnogobrojna pravila po broju upozorenja su: implicitno umanjanje veličine tipa prilikom konverzije (12,64%), elementi van globalnog namespace-a (8.95%) i tipografski slični identifikatori (7,38%). Preostala pravila daju manje od 5% upozorenja po pravilu.

Kako prijavljivanje velikog broja prekršaja korisniku nije uvek od koristi ugrađena je i mogućnost ograničavanja broja upozorenja po pravilu. Ograničavaljem na 1, 2 ili 5 upozorenja je dovoljno da se korisnik obavesti da kôd nije po standardu i koja su sve pravila prekršena. Zatim se može iterativno popravljati kôd i pokretati jedno po jedno pravilo bez ograničenja. Prilikom ograničenja na maksimalno 5 upozorenja po pravilu vreme potrebno da se prevede ceo LLVM test-suite je 11.2% duže od pokretanja bez ograničenja. Sa samo jednim upozorenjem dobija se sličan rezultat, dok ukoliko isključimo dijagnostiku za prijavljivanje upozorenja kako ne bi računali vreme potrebno za njihov ispis imamo uvećanje od 19,15%. Ova informacija nam bolje oslikava koliko se vremena troši na sam ispis dijagnostike, nasuprot vremenu potrebnom za celokupnu analizu koda našim kompajlerom. Ukoliko pišemo kôd po standardu i redovno vršimo provere možemo očekivati da ćemo imati mali broj upozorenja pa samim tim i manje vreme potrebno da se izvrši celokupna analiza koda sa našim kompajlerom.

VI. ZAKLJUČAK.

U okviru rada prikazano je da se značajan broj pravila može jednostavno i efikasno implementirati u Klengu. Sa novim dopunama, koje nemaju uticaja na već postojeću funkcinalnost, od već kvalitetnog kompajlera dobijamo novi alat sa velikim brojem statičkih analiza koda. Sve dodatne provere se mogu jednostavno (po potrebi) uključiti ili isključiti, pa ne moramo brinuti o tome da će naša verzija na bilo koji način poremetiti ili usporiti podrazumevani način upotrebe Klenga. Kako su sve izmene postojećeg koda minimalne, naša implementacija se može lako nadovezati na buduće verzije.

Sa svakom dodatnom proverom razmenjujemo procesorsko vreme za bezbedniji kôd. Na korisniku ostaje da odluči da li se isplati prosečno povećanje od 19,15% prilikom prevođenja za 164 dodatnih provera. Kako se sva analiza obavlja u prednjem delu, može se pokrenuti i bez potpunog prevođenja. Pametnom upotrebom analizatora sa ograničavanjem broja upozorenja, i samo nad jedinicama prevođenja koje korisnik u tom trenutku razvija, može se značajno uštedeti na vremenu.

A. Moguća unapređenja

Pored dodavanja novih pravila podjednako je važno i umanjivanje broja lažnih upozorenja. Zbog postojanja neodlučivih pravila nije uvek moguće dati precizan odgovor. Međutim, veliki broj lažnih upozorenja može zbuniti korisnika i obeshrabriti ga od upotrebe statičkih analizatora. Zato je testiranje na većim količinama koda i unapređivanje postojeće implementacije od velikog značaja.

ZAHVALNICA

Ovaj rad je delimično finansiran od strane Ministarstva za prosvetu, nauku i tehnološki razvoj Republike Srbije, na projektu brojevi: TR32041 i IO174021.

LITERATURA

[1] Milena Vujošević Janičić, Automatsko generisanje i proveravanje uslova ispravnosti programa, Doktorska teza, Matematički fakultet, Univerzitet u Beogradu, Decembar, 2013.
[2] AUTOSAR, Guidelines for the use of the C++14 language in critical and safety-related systems, 2018.
[3] LLVM, Clang C Language Family Frontend for LLVM, on-line at: https://clang.llvm.org/, 2019.
[4] MISRA, MISRA – The Motor Indistry Software Reliability Association, on-line at: https://www.misra.org.uk/, 2019.
[5] Patrick Cousot and Radhia Cousot, Abstract interpretation: a unified lattice model for static analysis of programs by construction or approximation of fi-xpoints, In Proceedings of the 4th ACM SIGACT-SIGPLAN symposium on Principles of programming languages, pages 238–252, ACM, 1977.
[6] AbsInt, Astree Runtime Error Analyzer, on-line at: https://www.absint.com/astree/index.htm, 2019.
[7] Synopsis, Coverty, on-line at: https://scan.coverity.com, 2019.
[8] MathWorks, Polyspace, on-line at: https://www.mathworks.com/products/polyspace.html
[9] Edmund M Clarke, Thomas A Henzinger, Helmut Veith, and Roderick Bloem, Handbook of model checking, Springer, 2018.
[10] E. Clarke, D. Kroening, and F. Lerda. A tool for checking ANSI-C programs. In Tools and Algorithms for the Construction and Analysis of Systems (TACAS), Springer, 2004, pp. 168-176.
[11] F. Merz, S. Falke, C. Sinz, LLBMC: Bounded Model Checking of C and C++ Programs Using a Compiler IR, in: Verified Software: Theories, Tools and Experiments (VSTTE), LNCS, Springer, 2012, pp. 146-161.
[12] M. Vujošević Janičić, V. Kuncak, Development and Evaluation of LAV: An SMT-Based Error Finding Platform, in: Verified Software: Theories, Tools and Experiments (VSTTE), LNCS, Springer, 2012, pp. 98-113.
[13] Roberto Baldoni, Emilio Coppa, Daniele Cono D'elia, Camil Demetrescu, and Irene Finocchi, A survey of symbolic execution techniques, ACM Computing Surveys (CSUR), 51(3):50, 2018.
[14] C. Cadar, D. Dunbar, and D. Engler, Klee: Unassisted and automatic generation of high-coverage tests for complex systems programs. In Proceedings of the 8th USENIX conference on Operating systems design and implementation (OSDI), USENIX Association Berkeley, 2008, pp.209-224.
[15] N. Tillmann and J. Halleux, Pex white box test generation for .NET, In Proc. of TAP 2008, volume 4966 of LNCS, pages 134-153. Springer, 2008.
[16] Perforce, AUTOSAR Compliance – Why AUTOSAR C++ Guidelines?, on-line at: https://www.perforce.com/resources/qac/autosar, 2019.
[17] RogueWave, AUTOSAR 18-03 Standard mapped to Klocwork C/C++ checkers, on-line at: https://docs.roguewave.com/en/klocwork/2018/autosar18idsmappedtoklocworkcandccheckers, 2019.
[18] LDRA, AUTOSAR C++ Coding Standards Compliance – LDRA, on-line at: https://ldra.com/automotive/standards/autosar-c-coding-standards-compliance/, 2019.
[19] Bruno Cardoso Lopes, Rafael Auler, Getting started with LLVM core libraries, Packt Publishing, 2014.
[20] Suyog Sarda, Mayur Pandey, LLVM essential, Packt Publishing, 2015.

ABSTRACT

Respecting coding standards is especially important in automotive industry because automotive software bugs can have fatal consequences. The paper presents an improvement of the Clang compiler by adding verification capabilities fulfillment of 164 rules from the MISRA/AUTOSAR coding standard. The paper presents the description and structure of the AUTOSAR coding standard, LLVM compiler infrastructure and Clang compiler, as well as implementation details and how to integrate the set of supported rules in the frontend of the LLVM infrastructure. Within the implementation, different types of extensions have been made to lexical, syntax and semantic analyzers. An evaluation was established that the implemented additional analyzes slow down the process of compilation for only 19.15%, which makes this implementation extremely efficient..

Improving Clang compiler with MISRA/AUTOSAR coding standard support

Đorđe Mićević, Mirko Brkušanin,
Milena Vujošević Janičić, Teodora Novković,
Petar Jovanović

Dodavanje podrške za arhitekturu nanoMIPS u alat za dinamičku analizu programskog koda Velgrind

Dimitrije Nikolić, Aleksandra Karadžić, Aleksandar Rikalo i Petar Jovanović

***Apstrakt*—Pojava nove arhitekure mikroprocesorkih sistema pored hardverskog čipa podrazumeva i određene softverske alate, poput kompajlera i emulatora. Upotreba nove arhitekture, odnsosno kreiranje složenih aplikacija za novu arhitekturu, povlači potrebu za softverskim alatima koji bi olakšali detektovanje nepravilnog rada programa i lakše pronalaženje grešaka, poput debagera i profajlera. Ovaj rad opisuje izmene koje su potrebne za dodavanje podrške jednom softerskom alatu za analizu programa, Velgrindu, za novu arhitekturu MIPS Techologies grupe – nanoMIPS.**

***Ključne reči*—Velgrind, MIPS, nanoMIPS**

I. Uvod

Traženje razloga nepravilnog rada sistema može potrajati dosta dugo, pogotovu ako se sistem sastoji iz desetina hiljada linija koda i desetine pa i stotine operacija alociranja i dealociranja memorije. Pod ovim podrazumevamo greške koje prevodioci ne prijavljuju, poput curenja memorije ili korišćenja neinicijalizovanog podatka. Za neke još složenije, višenitne sisteme, ove greške se mogu javiti u vidu neočekivanog pristupa deljenim podacima, odnosno utrkivanju za pristup istim. Jedan od alata koji pomažu u otkrivanju ovakvih grešaka jeste Velgrind. Velgrind predstavlja veoma koristan alat, pogodan za analizu svih nivoa memorije. Puštanje programa kroz Velgrind jeste značajno sporije (od 20 do 100 puta sporije), ali sa lakoćom može otkriti nepravilnosti prilikom korišćenja memorije.

Cilj ovog rada jeste omogućiti korisnicima, koji rade na najnovijoj MIPS arhitekturi - nanoMIPS, korišćenje Velgrinda i tako im olakšati razvoj sistema na ovoj arhitekturi. Kako bismo realizovali podršku Velgrind alata za nanoMIPS arhitekture, moramo, pored detaljnog upoznavanja načina rada Velgrinda, izučiti novi instrukcijski set nanoMIPS arhitekture, kao i uporediti novouvedeni ABI p32 sa do sada korišćenim ABI-jem o32, odnosno sličnosti i razlike između ova dva interfejsa.

Dimitrije Nikolić, Naučno istraživački institut RT-RK, Bulevar Milutina Milankovića 19b, 11000 Beograd, Srbija (telefon: 381-21-480-1145, e-mail: dimitrije.nikolic@rt-rk.com)

Aleksandar Rikalo, Naučno istraživački institut RT-RK, Bulevar Milutina Milankovića 19b, 11000 Beograd, Srbija (telefon 381-21-480-1145, e-mail: aleksandar.rikalo@rt-rk.com).

Aleksandra Karadžić, Naučno istraživački institut RT-RK, Bulevar Milutina Milankovića 19b, 11000 Beograd, Srbija (telefon 381-21-480-1145, e-mail: aleksandra.karadzic@rt-rk.com).

Petar Jovanović III_044009_1, Naučno istraživački institut RT-RK, Narodnog fronta 23a 21000 Novi Sad, Srbija (telefon: 381-21-480-1193, e-mail: petar.jovanovic@rt-rk.com).

II. Motivacija

Motivacija za pisanje ovog rada je višestruka:

- predstavljanje karakteristika nove arhitekture kompanije MIPS Technologies – nanoMIPS
- predstavljanje načina rada dinamičkog analizatora programskog koda Velgrind
- predstavljanje rešenja za dodavanje podrške nove arhitekture u dinamički analizator programskog koda Velgrind

III. Uvod u Velgrind

Valgrind predstavlja DBI radni okvir za Linuks/x86 platformu. Originalno, Valgrind je dizajniran da bude besplatan alat za detekciju grešaka prilikom korišćenja memorije za Linuks na x86 arhitekturi, ali je vremenom prerastao u okruženje za kreiranje alata za dinamičku binarnu analizu.

Trenutna verzija Velgrinda je 3.14.0 (objavljena 9. oktobra 2018. godine) i ona sadrži sledeće alate:

- *Memcheck* – alat za otkrivanje grešaka prilikom korišćenja memorije, i primarno je namenjen za C/C++ programe.
- *Cachegrind* – alat za profiliranje keš memorije.
- *Callgrind* – proširenje *Cachegrind*-a, može se koristiti za vizuelizaciju podataka dobijenih radom *Cachegrind*-a.
- *Massif* – alat za profiliranje dinamičke memorije (*heap*-a).
- *Helgrind* – alat za otkrivanje utrkivanja niti prilikom pristupa deljenom podatku.
- *DRD* – alat za analizu višenitnih program, koji zauzima mnogo manje memorije od *Helgrind*-a tokom svog rada.
- *Lackey* & *None* (*Nulgrind*)–alati koji postoje za testiranje Velgrinda i demonstrativne svrhe.

Podržane su sledeće arhitekture:

- Linux OS : x86, AMD64, ARM, MIPS32, MIPS64, TILEGX, PPC32, PPC64, PPC64LE, S390X
- Solaris OS: x86, AMD64
- Android OS: ARM (od 2.3.x), ARM64, x86(od 4.0), MIPS32
- Darwin OS: x86, AMD64 (Mac OS X 10.10, sa inicijalnom podrškom za 10.11)

IV. ANALIZA RADA VELGRINDA

Alati za dinamičku analizu koda se kreiraju kao dodatak, pisan u programskom jeziku C, na jezgro Velgrinda. Velgrind se najjednostavnije može predstaviti kao:

Osnova Velgrinda(jezgro+VEX) + dodatak u vidu alata = Velgrind alat

Vex (prevodilac JIT tipa) je zadužen za dinamičko prevođenje koda, dok je jezgro Velgrinda zaduženo za ostatak (raspoređivanje procesa, vođenje računa o deljenoj memoriji, itd.). Programska podrška Velgrinda je logički podeljena na dve suštinski različite celine: *host* (sprega između Velgrindove interne reprezentacije (IR) i ciljne arhitekture) i *guest* (sprega između klijentskog koda i Velgrindove interne reprezentacije).

Velgrind deli klijentski program u sekvence koje se nazivaju bazični blokovi. Bazični blok je deo programskog koda koji nema uskakanja u kod izuzev prve intrukcije i kraja bazičnog bloka i nema u sebi grananja, sem u poslednjoj instrukciji. Pre prve instrukcije bazičnog bloka nalaze se instrukcija/e skoka ili uskakanje/a u kod.[4] Dodatno ograničenje za veličinu bazičnog bloka kod Velgrinda je 60 mašinskih instrukcija. Na procesorima zasnovanim na MIPS arhitekturi, instrukcije skoka i grananja imaju tzv. „odloženo izvršavanje“, što znači da se prilikom njihovog izvršavanja izvršava instrukcija koja se nalazi neposredno iza instrukcije grananja ili skoka. Ukoliko je poslednja (60.) instrukcija bazičnog bloka instrukcija grananja ili skoka, Velgrind učitava i instrukciju koja se nalazi neposredno iza 60. instrukcije bazičnog bloka.

Programski kod svakog programa koji se analizira se ponovo prevodi na zahtev, pojedinačno na nivou bazičnog bloka, neposredno pre samog izvršavanja bazičnog bloka. Svaki IR blok sadrži listu iskaza koje predstavljaju operacije sa bočnim efektima, kao što su upis u registre/memoriju ili smeštanje vrednosti u privremenim promenljivama. Iskazi se sastoje iz izraza bez bočnih efekata, kao što su konstante, čitanje registara, dohvatanje podataka iz memorije, kao i aritmetičke operacije. Uzećemo za primer jednostavan iskaz za skladištenje podataka. On sadrži dva izraza, jedan za smeštanje adrese, a drugi za smeštanje vrednosti. Izrazi, proizvoljno, mogu biti organizovana u kompleksna stabla (IR stablo) ili u linearizovanu strukturu upotrebom iskaza koji među-vrednosti smeštaju u privremene promenljive.

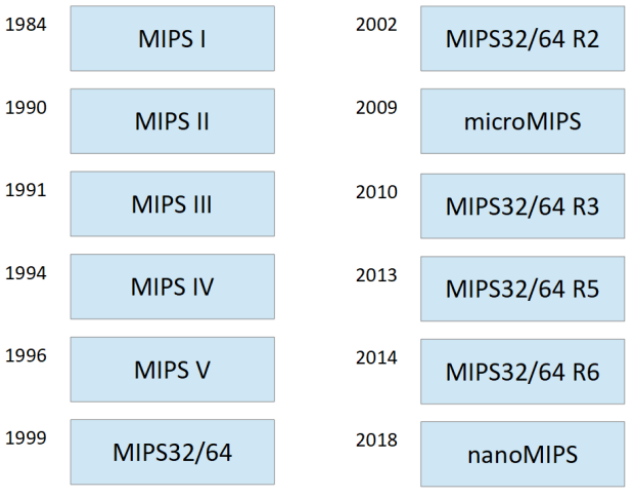
Iako je je ovaj skup primitivnih operacija dovoljan za skoro sve instrukcije, postoje neke instrukcije specifične za neku arhitekturu, koje je nemoguće "razbiti" na više standardnih IR operacija. U tom slučaju se koriste pomoćne funkcije, pisane u programskom jeziku C, koje emuliraju tu instrukciju.

Nakon pokretanja Valgrind-a, započinje translacija, te se kasnije izvršava novodobijeni klijentski program. U toku translacije, mašinski kod se disasemblira u Valgrind-ov međukod, ubacuju se delovi izabranog alata i na kraju asemblira u mašinski kod uz usputne optimizacije. Faza translacije se sastoji iz osam celina:

1. **Disasembliranje** - Disasembler konvertuje mašinski kod u stablo IR-a. Svaka instrukcija mašinskog koda se disasemblira nezavisno u jednu ili više instrukcija IR-a.
2. **Optimizacija IR-a** - Ovaj korak prestavlja prvu od dve optimizacija IR-a koje se rade prilikom translacije. Tokom ovog koraka, kod koji je bio organizovan kao stablo se linearizuje. Iz generisanog IR-a prilikom disasembliranja se uklanjaju redudantne get i put operacije, kao i mrtav kod. Pojednostaljuju se konstantni izrazi i vrši se jednostavno odmotavanje petlji koje se izvršavaju unutar jednog osnovnog bloka.
3. **Instrumentacija** - Instrumentacija ubacuje nove IR instrukcije u kod dobijen u prethodnom koraku. Dodate instrukcije ubacuje odabran alat i one služe za željenu analizu koda. Ove instrukcije ne narušavaju konzistentno izvršavanje originalnog programa. Kako bi se instrumentacija lakše izvršila, važno je da kod u ovom koraku bude linearizovan.
4. **Optimizacija IR-a** - Druga, jednostavnija, optimizacija dodatno pojednostavljuje konstantne izraze i uklanja mrtav kod. Takođe, uklanjaju se i nepotrebne provere definisanosti nastale u prethodnoj fazi. Ova optimizacija „olakšava život" alatu dozvoljavajući mu da uradi nešto na jednostavan način (neoptimizovano) znajući da će taj kod biti naknadno poboljšan.
5. **Gradnja stabla** - U ovom koraku, linearni IR se vraća u reprezentaciju stabla, pripremajući se za narednu fazu.
6. **Odabir instrukcija** - Ovaj korak je zavistan od arhitekture za koju radi Valgrind. Svaka IR instrukcija se prevodi u jednu ili više instrukcija ciljne arhitekture. Tokom ove faze instrukcije i dalje koriste virtuelne (tzv. guest) registre. Selektor instrukcija koristi jednostavan, "pohlepan" algoritam uparivanja stabla koji obrađuje stablo od vrha ka dnu.
7. **Dodela registara** -Virtuelni (guest) registri se zamenjuju registrima ciljne arhitekture (host). U slučaju da broj registra koji su na raspolaganju nije dovoljan, primenjuje se tehnika "presipanja" registra na stek (eng. register spill). Ova tehnika podrazumeva da se svi registri, čije se vrednosti moraju održavati kroz pozive funkcija, privremeno čuvaju na steku. Iako su instrukcije zavisne od arhitekture, alokator registra nije i koristi specifične funkcije da bi zaključio koji se registri koriste (čita ili upisuje u njih) prilikom svake instrukcije.
8. **Asembliranje** - Završna faza translacije, faza asembliranja, jednostavno prevodi odabrane instrukcije u mašinske i upisuje ih u blok memorije.[2]

V. NANOMIPS ARHITEKTURA

Iako trenutno manje zastupljeni, MIPS procesori su imali veliki uticaj na razvoj mikropocesorskih arhitektura. Primera radi, procesor MIPS R4000 je bio prvi mikroprocesor koji je koristio 64-bitnu magistralu podataka, mnogi procesori su se nalazili u ugrađenim sistemima poput Cisco rutera, uređajima za zabavu Sony Playstation, Nintendo, itd. Plod višedecenijskih istraživanja i usavršavanja mikroprocesorskih sistema je hronologijski prikazan na slici 1.



Slika 1 – Istorijat MIPS mikroprocesorskih arhitektura

NanoMIPS je najnoviji 32-bitni instrukcijski skup kompanije MIPS Technologies, objavljen 01. maja 2018. godine. Dizajniran za ugrađene uređaje, nanoMIPS je arhitektura sa promenljivom dužinom instrukcija, pružajući visoke performanse prilikom redukovanja veličine koda. Uz slične flegove kompajleru, nanoMIPS može redukovati veličinu koda do 40% u odnosu na MIPS32. Sa smanjenim brojem pristupa memoriji i efikasnijem korišćenjem instrukcijskog keša, nanoMIPS mikroprocesori takođe redukuju potrošnju snage sistema.

Kao što je već navedeno, nanoMIPS je arhitektura sa promenljivom dužinom instrukcija, dužine 16, 32 ili 48 bitova. Instrukcije dužine 16 bita su zapravo kraće verzije najčešće korišćenih 32-bitnih instrukcija sa ograničenim skupom registra (podskupom od 4, 8 ili 16 registara) koji se mogu koristiti kao operandi. Optimizacijom u kompajleru za nanoMIPS, prilikom dodeljivanja registara instrukcijama, forsiraju se registri iz pomenutih podskupova, te se 16-bitne instrukcije koriste uvek kada za to postoji mogućnost. Takođe, slična optmizacija je implementirana i u linkeru, kome je omogućeno da u objektnom fajlu zameni 32-bitne instrukcije 16-bitnim ekvivalentnim instrukcijama, ako su ispunjeni odgovarajući uslovi (registri prosleđeni kao operandi su deo odgovarajućeg podskupa). Instrukcije dužine 48 bita su novina u odnosu na prethodne MIPS arhitekture i pružaju mogućnost smeštanja 32-bitnog argumenta u samu instrukcijsku reč. Ovom novinom je eliminisana potreba za dodatnom instrukcijom koja bi isti argument smeštala u registar, pre izvršavanja instrukcije kojoj je taj argument potreban.

U zavisnosti od primene i složenosti, određene aplikacije mogu napraviti kompromis tako što će funkcionalnosti instrukcijskog skupa biti ograničene u korist manje "cene" njihove implementacije, odnosno potrošnje memorije i električne energije. Takav kompromis je omogućen postojanjem „nanoMIPS Subset (NMS)" podskupa nanoMIPS instrukcijskom skupa. Tačnije, mikroprocesor može podržati samo određen podskup instrukcijog skupa (NMS), zadržavajući u potpunosti mogućnosti celokupnog instrukcijskog skupa. Ušteda memorije i električne energije se postiže jednostavnijim hardverom koji je neophodan instrukacijama koje čine podskup NMS. [3]

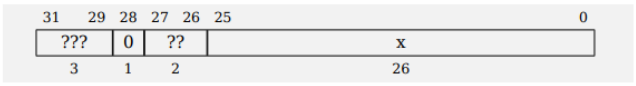
Uz sami instrukcijski set nanoMIPS, razvijan je i novi ABI sa oznakom „p32" kao i konvencija pozivanja potprograma. P32 ABI definiše skup registara od 32 registara opšte namene (GP registri) širine 32 bita ($r0-$r31) i 32 registara sa pokretnim zarezom (FP registri) širine 64 bita ($f0-$f31). Po osam registara od svake grupe se koristi za prenos argumenata potprograma ($r4-$r11, $f0-$f7), dok je kod o32 ABI-ja upola manji. Ostale karakteristike p32 ABI-ja, kao i poređenje sa najzastupljenijim MIPS32 ABI-jem (o32) prikazane su u tabeli 1.

	o32	n32
Korišćen kompajler	gcc/llvm	gcc
Model podataka	ILP32	ILP32
Konvencija poziva potprograma	o32	new
Broj(širina) GP registara	32(32bita)	32(32bita)
Broj GP registara za prosleđivanje argumenata	4($4..$7)	8($r4..$r11)
Broj(širina) FP registara	16/32 (32/64bita)	32(64bita)
Broj FP registara za prosleđivanje argumenata	4	8
Podržani instrukcijski skupovi	mips 1/2	mips 3/4
Poravnjanje steka	8 bajta	16 bajta

Tabela 1 - Poređenje o32 i p32 ABI

Celokupan instrukcijski skup nanoMIPS arhitekture je hijerarhijski podeljen na akumulacije instrukcija (engl. *instruction pools*) na osnovu dužine, odnosno formata instrukcije. Na slici 2 je prikazana akumulacija instrukcija P32 koji sadrži sve 32-bitne instrukcije (kao i akumulacija instrukcija P48I, koji sadrši 48-bitne instrukcije) koji se kasnije dalje dele na osvonu sekvenci bita [31..29,27..26], dok je za sve instrkcije zajednička vrednost bita 28 u instrukcijskoj reči (vrednost nula).

B.2 P32 pool



	???00	???01	???10	???11
000??	P.ADDIU	ADDIUPC[32]	MOVE.BALC	*
001??	P32A	*	P.BAL	*
010??	P.GP.W	P.GP.BH	PJ	*
011??	P48I	*	*	*
100??	P.U12	P.LS.U12	P.BR1	*
101??	*(CP1)	P.LS.S9	P.BR2	*
110??	*(MIPS64)	*	P.BRI	*
111??	P.LUI	*	*	*

Slika 2 - Primer akumulacije instrukcija

VI. IMPLEMENTACIJA PODRŠKE VELGRINDU ZA ARHITEKTURU NANOMIPS

S obzirom da je nanoMIPS arhitekura dizajnirana i razvijana "od nule", podrška u Velgrindu je implementirana shodno tome, iako ima dosta sličnosti sa prethodnim arhitekturama MIPS. Implementaciju podrške Velgrindu za nanoMIPS arhitekturu možemo podeliti na nekoliko najbitnijih celina:

- podrška za novi sistemski poziv u Velgrindu - *statx*
- podrška za prevođenje Velgrinda za nanoMIPS
- podrška za disasembliranje nanoMIPS mašinskog koda u (faza 1 rada Velgrinda)
- podrška za asembliranje Velgrindovog IR koda u nanoMIPS mašinski kod (faza 8 rada Velgrinda)

Sistemski poziv *statx* predstavlja noviju verziju sistemskih poziva *stat*, *fstat* i *lstat* (sistemski pozivi za dobijanje informacija o fajlu). Sastavni je deo Linuks operativnog sistema od verzije kernela 4.11, a podržan od strane biblioteka od glibc verzije 2.28.

S obzirom da u listi podržanih sistemskih poziva za arhitekturu nanoMIPS postoji *statx*, sastavni deo podrške za novu arhitekturu jeste i podrška ovog sistemskog poziva u Velgrindu. Implementacija podrške je realizovana emulacijom starijih verzija sistemskih poziva pomoću *statx* sistemskog poziva. Ukoliko sistemski poziv *statx* nije podržan na verziji kernela mašine koja pokreće Velgrind, sistemski poziv *statx* prijavi grešku o nepostojanju sitemskog poziva (*ENOSYS*), pa Velgrind poziva starije verzije ovog sistemskog poziva. Primer prethodno opisanog scenarija, odnosno emulacija sistemskog poziva *stat* pomoću sistemskog poziva *statx* je prikazana na sledećem listingu:

```
SysRes VG_(stat) ( const HChar* file_name, struct vg_stat* vgbuf )
{
   SysRes res;
   VG_(memset)(vgbuf, 0, sizeof(*vgbuf));

+#  if defined(VGO_linux)
+   /* On Linux, first try with statx. If that doesn't work out,
+      fall back to the stat64 or vanilla version. */
+   { struct vki_statx buf;
+     res = VG_(do_syscall5)(__NR_statx, VKI_AT_FDCWD, (UWord)file_name,
+                            0, VKI_STATX_ALL, (UWord)&buf);
+     if (!(sr_isError(res) && sr_Err(res) == VKI_ENOSYS)) {
+        /* Success, or any failure except ENOSYS */
+        if (!sr_isError(res))
+           TRANSLATE_statx_TO_vg_stat(vgbuf, &buf);
+        return res;
+     }
+   }
+#  endif
...
```

Listing 1 - Emulacija sistemskog poziva stat pomoću sistemskog poziv statx

Na osnovu vrednosti atributa --host(nanomipseb-linux-gnu ili nanomips-linux-gnu) prosleđenog prilikom /configure naredbe i izmena u make i configure fajlovima postavljamo određene CPP flegove za primarnu i sekundarnu platformu.

Pre pokretanja Velgrinda, poziva se funkcija sa potpisom static const char* select_platform(const char *clientname) koja određuje platformu za koju treba da se pokrene Velgrind. Rezultat ove funkcije se vezuje za naziv alata koji se pokreće (vrednost atributa uz –tool=atribut) i pokreće se aplikacija na putanji naziv_alata/naziv_alata-platforma (primer none/none-nanomips-linux). Detekcija nanoMIPS platofme prilikom pokretanja Velgrinda je prikazana na listingu 2.

```
static const char *select_platform(const char *clientname)
...
                    (header.ehdr32.e_flags & E_MIPS_ABI2)) {
                  platform = "mips64-linux";
               }
+              else
+              if (header.ehdr32.e_machine == EM_NANOMIPS &&
+                  (header.ehdr32.e_ident[EI_OSABI] == ELFOSABI_SYSV ||
+                   header.ehdr32.e_ident[EI_OSABI] == ELFOSABI_LINUX)) {
+                 platform = "nanomips-linux";
+              }
           }
```

Listing 2 - Detekcija nanoMIPS platforme prilikom pokretanja Velgrinda

Guest deo Velgrinda predstavlja spregu klijentskog koda (u ovom slučaju mašinski kod nanoMIPS aplikacije) i Velgrindovog IR koda, odnosno vrši disasembliranje mašinskog koda konkretne arhitekture u IR kod. Prilikom učitavanja mašinskog koda, Velgrind poziva arhitekturalno zavisnu funkciju na koju pokazuje *disInstrFn* – *disInstr_nanoMIPS*. Zahvaljujući hijerarhijskoj podeli instrukcijskog skupa nanoMIPS u akumulacija instrukcija, disasembliranje je implementirano shodno podeli instrukcija u instrukcijskom setu – jedna funkcija obrađuje jednu akumulaciju instrukcija. Na sledećem listingu je prikazan prvi nivo disasembliranja akumulacije instrukcija P.ADDIU, koji predstavlja podskup akumulacije P32 (prikazan na slici 2). Deo funkcije koji je prikazan u zavisnosti od dela mašinskog koda instrukcije *rt* vrši disasembliranje instrukcije (ADDIU rt, rs, u), odnosno dalje disasembliranje akumulacije instrukcija P.RI.

```
+static void dis_nanoMIPS32(DisResult *dres, UInt cins)
+{
+
+   switch (cins >> 26) {
+      case P_ADDIURI: {
+         UChar rt = (cins >> 21) & 0x1F;
+
+         if (rt != 0) {
+            UChar rs = (cins >> 16) & 0x1F;
+            UInt u = cins & 0xFFFF;
+            DIP("addiu r%u, r%u, %u", rt, rs, u);
+            putIReg(rt, binop(Iop_Add32, getIReg(rs), mkU32(u)));
+         } else {
+            nano_pri(dres, cins);
+         }
+
+         break;
+      }
....
```

Listing 3- Deo funkcije za disasembliranje mašinskog koda

Nakon disasembliranja, Velgrind izvršava optimizacije nad kreiranim IR kodom, instrumentalizaciju, gradnju stabla i odabir registara. Poslednja faza rada Velgrinda predstavlja *host* deo Velgrinda, odnosno asembliranje Velgrindovog IR koda u kod ciljne arhitekture. Poput izvrašavanja disasembliranja, i asembliranje je arhitekturalno zavisno i izvršava se pozivanjem funkcija na koju pokazuju *iselSB* i *emit* - *iselSB_NANOMIPS* i *emit_NANOMIPSInstr*. Prva od dve funkcije prevodi IR stablo u linearnu strukturu instrukcija koje treba generisati, dok druga "emituje" nanoMIPS instrukcije, odnosno generiše mašinski kod odabranih instrukcija. Funkcija *iselSB* (*iselStatementsBlocks*) prolazi

kroz iskaze u bazičnom bloku pozivajući funkciju *iselStmt* koja transformiše IR instrukciju u ekvivalentnu nanoMIPS instrukciju čiji mašinski kod kasnije generišemo. Deo funkcije je prikazan na slici 6. Funkcija koja generiše mašinski kod nanoMIPS instrukcija (*emit*) koristi informacije koje Velgrind postavlja tokom izvršavanja *isel* funkcije, formira mašinski kod i „emituje“ ga u izlazni bafer.

```
+static void iselStmt(ISelEnv * env, IRStmt * stmt)
+{
+   if (vex_traceflags & VEX_TRACE_VCODE) {
+      vex_printf("\n-- ");
+      ppIRStmt(stmt);
+      vex_printf("\n");
+   }
+
+   switch (stmt->tag) {
+      case Ist_Store: {
+         IRType tyd = typeOfIRExpr(env->type_env, stmt->Ist.Store.data);
+         HReg r_addr = iselWordExpr_R(env, stmt->Ist.Store.addr);
+
+         if (tyd == Ity_I8 || tyd == Ity_I16 || tyd == Ity_I32) {
+            HReg r_src = iselWordExpr_R(env, stmt->Ist.Store.data);
+            addInstr(env, NANOMIPSInstr_Store(sizeofIRType(tyd),
+                                              r_addr, 0, r_src));
+            return;
+         } else if (tyd == Ity_I64) {
+            HReg vHi, vLo;
+            iselInt64Expr(&vHi, &vLo, env, stmt->Ist.Store.data);
+            addInstr(env, NANOMIPSInstr_Store(4, r_addr, 0, vLo));
+            addInstr(env, NANOMIPSInstr_Store(4, r_addr, 4, vHi));
+            return;
+         }
+
+         break;
+      }
...
```

Listing 4 - Deo funkcije za transformaciju IR instrukcije u nanoMIPS instrukciju

VII. TESTIRANJE REŠENJA

Za potrebe testiranja je napisano nekoliko testova za nanoMIPS arhitekturu koja testira validnost dodatih delova disasembliranja i asembliranja prilikom rada Velgrinda. Ovim testovima je pokriven celokupan instrukcijski set nanoMIPS arhitekture. Testiranje je vršeno pomoću QEMU emulatora za nanoMIPS, usled neposedovanja odgovarajućeg hardvera. Trenutnu verziju rešenja podrške treba koristiti sa oprezom, s obzirom da je nepotpuna i nedovoljno testirana sa postojećim skupom Velgrind testova.

VIII. ZAKLJUČAK

U ovom radu opisan je ukratko način rada alata Velgrind,alata kao i nova arhitekura nanoMIPS. Prikazan je način implementacije podrške Velgrind alatu za novu arhitekturu (u ovom slučaju nanoMIPS), što predstavlja značajnu podršku korisniku prilikom testiranja aplikacija. Predstavljeno rešenje je potrebno dodatno testirati koristeći postojeći Velgrindov skup testova za verifikaciju. Cilj testiranja jeste uspešno izvršavanje celokupnog skupa pomenutih testova, te se tada implementacija može nazvati kompletnom.

ZAHVALNICA

Zahvaljujem se Petru Jovanoviću, šefu MIPS grupe, kao i celom Velgrind timu (Aleksandru Rikalu i Aleksandri Karadžić). Takođe, zahvaljujem se i naučno-istraživačkom institutu „RT-RK“ na pruženoj šansi da ovaj rad nastane kao proizvod višemesečnog rada na realnom problemu iz oblasti kojom se rad bavi.

LITERATURA

[1] Sweetman, Dominic. See MIPS Run. 2nd ed. San Francisco, Calif.: Morgan Kaufmann Publishers/Elsevier, 2007

[2] Nicholas Nethercote , Julian Seward, Valgrind: a framework for heavyweight dynamic binary instrumentation, Proceedings of the 2007 ACM SIGPLAN conference on Programming language design and implementation, June 10-13, 2007, San Diego, California, USA

[3] MIPS® Architecture Base: nanoMIPS32™ Instruction Set Technical Reference Manual, https://s3-eu-west-1.amazonaws.com/downloads-mips/I7200/I7200+product+launch/MIPS_nanomips32_ISA_TRM_01_01_MD01247.pdf (21.04.2019.)

[4] Zoran Jovanović „Instrukcijski nivo paralelizma“, http://rti.etf.bg.ac.rs/rti/ir4par/materijali/pdf/I%20Uvod.pdf (21.04.2019.)

ABSTRACT

Introducing of the new architectures of microprocessor systems in addition to a hardware chip implies certain software tools, such as compilers and emulators. The use of new architecture, the robust creation of complex applications for the new architecture, raises the need for software tools that would facilitate the detection of improper program performance and easier detection of errors, such as debugger and profilers. This paper describes the changes that are needed to add support to one software program analysis tool, Valgrind, for the new architecture of the MIPS Technologies Group - nanoMIPS.

Keywords-Valgrind, MIPS, nanoMIPS

Adding support for nanoMIPS architecture into Valgrind tool for dynamically analysis of binary code

Dimitrije Nikolić, Aleksandra Karadžić, Aleksandar Rikalo, Petar Jovanović

Unapređenje programskog prevodioca za jezik P4 sa podrškom za čitanje međukoda u formatu JSON

Jelena Vidaković, Enisa Hadžić, Miodrag Dinić, Dragan Samardžija

***Apstrakt*— P4 je programski jezik visokog nivoa koji je dizajniran da omogući programiranje prosleđivanja mrežnih paketa nezavisno od vrste protokola. P4 prevodilac je otvorenog koda koji održava neprofitna organizacija pod nazivom 'P4 Jezička Zajednica'. Jezik je prvobitno opisan u dokumentu 'SIGCOMM CCR' iz 2014. godine pod nazivom 'Programming Protocol-Independent Packet Processors' – odakle seže skraćenica P4. On radi zajedno sa kontrolnim protokolima softverski definisanog umrežavanja (SDN), kao što je '*OpenFlow*'. Koristimo apstraktni model prosleđivanja za definisanje jezika kako bi se opisao način na koji će se konfigurisati *switch*-evi i kako se paketi obrađuju.**

***Ključne reči*—P4, P4 prevodilac, simple_switch, BMv2, eBPF**

I. UVOD

S obzirom da je jezik namenjen za aplikacije čiji osnovni cilj jeste prosleđivanje paketa, jezik je dizajniran kako bi ispunio posebne ciljeve koji bi to omogućili. Prvi princip je 'Target independence' što predstavlja dizajniranje programa koji su nezavisni od implementacije, odnosno od uređaja na kome će se izvršavati. Ideja je da prevodilac uzima u obzir mogućnosti *switch*-a, kada pretvara program napisan u p4 jeziku u rezultat koji je zavistan od hardvera i koji se koristi za konfigurisanje *switch*-a. Drugi princip je nezavisnost od vrste protokola. *Switch* ne bi trebalo da bude usko vezan za određeni format paketa, već bi trebalo da postoji mogućnost da se specificira način parsiranja paketa za odvajanje polja zaglavlja i specificiranje tipova i grupa akcija i tabela koji obrađuju ova zaglavlja. Poslednji princip pruža mogućnost da se redefinišu već sačuvana pravila parsiranja paketa i procesiranja pojedinačnih polja.

Cilj ovog rada je upoznavanje sa P4 jezikom i radom P4 prevodioca, strukturama koje formira i načinom na koji čuva stanje programa u njima. Opisana je implementacija prolaza i metoda koje prevodilac poziva, odnosno sama arhitektura i organizacija. Nakon upoznavanja sa osnovnim principima uvodi se u detaljan opis unapređenja prevodioca u vidu podrške za čitanje međukoda u JSON formatu.

Jelena Vidaković, Naučno istraživački institut RT-RK, Bulevar Milutina Milankovića 19a, 11000 Beograd, Srbija (telefon: 381-21-483-1389, e-mail: jelena.vidakovic@rt-rk.com)

Enisa Hadžić, Naučno istraživački institut RT-RK, Bulevar Milutina Milankovića 19a, 11000 Beograd, Srbija (telefon 381-21-483-1491, e-mail: enisa.hadzic@rt-rk.com).

Miodrag Dinić, Naučno istraživački institut RT-RK, Bulevar Milutina Milankovića 19a, 11000 Beograd, Srbija (telefon: 381-21-483-1489, e-mail: miodrag.dinic@rt-rk.com).

Dragan Samardžija , Naučno istraživački institut RT-RK, Narodnog fronta 23a 21000 Novi Sad, Srbija (telefon: 381-21-480-1193, e-mail: dragan.samardzija@rt-rk.com).

II. UVOD U P4 TOK PODATAKA

[1]Pakete koji pristižu prvo obrađuje parser, za telo paketa se pretpostavlja da je skladišteno zasebno pa zbog toga nije dostupno za podudaranje bez prethodnog procesiranja. Parser prepoznaje i izdvaja polja zaglavlja i na taj način definiše protokole koje treba da podržava *switch*. Sam model ne pretpostavlja tip protokola, već mu je važno da postoji već isparsiran prikaz polja zaglavlja nad kojima rade tabele podudaranja i akcija.

Polja zaglavlja, koja su prethodno izdvojena iz istog, prosleđuju se gore pomenutim tabelama. Tabele podudaranja i akcija su podeljene izmedju ulaznog i izlaznog kontrolnog bloka. Dok obe kontrole mogu da modifikuju zaglavlje paketa, ulaz postavlja izlazni port i određuje u koji red će biti smešten paket. Na osnovu ulaznog procesiranja, paket može biti prosleđen, ponovljen, odbačen ili da predstavlja okidač kontrole protoka. Izlaz može izvršavati modifikacije u zaglavlju paketa za svaku instancu, kao što su kopije za višestruko slanje.

Paket može sa sobom da nosi dodatne informacije prilikom prolaska kroz različite faze, ti podaci se čuvaju u strukturi koja se naziva 'metapodatak' i koja se obrađuje na isti način kao i polja zaglavlja paketa. Neki primeri metapodataka uključuju izlazni port, odredište za slanje i red čekanja, kao i vremensku oznaku koja se koristi prilikom raspoređivanja paketa.

III. P4 KONCEPTI

Osnovne jezičke komponente P4 jezika sa svojim definicijama su [2]:

- *Zaglavlja* - predstavljaju sekvencu ili strukturu sačinjenu od niza polja, u kojima se čuva informacija o dužini polja i ograničenjima koja važe za ta polja.
- *Parseri* - specificiraju način na koji se prepoznaju i validaraju zaglavlja i polja unutar njih. Parser može imati više definisanih stanja. Verifikaciju započinje unutar stanja '*start*' i nastavlja dok ne naiđe na stanje '*stop*' ili stanje '*greške*'. Kada uđe u naredno stanje, na osnovu vrednosti određenih polja zadaje se stanje tranzicije[3].
- *Tabele podudaranja i akcija* - predstavljaju mehanizam

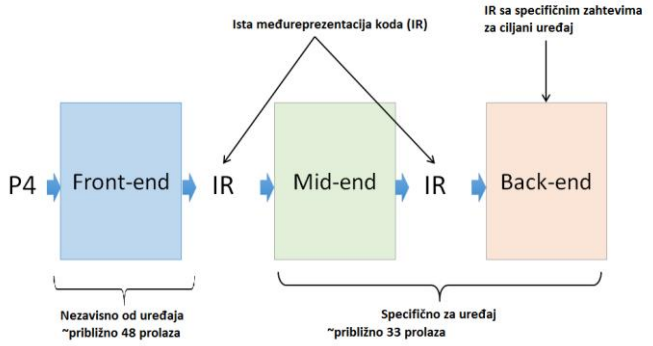
obrade paketa, a sam P4 program definiše polja nad kojima se ta pravila izvršavaju.

- *Akcije* - predstavljaju manipulaciju podacima koji se nalaze u poljima paketa i sadržinom metapodataka. U kontekstu P4 jezika metapodaci su informacije o paketu koje nisu dirketno rezultat parsiranja. P4 definiše metapodatke koji svi uređaji moraju da podrže, a postoji i opcija dodavanja novih korisnički definisanih za specifične uređaje.
- *Kontrole* - određuju redosled poziva akcija i tabela koje se izvršavaju nad podacima i regulišu tok kontrole.

IV. PREVOĐENJE P4 PROGRAMA

Prevodilac se koristi za sintaksnu i leksičku analizu programa, identifikaciju zavisnosti i traženje mogućnosti za obradu zaglavlja. On prevodi P4 program u određenu među reprezentaciju koda koju čuva u grafu zavisnosti koji dalje analizira kako bi utvrdio njihovu validnost i generiše kod za njihovo izvršavanje.

Prevodilac je otvorenog koda i sadrži lako proširivu arhitekturu, kako bi bilo omogućeno jednostavno dodavanje novih prolaza i optimizacija. Projekat prevodioca je implementiran u programskom jeziku $C++_{11}$. Modularna arhitektura projekta omogućava lako dodavanje novih funkcionalnosti i održavanje obe trenutno podržane verzije standarda jezika $P4_{14}$ i $P4_{16}$. Podrška za ove standarde je podeljena u 2 odvojena modula za prvu fazu procesiranja (eng. 'front-end'), pretvarač $P4_{14}$ u $P4_{16}$, narednu fazu optimizacije (eng. 'mid-end') i tri 'back-end' prototipa (sa skraćenicama eBPF, bmv2 i p4test). EBPF 'back-end' generiše .c kod , BMV2 generiše .json izlazni fajl koji se koristi za pokretanje 'simple-switch' mrežnog simulatora i poslednji 'p4-test' koji predstavlja 'lažni back-end', odnosno koristi se za testiranje, debagovanje i bolje razumevanje rada prevodioca.



Sl. 1 : Struktura prevodioca

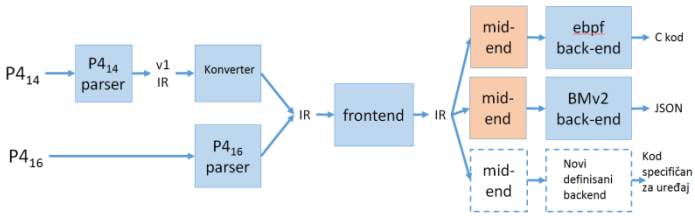
[4]Struktura prevodioca je grafički prikaza na slici 1, i sastoji se od tri osnovne faze pod nazivima Front-end, Mid-end i Back-end.

Front-end je u potpunosti nezavistan od arhitekture, unutar sebe poziva dodatne prolaze (ima ih približno 48) koji vrše validaciju programa, proveru tipova i dodatna promatranja:

- Parsiranje programa
- Validacija
- Razrešavanje imena
- Provera tipova
- Rešavanje bočnih efekata
- Optimizacija
- Razrešavanje 'Inline' funkcija
- Konvertovanje u P4 izvorni oblik

Mid-end je drugačiji za svaki uređaj, pokreće dodatne prolaze koji vrše optimizaciju, predikciju koda i eliminaciju suvišnih tipova. Poslednja faza prevođenja je deo modula pod nazivom Back-end, on je u potpunosti zavisan od uređaja, vrši alokaciju resursa i generisanje koda. Unutar prevodioca postoji jasna podela među prolazima odakle seže i svojstvo proširivosti, tj lakog dodavanja novih prolaza ili čak čitavog 'back-end'-a zavisnog od uređaja.

Tok prevođenja izvornog koda P4 programa može se grafički predstaviti slikom br.2.



Sl. 2 : Protok podataka

$P4_{14}$ izvorni kod prolazi kroz $P4_{14}$ parser, zatim kroz pretvarač kako bi se dobio isti međukod, koji daje $P4_{16}$ parser. Međureprezentacija koda prolazi kroz 'front-end' i zajednička je za svaki uređaj kome će biti prosleđena. 'Mid-end' i 'Back-end' su karakteristični za različite hardvere i za svaki je definisan drugačiji izlazni format koji generiše. Ovakvom organizacijom prevodioca omogućeno je dodavanje podrške za novi ciljani uređaj.

Na početku izvršavanja prevođenja poziva se parser, koji je implementiran koristeći generator leksičkih analizatora *Flex* i parser generator *Bison*. Nakon parsiranja fajla, pravi se instanca Front-End-a koji izvršava provere i čuva stanje programa, a nakon te faze i faza Mid-End u kojoj se nad datim programom izvršavaju i dodatne optimizacije. Najvažniji prolazi koji se moraju pozivati svaki put kada se promeni nešto u program su 'ResolveReferences' i 'TypeInference'. 'Evaluator' se poziva posle celokupne Front-End i Mid-End faze i koji formira hijerarhiju statički alociranih resursa. Navedeni 'ResolveReferences' prolaz popunjava strukturu 'ReferenceMap', pritom mapira i povezuje svako pojavljivanje promenljive sa njenom deklaracijom. Naredni 'TypeChecking' prolaz vrši proveru tipova, za njegovo izvršavanje je neophodna popunjena 'ReferenceMap'-a, a kao rezultat dobija se popunjena struktura 'TypeMap' koja je takođe neophodna za dalje prevođenje programa. 'TypeMap' predstavlja mapu, koja se popunjava za svaki čvor koji poseduje informaciju o izvornom tipu.

P4 prevodilac u sebi sadrži podršku za četiri različita 'back-end' prototipa. Jedan od njih je 'p4test', njegovo izvršavanje je znatno redukovano i služi samo za testiranje prva dva

prolaza prevođenja, tj Front-end i Mid-end prolaza. ‘P4test’ prevodi fajl i čuva P4 reprezentaciju programa nakon izvršavanja pomenutih faza i na taj način testira njihovu validnost. Implementacija ovog prototipa sadrži podršku za p4 programe napisane u verziji 14 i 16. Podržan je i ‘p4c-ebpf’, rezultat njegovog izvršavanja je .c kod, koji dalje može biti preveden u eBPF. BPF (eng. ‘The Berkeley Packet Filter’) podržava filtriranje paketa, dozvoljavajući korisničkom procesu da obezbedi filter koji specificira koje pakete želi da primi. ‘P4c-bm2’ u sebi sadrži dva ‘back-end’-a, ‘p4c-bm2-ss’ i ‘p4c-bm2-psa’. ‘Psa’ prototip nije u potpunosti realizovan, tj za njega se smatra da je u fazi implementacije. Celokupan ‘p4c-bm2 back-end’ kao izlaz prevođenja čuva stanje u .json formatu koji predstavlja ulazni fajl za ciljani uređaj ‘Behavioral Model version 2’, odnosno softverski switch nazvan ‘simple_switch’. ‘P4c-bm2-psa’ ima podršku za sve p4 programe koji su napisani za ‘psa.p4’ model u verziji 14 ili 16. PSA model generiše kod za arhitekturu prenosnog prekidača (eng. ‘The Portable Switch Architecture’) koja opisuje mogućnosti uređaja mrežnog prekidača koji obrađuje i prosleđuje pakete na više portova interfejsa. Drugi ‘p4c-bm2-ss’ takođe generiše kod za pomenuti mrežni simulator i prevodi $p4_{14}$ i $p4_{16}$ programe napisane koristeći ‘v1model.p4’. Navedeni JSON format je format otvorenog standarda, nezavistan od programskog jezika, koji koristi zapis čitljiv čoveku i kome se prenose različiti objekti podataka.

V. Unapređenje prevodioca

Realizacija rešenja opisanog u ovom radu proizvod je težnje da se arhitektura p4 prevodioca dodatno unapredi tako da se platformski nezavisan međukod može iznova koristiti za različite platforme, bez potrebe da se izvorni P4 program podvrgne svim fazama prevođenja. Do sada je u prevodiocu postojala mogućnost da se nakon Mid-End prolaza sačuva međureprezentacija koda. Prosleđivanjem argumenata komandne linije ‘--toJSON’ i ‘fajl.json’ , prevodilac će sačuvati međukod programa u navedenu .json datoteku.

Radi boljeg razumevanja rešenja i unapređenja opisanih u daljem radu, u nastavku su navedene komande prevodioca koje pokreću njegovo izvršavanje. Kao primer, navedene su komande za *bmv2/simple_switch* ‘back-end’.

(1) p4c-bm2-ss -o izlazniFajl.json izvorniFajl.p4 --toJSON fajl.json

(2) p4c-bm2-ss -o izlaz.json --fromJSON fajl.json

Grafički prikaz toka prevođenja uz unapređenje dat je na slici 3.

Sl. 3 : Struktura prevođenja

Prilikom generisanja sadržaja programa, podrazumevano ponašanje je da se informacije o izvornom kodu svakog čvora ne čuvaju u međureprezentaciji. Implementacija je proširena podrškom da se kao deo svakog čvora sada čuvaju i te informacije o izvornom kodu kako bi se dobili svi neophodni podaci za dalje rekonstruisanje.

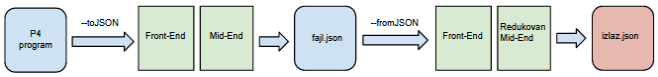
Prvi korak je dodavanje podrške za prepoznavanje novog argumenta komandne linije, ‘--fromJSON’ ‘fajl.json’. Nova opcija komandne linije prevodioca zaštićena je od nenamernih grešaka korisnika i testirana u kombinaciji sa svim ostalim do sada podržanim opcijama. Ukoliko je umesto očekivanog formata nakon *flag*-a(--fromJSON) naveden neki drugi, npr .p4 format, obezbeđeno je da prevodilac prijavi grešku. Ova podrška dodata je za sva četri delimično ili u potpunosti implementirana ‘back-end’-a. Prilikom kreiranja bilo kojeg od objekata koji predstavljaju strukturu u kojoj se čuva stanje svih navedenih opcija koje prevodilac treba da podrži i koji su zasebni za svaki ‘back-end’, definiše se novi ‘registerOption(--fromJSON, file)’, grafički prikaz na slici 4.

```
registerOption("--fromJSON", "file",
       [this](const char* arg) { loadIRFromJson = true; file = arg; return true; },
       "Use IR representation from JsonFile dumped previously,"\
       "the compilation starts with reduced midEnd.");
```

Sl. 4 : Dodavanje opcije komandne linije

Za rekonstruisanje stanja programa (unutar klase P4Program) zadužen je poseban modul (klasa JSONLoader) koji to čini pomoću informacija iz datoteke koja čuva međureprezentaciju u formatu JSON. Bilo je neophodno dodati podršku za kreiranje JsonObject-a (strukture u kojoj se čuva i formira jedan pročitan objekat) popunjavanjem njegovog polja SourceInfo (odnosno strukture koja nosi informaciju o izvornom kodu). Implementacija je postignuta proširenjem klasa novim konstruktorima, metodama i poljima koji čuvaju novonastalo stanje.

Prikaz faza prevođenja i toka dat je na slici 5.



Sl. 5 : Struktura prevođenja (sa i bez ‘--fromJSON’)

Ukoliko se komandom prevođenja navede ‘--fromJSON’ a prethodno je u nekom fajlu sačuvana međureprezentacija, umesto ponovnog parsiranja, čitaju se već optimizovani podaci i postupak prevođenja se nastavlja u znatno redukovanom obimu. Pošto su podaci izgenerisani nakon Mid-End prolaza, sve optimizacije nad čvorovima programa su već izvršene, pa ih nije potrebno ponovo pozivati. Prevođenje započinje čitanjem iz fajla i rekonstruisanjem stanja, a nakon toga poziva se redukovan Mid-End i Back-End prolaz. Mid-End je neophodno ponovo pokrenuti jer se strukture koje se popunjavaju u jednom delu njegovog izvršavanja prosleđuju kao argumenti Back-End-a. Redukovan prolaz čine ‘ResolveReferences’ koji će popuniti strukturu ‘refMap’ , ‘TypeChecking’ koji popunjava strukturu ‘TypeMap’ , ‘VisitFunctor’ koji poziva ‘evaluator’ prolaz i

popunjava strukturu ‘IR::ToplevelBlock’. Jedan od poziva unutar ‘Back-end’-a generiše se na osnovu strukture, koja se popunjava iz ulaznog fajla prilikom parsiranja. Pošto se ovim putem samo parsiranje izbegava, onda i ta struktura ostaje nepopunjena i time svako čitanje podataka iz nje postaje nevalidno. Stoga se ovaj prolaz (sa nazivom ‘ParseAnnotation’) dodaje prevođenju, samo ukoliko flag ‘--fromJSON’ nije prosleđen kao argument komandne linije.

Ispisivanju sadržaja u izlazni .json format, prethodile su izmene prilikom kreiranja polja JsonObject-a. Zbog izmenjenog ulaznog fajla i načina čitanja, time i drugačijeg načina formiranja pojedinih objekta, metoda ‘prepareSourceInfoForJSON’ , čiji zadatak je da modifikuje polja objekta u čitljiviji zapis, neće ispisivati validno vrednosti sačuvane unutar navedenog SourceInfo polja. Za rešenje ovog problema primenjena je simulacija kreiranja novog JsonObject-a, kome su kao parametri prosleđeni novosačuvani podaci. Nakon toga omogućeno je regularno ispisivanje ovog objekta u izlazni .json fajl.

Prvobitno rešenje sa redukovanim Mid-End prolazom u izlaznom fajlu nije sadržalo informacije o korisnički definisanim enum tipovima. Unutar koda prevodioca čuva se struktura ‘enumMap’ koja se popunjava prilikom izvršavanja prolaza ‘convertEnumPass’. Ovaj prolaz je sastavni deo Mid-End dela i njegov zadatak je da koristeći korisnički definisane smernice konvertuje enum tipove u tip bit<n>. Konverzija se vrši samo nad korisnički definisanim, odnosno ne odvija se nad enum tipovima koji su deo specifikacije arhitekture. Nakon obrađivanja čvora koji je tipa ‘Type_Enum’ vrši se popunjavanje navedene strukture ‘enumMap’ i njegovo uklanjanje iz vektora čvorova. Kada se nakon Mid-End-a izgeneriše međureprezentacija koda unutar nekog fajla, ‘convertEnumPass’ prolaz je već izvršio optimizacije i time uklonio sadržaj enum tipova iz programa. Nakon ponovnog prevođenja opcijom ‘--fromJSON jsonFile.json’, nedostajaće podaci o ovim tipovima koje je korisnik definisao prilikom pisanja .p4 programa. Rešenje koje je ponuđeno za prevazilaženje ovog problema je proširenje redukovanog Mid-End-a. Izmena koja je predstojila ovom proširenju, odnosno uvođenju novog prolaza, je čuvanje svih čvorova u programu nakon njihove obrade. Kada je to postignuto, zagarantovano je i validno čuvanje čitavog sadržaja programa u međureprezentaciji, a time i validno čitanje iz novog ulaznog fajla. ‘FillEnumMap’ je novi prolaz koji će prilikom obrade popuniti strukutru ‘enumMap’ i time obezbediti da se ispisivanje sadržaja u izlazni .json fajl obavi nepromenjeno.

Grafički prikaz redukovanog i finalnog Mid-End prolaza je dat na slici 6[5].

```
auto fillEnumMap = new P4::FillEnumMap(new PsaEnumOn32Bits("psa.p4"), &typeMap);
addPasses({
    new P4::ResolveReferences(&refMap),
    new P4::TypeChecking(&refMap, &typeMap),
    fillEnumMap,
    new VisitFunctor([this, fillEnumMap]() { enumMap = fillEnumMap->repr; }),
    evaluator,
    new VisitFunctor([this, evaluator]() { toplevel = evaluator->getToplevelBlock(); }),
});
```

Sl. 6: Redukovan Mid-End

VI. Testiranje rešenja

Unutar koda prevodica nalaze se testovi koji pokreću prevođenje i njihovu simulaciju. Komandom ‘*make check*’ moguće je pokrenuti njihovo izvršavanje i na standardnom izlazu ispratiti status i procenat uspešnosti. Nad unapređenjem navedenim u ovom radu, pokrenutu su svi testovi koji su podržani aktuelnom verzijom kompjalera u trenutku podizanja izmena, sa procentom uspešnosti od 100%.

Konkretno, sama realizacija rešenja je testirana poređenjem izlaznih fajlova koje svaki ‘back-end’ ponaosob izgeneriše (.json, .c, .graphs). Izlazni fajl koji je izgenerisao neizmenjen kod kompajlera komandom (1) u sebi sadrži iste informacije kao i izlazni fajl koji je izgenerisan komandom (2), odnosno unapređenim prevodiocem. Smatrajući to korektnim uslovom ispravnosti rada kompajlera, realizacija ovog unapređenja se može smatrati validnom[6].

VII. Zaključak

U ovom radu dat je uvod u p4 jezik i p4 prevodilac. Ukratko je opisana struktura i organizacija koda, kao i tok izvršavanja prevođenja za različite ciljane uređaje. Prikazano je unapređenje u vidu dodavanja podrške za čitanje međukoda u JSON formatu, što predstavlja značajnu podršku korisnicima. Uz efikasnost i brže reprodukovanje prevođenja, kao rezultat je dobijena i mogućnost da se isti međukod prosleđuje različitim uređajima na izvršavanje i poređenje.

Zahvalnica

Zahvaljujem se Miodragu Diniću, Draganu Samardžiji, kao i P4dev timu, Enisi Hadžić, na saradnji i pruženom znanju. Takođe, zahvaljujem se i naučno-istraživačkom institutu „RT-RK“ na pruženoj šansi da ovaj rad nastane kao proizvod višemesečnog rada na realnom problemu iz oblasti kojom se rad bavi. Ovaj rad je delimično finansiran od strane Ministarstva za prosvetu, nauku i tehnološki razvoj Republike Srbije, na projektu broj: TR36029.

Literatura

[1] Pat Bosshart, Dan Daly , Glen Gibb , Martin Izzard , Nick McKeown, Jennifer Rexford, Cole Schlesinger, Dan Talayco, Amin Vahdat, George Varghese, David Walker, ‘P4: Programming Protocol-Independent Packet Processors’, 2014

[2] The P4 Language Consortium, ‘P416 Language Specification’, https://p4.org/p4-spec/docs/P4-16-v1.1.0-spec.pdf (2018-11-30)

[3] P4 jezik, https://p4.org/

[4] P4 prevodilac, https://github.com/p4lang/p4c

[5] C++ programski jezik, http://www.cplusplus.com

[6] ‘Google Test’, https://github.com/google/googletest/

Unapređenje jezika P4 izrazima assume i assert kao pomoć u formalnoj verifikaciji

Enisa Hadžić, Jelena Vidaković, Miodrag Dinić, Miroslav Popović

***Apstrakt*— P4 je jezik otvorenog koda koji je dizajniran tako da se koristi za programiranje mrežnih uređaja. Razvoj P4 jezika uneo je veću fleksibilnost u mrežno programiranje i omogućio da uređaj postane rekonfigurabilan i nezavisan od protokola. Osnovni tok podataka u P4 jeziku može da se podeli u 4 osnovne faze. Prva faza je parsiranje mrežnih paketa. Prilikom prijema paketa, polja paketa moraju biti procesirana i sačuvana u reprezentaciji koja je čitljiva narednim fazama. Druga i treća faza je primena podudaranja i akcija tabela na ulaznom i izlaznom toku kontrole. Njihovo izvršavanje predstavlja mehanizam obrade parsiranih podataka i manipulaciju informacija sadržanih unutar polja paketa. Kako bi paket mogao da se dalje prosledi ciljanom uređaju, neophodno je da prođe kroz poslednju fazu deparsiranja, koja će emitovati obrađeno stanje paketa. U ovom radu biće predstavljen P4 jezik i način rada prevodioca, kao i način unapređenja novim izrazima assert i assume.**

***Ključne reči*—P4, Bmv2, prevodilac, data-plane**

I. Uvod

P4 [1] predstavlja jezik koji je namenjen za izražavanje načina na koji će se paketi obrađivati od strane *data-plane* programibilnog elementa kao što su mrežne kartice i ruteri. Programibilnost načina prosleđivanja olakšava operaterima da brzo implementiraju nove protokole i razvijaju mrežne usluge. Koristeći P4 jezik, moguće je specificirati kako treba manipulisati sa zaglavljima pristiglih paketa od strane različitih prosleđivačkih uređaja u infrastrukturi. Uprkos pomenutoj fleksibilnosti, ova paradigma takođe povećava šanse za pojavljivanje kvara u mreži. Traženje razloga nepravilnog rada mrežnih sistema može biti dug i skup proces, jer moderni uređaji podržavaju desetine različitih formata paketa i protokola. Dosadašnji pristup testiranju je da se izvrši iscrpno poređenje ulaznih paketa i očekivanih izlaznih paketa.

Enisa Hadžić, Naučno istraživački institut RT-RK, Bulevar Milutina Milankovića 19a, 11000 Beograd, Srbija (telefon 381-21-483-1491, e-mail: enisa.hadzic@rt-rk.com).

Jelena Vidaković, Naučno istraživački institut RT-RK, Bulevar Milutina Milankovića 19a, 11000 Beograd, Srbija (telefon: 381-21-483-1389, e-mail: jelena.vidakovic@rt-rk.com)

Miodrag Dinić, Naučno istraživački institut RT-RK, Bulevar Milutina Milankovića 19a, 11000 Beograd, Srbija (telefon: 381-21-483-1489, e-mail: miodrag.dinic@rt-rk.com).

Miroslav Popović, Univerzitet u Novom Sadu, Fakultet tehničkih nauka, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (telefon 381-21-480-104, e-mail: miroslav.popovic@rt-rk.uns.ac.rs).

Kako sada programer opisuje funkcionalnost uređaja koristeći P4 jezik, dolazi se do ideje za uvođenje *assert* i *assume* izraza koji se koriste prilikom testiranja P4 programa.

II. Uvod u P4 jezik

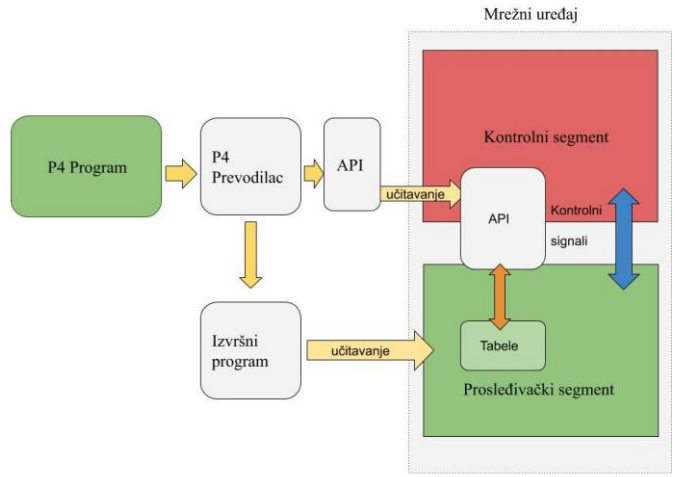
P4 je relativno jednostavan i statički tipiziran programski jezik, dizajniran da izrazi transformacije nad mrežnim paketima. Trenutna verzija $P4_{16}$ je 1.1.0, objavljena u novembru 2018. godine. Primarni ciljevi P4 jezika su da bude:

- Hardverski nezavisan
- Nezavisan od protokola
- Rekonfigurabilan

Osnovne apstrakcije koje pruža P4 jezik su:

- Zaglavlja (engl. *header*) – opisuju format svakog zaglavlja u mrežnom paketu
- Definisani meta podaci – skup podataka koji su povezani sa svakim paketom
- Parseri – opisuju dozvoljene sekvence zaglavlja unutar primljenog paketa i način na koji se identifikuju
- Akcije – fragmenti koda koji opisuju način na koji se manipuliše sa zaglavljima paketa
- Tabele – povezuju ključeve koje definiše korisnik sa odgovarajućom akcijom. Prilikom primenjivanja tabele izvšava se pretraživanje tabele sa ključem koji se dobije iz paketa i ako se desi podudaranje sa nekim od ulaza iz tabele koja se pretražuje, odgovarajuća akcija iz tog ulaza se izvršava
- Kontrolni blokovi – uključuju sekvencu pozivanja tabela
- Deparser – konstruiše izlazni paket od procesiranih zaglavlja ulaznog paketa

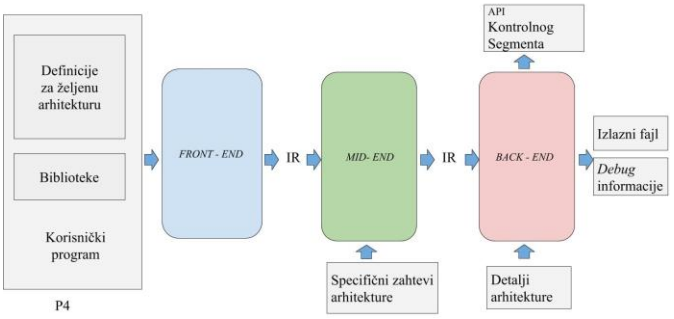
Na slici 1 nalazi se ilustrovani prikaz toka rada za programiranje mrežnog uređaja koristeći P4 jezik. Napisani P4 program prosleđuje se prevodiocu koji generiše izlaz u odgovarajućem formatu. Prevodilac ne samo da sintetiše funkcionalnosti prosleđvanja, već i API (skr. *application program interface)* između *control plane* i *data plane* jedinica.



Sl. 1. Programiranje mrežnog uređaja koristeći P4 jezik

III. ARHITEKTURA P4 PREVODIOCA

Da bi se implementirao P4 program, potreban je prevodilac za generisanje odgovarajuće konfiguracije za ciljni uređaj. P4 prevodilac [2] napisan je u C++ [3] programskom jeziku i dizajniran je tako da iz prosleđenog P4 programa generiše odgovarajuću reprezentaciju, izvršava više prolaza koji transformišu i optimizuju željeni program. Na samom početku rada prevodioca izvršava se sintaksna i leksička analiza P4 programa koji se prevodi. Parser koji se koristi generisan je od strane Bison [4] parser generatora i Flex [5] generatora leksičkih analizatora. Tok rada prevodioca posle sintaksne i leksičke analize može da se podeli na tri celine: *frontend*, *midend* i *backend*. Na slici 2 prikazan je tok rada prevodioca prilikom prevođenja ulaznog P4 programa.



Sl. 2. Faze prevođenja ulaznog P4 programa

Frontend prolaz vrši parsiranje izvornog programa i formira internu reprezentaciju (IR) koja odgovara tom izvornom kodu. Zatim se izvršava više prolaza koji vrše odgovarajuće provere koje P4 program mora da zadovolji kako bi bio semantički ispravan. Ukoliko u nekom od prolaza prevodilac utvrdi neregularnost programa koji se prevodi, biće prikazana odgovarajuća greška i proces prevođenja se završava neuspešno.

Za razliku od *frontend* prolaza, *midend* i *backend* se razlikuju za svaki ciljni uređaj za koji se vrši prevođenje. U *midend* prolazu, vrše se optimizacije interne reprezentacije koje obavljaju predikciju, uklanjaju nepotreban kod i pojednostavljuju izraze. Nakon *midend* prolaza, optimizovana interna reprezentacija programa se dodatno optimizuje u backendu i generiše se kod u zavisnosti od ciljnog *backend* prolaza. Trenutno u P4 prevodiocu postoje četiri različita *backend* prolaza:

- Bmv2 (*Behavioral model version 2* [6])
- Ebpf (*Extended Berkeley Packet Filter*)
- P4test
- Graphs

P4test služi za testiranje parsiranja, *frontend* i *midend* prolaza, dok *graphs* služi za generisanje grafičkog prikaza programa koji se prevodi. *Bmv2 backend* generiše izlazni fajl u json formatu, koji se koristi kao ulaz *Behavioral Model* simulatora. Simulator analizira json opis prevedenog P4 programa i konfiguriše tabele, parsere i akcije i formira primitive za svaku podržanu operaciju iz json fajla, koje se zatim izvršavaju i time simuliraju rad sistema. U json fajlu generiše se opis P4 programa u određenom formatu koji je čitljiv čoveku i omogućava prenošenje podataka u Bmv2 simulator. *Ebpf backend* kao izlaz generiše C program koji dalje može da se prevede u *ebpf* format, koristeći BCC ili CLANG, a koji se zatim izvršava pod Linux kernelom.

Svaki od pomenutih prolaza u P4 prevodiocu, realizovan je kao klasa koja je izvedena iz osnovnih klasa za definisanje novih prolaza – *Visitor* i *Transform* klase. Visitor je konstantni prolaz, koji samo obilazi čvor u IR stablu i sakuplja informacije o njemu, dok prolazi tipa *Transform* posećuju čvor i mogu da izvrše promene ili da ga zamene sa drugim čvorom. Ovi prolazi sadrže *preorder* i *postorder* metode za svaki tip IR čvora, koje mogu da se redefinišu. Ovako realizovana arhitektura P4 prevodioca omogućila je lako dodavanje novih prolaza i ciljnih backend-a.

IV. UNAPREĐENJE P4 JEZIKA

Realizacija unapređenja P4 jezika izrazima assert i assume nastala je kao potreba da se olakša način testiranja rada P4 programa. Tokom analiziranja ispravnosti programa, korisno je imati izraze:

assert(<boolean expression>)
assume(<boolean expression>)

Podrazumevano izvršavanje navedenih izraza je proveravanje prosleđenog uslova i ukoliko se ispostavi da je taj uslov netačan, odgovarajuća poruka o grešci se ispisuje i programer dobija informacije o fajlu, uslovu i broju linije zbog koje je došlo do greške. Razlika između ova dva izraza je u formalnoj verifikaciji. Dok se za uslov prosleđen assert izrazu nada da je tačan i proverava se da li je ikada netačan, za assume uslov programer čvrsto veruje da će prosleđeni uslov uvek biti tačan.

Prvi korak u realizaciji rešenja je dodavanje navedenih izraza tako da budu prepoznati od strane P4 prevodioca. Nove funkcije u P4 prevodiocu predstavljene su kao *extern* objekti, koji opisuju samo interfejs tog objekta, ali ne i njegovu implementaciju. Implementacija tih funkcija zavisi od ciljne arhitekture i generiše se u odgovarajućem *backend*-u.

Nalik na *assert* izraz u C++ programskom jeziku, željeno ponašanje je da se omogući i uklanjanje svih assert i assume

izraza iz ulaznog P4 programa ukoliko to programer želi. Da bi se to ponašanje obezbedilo, uveden je novi argument koji se prosleđuje prilikom pokretanja rada P4 prevodioca. Ukoliko se prilikom pokretanja procesa prevođenja P4 programa navede u komandnoj liniji argument *„--ndebug''*, vrednost istoimenog flag-a u P4 prevodiocu će biti postavljena na *true*. Kada je taj flag postavljen u *midend* prolazu se dodaje *RemoveAssertAssume* prolaz koji treba da ukloni sva pojavljivanja *assert* i *assume* izraza iz IR. Navedeni prolaz je implementiran tako da je izveden iz *PassManager* klase koja enkapsulira više prolaza koji će se izvršiti sekvencijalno. U okviru *RemoveAssertAssume* potrebno je da se izvrše dva prolaza: *TypeChecking* i *DoRemoveAssertAssume*. Prvi prolaz je potreban kako bi se izvršilo popunjavanje struktura, koje mapiraju svako pojavljivanje sa deklaracijom i svaki čvor sa tipom, koje su potrebne drugom prolazu. *DoRemoveAssertAssume* prolaz je izveden iz *Transform* klase koja predstavlja osnovnu klasu koja izvršava transformacije nad određenim IR čvorom. *DoRemoveAssertAssume* klasa redefiniše *preorder* metodu koja obilazi objekat IR čvora koji reprezentuje poziv metode ili funkcije u P4 programu. U okviru redefinisane metode izvršava se provera da li je IR čvor koji se obilazi tipa *assert* ili *assume* izraza i ukoliko jeste uklanja iz programa.

U okviru *backend* prolaza kada se utvrdi da čvor koji se obilazi predstavlja *assert* ili *assume* izraz vrši se generisanje koda. Kako *bmv2 backend* generiše fajl u json formatu koji se koristi kao ulaz *Behavioral model* simulatora, uvedene su nove primitive u simulatoru kako bi se ulazni json fajl uspešno pročitao. Deo json fajla koji se odnosi na primitivu *assert* prikazan je na slici 3. Kao što je prikazano, pored naziva primitive i uslova, generiše se i deo koda *source_info* koji sadrži informacije koje su potrebne za ispisivanje greške.

```
"primitives" : [
    {
      "op" : "assert",
      "parameters" : [
        {
          "type" : "expression",
          "value" : {
            "type" : "expression",
            "value" : {
              "op" : "b2d",
              "left" : null,
              "right" : {
                "type" : "bool",
                "value" : false
              }
            }
          }
        }
      ],
      "source_info" : {
        "filename" : "/home/vagrant/Desktop/basic.p4",
        "line" : 122,
        "column" : 1,
        "source_fragment" : "assert(false)"
      }
```

Sl. 3. Prikaz dela json fajla za *assert(false)* izraz

Za svaki IR čvor ulaznog programa generiše se odgovarajući blok koda u json formatu.

Primitive u simulatoru su realizovane kao objekti klase *ActionPrimitive*, koja predstavlja osnovnu klasu za sve primitive koje postoje u projektu. Prilikom parsiranja ulaznog json fajla, za svaku pročitanu operaciju se formira objekat odgovarajuće primitive i doda u red svih primitiva koje su pročitane iz fajla koji je generisao P4 prevodilac. U fazi izvršavanja svih pročitanih primitiva poziva se preklopljena metoda *operator()* svake primitive i izvršava se njeno telo. Kada je reč o *assert* ili *assume* operaciji u telu preklopljene metode izvršava se proveravanje uslova. Ako je uslov netačan to znači da je pretpostavka bila pogrešna i poziva se metoda koja ispisuje poruku o grešci i informacije o fajlu, liniji i uslovu koji je doveo do greške. Te informacije se dobijaju iz objekta klase *SourceInfo* koji je vezan za svaku operaciju i koji se dobija čitanjem json fajla. Nakon ispisivanja informacija o grešci, simulacija se prekida.

V. Testiranje rešenja

U fazi testiranja rada P4 prevodioca korišćeni su testovi koji dolaze uz prevodilac prošireni sa implementiranim izrazima. Testovi su podeljeni u više grupa u zavisnosti od *backend*-a i verzije P4 jezika. Prilikom pokretanja testova za bmv2 backend pokreće se i simulator koji simulira izvršavanje prosleđenog programa. Ukoliko bilo koji od testova generiše izlaz koji se razlikuje od očekivanog, test će da padne i potrebno je proveriti šta je uzrokovalo neregularnost programa.

VI. Zaključak

U ovom radu opisan je P4 jezik i arhitektura prevodioca, kao i simulator za pokretanje izlaza P4 prevodioca. Prikazano je unapređenje jezika izrazima *assert* i *assume* koji za cilj imaju da olakšaju testiranje programa napisanih u P4 jeziku.

VII. Zahvalnica

Zahvaljujem se Miodragu Diniću na pruženom znanju, kao i profesoru Miroslavu Popoviću. Takođe bih se zahvalila koleginici iz P4Dev tima, Jeleni Vidaković, na saradnji, kao i naučno-istraživačkom institutu „RT-RK“ na pruženoj šansi da ovaj rad nastane kao proizvod višemesečnog rada na realnom problemu iz oblasti kojom se rad bavi. Ovaj rad je delimično finansiran od strane Ministarstva za prosvetu, nauku i tehnološki razvoj Republike Srbije, na projektu broj: III44009-6.

Literatura

[1] P4 jezik, https://p4.org/
[2] P4 prevodilac, https://github.com/p4lang/p4c
[3] C++ programski jezik, http://www.cplusplus.com
[4] Bison generator parsera, https://www.gnu.org/software/bison/
[5] Flex leksički analizator, https://www.gnu.org/software/flex/
[6] Behavioral model simulator, https://github.com/p4lang/behavioral-model

Analiza vremenskih serija: Metode predviđanja buduće potražnje u veleprodaji

Aleksandar Stojčić, Nevena Radosavljević, Bratislav Predić, Marko Kovačević, Miloš Roganović

***Apstrakt*— Osnovna namena analize vremenskih serija jeste pojašnjenje korelacije i osnovnih karakteristika hronološki sortitanih podataka korišćenjem odgovarajućih matematičkih modela. Svoju primenu nalazi u najrazličitijim aspektima života i rada, te tako i u predviđanju buduće potražnje proizvoda, usluga, itd. Najzastupljeniji tip vremenske serije jeste onaj kod koga se opservacije uzimaju u jednakim vremenskim intervalima (dnevno, nedeljno, mesečno, itd.). Međutim, u ovom radu analizira se neuobičajena vremenska serija koja beleži trenutke kupovina nekog proizvoda od strane potrošača, i koja se, samim tim što ne postoje regularni periodi uzorkovanja, mora transformisati na odgovarajući način pre nego što se može krenuti sa tradicionalnim metodama analize. Nakon postupka transformacije, u modeliranju vremenske serije u okviru rada korišćen je opšte poznat ARIMA model za analizu nestacionarnih vremenskih serija, koji je odabran zbog izražene komponente sezonalnosti i pomoću koga se uspešno vrši predviđanje buduće potrošnje posmatranog proizvoda. Cilj ovakve analize jeste pravovremeno reklamiranje proizvoda potrošaču radi povećanja prodaje.**

***Ključne reči*— potražnja; analiza vremenskih serija; sezonalnost; ARIMA; prognoza.**

I. Uvod

Vremenske serije su sastavni deo života i rada svih nas. Shodno tome, njihova analiza je statistička disciplina koja beleži najdinamičniji razvoj poslednjih decenija. Kako je proces donošenja odluka često povezan sa predviđanjem budućih vrednosti promenljivih koje zavise od vremena, vremenske serije i njihova analiza predstavljaju pogodno i jako korisno sredstvo. U ovom kontekstu, predviđanje podrazumeva analizu istorijskih podataka i ekstrapolaciju istih u budućnosti uz upotrebu odgovarajućeg matematičkog modela.

Predmet analize ovog rada jeste neregularna vremenska serija koja beleži trenutak kupovine i količinu kupljenog proizvoda od strane potrošača. Neregularna je iz razloga što se uzorkovanje ne vrši u jednakim vremenskim intervalima (dnevno, nedeljno, itd.), već prilikom svake kupovine. Iz tog razloga, ona se ne može analizirati i modelovati klasičnim tehnikama, već je najpre potrebno izvršiti određene matematičke transformacije. Generalna ideja je da se na osnovu kupljene količine i perioda između dve kupovine izračuna prosečna dnevna potrošnja konkretnog proizvoda od strane jednog potrošača. Na taj način se dobija klasična vremenska serija prosečne potrošnje beležene na dnevnom nivou.

Nakon svođenja početne vremenske serije na regularnu, možemo se na dalje baviti problemom njene stacionarnosti. ARMA (*engl. Autoregressive Moving Average*) model pripada grupi stohastičkih modela koji mogu biti korišćeni u simulaciji i analizi stacionarnih vremenskih serija. Ukoliko je, kao u ovom slučaju, vremenska serija nestacionarna, ali se može svesti na stacionarnu, ARIMA model je odgovarajući model za analizu. Sezonski ARIMA (*engl. Autoregressive Integrated Moving Average*) model potiče od klasičnog ARIMA modela, ali kao bitan faktor uključuje i sezonsku komponentu vremenske serije. Zapravo, sezonski ARIMA model je moćno sredstvo u analizi vremenskih serija jer je prilagodljiv čitavom skupu vremenskih serija, bez obzira na regularnost njihove sezonske komponente. Box i Jenkins su sedamdesetih bili pioniri ove metodologije.

S obzirom na složenost matematičkih izračunavanja u izgradnji modela, ovaj rad će se oslanjati na postojeća statistička softverska rešenja – programski jezik R, u računanju i iscrtavanju povezanih grafika.

II. Analiza problema

Kada je reč o analizi kupovina kao i obradi ovih informacija, većina radova, kao i komercijalnih sistema se fokusira na takozvane sisteme preporuke (*engl. recommender system*). Ono što je karakteristično kod ovih sistema jeste to da na osnovu prethodnih kupovina generišu personalizovane preporuke postojećim kupcima. Domen ovih rešenja vezuje se za transakcije koje se vrše direktno između kompanija i potrošača (*engl. Bussines to consumer - B2C*). Problem koji se dosta ređe sreće u literaturi, i čiji je jedan deo obrađen u ovom radu, odnosi se na transakcije veleprodaje koje se vrše između kompanija (*engl. Bussines to bussines - B2B*). Neke od glavnih razlika između ove dve vrste poslovanja koje imaju uticaja na način modelovanja su prikazane su u tabeli 1.

Bratislav Predić – Elektronski fakultet, Aleksandra Medvedeva 14., 18106 Niš, Srbija (e-mail: bratislav.predic@elfak.ni.ac.rs).

Marko Kovačević – Elektronski fakultet, Aleksandra Medvedeva 14., 18106 Niš, Srbija (e-mail: marko.kovacevic@code3profit.com) .

Miloš Roganović – Fakultet zaštite na radu, Čarnojevića 10a Niš, Srbija (e-mail: milos.roganovic@code3profit.com).

Aleksandar Stojčić – Elektronski fakultet, Aleksandra Medvedeva 14., 18106 Niš, Srbija (e-mail: alexstojcic@elfak.rs).

Nevena Radosavljević – Elektronski fakultet, Aleksandra Medvedeva 14., 18106 Niš, Srbija (e-mail: r.nevena@elfak.rs).

Veleprodaja	Maloprodaja
kupovina 'na veliko'	kupovina par proizvoda
odluke o kupovini donosi komisija	odluke donosi pojedinac
duži period kupovine	kratak period
odluke su racionalne	odluke su emotivne
kupovina 'po potrebi'	kupovina vođena željom

Tabela 1. Poređenje karakterisika veleprodaje i maloprodaje [1]

Sve ovo upućuje na činjenicu da je B2B proces kupovine stabilniji. Racionalnost odluka, duži proces kupovina, kao i veće kupljene količine upućuju na proces koji prati određena pravila i samim tim, kao takav manje je sklon slučajnostima.

Većina istraživanja na ovu temu dolazi iz oblasti marketinga. Radovi koji se bave ovom tematikom, pored vremena sledeće kupovine, najčešće vrše predikciju i drugih vrednosti kao što su izbor brenda i predviđena količina.

Već postojeća rešenja u ovoj oblasti mogu se podeliti na dve veće celine. Prva grupa rešenja polazi od pretpostavke da iza svake kupovine stoji motiv, odnosno da je odluka koju je potrebno predvideti: *'da li kupiti?*'. Drugi pristup ne dovodi u pitanje okolnosti pod kojima je došlo do transakcije, već njih gleda iz perspektive vremenske distance između svake dve kupovine. Ovakav pristup za cilj ima da nađe odgovor na pitanje *'kada kupiti?*'.

A. Da li kupiti?

Kao što je već naglašeno, ovu grupu karakteriše motiv kupovine. Osnovna ideja ovog pristupa jeste ta da potrošač kupuje isključivo kada se za to javi potreba. Ova grupa bolje rezultate daje primenom na B2B model poslovanja. Razlog za ovo su racionalne, potrebom vođene kupovine koje su odlike ovakve vrste poslovanja. Jedan od predstavnika ove grupe obrađen je u [2]. Autori u ovom radu polaze od pretpostavke da se svaka kupovina obavlja kada zalihe padnu ispod određene granice. Modelovanje se vrši posebno za svaki segment (proizvod). Svakom uređenom paru kupac-proizvod dodeljuju se dve promenjive:

$\mathbf{CR^h}$ – stopa potrošnje kupca *h*
$\mathbf{INV^h_t}$ – stanje zaliha u trenutku *t* kupca *h*

Obe promenljive sa nekim koeficijentom utiču na verovatnoću kupovine. Stopa potrošnje ima pozitivan uticaj na verovatnoću i ideja je da kupci sa većom potrošnjom imaju više razloga da izvrše kupovinu. Stopa potrošnje kupca se računa u periodu incijalizacije i to: nalaženjem srednje vrednosti nedeljne potrošnje u ovom periodu. Nasuprot stopi potrošnje koja ima pozitivnu korelaciju sa verovatnoćom, stanje zaliha ima negativnu. Smatra se da kupci sa manjim zalihama imaju veću inicijativu da načine sledeću kupovinu. Vrednost ove promenljive računa se kao razlika kupljene količine i proizvoda proteklog vremena i stope potrošnje. Pored ovih, postoji i treća, podjednako značajna, promenljiva $\mathbf{CV^h_{tr}}$. Ona govori o vezanosti kupca za specifičan proizvod. Njena osnovna namena je upravljanje promenama u verovatnoći koje potiču od mogućnosti da kupac izabere proizvod drugog proizvođača.

Neke greške koje potencijalno nastaju ovakvim pristupom su sledeće:

- Stopa koja se izračuna u periodu inicijalizacije se nikad ne koriguje.
- Rast kupca, kao i sezonalnost proizvoda nigde nisu uključeni u model.

Jedna razlika ovog pristupa u odnosu na domen problema ogleda se u korišćenju $\mathbf{CV^h_{tr}}$ parametra. Kao što je već naglašeno, svrha ovog parametra je da unese konkurenciju u model. U suštini, to zapravo i nije neophodno. Od interesa za problem je svaka kupovina koja se desi, bez obzira na poreklo proizvoda. Razlog za ovo je potencijalno uticanje na kupca iz perspektive izbora dobavljača.

Pristup sličan ovome moze se naci u [3]. Autori u ovom radu objedinjuju izbor vremena kupovine i brenda u jedan model. Pretpostavka je da kupac najpre dolazi u situaciju gde je neophodno da izvrši kupovinu, a da nakon toga vrši izbor proizvođača. Kako je nama od značaja samo trenutak kada će kupovina biti načinjena, fokusiraćemo se samo na ovaj deo. U radu je definisana granica kupovine. Ova granica računa se preko dve promenljive:

$\mathbf{FREQ^i}$ -frekvencija kupovine
$\mathbf{LQ^i_t}$ – normalizovana količina poslednje kupovine

Slično kao što se u prethodnom radu stopa potrošnje kupca računa u periodu inicijalizacije i više se nikad ne menja, tako se i ovde frekvencija kupovina takođe izračuna samo u ovom periodu i važi za ceo životni vek kupca. $\mathbf{LQ^i_t}$ se koristi da unese efekte kupovina različitih količina proizvoda. Ova promenljiva se računa kao zadnje kupljena količina minus srednja kupljena količina u periodu inicijalizacije.

Granica kupovine se na dalje u radu koristi kao osnova za donošenje odluka o kupovini. Za svaki trenutak se svakom dobavljaču dodeljuje vrednost koja pokazuje 'poželjnost' tog dobavljača. Ova vrednost je zavisna od kolekcije marketinških promenljivih koje se u tom trenutku vezuju za svakog dobavljača ponasob. Do transakcija dolazi kada poželjnost nekog dobavljača pređe granicu neophodnu za kupovinu.

B. Kada kupiti?

Ovu grupu rešenja, kao sto je već napomenuto, odlikuje modelovanje vremena između dve kupovine. Generalni pristup zasniva se na nalaženju distribucije ili hazard funkcije koja najbolje opisuje među-transakcijska vremena, parametrizaciju ove distribucije/hazard funkcije za svaki par kupac-proizvod i onda primenu na konkretnu instancu.

Prvi pokušaj nalaženja distribucije viđen je u [4]. Ovde je iskorišćena eksponencijalna distribucija u cilju modelovanja. Negativna strana ovakvog pristupa je ta što, kod eksponencijalne distribucije, nikako nije moguće iskoristiti

informaciju da kupac još nije kupio proizvod. Odnosno nije moguće naći verovatnoću kupovine u narednom periodu pod uslovom da kupac još nije kupio. Jedan od ranijih pokušaja rešavanja ovog problema može se naći u [5]. Ovde autori za distribuciju biraju *Erlang-2*. Iako je ovako izabrana distribucija dala dobre rezultate, ona se ne može smatrati univerzalnom i naravno nije je moguće primeniti na svaki problem. Pregled karakteristike pojedinih distribucija koje su našle primenu u ovoj oblasti, kao i njihovih specifičnih slučajeva korišćenja dat je u [6]. Distribucije koje su u ovom radu obrađene su: *eksponencijalna, Erlang-2, Weibull, log-logistic* i *expo-power*.

Još jedan napredak na ovu temu viđen je u [7]. Većina radova koja polazi od drugog pristupa, odnosno modeluje vremena izmedju dve kupovine, koristi vreme zadnje kupovine kao početnu tačku. Kod ovakvog pristupa, isti model, izračunat u periodu inicijalizacije, se ponovo primenjuje svaki put nakon nove kupovine. Negativna strana ovakvog pristupa je ta što se istorijat kupovine ni u jednom trenutku ne uzima u obzir prilikom novog računanja. Pored ovog problema, ukoliko su diskretne mere vremena izražene u nedeljama ili mesecima, primenom ovakvih modela nije moguće predvideti dve ili više kupovina u istom periodu. Kao rešenje, autor predlaže korišćenje kalendarskog vremena umesto proteklog. Cilj je uključiti sve raspoložive informacije u model, od prve kupovine pa sve do trenutnog trenutka. Još jedan benefit korišćenja kalendarskog vremena je mogućnost dodavanja sezonalnosti. Da bi se ovo postiglo, koriste se četiri komponente: kalendarsko proteklo vreme, proteklo vreme između kupovina, funkcija kovarijanse i odstupanje od srednje vrednosti vremena između kupovina. Pored ovih komponenti, autor u svom modelu takođe dodaje *'at-risk'* promenjive za određivanje verovatnoća sukcesivnih kupovina kao i kupovina određenim danima. Posebna primena ovih promenljivih, pored kontrolisanja više kupovina u istom trenutku, je recimo dodeljivanje verovatnoća kupovinama u određenim trenucima. Na primer, u svetu veleprodaje, moglo bi biti od značaja postaviti verovatnoću kupovine subotama i nedeljama na nulu.

Bez obzira na izbor pristupa, važno je napomenuti da ni jedno od ovih rešenja nije univerzalno i da u velikoj količini zavise od specifičnog slučaja upotrebe. Ovo za posledicu ima da je prilikom primene nekog od rešenja na drugačiji skup podataka, neophodno izdvojiti dodatno vreme za istraživanje spesifičnog slučaja primene kao i parametrizaciju modela koji se koriste.

III. KONCEPT REŠENJA

Aplikacija je zamišljena kao Windows Desktop aplikacija čija je osnovna uloga predviđanje buduće potrošnje kupaca u sistemu veleprodaje. Naime, na osnovu istorije prethodnih kupovina, za svakog potrošača ponaosob je potrebno utvrditi da li će, i sa kojom verovatnoćom, u toku naredne sedmice od zadatog datuma (uglavnom današnjeg) potrošiti svu kupljenu količinu proizvoda. Ukoliko je predikcija za nekog kupca pozitivna, on će biti dodat na spisak potrošača kojima je potrebno preventivno reklamirati odgovarajući proizvod, a sve to u cilju održavanja konzistentnosti prodaje i eventualno, njenog povećanja. Ulazni parametri na koje korisnik ovog softverskog rešenja može uticati jesu datum u odnosu na koji je potrebno izvršiti predviđanje, kao i minimalna verovatnoća ostvarivanja predikcije u odnosu na koju se ona prihvata, odnosno odbacuje.

A. *Priprema podataka*

Na slici 1 prikazan je osnovni UML dijagram klasa aplikacije. Kao što je prethodno već napomenuto, s obzirom da podaci koji se analiziraju ne predstavljaju uobičajenu vremensku seriju, njih je potrebno najpre na odgovarajući način transformisati. Upravo iz tog razloga, prvi korak u procesu analize i predviđanja budućih vrednosti jeste *priprema podataka*.

Sagledavajući činjenicu da postoji više od jednog načina za obradu i transformaciju početnog skupa podataka u vremensku seriju, kreiran je interfejs *I_ARIMAPrepare*. Ideja je da se različiti metodi transformacije podataka enkapsuliraju u okviru posebnih klasa koje nasleđuju ovaj interfejs i implementiraju sve njegove osnovne funkcije. Osim što se omogućava ispitivanje efikasnosti različitih načina transformacije podataka, uvođenjem ovog interfejsa aplikacija postaje prilagodljiva širokom skupu ulaznih podataka, koji se mogu umnogome razlikovati od zabeleženih trenutaka kupovine potrošača.

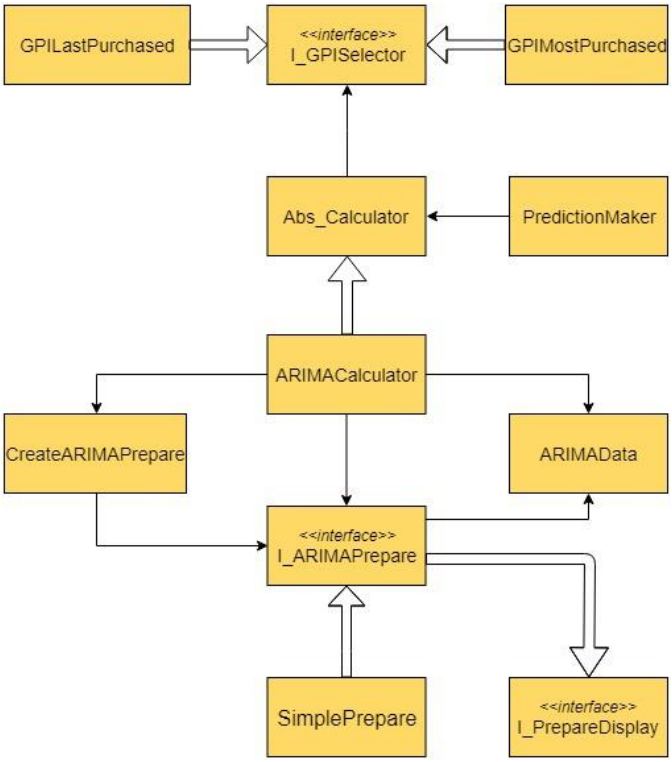
Metode za kreiranje regularne vremenske serije od početnog skupa podataka implementirane su u okviru klase *SimplePrepare*. Nakon učitavanja svih kupovina određenog potrošača do datuma unetog od strane korisnika aplikacije računa se prosečna dnevna potrošnja proizvoda. Prolazeći kroz niz kupovina, za svake dve kupovine se računa broj dana između njih, a zatim se deljenjem kupljene količine u prvoj sa brojem dana dobija prosečna dnevna potrošnja za period između te dve konkretne kupovine. Na ovaj način se od početnog skupa zabeleženih kupovina dobija regurlarna vremenska serija potrošnje proizvoda na dnevnom nivou. Ovaj proces je potrebno izvršiti za svaki proizvod koji potrošač kupuje, i dodatno, za svakog potrošača u sistemu. Osim ovog, direktno iz baze podataka učitava se još i niz prosečnih dnevnih potrošnji, ali na nivou proizvoda, odnosno dnevnih potrošnji proizvoda od strane svih potrošača u sistemu. Razlog za korišćenje ovog dodatnog niza je preciznija prognoza potrošnje kod potrošača sa oskudnom istorijom kupovina.

B. *Izračunavanje buduće potrošnje*

Nakon dobijanja regularne vremenske serije, može se krenuti sa njenom analizom i predikcijom budućih vrednosti pomoću ARIMA modela. Kao što je prethodno već naglašeno, sva izračunavanja vezana za modelovanje vremenske serije biće izvršena u R softverskom alatu.

Osnovna klasa u ovom delu procesa jeste apstraktna klasa *Abs_Calculator* koja slično kao interfejs *I_ARIMAPrepare*, omogućava lako proširenje aplikacije u budućnosti novim metodama za analizu vremenske serije. Trenutno

implementirana metoda korišćenjem ARIMA modela enkapsulirana je u okviru klase *ARIMACalculator*, koja kao što se sa slike 1 može primetiti, nasleđuje apstraktnu klasu *Abs_Calculator*. Uzimajući u obzir kompleksnost izračunavanja i količinu podataka koja se obrađuje, radi optimizacije čitavog procesa u sklopu *ARIMACalculator*-a kreirane su tri niti koje paralelno vrše obradu za pojedinačne potrošače. Funkcija ove klase je takođe i pravovremeno ažuriranje interfejsa radi informisanja korisnika aplikacije o količini obavljene i preostale analize podataka.



Slika 1. Osnovni UML dijagram klasa

Uzimajući u obzir činjenicu da se sama predikcija budućih vrednosti izvršava u R programskom jeziku, neophodno je bilo izvršiti integraciju tog softverskog alata i Windows Desktop aplikacije. To je jedna od osnovnih funkcija *PredictionMaker* klase koja u odgovarajućem trenutku poziva na izvršenje R skriptu sačinjenu od neophodnih naredbi za dobijanje prognoze.

Ulazni parametri R procedure jesu niz prosečnih dnevnih potrošnji za konkretnog potrošača, broj dana odnosno elemenata niza prosečnih dnevnih potrošnji konkretnog proizvoda od strane svih potrošača, kao i datum u odnosu na koji se vrši prognoza. Pozivanjem odgovarajućih R funkcija, uključujući i osnovnu *forecast* metodu, kao izlaz iz R skripte dobijamo prosečnu dnevnu potrošnju za narednih nedelju dana, na nivou konkretnog potrošača.

Na osnovu poslednje zabeležene kupljene količine konkretnog proizvoda, sa određenom verovatnoćom se zaključuje da li će potrošač u toku naredne nedelje potrošiti sve zalihe tog proizvoda. Ukoliko je to slučaj, onda se shodno tome taj potrošač dodaje na listu onih kojima je taj proizvod potrebno pravovremeno izreklamirati.

IV. Rezultati

Aplikacija je testirana nad realnim podacima proizvođača medicinske opreme.

A. Podaci

Podaci sa kojima smo radili sastoje se od preko tri miliona transakcija prodaje prikupljenih u periodu od dve godine i sedam meseci. Set čine nešto više od 9.000 jedinstvenih kupaca od kojih je oko 7.000 njih imalo ponovljene kupovine. Sa druge strane za vreme ovog perioda prodato je oko 6.300 jedinstvenih proizvoda. Za potrebe testiranja, predikcije su se vršile nedeljom i to za period od narednih nedelju dana. Prosečan broj transakcija na nedeljnoj bazi, u test periodu, bio je približno 23.000. Vremenski period kupovina bio je podeljen na:

1. Trening i validacioni set – 50%
2. Test set – 50%

Razlog za ovakvom podelom javio se iz podrebe za neuobičajeno velikim test setom. Naime, s obzirom da se kao izlaz iz sistema očekuju sve predikcije kupovine, a ne samo one koje se odnose na dobavljača čije podatke koristimo, bilo je neophodno odstraniti kupce i predikcije kupovina koje se više nikad nisu ponovile. Za ovakve slučajeve, vodimo se pretpostvakom da su se kupovine izvršile kod drugog dobavljača ili da je kupac našao neku drugu alternativu proizvodu. Kako ne možemo biti sigurni kada se kupovina u ovoj grupi desila kod drugog dobavljača, moramo je odstraniti iz statistike. Da bi bili sigurni da se kupovina nije desila, test period mora da bude dovoljno veliki da ukloni svaku sumnju.

B. Naivna metoda

Za potrebe poređenja korišćena je verzija starog sistema, u kome je bila implementirana naivna metoda dobijanja prognoze. Ova metoda trenutak sledeće kupovine nalazila je računanjem tri parametra za svaki par kupac-proizvod: srednje, minimalno i maksimalno vreme između dve kupovine. Pomoću ovih parametara vreme sledeće kupovine dobijalo se dodavanjem srednjeg vremena između dve kupovine na trenutak zadnje kupovine, pod uslovom da nema većih odstupanja od minimalnog i maksimalnog vremena. Dozvoljena razlika između ova dva parametra dobijana je pravljenjem balansa između preciznosti predikcija i broja predikcija koju je potrebno načiniti. Ukoliko bi se koristila velika razlika, preciznost bi drastično opala. Ukoliko bi ta razlika pak bila previše mala, ne bi se načinilo dovoljno predikcija za sledeću nedelju. Ovakvo rešenje davalo je prilično loše rezultate, srednja preciznost se iz nedelje u nedelju kretala oko 6,9% dok je pokrivenost slučajeva (broj tačnih predikcija / broj ostvarenih kupovina) iznosila 0,53%.

Postoji više razloga zašto je ovakav pristup davao loše rezultate, a među najvažnijima su:

1. Prethodno kupljena količina uopšte nije uzimana u obzir prilikom izračunavanja
2. Favorizuju se kupci sa manjim brojem kupovina

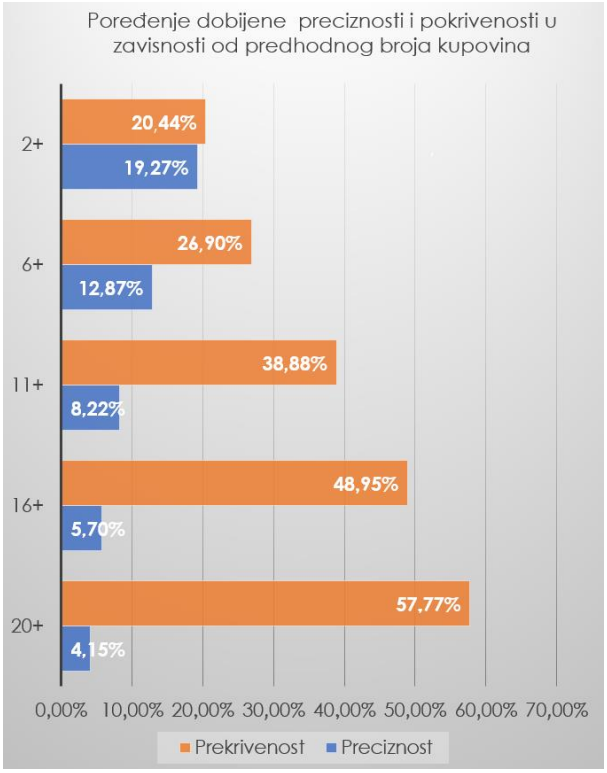
Razlog zašto je dolazilo do favorizovanja kupaca sa manjim brojem kupovina je taj što je, pri velikom broju kupovina, razlika između minimalnog i maksimalnog vremena između dve transakcije statistički mogla samo da se povećava. Ovakve predikcije bi pre bile odbačene kao neopouzdane što je suprotno željenom ponašanju.

C. *Rezultati testiranja*

Kao što je već spomenuto za računanje preciznosti predikcija, korišćene su samo kupovine koje su se makar jednom realizovale u test periodu. Kada je reč o pokrivenosti realizovanih kupovina, iz ukupnog broja transakcija izuzete su one koje se kao par kupac-proizvod javljaju po prvi ili po drugi put. Razlog za ovo je delimična nepredvidljivost ovakvih kupovina.

Rezultati su dati na na slici 2. Može se uočiti da za parove kupac-proizvod postoji pozitivna korelacija između broja transakcija i preciznosti predikcije. Kada bi se predikcija vršila za sve moguće parove, preciznost bi iznosila 20,44%, dok za parove koji imaju više od 20 načinjenih kupovina, ta vrednost dostiže skoro 60%. Još jedna značajna stvar je da uvođenjem jednog ovakog sistema, preciznost bi se samim delovanjem samo još vise povećala.

Nasuprot preciznosti, pokrivenost realizovanih kupovina ima negativnu korelaciju sa povećanjem broja neophodnih transakcija za kreiranje predikcije.



Slika 2. Poređenje dobijene preciznosti i pokrivenosti u zavisnosti od prethodnog broja kupovina

D. *Performanse*

Celokupna obrada podataka rađena je na računaru sa procesorom marke Intel model i7-6700T i 8GB ram memorije. Za obradu je bilo neophodno između 15 i 17 sati. Kompleksnost algoritma je u lineranoj zavisnosti sa brojem parova kupac-proizvod, a ovo ima povoljne ishode na skalabilnost. Takodje rešenje se može u velikoj meri paralelizovati i ima elemente računice koja potencijalno može da se izvršava na grafičkom procesoru, čime bi se dodatno povećale performanse.

V. Zaključak

Modeli vremenskih serija predstavljaju vrlo moćno sredstvo za predviđanje budućih vrednosti i donošenje odluka u različitim oblastima života i rada: ekonomiji, poljoprivredi, industriji, medicini, itd. U praksi se pokazalo da veliki broj pojava zavisnih od vremena mogu da se modeliraju pomoću stohastičkih procesa. Rasprostranjenost i atraktivnost ovih modela je posledica njihove same strukture koja je lako razumljiva i prilično intuitivna. Za razliku od nekih drugih modela koji razmatraju vezu između dve ili više različitih pojava, modeli vremenskih serija, uključujući i ARIMA model, ispituju uticaj istorijskih vrednosti jedne pojave na njenu sadašnju i buduću vrednost. Ovakav pristup omogućava proučavanje ponašanja date pojave u vremenu i daje dobre rezultate, naročito ukoliko je dostupan veliki broj istorijskih podataka.

Naravno, kao i svi modeli, ni ARIMA model, korišćen u ovom radu, ne predstavlja savršen prikaz stvarnog stanja, ali uprkos tome, ovaj model i dalje omogućava brz i efikasan postupak koji sa dosta preciznosti daje prognozu budućih vrednosti.

Literatura

[1] http://www.ijstr.org/final-print/sep2014/A-Comparative-Study-On-B2b-Vs-B2c-Based-On-Asia-Pacific-Region.pdf

[2] Bucklin, Randolph E., and Sunil Gupta. 1992. “Brand Choice, Purchase Incidence, and Segmentation: An Integrated Modeling Approach.” Journal of Marketing Research 29 (2): 201–215.

[3] Zhang, Jie, and Lakshman Krishnamurthi. 2004. “Customizing Promotions in Online Stores.” Marketing Science 23 (4): 561–578.

[4] Ehrenberg, A.S.C. 1959. “The Pattern of Consumer Purchases.” Applied Statistics 8: 26–41.

[5] Gupta, Sunil. 1988. “Impact of Sales Promotions on When, What, and How Much to Buy.” Journal of Marketing Research 25 (4): 342–355.

[6] Seetharaman, P.B., and Pradeep K. Chintagunta. 2003. “The Proportional Hazard Model for Purchase Timing: A Comparison of Alternative Specification.” Journal of Business & Economic Statistics 21 (3): 368–382.

[7] Bijwaard, Govert E. 2006. “Regularity in Individual Shopping Trips: Implications for Duration Models in Marketing.” Working paper, Econometric Institute, Erasmus University, Rotterdam.

Abstract

Time series analysis is to explain correlation and the main features of the data in chronological order by using appropriate statistical models. It’s being used in various aspects of life and work, as well as in forecasting future product demands, service demands, etc. The most common

type of time series data is the one whose observations are taken in equally distributed time intervals (daily, weekly, monthly, etc.). However, in this paper, we analyze unusual time series that represents product purchase moments. Thus there isn't regular obesrvation periods, this irregular time series must be transformed in some way before traditional methods of analysis can be applied. After data transformation is complete, next step is modeling nonstationary time series using commonly known ARIMA model, which has been chosen for accentuated seasonality and fairly easy and successful forecasting process. Goal of this analysis is timely product advertising to a customer in order to increase sale.

Time Series Analysis: Forecasting wholesale future demands

Aleksandar Stojčić, Nevena Radosavljević, Bratislav Predić, Marko Kovačević, Miloš Roganović

Arhitektura i implementacija softverskog sistema za fleksibilno sprovođenje korisnički definisanih anketa

Ognjen Milošević, Marko Mišić, *Member, IEEE*, Jelica Protić

***Apstrakt*—Sprovođenje korisnički definisanih anketa je čest proces u današnjem poslovanju. Bilo da su u pitanju akademske ili marketinške potrebe, postavlja se pitanje na koji način ankete sprovesti tako da se na brz i jednostavan način dopre do ispitanika bez suvišnih koraka koji bi ga eventualno odvratili od popunjavanja zadatog upitnika. Sa druge strane, prikupljanje podataka treba da bude jednostavno i sa stanovišta onoga ko ispituje neku temu, bez potrebe za korišćenjem komplikovanih sistema. Rasprostranjenost mobilnih uređaja u današnjem svetu olakšava sprovođenje anketiranja sa strane ispitanika, a u radu su prikazani arhitektura i implementacija jednog sistema za fleksibilno sprovođenje korisnički definisanih anketa. Sistem se sastoji od mobilne, veb i serverske aplikacije, a ceo proces se sprovodi bez korišćenja klasične baze podataka i sa relativno skromnom upotrebom resursa.**

***Ključne reči*—korisničke ankete; veb programiranje; mobilna aplikacija.**

I. Uvod

SVE veća dostupnost Interneta u poslednje dve decenije je značajno promenila način sprovođenja anketa (upitnika, intervjua). Anketiranje je poseban metod prikupljanja informacija pomoću kojeg se dolazi do podataka o stavovima i mišljenjima ispitanika na zadatu temu. Umesto tradicionalnih načina kao što su anketa licem u lice, telefonska anketa, anketa na papiru, sve je više zastupljeno anketiranje elektronskim putem [1].

Anketiranje elektronskim putem se danas najčešće sprovodi putem Interneta. Bilo kakvo ozbiljnije ispitivanje javnog mnjenja bez upotrebe modernih tehnologija je danas gotovo nezamislivo. Ovakav način anketiranja štedi vreme, energiju, kao i materijalne resurse anketara. Takođe, s obzirom na dostupnost interneta danas, ova vrsta anketiranja pruža velike mogućnosti za ispitivanje statistički značajnog uzorka, jer se ispitanici lakše odlučuju da posvete vreme ovakvoj vrsti upitnika zbog njegove jednostavnosti.

Razvoj interneta je najpre doneo elektronske upitnike putem veba i računara. Međutim, sa povećanjem mogućnosti pametnih telefona, teži se potpunoj zameni računara. Pametni telefoni danas pružaju korisniku mogućnost da gotovo svaki posao, za koji je do skoro računar bio neophodan, obave u bilo kom trenutku i na bilo kom mestu.

Mobilna aplikacija za anketiranje je iz tog razloga sagledana kao vrlo praktičan način anketiranja. Osim dostupnosti na svakom mestu, mobilni telefon ima i druge prednosti, kao što je mogućnost skeniranja različitih kodova putem kamere (bar kod, QR kod). Pomenute prednosti mobilne aplikacije za anketiranje motivisale su projektovanje i izradu softverskog sistema koji kroz mobilnu aplikaciju u samo nekoliko koraka omogućava pristupanje anketi preko QR koda, popunjavanje i slanje odgovora na server gde će se oni čuvati za dalju upotrebu.

Da bi softverski sistem bio potpun, osim pomenute mobilne aplikacije, bilo je potrebno realizovati i veb aplikaciju za administriranje anketa. Sa veb aplikacije moguće je kreiranje novih anketa, definisanje kodova za pristup aneketi, ali i pregled statistike za postojeće ankete, kao i generisanje izlaznog fajla sa prikupljenim odgovorima u tabelarnom formatu radi lakšeg pregleda odgovora. Treći deo ovog sistema je serverska aplikacija, koja služi za manipulaciju podacima kao što su čuvanje anketa, provere kodova sa mobilne aplikacije, slanje odgovarajuće ankete ka mobilnoj aplikaciji i slično. Da bi sistema bio skalabilan i fleksibilan za korišćenje od strane, kako anketara, tako i ispitanika, zamišljen je da bude sa malim režijskim troškovima (eng. *lightweight*) i naporom za korišćenje. U nastavku rada će biti data motivacija za njegovu implementaciju i opisane donesene projektne odluke.

Rad je podeljen na nekoliko sekcija. U okviru drugog poglavlja data je funkcionalna specifikacija sistema sa osvrtom na bitne detalje za implementaciju. Treće poglavlje kratko opisuje korišćene tehnologije, dok su u četvrtom poglavlju opisane arhitektura i implementacija sistema. U finalnom poglavlju je dat kratak zaključak i pravci za dalji razvoj.

II. Funkcionalna specifikacija

Cilj ovog rada bilo je kreiranje softverskog sistema za sprovođenje korisnički definisanih anketa. Sistem je trebalo projektovati tako da obuhvata sve bitne aspekte: od kreiranja ankete, preko administriranja ankete, do sakupljanja odgovora uz naglašenu jednostavnost upotrebe i što je moguće manju upotrebu resursa za smeštanje podataka.

Ognjen Milošević – Elektrotehnički fakultet, Univerzitet u Beogradu, Bulevar Kralja Aleksandra 73, 11020 Beograd, Srbija (e-mail: milosevic.etf@gmail.com).

Marko Mišić – Elektrotehnički fakultet, Univerzitet u Beogradu, Bulevar Kralja Aleksandra 73, 11020 Beograd, Srbija (e-mail: marko.misic@etf.bg.ac.rs).

Jelica Protić – Elektrotehnički fakultet, Univerzitet u Beogradu, Bulevar Kralja Aleksandra 73, 11020 Beograd, Srbija (e-mail: jelica.protic@etf.bg.ac.rs).

Iz tog razloga sistem je zamišljen iz tri dela, kao distribuirana aplikacija, koji rade kao jedna celina. Delovi koji čine ovaj sistem, kao i skup njihovih funkcionalnosti navedeni su u nastavku:

- Veb aplikacija za kreiranje i administriranje anketa;
- Android aplikacija za popunjavanje anketa i slanje odgovora na server;
- Serverska aplikacija koja posreduje između dve prethodno navedene, distribuira ankete do korisnika i čuva sve potrebne informacije i odgovore.

Nadalje biće detaljnije opisani zahtevi za svaku od pomenutih aplikacija.

A. Veb aplikacija

Zahtev koji je postavljen u vezi sa veb aplikacijom je jednostavan interfejs za administriranje korisnički definisanih anketa od strane korisnika koji sprovodi anketu. Pod administriranjem se podrazumevalo kreiranje same ankete, generisanje kodova za pristup anketi, kao i pregled postojećih anketa. Potrebno je bilo omogućiti generisanje kako tekstualnih, tako i generisanje grafičkih kodova, poželjno QR (eng. *Quick Response*) kodova.

Prilikom kreiranja ankete administrator unosi naziv ankete, opis, broj kodova koji će biti generisani, datum i vreme isteka ankete, kao i pitanja koja mogu biti različitih tipova. Podržana su tri najčešće korišćena tipa pitanja u anketama zatvorenog tipa: pitanje sa jednostrukim izborom, pitanje sa višestrukim izborom i Likertova skala. Kod Likertove skale su dati ponuđeni odgovori u opsegu od 1 do 5.

Što se pregleda i administracije postojećih anketa tiče, tu su podaci o broju poslatih odgovora do trenutka proveravanja, broju generisanih kodova za izabranu anketu, datumu kreiranja i isteka, mogućnost da se kreiraju dodatni kodovi za pristup anketi i da se generiše izveštaj u MS Excel (XLSX) formatu od trenutno primljenih odgovora.

B. Mobilna aplikacija

Mobilna aplikacija treba da pokriva tri osnovne funkcionalnosti. To su učitavanje QR koda za pristup anketi koje se izvršava preko kamere, popunjavanje odgovara na postavljena pitanja koja server vraća nakon slanja koda na serversku stranu i slanje popunjenih odgovora na server.

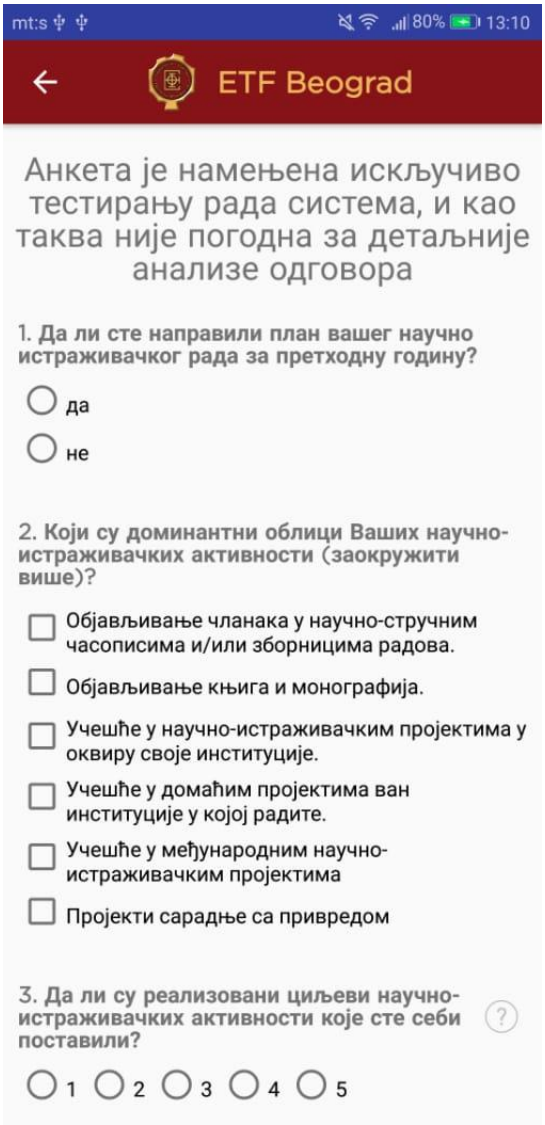
Jednostavnost upotrebe se pre svega ogleda u odsustvu bilo kakve forme registracije ili autentifikacije korisnika. Dakle, dovoljno je da korisnik instalira aplikaciju na pametni uređaj, pokrene je, unese dobijeni kod i može odmah pristupiti popunjavanju ankete, kao što se može videti na Slici 1.

C. Serverska aplikacija

Zadatak serverske aplikacije je da opslužuje dve prethodno pomenute aplikacije podacima sa servera i skladišti podatke od interesa. U okviru serverske aplikacije odvijaju se sve obrade nad podacima, kao što je provera validnosti QR koda i slanje na mobilnu aplikaciju odgovarajuće ankete ukoliko je kod validan i anketa aktivna. Takođe, serverska aplikacija vrši prijem podataka sa mobilne aplikacije u vidu odgovora na anketu i čuvanje istih.

Čuvanje ankete koja se kreira preko veb aplikacije takođe se vrši preko serverske aplikacije, ali i druge obrade kao što su invalidacija koda nakon potvrde završetka ankete, inkrementiranje broja dosadašnjih odgovora na odgovarajuću anketu, generisanje novih kodova.

Zbog jednostavnosti i fleksibilnosti, zahtevano je da na serverskoj strani ne postoji klasična baza podataka, već da se podaci čuvaju u odgovarajućem međuformatu koji se lako može izvesti u tabelarni format pogodan za dalju obradu. U tom smislu, čuvanje podataka o isteklim anketama nije ni od kakvog interesa za sam sistem.



Sl 1. Stranica za popunjavanje ankete u okviru mobilne aplikacije

III. KORIŠĆENE TEHNOLOGIJE

S obzirom na zahteve za realizacijom distribuiranog sistema koji se sastoji od tri zasebne aplikacije koje se izvršavaju u tri različita okruženja, pažljivo se pristupilo odabiru korišćenih tehnologija za njihovu izradu.

Za programiranje mobilne aplikacije za Android uređaje, postojala je dilema između upotrebe programskog jezika Java

i njegove nadogradnje Kotlin koja pruža nešto jednostavniji i koncizniji način pisanja mobilnih aplikacija [2]. S obzirom na bogatiji izbor biblioteka za različite obrade, platformsku nezavisnost, činjenice da je programski jezik Java jedan od najzastupljenijih programskih jezika, Java je izabrana za realizaciju svih bitnih aspekata sistema.

Kada je reč o razvoju internet aplikacija, tendencije su takve da se sve više prelazi na skriptne jezike, bazirane na *Javascript*-u, kao i prelaz na SPA (eng. *Single Page Application*) aplikacije, kada one zadovoljavaju korisničke zahteve. SPA, su aplikacije koje ne zahtevaju ponovno očitavanje stranica tokom upotrebe, jer se većina resursa učitava jednom tokom životnog ciklusa aplikacije [3]. Manipulacija podacima koji se šalju na server i uzimaju sa servera su jedina očitavanja koja se očekuju kasnije u toku rada. Iz tog razloga, za veb aplikaciju realizovanu u ovom radu, SPA je veoma pogodan i efikasan izbor koji je realizovan putem *Angular* radnog okvira i *Typescript* jezika.

U ovom softverskom sistemu, razmena podataka između aplikacija realizovana je preko HTTP protokola, a što se tiče izbora servisa za API pozive, prvi izbor u izradi ovog rada su bili REST (eng. *Representational state transfer*) servisi, a razlozi za to su mnogobrojni.

Nezavisnost između klijentske i serverske aplikacije, uz komunikaciju jedino preko zahteva i odgovora je jedna od odlika REST servisa koja je bila pogodna za ovakvu arhitekturu sistema, s obzirom da jedna serverska aplikacija opslužuje dve klijentske. REST servisi su realizovani kroz Java programski jezik i radni okvir *Spring boot*. Razlozi za to su jednostavnost prilikom podešavanja okruženja i pokretanja aplikaciije na serveru, kao i dovoljan skup funkcionalosti i alata da se pokriju svi zahtevi iz ovog projektnog zadatka.

Za razmenu i čuvanje podataka je odabran JSON format. JSON format je čitljiv ljudima, ima izraženu hijerarhiju, čime se mogu predstaviti vrednosti unutar objekata, može se parsirati i koristiti iz mnogih programskih jezika. Alternativa je bila upotreba XML formata, ali JSON daje dosta kraći kod, pa je samim tim lakši za „ručno“ pisanje i čitanje, a najveća prednost koju treba izdvojiti za kreiranje ovog sistema je to što je za XML potreban odgovarajući parser da bi se koristio u skriptnim jezicima, dok se JSON kao takav može koristiti u skriptnim jezicima kao što je *Typescript* u kom se u ovom slučaju koristi [4].

IV. Opis rešenja

Ovo poglavlje je posvećeno detaljnijem objašnjenju realizacije sistema. Sistem je bliže predstavljen prolaskom kroz neke od najvećih problema prilikom realizacije u narednim podglavama i objašnjenje njihovih rešenja.

A. Format JSON fajlova

Sa odlukom da kompletan softverski sistem realizuje bez klasične baze podataka i da se svi podaci čuvaju u tekstualnim fajlovima u JSON formatu, bilo je potrebno pažljivo osmisliti u koliko fajlova smestiti sve podatke, kako ih rasporediti, ali i kako ih održavati i dopunjavati.

Prilikom kreiranja ankete, formira se fajl sa identifikacionim brojem ankete kao nazivom, u koji se upisuju pitanja i ponuđeni odgovori. Format fajla za čuvanje ankete prikazane na Slici 1. prikazan je na Slici 2.

```
{
   "survey":{
      "id":"11",
      "description":"Anketa je namenjena isključivo
testiranju rada sistema, i kao takva nije pogodna za
detaljnije analize odgovora",
      "numOfQuestions":"3",
      "questions":[
         {
            "num":"1",
            "text":"Da li ste napravili plan vašeg
naučno istraživačkog rada za prethodnu godinu?",
            "type":"single",
            "answers":[
               "da",
               "ne"
            ]
         },
         {
            "num":"2",
            "text":"Koji su dominantni oblici Vaših
naučno-istraživačkih aktivnosti (zaokružiti više)?",
            "type":"multi",
            "answers":[
               "Objavljivanje članaka u naučno-
stručnim časopisima i/ili zbornicima radova.    ",
               "Objavljivanje knjiga i monografija.
",
               "Učešće    u    naučno-istraživačkim
projektima u okviru svoje institucije.",
               "Učešće u domaćim projektima van
institucije u kojoj radite. ",
               "Učešće    u    međunarodnim    naučno-
istraživačkim projektima    ",
               "Projekti saradnje sa privredom    "
            ]
         },
         {
            "num":"3",
            "text":"Da li su realizovani ciljevi
naučno-istraživačkih aktivnosti koje ste sebi
postavili?",
            "type":"likert",
            "answers":null
         },
      ]
   }
}
```

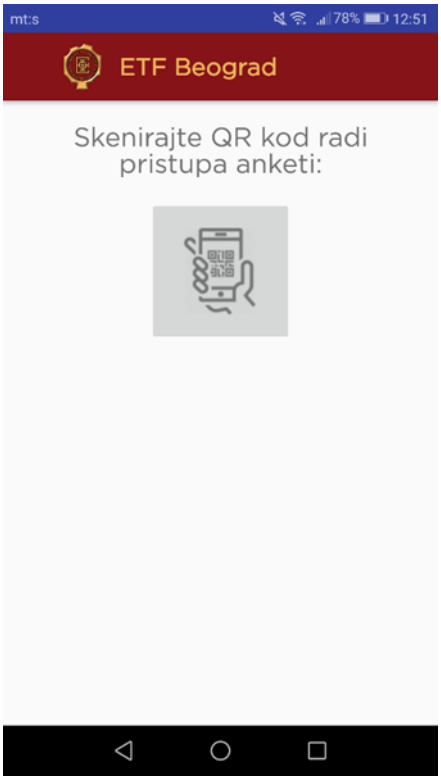
Sl. 2. Primer JSON fajla za čuvanje anketa i pitanja

Osim prikazanog fajla, pri kreiranju ankete se ažuriraju i postojeći fajlovi u kojima se čuvaju osnovni podaci o svim postojećim anketama, kao i podaci o kodovima za anketu. Fajl u kojem se čuvaju odgovori se takođe kreira prilikom kreiranja ankete, sa praznom listom JSON objekata i ažurira se prilikom slanja svakog odgovora na server. U njemu se osim odgovora na pitanja čuva i odgovarajući kod sa kojim je pristupljeno anketi. Objekat za odgovor na svako pitanje ima isti format, radi lakšeg formiranja izlazne tabele, a u zavisnosti od tipa pitanja neka polja u objektu ostaju prazna, tačnije imaju *null* vrednost.

B. *Skeniranje kodova*

Da bi skenirali kod preko Android aplikacije potreban, ali ne i dovoljan uslov je pristup kameri telefona koji je omogućen dodavanjem odgovarajuće linije koda u *AndroidManifest.xml* fajlu u sklopu aplikacije. Osim toga, potreban je i alat koji može da skenira QR kodove i da ih pretvori u odgovarajući link ili niz karaktera. Kako je sama realizacija ovakvog alata veliki posao, iskorišćena je već postojeća aplikacija *Barcode Scanner* kojoj se pristupa preko naše aplikacije [5]. Naravno, ovo nije jedina aplikacija za ovu namenu, ali kao besplatna aplikacija, pogodna za integraciju unutar drugih aplikikacija, pokazala se kao efikasno rešenje.

Skeniranje se pokreće pritiskom na odgovarajuće dugme u okviru forme mobilne aplikacije (Slika 3.), a zatim se u posebnoj metodi obrađuje rezultat skeniranja. U zavisnosti od koda rezultata, ili se korisniku ispisuje poruka o neuspešnom skeniranju i traži novi pokušaj, ili se sa dobijenim kodom poziva metoda za dohvatanje ankete, čiji je argument učitani kod, u kojoj će biti poslat zahtev serveru za dobijanje ankete koja odgovara datom kodu.



Sl. 3. Ekran mobilne aplikacije u okviru koga se vrši skeniranje kodova

C. *Tipovi REST servisa i njihova upotreba*

Postoji više različitih tipova HTTP zahteva koji se mogu slati ka serveru, u zavisnosti od namene servisa. Kod REST API servisa uobičajno je da se koriste *get*, *put*, *post* i *delete* metode [6].

Za skup funkcionalnosti koje su bile potrebne za realizaciju ovog softverskog sistema, bilo je dovoljno korišćenje *get*, *post* i *put* moetoda te će samo one u ovom radu biti detaljnije objašnjene, a zatim će biti izneti primeri korišćenja ovih zahteva u softverskom sistemu.

Get metode namenjene su isključivo čitanju podataka sa serverske strane. Koriste se za čitanje određenog podatka ili cele kolekcije podataka. Zbog toga mogu imati parametre unutar putanje, koji mogu biti identifikator pojedinačnog objekta iz kolekcije koji će biti vraćen ili neki drugi kriterijum po kom će se filtrirati kolekcija da bi se dobio odgovarajući odgovor od servera. U skladu sa tim, u ovom projektu su *get* metode korišćene za pronalaženje ankete preko učitanog koda gde je kod parametar u zahtevu, kao i za vraćanje svih anketa za ispis na stranici sa anketama - zahtev bez parametara i vraćanje odgovora za određenu anketu preko identifikacionog broja gde je *id* parametar u zahtevu.

Post i *put* metode služe za umetanje elementa u postojeće kolekcije ili za kreiranje kolekcija entiteta. Po preporuci se *post* metode koriste za kreiranje novih entiteta ili kolekcija, dok *put* metode služe za izmene na postojećim. Shodno tome, *put* metode se mogu pozvati slično kao *get*, sa parametrima unutar putanje, dok je kod *post* metoda najčešće praksa da osim opcionih parametara unutar putanje, postoji i određeni entitet umetnut u zahtev koji se prosleđuje ka serveru.

Post metode su korišćene za pozive serveru prilikom kreiranja nove ankete i prilikom slanja odgovora na server, gde se prosleđuju kompletni entiteti. *Put* metode su iskorišćene kod metoda u kojima se menjaju postojeći entiteti, a to su metode za generisanje novih kodova za pristup anketi i metoda za promenu vremena isteka ankete. Sve metode su napisane uz korišćenje odgovarajućih *Spring boot* anotacija.

D. *Dinamičko kreiranje komponenti na veb stranici*

Kod stranice za kreiranje anketa na veb aplikaciji, postojala je potreba za dinamičkim kreiranjem komponenti, s obzirom da nije unapred određen broj pitanja u anketi, niti broj ponuđenih odgovora za pojedinačna pitanja.

Stoga se incijalno se kreira komponenta koja sadrži prazan kontejner na koji se mogu dodvati odgovarajuće komponente, da bi se kasnije korišćenjem klase *ViewContainerRef* kreirale i dodavale komponente na njega. Na gotovo identičan način je rešen i problem dodavanja komponenti za odgovore na komponentu za pitanja. Sve ovo je izvedeno kroz mogućnosti koje pruža radni okvir *Angular* [7].

E. *Izmene postojećih JSON fajlova*

Za manipulaciju JSON fajlovima iz serverske Java aplikacije korišćene su klase *JSONParser*, *JSONObject* i *ObjectMapper*.

Prve dve pomenute klase su iz *org.json.simple* paketa. *JSONParser* klasa služi za parsiranje JSON teksta u Java objekat koji se (ukoliko je JSON tekst validan) može eksplicitno konvertovati u objekat tipa *JSONObject*. *JSONObject* klasa nasleđuje *HashMap* klasu, *java.util* paketa. Za razliku od standarnog funkcionisanja heš mape, ova klasa ima i dodate metode za konvertovanje parova ključ-vrednost u JSON stringove, što je čini pogodnom za potrebne operacije pristupanja postojećim JSON objektima i generisanja izmena na njima.

Što se tiče klase *ObjectMapper*, ona pripada *com.fasterxml.jackson.databind* paketu, njena namena je čitanje i pisanje JSON formata, kako iz Java entiteta, tako i iz samih JSON fajlova, što je korišćeno prilikom dodavanja odgovora u postojeći fajl. S obzirom na potrebu da se odgovori umetnu u postojeću listu u okviru JSON objekta, bilo je potrebno prvo pristupiti toj listi, što je urađeno odmah nakon parsiranja i konvertovanja objekta u *JSONObject*. U tu listu nije moguće umetnuti objekat onakav kakav je prosleđen kroz zahtev serverskoj aplikaciji, već je neophodno formirati odgovarajući JSON objekat. Tako se kreiran novi JSON objekat u koji se podaci upisuju kao u mapu. Kada se taj objekat doda u listu, potrebno je samo ažurirani objekat ispisati preko mapera. Takođe, prilikom dodavanja odgovora invalidira se kod koji je poslala klijentska aplikacija i vraća joj se odgovor da je čuvanje odgovora uspešno završeno.

F. Generisanje izlaza u XLSX formatu

Jedna od fukncionalnosti ovog softverskog sistema je generisanje izlaza u XLSX formatu, tačnije ispis tabele iz JSON fajla sa odgovorima za odgovarajuću anketu, kako bi bili jednostavniji za dalju obradu.

Za parsiranje i ispisivanje samog fajla korišćena je *xlsx* biblioteka, dok je za čuvanje fajla na korisničkoj strani korišćena *file-saver* biblioteka.

Same fajlove u kojima su čuvani odgovori je bilo neophodno dodatno obraditi kako bi se dobio odgovarajući izlaz. Razlog leži u tome što se metoda *json_to_sheet* nije pokazala kao dobro rešenje za konvertovanje niza složenijih objekata koji sadrže nizove, te je u petlji formiran niz prostih objekata u kojima su sva polja tipa *string*, kako bi ih ispisao na adekvatan način u XLSX fajlu.

V. Zaključak

Korišćenjem pametnih telefona ili tablet računara za popunjavanje anketa, može se uprostiti kako kontrola pristupa anketi pomoću koda, tako i samo ispitivanje, koje se obavlja u par jednostavnih koraka i što je možda još bitnije uz vrlo malo utrošenog vremena. Veb aplikacija za kreiranje i administraciju anketa je realizovana pomoću radnog okvira *Angular*, veoma je intuitivna, jednostavna za korišćenje, a zahvaljujući odluci da se realizuje kao SPA i performanse su na zadovoljavajućem nivou, te i deo posla oko administracije ne zahteva previše vremena. Kada je reč o serverskoj aplikaciji, takođe su postignuti zadovoljavajući rezultati. Pri testiranju sa lokalnog servera, serverska aplikacija je efikasno opsluživala obe aplikacije, te nije bilo značajnijeg kašnjenja prilikom pozivanja servisa i čekanja odgovora sa servera.

Postoje dva glavna pravca za proširenje i poboljšanje ovog sistema. Jedan se tiče aspekta bezbednosti. Prilikom testiranja na lokalnom serveru, nije realizovana autentifikacija prilikom poziva metoda, te su zahtevi ka serveru dostavljani bez kredencijala u zahtevima. Sa proširenjem primene van lokalnog servera, bezbednosni mehanizmi bi morali da postoje. Slično važi i za veb aplikciju, u kojoj takođe u nekoj od narednih verzija treba zahtevati autentifikaciju korinika pre pristupa stranicama za administriranje.

Drugo unapređenje bi se odnosilo na uvođenje kompletne baze podataka na serverskoj strani aplikaciji. U tom slučaju, dalji razvoj bi išao ka dodavanju komponente za analizu rezultata, što se u ovom slučaju želelo izbeći. Za testiranje ovakvog sistema, kao i za rad sa umerenim brojem ispitanika, JSON fajlovi su se pokazali kao dovoljno obuhvatni i efikasni za čuvanje svih podataka, međutim za realne primene ovakve aplikacije sa velikim brojem ispitanika i analizom rezultata, rad bez baze podataka je gotovo nezamisliv. Kao međurešenje, takođe bi se moglo razmisliti o uvođenju neke nerelacione baze podataka, poput MongoDB, za smeštanje samih JSON fajlova, umesto korišćenja fajl sistema.

Zahvalnica

Ovaj rad je delimično finansiran od strane Ministarstva prosvete, nauke i tehnološkog razvoja Republike Srbije, projekti broj III44009 i TR32047, 2019. Autori izražavaju zahvalnost na finansijskoj podršci.

Literatura

[1] A. R. Singleton, B.C. Straits B.C. *Approaches to Social Research*, New York, Oxford University Press, 2005.
[2] Kotlin vs Java: key differences between Android's officially-supported languages, https://www.androidauthority.com/kotlin-vs-java-783187/, pristupano 20.11.2018.
[3] Single-page application vs. multiple-page application, https://medium.com/@NeotericEU/single-page-application-vs-multiple-page-application-2591588efe58, pristupano 13.03.2019.
[4] JavaScript, JSON vs XML, https://www.w3schools.com/js/js_json_xml.asp, pristupano 15.03.2019.
[5] Use the ZXing Barcode Scanner in an Android App, https://tekeye.uk/android/examples/scan-barcode-from-android-app, pristupano 09.09.2018.
[6] REST Documentation, https://spring.io/understanding/REST, pristupano 15.10.2018. - 10.02.2019.
[7] How to Dynamically Create a Component in Angular, https://dzone.com/articles/how-to-dynamically-create-a-component-in-angular, приступано 12.02.2019.

Abstract

Conducting user-defined surveys is a common process in modern business. Whether academic or marketing needs are concerned, the question arises as to how to conduct surveys in a quick and easy way. Respondents need to be reached without any redundant steps that might deter them from filling in the given questionnaire. On the other hand, data collection should be easy from the point of view of interviewer who examines a topic, without the need to use complicated systems. The prevalence of mobile devices in modern world makes thing easier for the respondents to fill in the survey. The paper presents an architecture and implementation of a system for the flexible implementation of user-defined surveys. The system consists of a mobile, web and server application, and the entire process is implemented without the use of a database and relatively modest resource use.

Architecture and Implementation of Software System for Conducting User-Defined Surveys

Ognjen Milošević, Marko Mišić, Jelica Protić

Jedno rešenje daljinskog upravljanja STB platforme putem REST protokola

Milan Gvero, Ilija Bašičević, Nikola Špirić

***Apstrakt*—Ovaj rad predstavlja jedno rešenje daljinskog upravljanja set top box (STB) platforme putem REST protokola. Rešenje je realizovano za već postojeću DTV aplikaciju te predstavlja njeno unapređenje. Rešenje omogućava korisniku da nastavi gledanje sadržaja sa mobilnog telefona, tableta i sličnih uređaja na STB uređaju tako što korisnik prevuče prstom na gore (eng. swype) dok konzumira multimedijalni sadržaj. Multimedijalni sadržaj se tada pauzira na telefonu, a nastavlja da se reprodukuje na STB uredjaju. Takođe rešenje dodaje podršku za simulaciju daljinskog upravljača, stoga korisik može da kontroliše STB putem mobilnog uređaja.**

***Ključne reči*—Android, Digitalna televizija, Set-top box (STB), Rest Protokol, UPnP Protokol**

I. UVOD

Za razumevanje ovog rešenja, potrebno je razumeti osnove digitalne televizije i android operativnog sistema. Takođe potrebno je razumeti način funkcionisanja postojeće DTV aplikacije za koju je ovo rešenje realizovano, kao i protokole korišćene za realizaciju ovog rešenja. Stoga je ovaj rad podeljen u 8 celina.

U prvoj celini dat je uvod. U drugoj celini je ukratko opisana digitalna televizija, android operativni sistem kao operativni sistem na kome se izvršava postojeća aplikacija, kao i par reči o json fajlovima. Treća celina ukratko opisuje način rada postojeće aplikacije da bi moglo da se razume šta je sve dodato u postojeću aplikaciju da bi se realizovalo ovo rešenje. Pošto je rešenje realizovano za STB uređaj, jedno od ključnih i osnovnih pitanja je pitanje vidljivosti i povezivanja STB uređaja i uređaja sa kojih se vrši akcija prebacivanja sadržaja na STB. Stoga je četvrta celina posvećena otkrivanju i povezivanju. U petoj celini će biti reči o REST arhitekturi na kojoj se bazira server, a koji predstavlja i način komunikacije mobilnog uređaja i STB-a kao i osvrt na sličan rad koji se bavio rest arhitekturom. Šesta celina je posvećena udaljenoj kontroli u kojoj je opisan sam server kao i resursi koje on podržava, dok se u sedmoj celini govori o načinu i rezultatima testiranja. U osmoj celini dat je kratak zaključak.

U ovom delu treba dati isključivo podatke o afilijaciji. Molimo NE unosite ovde zahvalnice za finansiranje. Koristite specijalno nenomerisanu sekciju Zahvalnica na kraju članka, neposredno pre liste referenci.

Kada navodite trenutnu afilijaciju autora, molimo dajte punu poštansku adresu i elektronsku poštu za sve autore.

Milan Gvero – Institut RT-RK, Narodnog fronta 23a, 21000 Novi Sad, Srbija (e-mail: milan.gvero@rt-rk.com).

Ilija Bašičević – Fakultet tehnickih nauka, Trg Dositeja Obradovica 6,, 21000 Novi Sad, Srbija (e-mail: ilibas@uns.ac.rs).

Nikola Špirić – Institut RT-RK, Narodnog fronta 23a, 21000 Novi Sad, Srbija (e-mail: nikola.spiric@rt-rk.com).

II. TEORIJSKE OSNOVE

A. Digitalna televizija

Predstavlja dominantan izvor zabave u domaćinstvima prema brojnim istraživanjima [1] sprovedenim u svetu sa udelom od skoro 80%. Razvojem tehnologije do tada dominantna analogna televizija svoje mesto je ustupila digitalnoj koja je donela značajne benefite u pogledu kvaliteta slike i zvuka, uštede električne energije, prijem slike i zvuka u pokretu, kao i bolju iskorišćenost frekvencijskog opsega. Razvojem interneta, sa stanovništva korisnika dolazi do revolucije u digitalnoj televiziji, jer televizija postaje vid dvosmerne komunikacije između digitalne televizije i njenih interaktivnih servisa sa jedne strane i korisnika sa druge. Samim tim korisnik dobija mogućnost da bira šta će da želi i kada. Takodje dolazi do novog načina distribuiranja signala, kroz već postojeću mrežnu infrastrukturu koji je definisan IPTV (eng. Internet protocol television) tehnologijom. Ovo je omogućilo još bržu ekspanzuju digitalne televizije jer se koristi već postojeća arhitektura. Postojeća aplikacija za pristup sadržaju koristi OTT (eng. Over the top) protokol koji omogućava pristup sadržaju preko internet konekcije, bez potrebe da se korisnik nalazi u mreži operatera. Takođe se koristi i DASH (eng. Dynamic Adaptive Streaming over HTTP) protokol, kao protokol koji određuju kvalitet video sadržaja u zavisnosti od brzine internet konekcije.

B. Android operativni sistem

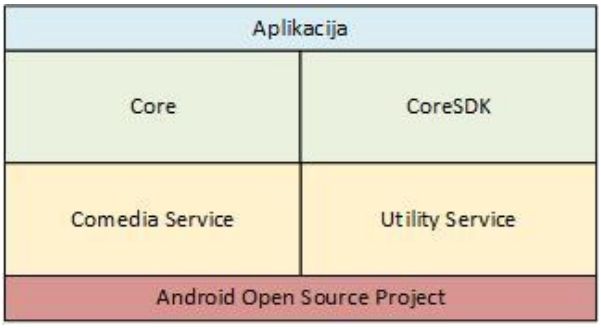
Android operativni sistem je operativni sistem razvijen od strane kompanije Gugl (eng. Google), prema kojoj je operativni sistem dostupan na preko dve milijarde uređaja [2]. Prevashodno namenjen za mobilne uređaje, trenutno je zastupljen u skoro svakom uređaju potrošačke elektornike, počevši od pametnih satova preko pametnih kuća, a u zadnje vreme se pojavljuje čak i u automobilima. Razlog za njegovu rasprostranjenost leži u činjenici da je to operativni sistem otvorenog koda (eng. Open source) dostupan pod Apache licencom. Takođe postoji velika zajednica ljudi koja održava i menja izvorni kod android-a. Programski kod android-a je dostupan preko Android Open Source Project-a, i postojeća aplikacija se upravo zasniva na AOSP projektu, zbog čega u njoj nisu implementirani servisi kompanije Gugl.

C. JSON fajl

JSON (eng. Javascript object notation) predstavlja otvoreni standard koji definiše razmenu poruka razumljivu ljudima [3].

Često se koristi za serijalizaciju i prenos struktuiranih podataka preko mrežne veze. Sastoji se od parova ključ:vrednost.

III. OPIS DTV APLIKACIJE



Slika 1. Arhitektura DTV aplikacije

Sa slike 1, može se videti, da se postojeća aplikacija sastoji iz 3 dela:

- Aplikativni deo
- Comedia servis i Utility servis, koje služe za reprodukciju MPEG_DASH sadržaja i promenu sistemska podešavanja ploče na kojoj se izvršava aplikacija.
- AOSP deo

Sam aplikativni deo je podeljen na tri dela. UI aplikacija, koja predstavlja glavni deo aplikacije koja je zadužena za iscrtavanje sadržaja po ekranu, u njoj je definisan ceo izgled aplikacije. Oslanja se na core i na coreSDK biblioteke definisane ispod nje.

Core biblioteka je biblioteka u kojoj se nalazi skup interfejsa (eng. Interfaces) i rukovaoca(eng. Handlers).

CoreSDK biblioteka je biblioteka u kojoj su realizovane implementacije interfejsa u core biblioteci, ali takođe ovde se nalazi i skup specifičnih interfejsa i njihovih implementacija koji nisu mogli biti realizovani u okviru core biblioteke.

Za potrebe ovog rešenja realizovane su sledeće klase koje implementiraju interfejs rukovaoca:

1. HttpServerHandler
2. SsdpServerHandler

Kao i klase koje predstavljaju serverske resurse.

3. CurrentMediaResponse
4. StbStatusResponse
5. remoteControllerReponse

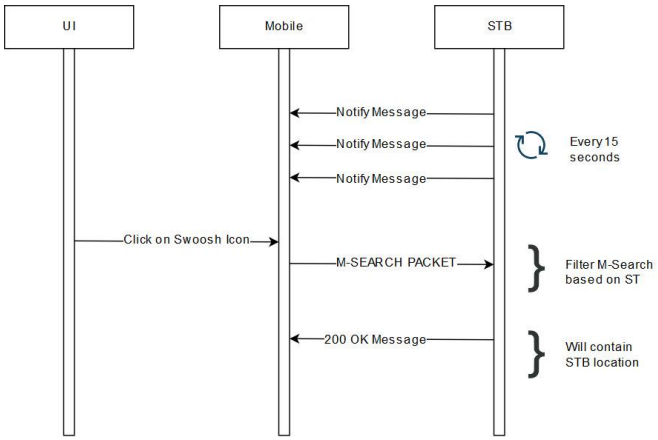
Više o ovim klasama, kao i njihovoj implementaciji biće rečeno u sledećim poglavljima.

IV. OTKRIVANJE I POVEZIVANJE UREDJAJA

Tokom reprodukcije digitalnog sadržaja, ukoliko se korisnik nalazi u istoj mreži kao i STB uređaj, na njegovom mobilnom uređaju se pojavljuje ikonica, koju ukoliko korisnik odabere dobija spisak svih STB uređaja koji se nalaze u njegovoj mreži. Klikom na STB uređaj iz liste, korisnik se povezuje na željeni STB. Povezivanje i otkrivanje je obavljeno putem UPnP protokola. UPnP protokol je skup mrežnih protokola koji omogućava automatsko pronalaženje, povezivanje i korišćenje uređaja koji su priključeni na računarsku mrežu [4]. Način povezivanja i otkrivanja je sledeći.

Nakon pokretanja STB uređaja, u fazi inicijalizacije se inicijalizuju SsdpServerHandler, koji predstavlja rukovaoc za otkrivanje i povezivanje, kao i HttpServerHandler, koji predstavlja rukovaoc za HTTP server koji se nalazi na STB uređaju. Nakon toga, STB uređaj na svakih 15 sekundi šalje NOTIFY pakete na adresu 239.255.255.250:1900, koja predstavlja multicast adresu.

U okviru svakog NOTIFY paketa nalaze se podaci vezani za STB uređaj, kao što su verzija HTTP protokola preko kojeg se šalju paketi, adresa i port na koju se šalju poruke, kao i ostali podaci definisani UPnP standardom. Sa druge strane mobilni uređaj će osluškivati NOTIFY pakete u pozadini, i ukoliko prepozna da NOTIFY paketi dolaze od strane STB uređaja, mobilni uređaj će da pošalje M-SEARCH paket na multicast adresu koju STB osluškuje. Kada se M-SEARCH paket pojavi na mreži, STB filtrira M-SEARCH paket na osnovu search-target (ST) polja u okviru M-SEARCH paketa, i ukoliko se polje podudara sa unapred definisanim stringom, STB uređaj odgovara sa 200 OK paketom u kome se nalazi adresa i resurs na koju mobilni telefon može da šalje komande. Dijagram otkrivanja i povezivanja dat je na slici 2.



Slika 2. Dijagram otkrivanja i povezivanja

V. REST ARHITEKTURA

Rest (eng. Representational State) predstavlja jednu od navažnijih i najpopularnijih arhitektura za kreiranje web aplikacija i web usluga [5]. Rest arhitektura je zvanično predstavljena u doktorskoj disertaciji Roja Fildinga 2000. godine[6]. Neki od principa Rest arhitekture su:

1. Zasnovan na resursima, gde svaki resurs ima svoj identifikator
2. Nema stanja
3. Klijent-Server arhitektura

Rest se oslanja na HTTP protokol i koristi metode HTTP protokola za pristup resursima. Neki od tih metoda su:

1. GET – Prikaz, odnosno čitanje određenog resursa.
2. POST – Kreiranje novog resursa
3. PUT – Ažuriranje ili zamena resursa.

Tokom izrade ovog rešenja korišćena je rest arhitektura kao način na koji će mobilna aplikacija da šalje zahteve na server STB uređaja, i način na koji će STB uređaj da odgovara aplikaciji.

Jedna od stvari koja se posebno ističe kod rest arhitekture, i samih restful servisa je da su veoma intuitivni kao i da se oslanjaju na već postojeću HTTP infrastrukturu. Takođe su veoma prilagodljvi jer način komunikacije ne mora da bude samo JSON fajl, a zbog mogućnosti keširanja povećava efikasnost čitavog sistema. O prednostima REST arhitekture postoji mnogo radova a naveo bih rad mojih kolega [8], koji su imali zadatak da objedine podatke iz DTV transportnog toka sa podacima sa interneta, te da te objedinjene podatke prikazuju klijentu. Oni su kao rešenje ponudili RESTful API koji se bazira na REST arhitekturi, i u svom radu su zaključili da je upravo REST arhitektura zbog svih njenih navedenih prednosti bila bolji izbor od npr. WSDL (eng. Web Service Description Language) kombinovanog sa SOAP (eng. Simple Object Access Protocol) protokolom.

VI. UDALJENA KONTROLA

Samo otkrivanje i povezivanje je opisano u poglavlju 4. Nakon toga, ukoliko korisnik izvrši swoosh akciju (koja predstavlja prevlačenje prstom na gore dok reprodukuje multimedijalni sadržaj), tada mobilni uređaj na IP adresu STB uređaja, a na resurs currentMediaResponse putem POST metode šalje JSON fajl koji u sebi sadrži sledeće podatke.

1. item_type – Predstavlja tip multimedijalnog sadržaja koji korisnik želi da reprodukuje. Može biti tv kanal, ili video na zahtev.
2. Playback_type – Predstavlja tip reprodukcije. U slučaju tv kanala, može da bude emisija emitovana u prošlosti (eng. catchUP), trenutna emisija ali pokrenuta ispočetka (eng. startOver), ili trenutna emisija (eng. Live). U slučaju videa na zahtev ovaj parametar se ignoriše.
3. programID – Predstavlja jedinstveni broj emisije za koju je zatražena reprodukcija.
4. fileID – Jedinstveni identifikator fajla kojeg želimo da pustimo na bekendu.
5. MediaMark – Broj koji predstavlja od koje sekunde treba da nastavimo reprodukciju. Samo za video na zahtev.
6. UserID – Jedinstveni identifikator korisnika na bekend strani.

Server koji se koristi u izradi ovog rešenja je definisan u HttpServerHandler klasi. Sama podrška za server se ispod haube oslanja na restlet biblioteku[9]. U HttpServerHandler klasi se samo definiše novi server kojeg implementira ova klasa, kao i dostupni resursi. Za svaki resurs koji server podržava definisana je odgovarajuća klasa.

Zadatak currentMediaResponse klase je da odreaguje na POST metodu koju će da prozove mobilni uređaj. Potrebno je isparsirati json fajl, proveriti validnost podataka u njemu, npr. Da li je trenutni id ulogovanog korisnika isti kao i userId koji je stigao u json fajlu, da li id fajla postoji na bekendu, itd. Zatim je potrebno proveriti da li se STB nalazi u stanju iz kojeg je moguće pustiti medija fajl. Ukoliko sve provere prođu uspešno, potrebno je pustiti traženi fajl i vratiti status kod 201, a u slučaju greške je potrebno vratiti status kod 401 i odvarajući json fajl sa porukom o grešci.

Još jedan od resursa koje imamo na serveru je stbStatusResponse. Naime, mobilni uređaj će svake sekunde slati zahtev za statusom STB uređaja, pa je njegov zadatak da odreaguje na ove zahteve. U status STB uređaja ulaze sledeći podaci:

1. userId
2. deviceId
3. playback – play/pause/none
4. assetId
5. subtitles – lang/none
6. audiolang – lang/none
7. epgstatus – epgEventId
8. mediaMark
9. parentalStatus
10. tvVolume

Poslednji resurs dostupan na serveru je remoteControllerReponse. On je zadužen za simulaciju daljinskog upravljača, tako što mobilni telefon preko rest protokola šalje json fajl sa željenim tasterom koji se simulirao na mobilnom telefonu. Zadatak ovog resursa je da odreaguje na POST zahtev, te da simulira pritisak određenog tastera na daljinskom upravljaču. Ovo je urađeno tako što se prvo isparsira parametar iz json fajla, i mapira na određenu vrednost iz KeyEvent klase. Nakon toga se koristi metoda *sendKeyDownUpSync* iz *Instrumentation* klase kojom se simulira odgovrajuće dugme na daljinskom upravljaču.

VII. NAČIN I REZULTATI TESTIRANJA

Testiranje je obuhvatalo dve faze. U prvoj fazi se obavljalo testiranje otkrivanja i povezivanje mobilnog uređaja i STB uređaja, dok je u drugoj fazi testirano izvršavanje komandi sa mobilnog uređaja, koje je obuhvatalo i puštanje multimedijalnog sadržaja i daljinsku kontrolu STB uređaja. tj simulaciju daljinskog upravljača. U prvoj fazi, rezultati pokazuju da se mobilni uređaj svaki put uspešno poveže na STB uređaj. Izvršeno je preko 20 merenja, i svaki put kada bi se mobilni uređaj povezao na istu mrežu kao STB pojavila bi se ikonica za swoosh opciju, i nakon klika na swoosh ikonicu, mobilni uređaj bih poslao M-SEARCH paket, koji bi STB prepoznao i odgovorio sa 200 OK porukom. Za posmatranje paketa kroz mrežu korišćen je wireshark program [6]. U drugoj fazi testirana je većina slučajeva kad STB može i kad STB ne može da pusti multimedijalni sadržaj. Nisu pokriveni svi slučajevi jer je varijacija stanja u kojima STB može da se nađe zaista prevelika. Za testirane slučajeve STB je svaki put vratio odgovarajući kod (201 ili 401). Da bismo bili sigurni da rešenje radi svaki put, testirali smo i sa pogrešno poslatim podacima, gde npr. korisnički id broj sa mobilne aplikacije nije odgovara identifikacionom broju korisnika koji je ulogovan na STB, kao i slanjem drugih netačnih podataka, ili čak praznih podataka. Ovim je potvrđeno da je STB u stanju da odreaguje na pogrešno poslate podatke, iako će mobilni uređaj svaki put slati prave podatke.

VIII. Zaključak

U ovom radu predstavljeno je unapređenje već postojeće DTV aplikacije podrškom za daljinsku kontrolu koja u velikoj meri poboljšava korisničku iskustvo i prelazak sa uređaja manjeg ekrana na uređaje većeg ekrana sa kojih je mnogo bolji doživljaj konzumacije digitalnog sadržaja.

Aplikacija sa uređaja manjeg ekrana i STB aplikacija se otkrivaju i povezuju putem UPnP protokola, dok se sama razmena podataka odvija preko json poruka koje klijent šalje na odgovarajući resurs na serveru, a server odgovara u skladu sa primljenim porukama i lokalnom obradom podataka.

Rezultati pokazuju da se aplikacija ponaša u skladu sa zahtevima specifikacije.

Mogućnosti za dalji rad na ovom proširenju su velike, jer je vrlo lako proširiti serversku stranu dodatnim resursima.

Zahvalnica

Ovaj rad je delimično finansiran od strane Ministarstva za prosvetu, nauku i tehnološki razvoj Republike Srbije, na projektu broj: TR32030.

Literatura.

[1] Istraživanje – televizija kao dominantan izvor zabave, https://www.digitalcenter.org/web-insight-views-aboutentertainment-sources-2010/, 2010.
[2] Procena kompanije Gugl o broju uređaja koji koriste Android OS, https://www.theverge.com/2017/5/17/15654454/android-reaches-2billion-monthly-active-users, 2017.
[3] Javascript object notation, http://www.json.org/, april 2019.
[4] UPnP Device Architecture 2.0, UPnP Forum, 20. Februar, 2015
[5] Rest Arhitektura, https://restfulapi.net/rest-architectural-constraints/, april 2019
[6] Roy Thomas Fielding, Architectural Styles and the Design of Network-based Software Architectures, doktorska disertacija, University of California, Irvine, SAD, 2000
[7] Wireshark alat, https://www.wireshark.org/, april 2019.
[8] S. Pijetlović, N. Jovanov, V. Vukobrat and I. Basicevic, "One solution of a RESTful API for a cloud based DTV content provider," *2014 IEEE Fourth International Conference on Consumer Electronics Berlin (ICCE-Berlin)*, Berlin, 2014, pp. 384-387.
[9] Restlet biblioteka, https://restlet.com/, april 2019.

Abstract

This paper presents one solution for the remote control of the STB platform through the REST protocol. The solution has been implemented for the already existing DTV application and represents its improvement. The solution allows the user to continue watching content from a mobile phone, tablets and similar devices on the STB device by swiping up while consuming multimedial content. Multimedial content is then paused on the phone, and continues to play on STB device. Also, the solution adds support for simulation of the remote control, so that user can control the STB via a mobile device.

One solution for remote control of STB platform via REST protocol

Milan Gvero, Ilija Bašičević, Nikola Špirić

Unapređenje arhitekture sloja za apstrakciju fizičke arhitekture DTV srednjeg sloja

Lara Milovanović, Miroslav Bako, Milan Savić

***Apstrakt*—DTV srednji sloj je najvažniji deo programske podrške za TV prijemnike, koji sadrži gotovo kompletnu DTV funkcionalnost i omogućava izvršavanje aplikacija. Zavisnost DTV srednjeg sloja od fizičke arhitekture ciljane platforme oduvek je predstavljala barijeru u rayvoju DTV srednjeg sloja. U ovom radu opisan je problem prenošenja DTV srednjeg sloja, odnosno sloja za apstrakciju fizičke arhitekture, kao i unapređenje arhitekture ovog sloja. Datim unapređenjem omogućeno je da DTV srednji sloj podržava više HAL API-ja istovremeno, što u velikoj meri doprinosi portabilnosti DTV srednjeg sloja, i omogućava razvijanje hibridnog STB (Set-top box) uređaja, koji može da emituje OTT sadržaj, kao i sadržaj dobijen iz DVB transportnog toka.**

***Ključne reči*—STB; DTV; Android; DTV srednji sloj; OTT; HAL; SoC; Media Codec; Hibridni STB;**

I. Uvod

DTV srednji sloj je je najvažniji deo programske podrške za TV prijemnike, koji sadrži gotovo kompletnu DTV funkcionalnost i omogućava izvršavanje aplikacija.

Važna osobina DTV srednjeg sloja je da on apstrahuje funkcionalnosti DTV uređaja, fizičke arhitekture i operativnog sistema, i na taj način obezbeđuje da proizvođač aplikativne DTV programske podrške ne mora da bude upoznat sa specifičnostima fizičke arhitekture.

Različiti proizvođači hardverskih komponenata imaju različit mehanizam upravljanja (spregu) nad ovim komponentama. Iz tog razloga se javlja glavni problem pravljenja univerazalnog DTV srednjeg sloja koji može da funkcioniše na svim platformama.

Kako bi se omogućila DTV funkcionalnost na određenoj platformi, neophodno je preneti sloj za apstrakciju fizičke arhitekture (eng. Hardware Abstraction Layer - HAL) srednjeg sloja na ciljanu platformu.

Sloj za apstrakciju fizičke arhitekture predstavlja programski sloj u kome se vrši apstrakcija pristupa ka rukovaocima hardverskih komponenti (eng. drivers). Prilikom prenošenja DTV srednjeg sloja na određenu platformu, korišćenjem HAL-a se postize da DTV srednji sloj ne mora da se menja, već se sve promene koje su vezane za ciljanu platformu dodaju u HAL.

S obzirom na prethodno pomenute probleme, cilj ovog rada je opis problema prenošenja DTV srednjeg sloja, odnosno sloja za apstrakciju fizičke arhitekture, kao i unapređenje arhitekture ovog sloja. Datim unapređenjem omogućeno je da DTV srednji sloj podržava više HAL API-ja istovremeno, što u velikoj meri doprinosi portabilnosti DTV srednjeg sloja, i omogućava razvijanje hibridnog STB uređaja, koji može da emituje OTT sadržaj, kao i sadržaj iz DVB transportnog toka.

Lara Milovanović, Nacionalni institut za nauku i razvoj RT-RK, Narodnog fronta 23A, 21000 Novi Sad, Srbija (e-mail: Lara.Milovanovic@rt-rk.com).

Miroslav Bako, Nacionalni institut za nauku i razvoj RT-RK, Narodnog fronta 23A, 21000 Novi Sad, Srbija (e-mail: Miroslav.Bako@rt-rk.com).

Milan Savić, Nacionalni institut za nauku i razvoj RT-RK, Narodnog fronta 23A, 21000 Novi Sad, Srbija (e-mail: Milan.Savic@rt-rk.com).

II. Teorijske osnove

A. Android TV

Android TV je proširenje osnovne Android platforme razvijeno od strane Google-a, posebno za televizijske i samostalne multimedijalne uređaje [1].

U okviru Android TV platforme, Google nudi skup GTVS (eng. Google TV Services) [2] aplikacija. GTVS aplikacije su skup aplikacija razvijenih od strane kompanije Google sa ciljem da se poboljša iskustvo korišćenja Android TV platforme. Neki od pripadnika GTVS aplikacija i servisa su unapred instalirane na sistemsku particiju Android TV-a, dok korisnik može da instalira i ostale koje se nalaze u Google-ovoj prodavnici aplikacija (eng. Google Play Services). Neke od aplikacija razvijanih od strane GTVS-a, koje su unapred instalirane (preintegrisane) na sistemsku particiju Android TV-a su:

- "Leanback Launcher" - pokretač aplikacija ya Android TV
- "Google Asssistant" - pomoć i pretraga
- "Play Store" - zvanična Google prodavnica Android aplikacija
- "Play Movies" - aplikacija za gledanje filmova i drugih TV emisija
- "You Tube" - aplikacija za promovisanje video sadržaja

Da bi proizvođači Android uređaja dobili dozvolu da preintegrišu GTVS aplikacije, Android TV platforma mora da prođe Google-ov sertifikacioni proces.

B. Android programska podrška za TV prijemnike

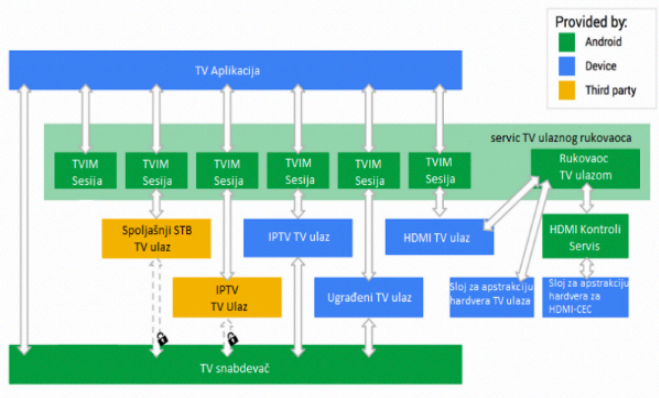
Android programska podrška za TV prijemnike (eng. TV Input Framework – TIF) [3] definiše aplikativnu programsku spregu i standardizuje način implementacije dopremanja emitovanog sadržaja do TV aplikacije. TIF omogućuje implementaciju TV ulaza preko kog je moguće reprodukovati sadržaj sa mrežnog poslužioca i/ili iz DVB transportnog toka

podataka, kao i pretragu televizije uživo. Programska podrška ne teži ka tome da implementira TV standarde ili regionalne zahteve, ali proizvođačima uređaja omogućava da lakše ispune regionalne digitalne TV standarde bez ponovne implementacije.

C. Komponente Android programske podrške za TV prijemnike

TIF se sastoji od:

- TV aplikacija (eng. TV Application) – aplikacija pomoću koje korisnik rukuje hibridnom programskom podrškom za TV prijemnike
- TV snabdevač (eng. TV Provider) – baza podataka sa kanalima, programima I pratećim dozvolama
- TV ulaz (eng. TV Input) – aplikacija koja predstavlja jedan izvor TV sadržaja
- Ulazna TV fizička arhitektura (eng. TV Input Hardware Abstraction Layer) – fizička definicija arhitekture koja omogućuje sistemskim TV ulazima pristup fizičkoj arhitekturi specifičnoj za TV prijemnike
- Roditeljska kontrola (eng. Parental Control) – tehnologija koja omogućava blokiranje kanala I programa
- HDMI-CEC – tehnologija koja omogućava daljinsku kontrolu raznih uređaja putem HDMI-CEC poruka



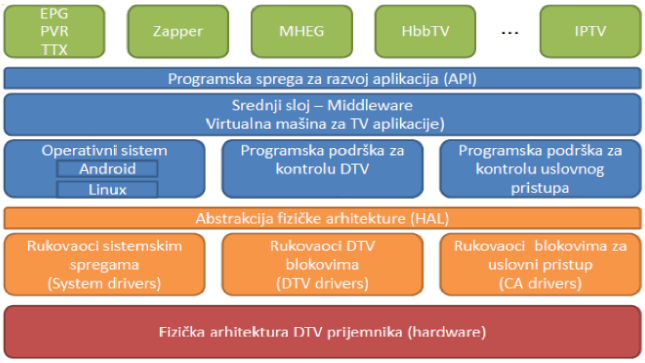
Sl. 1. TIF arhitektura

Na osnovu arhitekture sa slike se može zaključiti:

- Korisnik može da vidi i bude u direktnom kontaktu sa TV aplikacijom
- TV aplikacija prikazuje sadržaj sa TV ulaza
- TV aplikacija nije u mogućnosti da direktno komunicira sa TV ulazima, već je komunikacija ostvarena pomoću rukovaoca TV ulazom koji identifikuje stanje TV ulaza za TV aplikaciju

D. Programska podrška DTV prijemnika i koncept DTV srednjeg sloja

Arhitektura programske podrške za TV prijemnike je slojevita, gde svaki sloj ima posebnu naemenu. Primer uobičajenog rasporeda programskih slojeva TV prijemnika može se videti na Sl. 2.



Sl. 2. Programska podrška DTV prijemnika

Najvažniji deo programske podrške, koji sadrži kompletnu DTV funkcionallnost i omogućava izvršavanje aplikacija je DTV srednji sloj (eng. Middleware). Uloga srednjeg sloja je da realizuje najvažnije operacije kao što su raščlanjavanje DVB podataka, kontrola pristupa i organizacija servisa, prikupljanje podataka o programima (eng. Event Information Table – EIT), podrška za snimanje, kontrola podsetnika, reprodukcija (dekodovanje) multimediijalnih sadržaja, kao i da kontroliše fizičku arhitekturu kroz komunikaciju sa nižim slojevima i da obezbedi potrebne programske sprege ka višim programskim slojevima.

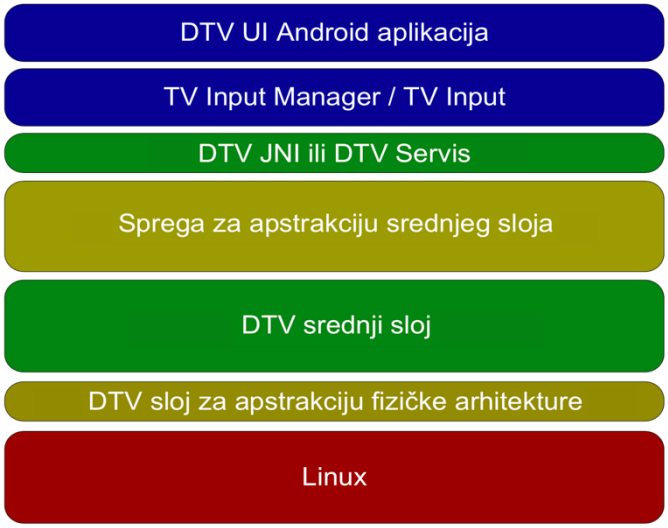
Važna osobina srednjeg sloja je da on apstrahuje funkcionalnosti DTV uređaja, fizičke arhitekture i operativnog sistema, i na taj način obezbeđuje da proizvođač aplikativne DTV programske podrške ne mora da bude upoznat sa specifičnostima fizičke arhitekture. Drugim rečima, srednjii sloj formira virtuelnu celinu na kojoj se izvršava aplikativni deo programske podrške. Ovo takođe garantuje identično izvršavanje iste aplikacije na različitim fizičkim arhitekturama ukoliko se koristi isti srednji sloj.

E. Arhitektura DTV srednjeg sloja

DTV srednji sloj sastoji se iz tri osnovne celine [4]:

- Jezgro srednjeg sloja – zaduženo za realizaciju osnovnih DTV funkcionalnosti i nezavisno je od fizičke platforme
- Sloj za apstrakciju srednjeg sloja (eng. Middleware Abstraction layer, MAL) – vrši enkapsulaciju svih funkcionalnosti srednjeg sloja kroz programsku spregu koja omogućava laku integraciju sa slojevima višeg nivoa
- Sloj za apstrakciju fizičke arhitekture (eng. Hardware Abstraction Layer, HAL) – sadrži funkcionalnost koja je direktno vezana za fizičku arhitekturu. Da bi se omogućila funkcionalnost DTV srednjeg sloja na određenoj platformi, potrebno je ovaj sloj preneti na ciljanu platformu.

F. *Veza Android aplikacije i DTV srednjeg sloja*



Sl. 3. Veza Android aplikacije i DTV srednjeg sloja

- Funkcionalnosti srednjeg sloja uobličene su u okviru DTV klase koja obezbeđuje standardizovan Java API za pisanje DTV aplikacija.
- U okviru implementacije TV Input klase, koriste se pozivi DTV klase koja daje na raspolaganje sve osnovne elemente DTV srednjeg sloja
- DTV JNI povezuje Java sloj sa nativnim kodom (korišćenjem standardnih Java Native Interface mehanizama).
- Implementacija DTV klase u Javi poziva odgovarajuće C funkcije.
- DTV JNI omogućava pozivanje svih MAL API funkcija.
- MAL API omogućava pristup svim funkcionalnostima srednjeg sloja.
- Srednji sloj realizuje DTV funkcionalnosti.
- Sprega za apstrakciju fizičke arhitekture realizuje konkretnu komunikaciju sa fizičkim rukovaocima DTV blokovima.

G. SoC API i *Media Codec*

SoC (eng. System on a chip) API predstavlja spregu za upravljanje rukovaocima fizičkih komponenti (drajverima) koji je svaki proizvođač fizičke arhitekture dužan da obezbedi kako bi DTV srednji sloj mogao biti prilagođen za izvršavanje na određenoj platformi. SoC API je uglavnom pisan u jezicima niskog nivoa i zahteva da aplikacije koje ga pozivaju takođe budu niskog nivoa.

Media Codec [5] predstavlja modul univerzalne Android programske sprege. Omogućava pristup za rukovanje komponentama fizičke arhitekture ali putem prethodno pomenute univerzalne Android programske sprege koja je dostupna na svim Android uređajima. Proizvođači hardverskih komponenti su dužni da omoguće da ova programska podrška radi na njihovim platformama. Komponente kojima se može upravljati kroz Media Codec su A/V dekoderi, enkoderi, i slično. Pored upravljanja ovim komponentama, Media Codec takođe nudi spregu sa hardverskim blokovima za dekripciju sadržaja tako da time podržava ne samo rad sa nezaštićenim, već i sa zaštićenim multimedijalnim sadržajem.

III. Opis problema

Da bi se omogućilo izvršavanje aplikacija i kompletna DTV funkcionalnost na određenoj platformi, potrebno je preneti sloj za apstrakciju fizičke arhitekture DTV srednjeg sloja na ciljanu platformu. Svaki proizvođač fizičke platforme dužan je da obezbedi API za rukovanje fizičkom arhitekturom. Sa druge strane, da bi se omogućilo rukovanje fizičkom arhitekturom, HAL sloj DTV srednjeg sloja dužan je da implementira API dobijen od strane proizvođača, bez potrebe za poznavanjem detalja realizacije te arhitekture.

Postoje različite promene DTV srednjeg sloja shodno metodi dopremanja TV sadržaja. Na primer, moguće je da na jednom STB uređaju DTV srednji sloj služi za reprodukciju DVB sadržaja, dok na drugom STB uređaju služi za reprodukciju sadržaja dobijenog preko mreže. STB uređaji sa pristupom internetu omogućili su korišćenje OTT sadržaja poput:

- TV-sadržaja koji se doprema posredstvom interneta (stream ili download) - Internet TV
- Socijalnih internet servisa
- Web servisa opšte namene

Omogućavanjem pristupa internetu bilo je potrebno obezbediti adekvatnu zaštitu datog sadržaja radi dalje distribucije. Kako bi se takav sadržaj mogao reprodukovati, potrebno ga je dekriptovati.

Upravljanje digitalnim pravima (eng. Digital Rights Managment – DRM) [6] je sistematski pristup zaštiti autorskih prava za digitalne medije. Svrha DRM-a je da spreči neovlaštenu redistribuciju digitalnih medija i ograniči načine na koje korisnici mogu kopirati sadržaj koji su kupili.

Drajveri nemaju direktnu podršku za dešifrovanje zaštićenog sadržaja,zbog čega bi se za svaku arhitekturu platforme morao praviti poseban mehanizam za dešifrovanje. Međutim, Android je pružio univerzalno rešenje za dešifrovanje zaštićenog sadržaja u vidu *Media Codec* API-ja koji je nezavisan od fizičke arhitekture platforme.

IV. Koncept rešenja

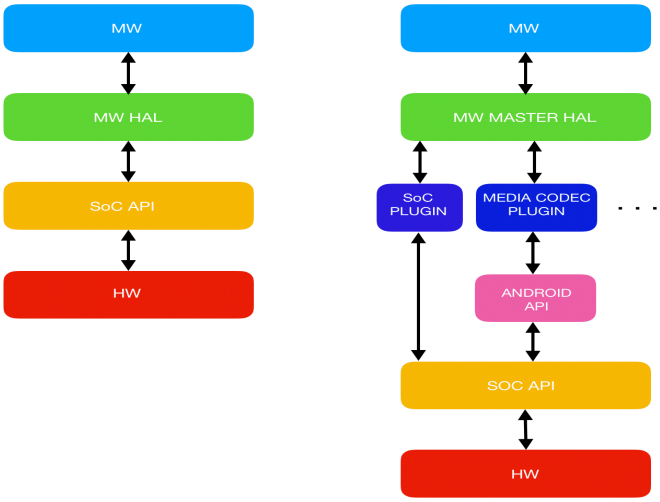
Sa pojavom Android-ovog *Media Codec* mehanizma, proces razvijanja OTT uređaja znatno je ubrzan, jer je za apstrakciju HAL sloja korišćen Android-ov API.

Media Codec API je univerzalan za sve fizičke platforme, što znači da DTV srednji sloj koji koristi *Media Codec* mehanizam za dopremanje TV sadržaja predstavlja univerzalno rešenje OTT middleware-a, nezavisno od fizičke arhitekture.

Sledeći izazov u unapređenju DTV aplikacija podrazumeva razvoj hibridnog STB uređaja koji koristi Android-ov *Media Codec* API za OTT sadržaj, dok za dobavljanje i reprodukciju sadržaja iz DVB transportnog toka koristi SoC API za datu platformu.

Ovo je postignuto podelom arhitekture DTV sloja za apstrakciju fizičke arhitekture na dva sloja, jedan ispod DTV srednjeg sloja koji služi kao glavni HAL modul, i ostali HAL moduli u vidu umetka (eng. plugin) koje kontroliše glavni

HAL. Korišćenjem ovakve arhitekture, za svaki SoC API se može nezavisno napraviti i testirati HAL plugin.



Sl. 4. Ilustracija stare i nove arhitekture DTV sloja za apstrakciju harvera

Plugini su realizovane kao deljene programske biblioteke. Glavni HAL modul pretražuje sve dostupne plugine u sistemu i dinamički ih učitava u radnu memoriju u toku izvršavanja DTV srednjeg sloja. Nakon učitavanja plugina u radnu memoriju mapiraju se pozivi ka funkcijama na određene memorijske adrese u okviru učitanih biblioteka. Tokom rada, DTV srednji sloj identifikuje koji sadržaj se reprodukuje i na osnovu toga se pozivaju zahtevane funkcije iz HAL-a koji se trenutno koristi za reprodukciju sadržaja. Ovakav mehanizam, pored osnovnog zadatka DTV srednjeg sloja koji je reprodukcija sadržaja, podržava i ostale napredne funkcionalnosti kao što su PVR, EPG. Pored toga, DTV srednji sloj nije ograničen da u jednom trenutku radi samo sa jednim HAL-om, nego je moguće da radi istovremeno sa svim učitanim HAL pluginima i jedino ograničenje predstavlja sama dostupnost fizičkih resursa ciljanog uređaja.

Ovakva arhitektura koja omogućava postojanje više HAL plugina istovremeno, pružila je rešenje za razvoj hibridnog STB uređaja koji koristi *Media Codec* plugin za OTT sadržaje, dok za DVB sadržaj razvija *SoC* plugin u zavisnosti od fizičke arhitekture ciljane platforme.

V. Zaključak

Zavisnost DTV srednjeg sloja od fizičke arhitekture ciljane platforme oduvek je predstavljala barijeru u razvoju DTV srednjeg sloja. Nova arhitektura DTV sloja za apstrakciju fizičke arhitekture koja je omogućila je korišćenje plugin-a, u velikoj meri doprinela je portabilnosti DTV srednjeg sloja, jer prenošenje na drugu platformu zahteva samo razvijanje SoC plugin-a za datu arhitekturu, bez potrebe za drugim izmenama u DTV srednjem sloju. Takođe, STB koji koristi *Media Codec* plugin, se vrlo brzo može transformisati u hibridni STB dodavanjem HAL plugin-a za SoC API.

VI. Zahvalnica

Ovaj rad je delimično finansiran sredstvimaMinistarstva za nauku, tehnologiju i razvoj Republike Srbije preko projekta TR32041.

Literatura

[1] Android TV, https://www.android.com/tv, april 2019
[2] Android –Google TV Services, http://www.androd.com/gtvs/, april 2019
[3] Tv Input Framework, https://source.android.com/devices/tv, april 2019
[4] Predavanja iz predmeta Programska podrška u televiziji i obradi slike 2, http://www.rt-rk.uns.ac.rs/studijski-program-2009/ix-2009/pputvios2/761-pputvios-2-predavanja
[5] Android – MediaCodec https://developer.android.com/reference/android/media/MediaCodec, april 2019
[6] DRM - https://www.geoguard.com/drm/, april 2019

Abstract

The DTV middleware is the most important part of the software for TVs, which contains almost complete DTV functionality and allows execution of applications. The dependence of the DTV middle layer on the physical architecture of the target platform has always been a barrier in the DTV medium layer.

The aim of this paper is to describe the problem of porting the DTV middleware, precisely hardware abstraction layer, as well as the improvement of the architecture of this layer. By enhancing this feature, DTV middleware supports multiple HAL APIs at the same time, which greatly contributes to the portability of the DTV middle layer, and allows the development of a hybrid STB (Set-top box) device that can transmit OTT content as well as the content obtained from the DVB transport stream.

Improvement of the architecture of hardware abstraction layer in DTV middleware

Lara Milovanović, Miroslav Bako, Milan Savić

Primena tehnologije Google Assistant u interaktivnoj digitalnoj televiziji

Aleksandar S. Lazić, Milan Z. Bjelica, *Member, IEEE*, Veljko Lj. Ilkić i Dejan Đ. Nađ

***Apstrakt*— U današnje vreme, sve više i više korisnika koristi usluge virtuelnih asistenata na različitim platformama (Android, Windows, iOS) i oni postaju sve zastupljeniji u svakodnevnim životnim aktivnostima, bilo da je to asistencija u kupovini namirnica, u vožnji do posla ili pak, u svrhu zabave korisnika. U okviru ovog rada istražena je i realizovana mogućnost primene tehnologije Google asistent u interaktivnoj digitalnoj televiziji na prijemniku zasnovanom na operativnom sistemu Android. Jedan od ciljeva je unapređenje korisničkog iskustva u vidu komplementiranja osnovnih funkcionalnosti poput izmene kanala ili nivoa jačine zvuka glasovnim komandama, i dodatnim mogućnostima izbora tačnog kanala, događaja ili video sadržaja na zahtev sa mogućnošću trenutnog ili naknadnog gledanja, zakazivanja snimanja i dr. Velika većina alata za prepoznavanje glasovnih komandi generiše rezultate naredbe u vidu slobodnog teksta ili struktuiranog objekta. Kako bi se taj rezultat iskoristio u postojećim TV aplikacijama, neophodno je detektovati šablone koji odgovaraju gore pomenutim naredbama.**

***Ključne reči*—Android, Google Assistant, Digitalna televizija, glasovna komanda, integracija, korisničko iskustvo, Actions on Google, Dialogflow, Set-top box (STB).**

I. Uvod

Digitalna televizija i video sadržaj su danas među najzastupljenijim vidovima medijske komunikacije sa korisnicima [1]. Sama televizija jeste komunikacioni medijum za slanje i prihvatanje pokretnih slika i zvuka. U poslednje dve decenije doživljava veliku ekspanziju u svetu, pre svega slanjem analognog TV prijemnika u istoriju, a zatim i razvojem softverske podrške i realizacijom prvog TV uređaja sa Android operativnim sistemom [2] u junu 2014. godine. Prema trenutnim istraživanjima, televizija uz Internet predstavljam dominantan izvor zabave u svetskoj populaciji [3], pre svega kada je reč o korisnicima srednjeg i starijeg životnog doba. Ovome doprinose i novi operativni sistemi na korisničkim uređajima, kao što je Android.

Android je najrasprostranjeniji operativni sistem u svetu za namenske uređaje poput pametnih telefona, tableta, satova, televizora i drugih kućnih uređaja. Glavna Androidova karakteristika je da je to sistem „otvorenog koda“ [4], tj. postoji zajednica programera gde svako može da ga izmeni, unapredi prema svojim potrebama i podeli sa drugim inženjerima. Arhitektura Androida može se podeliti u 5 celina, od najniže ka najvišoj:

- Sloj jezgra (Kernel);
- Sistemske (nativne) biblioteke;
- Izvršno radno okruženje;
- Okruženje za razvoj aplikacija;
- Aplikativni sloj;

Za realizaciju u ovom radu, poslednja dva sloja su bitna jer je to ono što korisnik vidi dok upravlja TV prijemnikom. Između ta dva sloja, odnosno između srednjeg sloja (*Middleware*) i aplikativnog sloja, nalazi se, specijalno dizajnirano za Android u televiziji, radno okruženje poznatije kao TV Input Framework (TIF) [5]. TIF omogućava manipulisanje bazom podataka u kojoj su smešteni TV servisi, standardizuje dopremanje sadržaja do TV aplikacije [6], omogućava pretragu kanala uživo, njihovo reprodukovanje iz DVB transportnih tokova itd. Programska podrška ne teži ka tome da implementira nove TV standarde, ali proizvođačima omogućava da lakše ispune regionalne digitalne TV standarde, bez ponovne implementacije.

Na današnjim aplikativnim platformama važnu ulogu zauzimaju virtuelni asistenti. Za svaku od platformi je karakterističan neki virtuelni asistent. Na primer, ukoliko korisnik koristi *Apple*-ove uređaje, asistent će mu biti *Siri*, Windows korisnicima je na raspolaganju *Cortana*, dok je za Android (a može se koristiti i na drugim platformama) namenjen Google asistent [7]. On poseduje veštačku inteligenciju i omogućava dvosmernu komunikaciju sa korisnikom u vidu govorne komunikacije ili pisanim putem. Izuzetno je dobro obučen, ima pristup Internetu, te konverzacija sa njim ostavlja utisak razgovora sa stvarnim čovekom. Osim zanimljivih konverzacija, Google asistent omogućava još pregršt korisnih funkcionalnosti posebno kada je reč o pametnim kućama (*Internet of things* [8]) gde je korisniku omogućeno da upravlja osvetljenjem, temperaturom i drugim parametrima u svom domu, koristeći glasovne komande. Po ugledu na to, slične akcije su omogućene i u integraciji asistenta u TV aplikaciji. Google asistent poseduje posebne platforme koje omogućavaju korisniku da osmisli bilo kakvu akciju, uz minimalno obučavanje asistenta i obradu podataka da bi se akcija izvršila.

Ključne reči za pozivanje asistenta su „*Hey Google“* ili *„Ok Google“*, dok, ukoliko korisnik želi da koristi mogućnosti implementirane u ovom radu, potrebno je da započne konverzaciju sledećom frazom – *„Talk to Stage application“*. Ova fraza je podložna promenama u zavisnosti

Aleksandar Lazić – Institut RT-RK, Narodnog fronta 23a, 21000 Novi Sad, Srbija (e-mail: aleksandar.lazic@rt-rk.com).

Milan Z. Bjelica – Institut RT-RK, Narodnog fronta 23a, 21000 Novi Sad, Srbija (e-mail: milan.bjelica@rt-rk.com).

Veljko Ilkić – Institut RT-RK, Narodnog fronta 23a, 21000 Novi Sad, Srbija (e-mail: veljko.ilkic@rt-rk.com).

Dejan Nađ – Institut RT-RK, Narodnog fronta 23a, 21000 Novi Sad, Srbija (e-maill:dejan.nadj@rt-rk.com).

od korisničkih podešavanja ili imena same TV aplikacije.

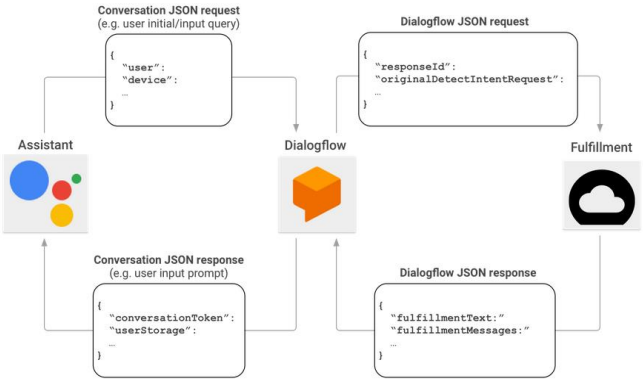
U okviru rada realizovane su komande za prikazivanje liste dostupnih kanala i filmova, komande za podatke o trenutnom kanalu, događaju na trenutnom kanalu, narednom događaju, prikaz događaja na određenom kanalu u određeno vreme kao i mogućnosti zakazivanja snimanja i podsetnika za gledanje događaja.

II. Prethodni rad i istraživanja

Upotreba Google asistenta i njegova integracija u aplikacije spada u rešenja prisutna prethodnih nekoliko godina. Kako je njegova popularnost uhvatila zamah, tako su inženjeri, ali i obični entuzijasti počeli da isprobavaju mogućnosti Google asistenta, uglavnom u svrhu zabave. Takođe, Google asistent je mnogo bolje integrisan i funkcionalniji u uređajima poput pametnih telefona, gde nudi mnoštvo raznolikih razgovora o brojnim temama iz svih sfera života i drugih mogućnosti, nego u pametnim televizorima ili STB uređajima, gde je konverzacija ograničena mogućnostima TV uređaja, te je i podrška zajednice za asistenta u TV aplikacijama utoliko „siromašnija" nego za telefone ili tablete.

Kako se Google asistent razvijao, tako su se razvijale i platforme koje omogućavaju svakom čoveku da napravi neku vrstu konverzacije sa Google Assistant-om. Postoje nekoliko platformi koje se koriste za to, u zavisnosti od potreba, složenosti naredbi, na raspolaganju su: *Actions on Google* i *Dialogflow* [9].

U prošlosti je na sličnu temu pisan naučni rad [10], gde je akcenat bio na izvršavanju zadate akcije, dok deo vezan za interakciju sa korisnikom nije bio od vitalne važnosti. Google asistent je obezbeđivao deo odgovora koji je statički, dok se drugi deo menjao u zavisnosti od zadate korisničke komande. U ovom radu se napravio iskorak, te akcenat nije više samo na izvršavanju akcije, već i na interaktivnom dijalogu između korisnika i asistenta. Osim osnovnih korisničkih naredbi za promenu kanala ili jačine zvuka, korisnik ima mogućnost da asistentu postavi pitanje, zatraži podatke o događajima, odnosno da se uspostavi interakcija između istih. Drugim rečima, asistent nije više prividno obučen za konverzaciju, već dobija veštačku inteligenciju i u pozadini rukuje željenim podacima, vraća korisniku smisleni odgovor ili opcije koje može da izabere, ali i izvršava željenu akciju i rezultat se može videti u samoj aplikaciji, prikazivanjem odgovarajućeg sadržaja. Iz tog razloga, za ovaj rad se nije koristila samo platforma *Actions on Google,* već i *Dialogflow* simultano, a grafički prikaz toga može se videti na Slici 1.



Slika 1 – Tok podataka od Google Assistant-a preko DialogFlow do Actions on Google platforme

III. Opis rešenja

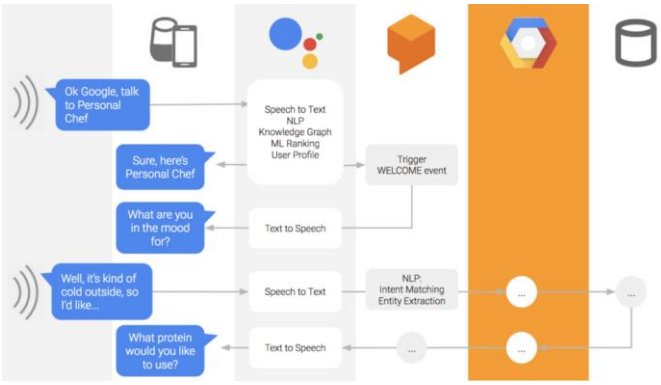
U ovoj sekciji dat je opis jednog rešenja obuke i integracije Google asistenta za složenije korisničke zahteve. Kao i u prethodno pomenutom naučnom radu [10], gde je bilo potrebno obučiti asistenta za jednostavnije komande, potrebno ga je obučiti i za složenije. Uporedni prikaz podržanih komandi u ovom radu u odnosu na prethodno pomenuti je tabelarno prikazan u Tabeli 1.

Table 1 - Uporedni prikaz podržanih komandi u naučnim radovima

Komande podržane u prethodnom radu	Komande podržane u ovom radu
Promena kanala unapred	Prikaz informacija o trenutnom kanalu
Promena kanala unazad	Prikaz informacija o trenutnom događaju na trenutnom/nekom drugom kanalu
Pojačavanje tona	Prikaz informacija o sledećem događaju na trenutnom/nekom drugom kanalu
Stišavanje tona	Prikaz informacija o događaju na određenom kanalu u određeno vreme
Potpuno utišavanje tona	Prikaz svih dostupnih kanala
Vraćanje na jačinu tona pre potpunog utišavanja	Prikaz dostupnih filmova (VOD)
Prebacivanje na kanal sa određenim brojem	Dodavanje/ brisanje podsetnika za neki događaj
	Dodavanje/brisanje događaja u listu za snimanje

Pored obuke, potrebno je generisati i odgovor korisniku i obaviti željenu akciju, ukoliko je to moguće, što je predstavljalo i najveći izazov u ovom radu. Korisnik više ne dobija odgovor u vidu obaveštenja da je željena radnja uspešno obavaljena, već uz dobavljeni TV sadržaj nastavlja

interakciju sa korisnikom nudeći mu dodatne opcije i nastavak dijaloga. Ceo princip rada dodatno je pojašnjen slikom 2.



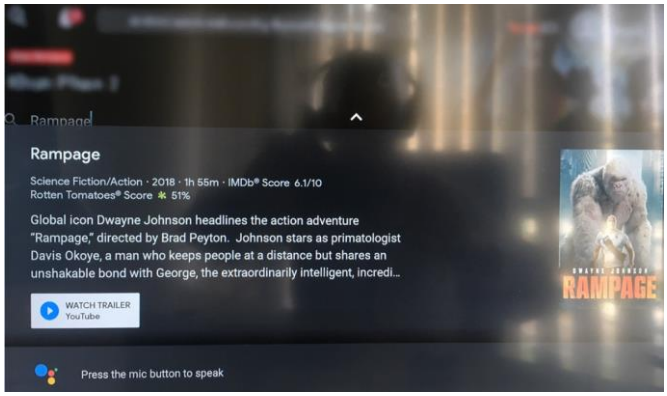
Slika 2 - Primer korisničkog zahteva

Na slici 2 dat je primer pod sistema od korisničkog zahteva, koji se unosi kao glasovna komanda i pokreće samog asistenta. Nakon toga asistent pokreće Dialogflow u potrazi za početnom akcijom (*Welcome intent)*. U nastavku korisnik zadaje komandu, a asistent pomoću Dialogflow i Actions on Google platforme šalje zahtev u aplikaciju. Aplikacija dobavlja neophodne podatke i istim putem ih vraća do korisnika, uz izvršavanje željene komande u pozadini.

Da bi obuka bila što efikasnija, poželjno je za svaku akciju (*Intent*) uneti pozamašan broj reči ili fraza koje korespondiraju željenoj akciji. Intuitivno, može se naslutiti da je ovo najlakša faza u ovom radu, ali je istovremeno iziskivala i određeno vreme da se u bazu mogućih korisničkih pitanja ili zahteva doda veliki broj TV servisa, TV događaja, filmova, serija i drugih sadržaja.

Druga, najizazovnija faza rada, je generisanje odgovora za korisnika, odnosno nastavak smislene konverzacije. Osnovni problem je bio što odgovor ne sme biti jednoznačan, već asistent mora posedovati veštačku inteligenciju. To se prevazilazi generisanjem različitih statičkih odgovora u zavisnosti od mogućih situacija (što nije bio slučaj u ovom radu) ili generisanje dinamičkih odgovora, koji u sebi treba da sadrže podatke o željenim TV servisima. Ono što jeste bila prepreka je činjenica da se podaci nalaze u aplikaciji, a da posluživač mora nekako dobaviti te podatke i prikazati ih u svom odgovoru korisniku. Za tu svrhu je korišćena baza podataka u kojoj su bili smešteni sadržaji iz aplikacije, kojima je posluživač pristupao i koristio ih. Korišćenjem baze podataka je rešen problem razmene podataka o TV servisima, između asistenta i TV aplikacije, tokom konverzacije sa korisnikom, jer je posluživač imao odakle da obezbedi podatke korisniku.

Kada je reč o akcijama prikazivanja TV servisa u samoj aplikaciji, uporedo su korišćeni Dialogflow, obezbeđen od strane Google-a i Android-ovi *intent*-i, tačnije namera za pretragu, što se može uočiti na slici 3.



Slika 3 - Primer pretrage filma iz aplikacije

Ukoliko asistent prepozna korisničku naredbu, započinje pretragu po bazi podataka. Nakon što je željeni TV servis pronađen, u zavisnosti od toga da li se traži neka specifična ili uopštena informacija, podaci se prikazuju ili se pušta određeni sadržaj.

Jedno od ograničenja ove implementacije je to što pokretanjem Google asistenta, odnosno njegovog agenta za akcije vezane za TV aplikaciju, moguće je komunicirati sa njim samo onoliko koliko je obučen. Drugi rečima, asistent nema pristup Internetu i ne mogu se dobiti druge informacije koje nisu vezane za aplikaciju. Konverzacija sa agentom se prekida jednostavnom komandom „Bye“, i tek tada se Google Assistant vraća u svoju osnovnu formu i može da pretražuje sve ostale podatke i informacije posredstvom Interneta.

IV. TESTIRANJE I VERIFIKACIJA

Zahvaljujući činjenici da je TV aplikacija sama po sebi robusna, kao i da je trebalo pokriti mnoštvo testnih slučajeva, izvršen je niz uporednih testova. Tri nezavisna korisnika su zadavali glasovne komande na osnovu kojih su vršena merenja.

. U cilju potvrđivanja funkcionalnosti realizovanog rešenja izvršeno je nekoliko različitih ispitivanja:

- Ispitivanje prepoznavanja unetih komandi;
- Ispitivanje uspešnosti izvršenja naredbi;
- Ispitivanje procenta uspešnosti prepoznavanja komandi od strane Google asistenta;
- Ispitivanje vremena odziva Google asistenta

Testovi su vršeni u normalnim uslovima, pri određenim šumovima, na engleskom jeziku. Svi slučajevi korišćenja sa rezultatima testiranja su prikazani u Tabeli 2.

Tabela 2 - Ispitivanje uspešnosti prepoznavanja i izvršavanja komandi

Ispitni slučaj	Rezultat
Prikaz informacija o trenutnom kanalu	USPEŠNO ZAVRŠEN
Prikaz informacija o trenutnom/sledećem događaju na trenutnom kanalu	USPEŠNO ZAVRŠEN
Prikaz informacija o trenutnom/sledećem događaju na nekom drugom kanalu	USPEŠNO ZAVRŠEN
Prikaz informacija o događaju na određenom kanalu u određeno vreme	USPEŠNO ZAVRŠEN
Prikazivanje liste dostupnih TV kanala	USPEŠNO ZAVRŠEN
Ispitivanje liste dostupnih filmova (VOD)	USPEŠNO ZAVRŠEN
Zakazivanje snimanja određenog TV sadržaja	USPEŠNO ZAVRŠEN
Brisanje događaja iz liste za snimanje	USPEŠNO ZAVRŠEN
Zakazivanje podsetnika za određeni TV sadržaj	USPEŠNO ZAVRŠEN
Brisanje događaja iz liste podsetnika za određeni TV sadržaj	USPEŠNO ZAVRŠEN

Iz Tabele 2 može se videti da su svi testni slučajevi pokriveni te da su uspešno izvršeni. Za svaki od ovih 10 testnih slučajeva je izvršeno u proseku 25 komandi. Proces ispitivanja prepoznavanja unetih glasovnih (pisanih) komandi je umnogome olakšan i Google-ovim simulatorom gde se jednostavnom može ispitati reakcija asistenta na korisničke pobude.

Kada je reč o procentu uspešnosti prepoznavanja korisničkih komandi od strane asistenta, uspešnost prepoznavanja je visokih 92%. Najbitniji faktor je izgovor engleskog jezika (srpski još nije podržan). Zahvaljujući visokom stepenu obučenosti asistenta i dobrom izgovoru komandi, uspešnost prepoznavanja komandi je na zavidnom nivou.

Ispitivanje vremena odziva u aplikaciji od trenutka izdavanja komande do trenutka odgovora ili izvršavanja iste. Za 10 ispitnih slučajeva je vršeno po 10 merenja za različite pod slučajeve. U zavisnosti od toga da li asistent treba da obezbedi samo podatke od TV sadržaju ili treba da odradi i akciju, vreme izvršavanja varira od 2 do 5 sekundi. Iz ugla korisničkog iskustva, može se diskutovati da li je vreme odziva zadovoljavajuće ili nije. U Tabeli 3 dat je uporedni prikaz vremena odziva za gore pomenuta 4 testna slučaja.

Tabela 3 - Ispitivanje vremena odziva na korisničku pobudu

Ispitni slučaj	Vreme odziva
Prikaz informacija o kanalu	~1,388
Prikaz informacija o trenutnom/narednom događaju	~1,913
Prikaz informacija o događaju u određeno vreme na određenom kanalu	~1,888
Prikaz svih kanala	~3,036
Prikaz svih filmova	~4,173
Brisanje događaja iz liste podsetnika	~2,098
Brisanje događaja iz liste za snimanje	~2,100
Zakazivanje snimanja određenog TV sadržaja	~2,734
Zakazivanje podsetnika za određeni TV sadržaj	~2,525

Vreme odziva je u proseku od 2 do 4 sekunde u zavisnosti od komande, zahvaljujući optimizovanoj pretrazi pojmova i ključnih reči. Najviše vremena iziskuje prikaz TV kanala, odnosno filmova, iz razloga što je potrebno prikazati mnogostruko više rezultata korisniku nego kada se zahtevaju informacije o samo jednom događaju.

V. Zaključak

U ovom radu predstavljeno je unapređenje postojećeg rešenja [10] integracije Google asistenta u TV aplikaciju na operativnom sistemu Android. Pored osnovnih komandi za upravljanje kanalima i jačinom tona, poboljšano je rešenje sa aspekta korisničkog iskustva, te je sada moguće imati prirodniju komunikaciju sa asistentom kao i njegovo izvršavanje složenijih komandi. Potrebno je napomenuti da je implementacija i integracija asistenta u aplikaciju potpuno autonomna, te je moguće ovo rešenje iskoristiti i u drugim aplikacijama za gledanje televizije uz minimalne izmene i prilagođavanja.

Rezultati dobijeni prilikom ispitivanja su zadovoljavajući, s tim da je cilj u narednom periodu da se dodatno poboljšaju koliko god da je to moguće.

Asistent je dodatno obučen za interaktivnu komunikaciju, iako nisu pokriveni svi slučajevi korišćenja u aplikaciji, jer ih ima mnogo. U tom svetlu, budući rad bi mogao da se bazira da dodatnom proširenju funkcionalnosti aplikacije prilikom korišćenja asistenta i omogućavanje navigacije kroz istu posredstvom isključivo glasovnih naredbi.

ZAHVALNICA

Ovaj rad je delimično finansiran od strane Ministarstva za prosvetu, nauku i tehnološki razvoj Republike Srbije, na projektu broj: TR32041.

LITERATURA.

[1] M. Z. Bjelica, N. Teslić and V. Mihić: „Softver u televiziji i obradi slike 1“, Faculty of Technical Sciences, Novi Sad, Serbia, 2017.
[2] I. Pap and N. Lukić: „ Projektovanje namenskih računarskih sistema 1“, Faculty of Technical Sciences, Novi Sad, Serbia, 2016.
[3] Center for the Digital Future , „Web Insight: views about sources of entertainment“, Available: http://www.digitalcenter.org/web-insight-views-about-entertainment-sources-2010/ [Accessed: May 5,2018].
[4] Android Open Source Project (AOSP), available: https://source.android.com/ [Accessed: Apr.3,2019]
[5] Source, „TV Input Framework - TIF“, available: https://source.android.com/devices/tv/ [Accessed: Apr. 24,2018].
[6] P. Treblicox-Ruiz,: „Android TV Apps Development,”, ISBN: 978- 1-4842-1784-9, Apress,2016.
[7] G. López, L. Quesada and L.A. Guerrero: “Alexa vs. Siri vs. Cortana vs. Google Assistant: A Comparison of Speech-Based Natural User Interfaces”. In: Nunes I. (Advances in Human Factors and Systems Interaction. AHFE 2017. Advances in Intelligent Systems and Computing, vol 592. Springer, ChamM. Young, *The Techincal Writers Handbook*. Mill Valley, CA: University Science, 1989.
[8] B. Rajkumar, B. James, A Goscinski: „Cloud Computing Principles and Paradigms“, Hoboken, New Jersy, 2011.
[9] Developers, „Actions on Google“, available: https://developers.google.com/actions/ [Accessed: May 15,2018]
[10] A.Lazić, M.Bjelica, D.Nađ: „Integracija podrške za Google Assistant u aplikaciji za gledanje televizije na operativnom sistemu Android“, University of Novi Sad, Faculty of technical sciences, Novi Sad, 2018.
[11] M. Robin and M. Poulin: „Digital television fundamentals – Design and Installation of Video and Audio Systems”, McGraw- Hill, 1997
[12] E. Nan: „Upravljanje pametnom kućom uz pomoć Google asistenta”, University of Novi Sad, Faculty of Technical Sciences, Novi Sad, 2017.
[13] J. Bloch: „ EffectiveJava (2nd edition)“, Addison-Weasly, 2008.
[14] M. Vidakovic, N. Teslic, T. Maruna and V. Mihic: „Android4TV: a proposition for integration of DTV in Android devices“, IEEE 30th International Conference on Consumer Electronics (ICCE), Las Vegas, January 2012.
[15] S. Ghosh and J. Pherwani: „Designing of a natural voice assistants for mobile through user centered design approach“. In: Kurosu, M. (ed.) Human–Computer Interaction: Design and Evaluation, pp. 320–331. Springer International Publishing, Cham (2015).

ABSTRACT

Nowadays, more and more people are using virtual assistant services on various platforms (Android, Windows, iOS). They are becoming present in everyday life such as buying food in market, driving to the job, or for the purpose of an entertainment. This paper presents a software architecture that supports Google Assistant integration in TV application, designed for Android operating system. The aim is to enhance user experience, so previously added intents such as “channel up/down”, “volume up/down” are extended with some more complex commands such as finding TV channel by name and zapping to it or choosing some VOD and schedule it for recording or watching later, by using only a voice command. The majority of existing speech recognition tools provide the result of the speech processing in a free form textual output or structured form textual output. In order to use obtained outputs in an existing TV applications, it is necessary to detect patterns that correspond to mentioned commands.

Use of Google Assistant technology in the interactive digital television

Aleksandar Lazić, Milan Z. Bjelica, Veljko Ilkić, Dejan Nađ

Proširenje TV Input radnog okvira funkcionalnostima paketa Google Assistant u Android okruženju

Radenko Banović, Milan Z. Bjelica, Darko Dejanović, Milan Gvero

***Apstrakt* — U ovom radu je predstavljeno jedno rješenje proširenja TV input radnog okvira funkcionalnostima paketa Google Assistant. TV Input predstavlja programsku reprezentaciju jednog izvora odakle se prima TV sadržaj, te popunjava sadržaj u bazu podataka TV poslužioca. Proširenje TV Input radnog okvira se odnosi na prilagođavanje pretrage i prikaza EPG (electronic program guide) podataka smještenih u bazi podataka TV poslužioca tako da se oni nađu u Google Assistant rezultatima pretrage.**

***Ključne riječi*— Android, Digital TV, Google Assistant, DTV**

I. Uvod

U poslednje vrijeme raste popularnost upravljanja uređajima govorom [1], što je posebno interesantno u oblasti digitalne televizije, jer digitalna televizija nudi sve više sadržaja. Unos podataka daljinskim upravljačem otežava korištenje svih servisa koje nudi digitalna televizija.

Android TV je proširenje operativnog sistema Android dodatnim funkcionalnostima za TV uređaje (set-top box i smart TV). Android TV zadržava mogućnosti kao što su instaliranje aplikacija sa Google Play Store (npr. video igre), ili korištenje ostalih Google servisa kao što je Google Asistent [2].

Google Asistent je virtuelni lični pomoćnik razvijen od strane Google-a koji je dostupan i na Android TV uređajima od 2017. godine [3], a omogućuje dvostranu komunikaciju. Korisnici primarno komuniciraju sa Google Asistentom prirodnim glasom, ali je i unos preko tastature takođe podržan. Pretražuje internet i aplikacije koje su lokalno instalirane u slučaju da su aplikacije prilagođene i dozvoljavaju globalno pretraživanje njihovog sadržaja. Takođe, moguće je podešavati fizičke komponente uređaja, dodavati događaje u kalendaru itd.

Zadatak ovog rada je da omogući Google Asistentu pretraživanje baze podataka TV poslužioca proširenjem TV Input radnog okvira. Ukoliko se pronađe traženi sadržaj (npr. EPG (electronic program guide) sadržaj – naziv emisije) potrebno je iz rezultata pretrage omogućiti reprodukovanje živog TV sadržaja kanala za koji je pronađeni program vezan.

Ranije je pretragu TV baze podataka bilo moguće pozvati i rezultate pretrage prikazati samo u aplikaciji koja i skladišti podatke. Omogućujući Google Asistentu pretragu baze podataka TV poslužioca korisniku je omogućena lakša pretraga TV sadržaja, i lakši pristup željenom sadržaju.

Ovaj Rad je sačinjen od 5 poglavlja.

U drugom poglavlju se nalaze teorijske osnove neophodne za razumijevanje digitalne televizije, Android platforme, dostavljača sadržaja i Google Asistenta.

Treće poglavlje sadrži osnovne informacije o dvema aplikacijama koje je potrebno proširiti pri izradi ovog rada, opis ciljne platforme, kao i informacije o programskom okruženju i implementaciji rješenja u obje aplikacije.

Četvrto poglavlje sadrži opis testova koji su obavljeni nakon izrade rješenja radi validacije istog, i diskutovani su dobijeni rezultati.

Peto poglavlje sadrži kratak pregled onoga što je obavljeno u ovom radu, i dalje pravce razvoja implementiranog rješenja.

II. Teorijske osnove

A. *Digitalna televizija*

Televizija u širem smislu podrazumijeva prenos slike i zvuka od proizvođača, preko emitera, do gledalaca. Televizijski signal primarno služi da se prenese pokretna slika, zvuk i dodatne informacije važne za reprodukciju. Nerijetko se u okviru signala prenose i sadržaji koji nisu audio-video tipa, poput digitalnog teleteksta. Analogni televizijski signal je vremenom zamijenjen digitalnim TV signalom, a ta promjena je donijela poboljšanja u nekoliko segmenata, veći kvalitet audio i video signala, digitalnim kanalima je moguće prenijeti mnogo veše televizijskih programa i moguće je prenijeti dodatne sadržaje kao što je EPG koji će u ovom radu biti pretraživan u bazi podatak TV poslužioca.

Prednost digitalnog signala nad analognim jeste u primjeni digitalnih diskretnih algoritama obrade signala, pogotovo algoritama kodovanja (kompresije) video signala. Tek nakon efikasne primjene algoritama kompresije digitalni prenos dolazi u prednost u odnosu na analogni. Algoritmi kompresije omogućuju da se sa manje informacija prenese identičan video i audio sadržaj [4].

Efekat digitalne litice jedna je od najvažnijih osobina u

Radenko Banović – Fakultet Tehničkih Nauka, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail: Radenko.Banovic@rt-rk.com).
Milan Z. Bjelica – Fakultet Tehničkih Nauka, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (e-mail: Milan.Bjelica@rt-rk.com).
Darko Dejanović – RT-RK d.o.o., Narodnog Fronta 23a, 21000 Novi Sad, Srbija (e-mail: Darko.Dejanovic@rt-rk.com).
Milan Gvero – RT-RK d.o.o., Narodnog Fronta 23a, 21000 Novi Sad, Srbija (e-mail: Milan.Gvero@rt-rk.com).

digitalnoj televiziji. U slučaju pogoršanja kvaliteta prijema signala, kvalitet slike i zvuka se ne degradira postepeno (kao u slučaju analogne televizije), već su kvalitet slike i zvuka konstantni dok signal ne bude toliko loš da se ne može demodulisati, kada prijem u potpunosti prestaje.

B. Elektronski programski vodič - EPG

Elektronski programski vodič (eng. Electronic program guide) predstavlja skup informacija o programima iz određenog vremenskog intervala sa kanala dovedenih na TV Input. Informacije koje uobičajno čine EPG podatke su: naziv programa, kratak opis, vrijeme početka i kraja programa, naziv kanala za koji je program vezan, te nivo roditeljske zaštite.

C. Android platforma

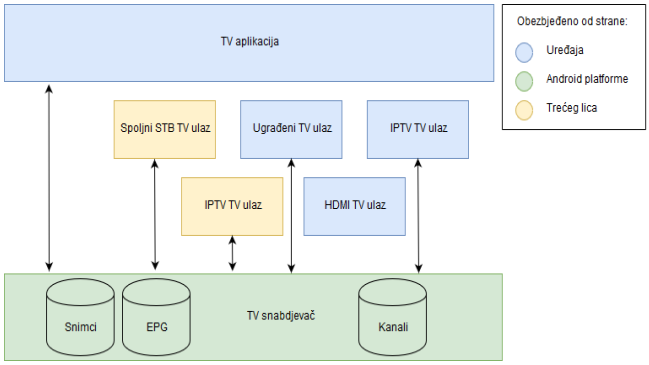
Android je slojeviti softver otvorenog koda (eng. open source) baziran na Linux jezgru namijenjen različitim uređajima. Ovaj rad je realizovan na Android platformi. Korištenje Linux jezgra omogućuje Androidu da iskoristi glavne sigurnosne karakteristike kao i proizvođačima uređaja da razviju hardverske upravljače za dobro poznato jezgro [5].

D. Android programska podrška za TV uređaje - TIF

Android programska podrška za TV uređaje definiše aplikativnu programsku spregu i standardizuje način implementacije dopremanja emitovanog sadržaja do TV aplikacije. TIF omogućuje implementaciju TV Input radnog okvira preko kojeg je moguće reprodukovati sadržaj sa mrežnog poslužioca i/ili iz DVB transportnog toka podataka kao i pretragu televizije uživo, i preporuke putem meta podataka objavljene od strane TV izvora [6].

TIF se sastoji od:

- TV aplikacija (eng. TV Application) - aplikacija koja rukuje zahtjevima korisnika;
- TV ulazni rukovalac (eng. TV Input Manager) – omogućuje TV ulazima da komuniciraju sa TV aplikacijom;
- TV Input – aplikacija koja predstavlja jedan izvor TV sadržaja;
- TV snabdjevač (eng. TV Provider) – baza podataka sa kanalima, programima i pratećim dozvolama;



Slika 1 - TV snabdjevač

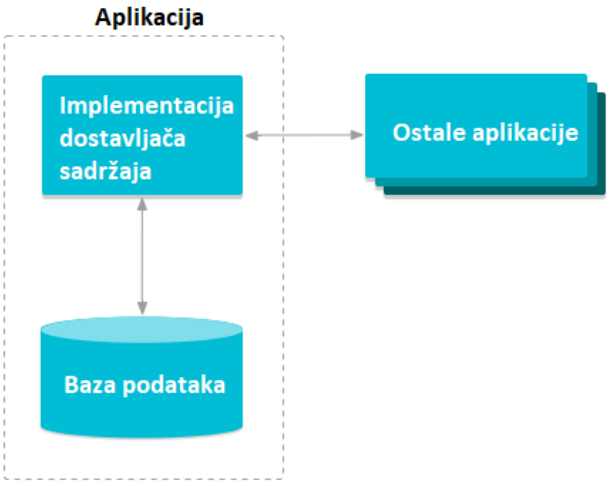
E. Google Asistent

Google Asistent je virtuelni lični pomoćnik razvijen od strane Google-a koji je dostupan i na Android TV uređajima od 2017. godine, a omogućuje dvostranu komunikaciju. Korisnici primarno komuniciraju sa Google Asistentom prirodnim glasom koristeći Google-ov algoritam za obradu prirodnog glasa, ali je i unos preko tastature takođe podržan. Pretražuje internet i aplikacije koje su lokalno instalirane ako su aplikacije prilagođene i dozvoljavaju globalno pretraživanje njihovog sadržaja, može podešavati hardverske komponente na uređaju, dodavati događaje u kalendar itd. Rezultati pretrage su predstavljeni kao kartice čijim se klikom ili dodirom otvara povezana stranica.

F. Dostavljač sadržaja (eng. Content Provider)

Dostavljači sadržaja omogućuju aplikacijama da upravljaju pravima pristupa podacima sačuvanim od strane njih samih, sačuvanih od strane drugih aplikacija i da omogući dijeljenje podataka sa drugim aplikacijama. Oni omogućuju mehanizam za definisanje sigurnosti podataka. Implementacija dostavljača sadržaja ima mnoge prednosti. Najvažnije je to što se može konfigurisati dostavljač sadržaja tako da se dozvoli drugim aplikacijama da bezbjedno pristupe i mijenjaju podatke aplikacije.

Dostavljači sadržaja predstavljaju odličan način apstrakcije, na primjer omogućavaju da se promijeni baza podataka u koju se smještaju podaci aplikacije, bez da to utiče na druge aplikacije koje se oslanjaju na pristup podacima. U tom slučaju moraju se napraviti izmjene samo u dostavljaču sadržaja, dok aplikacije koje pristupaju podacima ostaju nepromijenjene.



Slika 2 - Dostavljač sadržaja

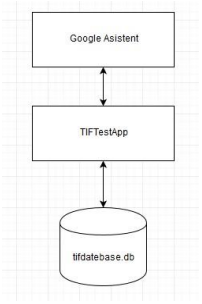
III. OPIS REALIZACIJE

TV Input radni okvir je potrebno proširiti tako da se omogući da se podaci iz baze podataka TV poslužioca nađu u rezultatima pretrage Google Asistenta.

Pored proširenja TV Input radnog okvira potrebno proširiti i testnu aplikaciju koja prikazuje sadržaj korisniku. Testna

aplikacija se oslanja na TV Input radni okvir. Ovo je neophodno da bi podaci iz baze podataka TV poslužioca bili globalno pretraživi, te da se iz rezultata pretraživanja jednim klikom može reprodukovati živi sadržaj kanala vezanog za program dobijen kao rezultat pretraživanja. Kao početna tačka izrade programskog rješenja korištene su dvije gotove aplikacije. Njih je potrebno nadograditi da bi omogućili Google Asistentu da prilikom pretrage u istu uključi i EPG sadržaj kanala dovedenih na TV Input, i to TIFTestApp i jednu implementaciju TV Input radnog okvira.

U toku istraživanja metoda za izradu programskog rješenja obavljen je jednostavan testni primjer najvažnijeg segmenta rješenja koji je služio za dokazivanje koncepta rješenja. Svrha ovakvog pristupa je da dokaže da je problem moguće riješiti na određen način. Testni primjer se odnosio na implementaciju nove testne baze podataka u okviru aplikacije TIFTestApp, te prilagođenja aplikacije Android pretražvačkoj sprezi koja treba da pročita informacije iz testne baze podataka i dostavi rezultate pretrage korisniku (Slika 3).



Slika 3 - Prototip testne aplikacije

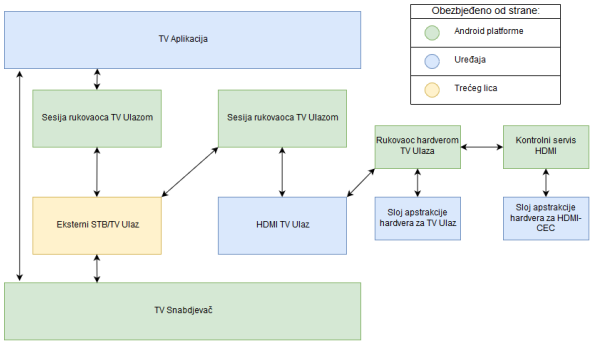
A. *Programsko okruženje i ciljna platforma*

Okruženje za izradu je Android Studio u programskom jeziku Java. Programski jezik Java prilagođen je potrebama i standardima Android operativnog sistema. Android studio razvojno okruženje omogućava laku ugradnju potrebnih alata i pruža širok spektar mogućnosti u realizaciji Android aplikacija, dok je SDK kolekcija alata koji pomaže u kreiranju Android aplikacija.

DTV prijemnik korišten pri izradi ovog rada je Android platforma bazirana na Marvell BG5CT čipsetu.

B. *Jedno rješenje TV Input radnog okvira*

TV Input popunjava TV sadržaj u bazu podataka TV poslužioca. TV Input predstavlja programsku reprezentaciju jednog izvora odakle se prima TV sadržaj (Slika 4). Može biti prisutno više TV Input radnih okvira.



Slika 4 -TV Input kreiran od treće strane

Rješenje TV Inputa koje je bilo potrebno nadograditi je imalo implementirano rukovanje tokovima audio i video podataka sa izvora TV sadržaja, skladištenje podataka o TV kanalima, rukovanje prevodima, itd. Dakle, imalo je skoro kompletnu implementaciju, ali nije imalo implementaciju skladištenja EPG podataka u bazu podataka TV poslužioca.

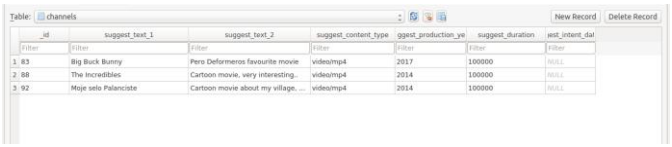
C. *TIFTestApp*

Druga aplikacija koju je potrebno izmijeniti jeste TIFTestApp, čija je funkcija da prezentuje TV sadržaj korisniku. U njoj korisnik može da bira TV Inpute (ukoliko ih ima više), da mijenja kanale, pregleda EPG sadržaj, te da gleda i sluša audio i video tokove izabranog kanala. Proširanje ove aplikacije podrazumijeva omogućavanje pretraživanje aplikacije od strane Google Asistenta, te korištenje API-ja TV Inputa radi dobavljanja informacija o trenutnom EPG sadržaju kanala na ulazu.

D. *Prototip kao dokaz validnosti koncepta rješenja*

Android TV koristi Android pretraživačku spregu da dobavi sadržaj od aplikacija i dostavi rezultate pretrage korisniku. Svaka aplikacija može biti prilagođena, i sadržaj aplikacije može biti uključen u rezultate pretrage. Aplikacija mora implementirati dostavljač sadržaja (eng. Content Provider) koji servira rezultate, kao i *searchable.xml* konfiguracioni fajl koji opisuje implementirani dostavljač sadržaja i ostale vitalne informacije Android TV-u. Izrada prototipa je prvi korak ka izradi programskog rješenja. Ideja je da se TIFTestApp aplikacija prilagodi na način da se omogući globalno pretraživanje kroz aplikaciju, i da se pripremi testna baza podataka sa nazivima kolona identičnim nazivima polja podataka pretraživačkog rukovaoca (eng. Search Manager). Kreirana je testna baza podataka za potrebe izrade prototipa pod nazivom *tiftestbase.db* (Slika 5).

Polja nose ova ime iz razloga što Android pretraživačka sprega očekuje baš takve nazive, pa je najlakše mapirati na nazive kolona koji se upravo tako zovu.



Slika 5 - Primjer testne baze podataka

Programsko rješenje se sastoji iz dva dijela, prvi dio predstavlja proširenje jednog rješenja TV Inputa implementacijom preuzimanja, parsiranja i skladištenja EPG podataka u bazu podataka TV poslužioca. Drugi dio predstavlja proširenje TIFTestApp aplikacije na način da, za razliku od prototipa, poslužuje EPG podatke koji su smješteni u bazi podataka TV poslužioca.

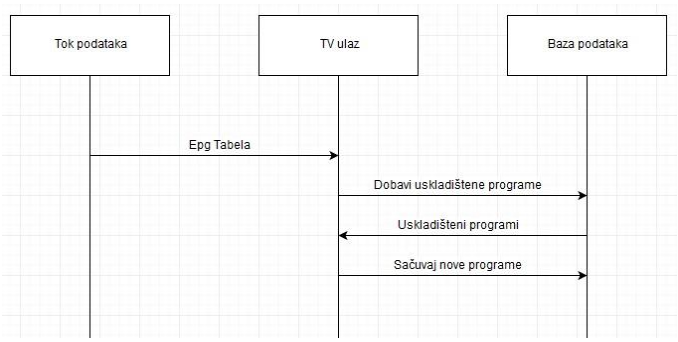
E. Proširenje TV Inputa

U TV Inputu treba da se iz ulaznog toka podataka izdvoje EPG podaci (naslov programskog događaja, opis, vrijeme početka i vrijeme završetka programskog događaja, kanal za koji je vezan događaj).

Da bi smjestili nove EPG podatke u bazu podataka TV poslužioca, neophodno je iz baze podataka dobaviti sve programske događaje u određenom vremenskom opsegu (u ovom slučaju 7 dana unaprijed i 36 sati unazad). Ovo je neophodno uraditi jer EPG podaci stižu periodično, i nema potrebe u bazu podataka smještati EPG podatke kada god tabela sa njima stigne, već samo onda kada stigne tabela sa događajima koji do tad nisu smješteni u bazu.

Kreira se upit bazi podataka kojim se dobavljaju uskladišteni programski događaji u određenom vremenskom opsegu, te se za svaki od novozaprimljenih događaja provjerava da li postoje u listi koja je popunjena podacima iz baze podataka, ako nije, događaj se stavlja u novu listu (array) koja će biti smještena u bazu podataka (Slika 6).

Ukoliko neki od programskih događaja postoji u bazi podataka, ali se sadržaj novopridošlog i programskog događaja smještenog u bazi podataka razlikuju, događaj se ažurira (umjesto uskladištenog sadržaja, u bazu podataka se smješta najnoviji sadržaj).



Slika 6 - Sekvencni dijagram skladištenja novih EPG podataka

F. Proširenje TIFTestApp

Aplikacija mora dostaviti Android TV-u polja podataka iz kojih se generišu predloženi rezultati pretrage nakon što korisnik ukuca karakter u dijalog za pretraživanje ili kaže ključne riječi Google Asistentu.

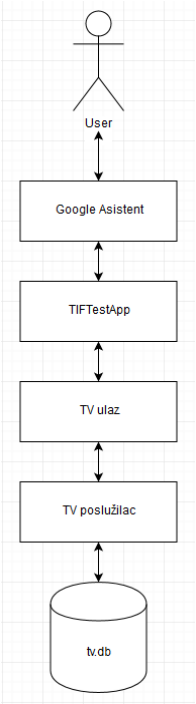
Da bi to omogućila, aplikacija mora implementirati dostavljač sadržaja (eng. Content Provider) koji servira prijedloge. Potrebno je kreirati i searchable.xml konfiguracioni fajl koji opisuje dostavljač sadržaja kao i ostale vitalne informacije Androd TV-u.

Pretraživački rukovaoc opisuje polja podataka koja očekuje, reprezentujući ih kao kolone lokalne baze podataka. Vrlo često su kolone u bazi podataka nazvane drugačije od polja pretraživačkog rukovaoca, pa je neophodno mapriati imena kolona lokalne baze podataka na nazive polja pretraživačkog rukovaoca.

Dostavljač sadržaja vraća rezultate pretrage u Android TV dijalog pretrage. Sistem šalje upit dostavljaču sadržaja svaki put kada Google Asistent bude prozvan. Dostavljač sadržaja pretražuje podatke i vraća pokazivač koji pokazuje na kolone izabrane kao prijedloge rezultata pretrage.

U *AndroidManifest.xml* datoteci dostavljač sadržaja ima poseban tretman, jer nije opisan tag-om kao ostali activity-ji već posebnim tag-om. Provider sadrži atribut koji se odnosi na ime klase u kojoj je implementiran dostavljač sadržaja, zatim atribut koji sistemu dobavlja informaciju o imenskom prostoru opisanog dostavljača sadržaja, a potrebno je i postaviti atribut kojim se dozvoljava da Android global search može koristiti podatke dobijene iz dostavljača sadržaja.

Da bi se moglo rukovati rezultatima pretrage, aplikacija mora da sadrži *searchable.xml* datoteku radi podešavanja prijedloga rezultata pretrage. Mora da sadrži atribut koji govori sistemu koji je imenski prostor dostavljača sadržaja, i on mora biti identičan kao isti takav atribut u *AndroidManifest.xml* datoteci. Prikaz komunikacije različitih elemenata sistema dat je na slici 7.



Slika 7 – Prikaz komunikacije različitih elemenata sistema

IV. TESTIRANJE

Prilikom izrade ovog rada obavljena su ispitivanja realizovanih modula i funkcionalnosti. Obavljene su dvije vrste ispitivanja: JUnit nezavisni testovi i ručno ispitivanje. JUnit ispitnim slučajevima su provjereni svi moduli opisani u ovom radu. JUnit ispitni slučajevi ne mogu da daju kompletan prikaz ispravnosti realizovanih funkcionalnosti zbog toga što ispituju da li je ispravna očekivana funkcionalnost na programskom nivou, i ne mogu da provjere da li se prikazala slika i slično, te su iz tog razloga obavljeni ručni ispitni slučajevi. Tabela JUnit ispitnih slučajeva se nalazi u tabeli 1.

TABELA 1
Uspješnost JUnit testova

Ispitni slučaj	**Rezultat**
Skladištenje EPG podataka	Prošao
Skladištenje EPG podataka koji već postoje u bazi podataka	Prošao
Upit po nazivu	Prošao
Upit po opisu	Prošao
Upit po nazivu za nepostojeće EPG podatke	Prošao

Ispitivanja vezana za Google asistent su obavljena ručno, odnosno Google asistent je pobuđivan ključnim riječima (nazivom ili kratkim opisom) EPG programa, što bi u rezultatima pretrage trebalo da prikaže sadržaj aplikacije, i da ponudi pokretanje aplikacije na kanalu za koji je EPG program vezan.

Zbog toga što platforma na kojoj su vršena testiranja nema mogućnost obrade govora, testiranja su vršena isključivo pokretanjem adb komande, unošenjem različitih upita kako naziva programskog sadržaja, tako i kratkog opisa programskog sadržaja. Ukoliko se nakon pokretanja adb komande sa upitom za programski sadržaj koji postoji u bazi podataka u rezultatima pretrage pojavi sadržaj čijim pokretanjem dolazi do pokretanja kanala koji je vezan za dati sadržaj, smatra se da je ispitni slučaj uspješno završen. Takođe, ukoliko se unese upit za sadržaj koji ne postoji u bazi podataka, i rezultati pretrage ostanu prazni, znači da je ispitni slučaj uspješno završen. Prikaz uspješnosti ručnih ispitnih slučajeva dat je u tabeli 2.

TABELA 2
Uspješnost ručnih testova

Ispitni slučaj	**Rezultat**
Football beast	Prošao
RT News	Prošao
News	Prošao
Dnevnik	Prošao
Sportski dnevnik	Prošao

V. ZAKLJUČAK

U ovom radu je realizovano skladištenje EPG podataka u bazu podataka TV poslužioca, te proširena aplikacija TIFTestApp radi omogućavanja globalne pretrage Google Asistenta kroz aplikaciju radi dobavljanja EPG podataka. Omogućena je reprodukcija živog programskog sadržaja kanala vezanog za programski sadržaj dobijen u dijalogu rezultata pretrage. Rješenje je realizovano na Marvell BG5CT platformi oslanjajući se na Android operativni sistem sa Android programskom podrškom za TV uređaje (TIF). Rješenje je ispitivano na ciljnoj platformi, i testiranjem je utvrđena funkcionalnost rješenja. Razvoj Android platforme i proširenje skupa njenih funkcionalnosti pruža mogućnost za dalju nadogradnju i poboljšanje datog rješenja. Jedan od pravaca daljeg razvoja programskog rješenja bi moglo biti proširenje obje aplikacije na način da se omogući skladištenje živog TV programskog sadržaja, te nakon pretraživanja EPG podataka, pokretanje baš tog programskog sadržaja koji je prethodno emitovan i uskladišten pomoću snimača kanala.

ZAHVALNICA

Ovaj rad je djelimično finansiran od strane Ministarstva za prosvetu, nauku i tehnološki razvoj Republike Srbije, na projektu broj: TR36029.

LITERATURA

[1] Kenneth Olmstead, Nearly half of Americans use digital voice assistants, mostly on their smartphones, Jun 2018. [online]. http://www.pewresearch.org/fact-tank/2017/12/12/nearly-half-of-americans-use-digital-voice-assistants-mostly-on-their-smartphones/

[2] Android TV, Jun 2018. [online]. https://en.wikipedia.org/wiki/Android_TV

[3] Google Assistant, Jun 2018. [online]. Available: https://en.wikipedia.org/wiki/Google_Assistant

[4] Milan Bjelica, Nikola Teslić, Velibor Mihić, "Softver u digitalnoj televiziji 1", 2017.

[5] Ištvan Pap, Nemanja Lukić, "Projektovanje namenskih računarskih sistema 1", 2016.

[6] TV Input Framework, Jun 2018. [online]. Available: https://source.android.com/devices/tv/

ABSTRACT

This paper presents one solution of extending the TV Input Framework with functionalities of the Google Assistant package. TV Input is a program representation of a single source from which it receives TV content, and fills content into a TV provider's database. The expansion of the TV Input refers to customizing the search and display of the EPG (electronic program guide) data stored in the TV provider database so that they appear in Google Assistant search results.

Extension of TV Input framework with Google Assistant functionalites in Android environment

Radenko Banović, Milan Z. Bjelica, Darko Dejanović, Milan Gvero

Jedno rešenje reprodukcije multimedijalnog sadržaja na Android Things platformi

Marijana Gligorić, Vladimir Nešić, Nikola Vranić, Miloš Subotić, Đorđe Glišić

Apstrakt— **U poslednjih deset godina širi se upotreba digitalne televizije, razvijaju tehnologije i pojavljuju novi uređaji. Takođe se pojavljuju i zahtevi za novim funkcionalnostima od strane korisnika. Razvijanje aplikacije za gledanje televizije na set top box uređaju je nešto što smo do sada već videli. Pored toga, to izaziva određene komplikacije ako je u pitanju samostalan inženjer zbog nedostupnosti uređaja na kome bi se razvijale aplikacije ili nedostatka root prava na istim. Kao odlično inovativno rešenje u ovakvoj situaciji nameće se platforma za razvijanje Internet of Things aplikacija koju je ponudio sam Google: Android Things. Ovaj rad predstavlja konkretno rešenje reprodukcije televizijskog sadržaja korišćenjem pomenute platforme. Za implementaciju rešenja korišćen je uređaj Rasberry Pi 3 Model B.**

Ključne reči **— Raspberry Pi, Android, Digital Video Broadcast, Dynamic Adaptive over HTTP, Internet of Things.**

I. Uvod

Vreme u kojem živimo danas veoma je teško zamisliti bez mobilnih uređaja, posebno bez mobilnih telefona. Njihovo postojanje beleži nešto manje od pola veka, ali njihov razvoj pokazuje ogroman napredak za kratko vreme. Predviđa se da će do 2020. godine čak devedeset posto populacije posedovati mobilni telefon. Dva trenutno najaktuelnija operativna sistema koja se koriste za razvoj mobilnih aplikacija su Android i IOS. Za razvoj Android aplikacije se odlučilo 47% programera, dok IOS preti da ga prestigne sa 31%. Što se tiče procenta korisnika koji na svom uređaju poseduju Android sistem u odnosu na one koji koriste IOS on okvirno iznosi 70 prema 30 respektivno [1].

U poslednje vreme sve aktuelnija tema jeste Internet of things. Internet of things uređaji predstavljaju fizičku arhitekturu sa delom programske podrške koja je ugrađena u njih i koja ima pristup internetu. Jedna od ključnih razlika izmedju Internet of Things uređaja i tradicionalnih uređaja jeste u tome što pametnim uređajima mi zadajemo funkciju koju želimo da nam izvrše i na taj način vršimo interakciju sa samim uređajem. Internet of things uređaji sami prikupljaju podatke, selektuju ih, šalju preko interneta na dalju obradu i na taj način vrše funkciju za koju su namenjeni. Neki od najpopularnijih projekata iz ove kategorije su: pametni gradovi, pametni domovi, briga za stara lica i slično.

Marijana Gligorić, Istraživačko razvojni institut RT-RK, Narodnog Fronta 23a, Novi Sad, Srbija (email: marijana.gligoric@rt-rk.com)
Vladimir Nešić, Istraživačko razvojni institut RT-RK, Narodnog Fronta 23a, Novi Sad, Srbija (email: vladimir.nesic@rt-rk.com)
Nikola Vranić, Istraživačko razvojni institut RT-RK, Narodnog Fronta 23a, Novi Sad, Srbija (email: nikola.vranic@rt-rk.com)
Miloš Subotić, Istraživačko razvojni institut RT-RK, Narodnog Fronta 23a, Novi Sad, Srbija (email: milos.subotic@rt-rk.com)
Đorđe Glišić, Istraživačko razvojni institut RT-RK, Narodnog Fronta 23a, Novi Sad, Srbija (email: djordje.glisic@rt-rk.com)

Jedna od stvari koje su neophodne za pravljenje programa jeste operativni sistem. Google je ponudio rešenje u vidu Android Things platforme i ponudio je uređaje na kojima ovaj sistem može da se koristi. Trenutno najprodavaniji i najčešće korišćeni uređaj jeste Raspberry Pi. Ovaj uređaj će se koristiti i za realizaciju rešenja predstavljenog u ovom radu.

Još jedan pojam sa čijom se upotrebom susrećemo duže od mobilnih uređaja jeste televizija. Ranije je gledanje televizijskog sadržaja bilo usko povezano sa televizorom kao uređajem preko kojeg je to moguće ostvariti. Danas je poznato da je televizijski prenos moguć i korišćenjem mobilnih telefona. Velika prednost jeste to što se prilikom potpisivanja ugovora sa odgovarajućim operaterom i dogovora o pretplati na određen broj kanala dobija korisničko ime i lozinka kojom je moguć pristup pretplaćenom sadržaju sa bilo koje platforme čije korišćenje omogućava operater. Upotreba Android platforme na namenskom uređaju podrazumeva prilagođenje potrebama i ograničenjima konačnog proizvoda, i predstavlja složen zadatak koji zahteva razumevanje unutrašnje arhitekture Androida, prilagođavanje slojeva operativnog sistema mogućnostima platforme kao i Linux jezgra na koji se Android oslanja.

Do sada su istraživanja rađena u sferi TV industrije ka korišćenju STB uređaja koji bi koristili Android kao operativni sistem [2]. Razvoj je proširen na internet ka streaming servisima i IPTV rešenjima. U okviru Internet of Things primena, jedan uži skup ali izuzetno aktivan predstavljaju pametne kuće i integracija kućnih senzora i uređaja u mrežu preko koje mogu da razmenjuju informacije. Podskup istraživanja je išao u smeru integracije TV prijemnika u IoT mrežu radi kontrole aparata, osvetljenja, vrata i prozora, kućnih aparata i sl [3]. U primeni IoT rešenja, postoji čitav niz uređaja na tržistu, najrazličitijih namena, od senzora koji se napajaju običnim baterijama do rutera koji mogu da prikupljaju podatke sa senzora i kontrolišu iste. Jedna popularna platforma za razvoj IoT rešenja je i Raspberry Pi. Kompanija Google je prepoznala značaj i mogućnosti ovih modela i njihovu primenljivost u IoT industriji kada je učinila dostupnim Android Things operativni sistem za ove uređaje. Ovaj rad za cilj ima da dokaže da je već sada moguće dobiti uređaj koji se može koristiti kao digitalni TV prijemnik, od IoT uređaja na kome se izvršava Android operativni sistem.

U nastavku će biti dat pregled arhitekture i odlika sistema, sledi opis rešenja, zatim testiranje i na kraju zaključak i literatura.

II. ANDROID OS

Kao vodeći operativni sistem za mobilne uređaje na tržištu, ali i za ostatak potrošačke elektronike, Android OS sadrži specifičnu arhitekturu koja je kompleksna i podeljena u nekoliko suštinskih slojeva.

Arhitektura Android Sistema se sastoji od četiri sloja i to su:

1. Sloj jezgra
2. Radni sloj
3. Sloj programskih okvira
4. Aplikativni sloj

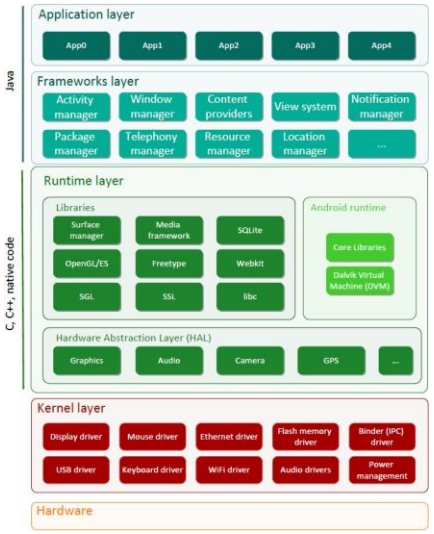
Sloj jezgra je zadužen za sledeće funkcionalnosti: rukovanje fizičkim i logičkim resursima ciljne platforme, rukovanje i pokretanje programskih procesa i obezbeđivanje konteksta izvršavanja, rukovanje sistemom datoteka i realizacija prava pristupa. Jezgro Androida je Linux, uz modifikacije specifične za Android, koje obezbeđuje dodatne funkcionalnosti neophodne za ispravan rad ostatka Androida. Modifikacije su zapravo skup zakrpa koji se primenjuje na izvorni kod Linux jezgra.

Radni sloj predstavlja skup različitih komponenata koje se izvršavaju u formi mašinskog koda na ciljnoj platformi. Namena ovog sloja jeste da obezbedi funkcionalnosti za više slojeve, obezbeđujući optimalne performanse.

Ovaj sloj se sastoji od sloja za apstrakciju i prilagođenje sistema, skupa biblioteka u izvršnom obliku i jezgra za izvršavanje Android aplikacija. Namena podsloja za apstrakciju sistema je obezbeđivanje uniformnog programskog sloja višim programskim slojevima prilagođavanjem sistemskih funkcija. Cilj uvođenja ovog sloja je izolacija viših programskih slojeva od ciljne platforme, oslanjajući se isključivo na obezbeđenu programsku spregu. U radnom sloju je enkapsuliran i značajan broj biblioteka čija je zajednička osobina to što su direktno izvršive na ciljnoj platformi. U ovaj sloj se ubacuju i biblioteke koje obezbeđuje proizvođač platforme i koje obavljaju funkcije karakteristične za datu platformu. Poslednja komponenta u radnom sloju je okruženje za izvršavanje Android aplikacija. Android je do verzije 5.0 posedovao svoju implementaciju Java virtuelne mašine Dalvick, koju je kasnije zamenio Android Runtime – ART. Glavnu motivaciju za razvoj ART-a predstavlja veće iskorišćenje resursa hardvera i stvaranje novog programskog okvira za držanje koraka sa daljim razvojem hardvera, pre svega prelaskom na 64-bitne procesore [4].

Sloj programskih okvira predstavlja spregu ka korisničkim aplikacijama. Njegov zadatak je da obezbedi adekvatnu predstavu funkcionalnosti sistema u formi Java klasa i programskih sprega. Sloj definiše programsku spregu ka aplikacijama kao skup Java klasa. Implementacija tih klasa je većinom jednostavna i brzo se završava pozivom funkcije iz radnog sloja. Mehanizam koji se koristi za integraciju sa izvršnim kodom u radnom sloju je JNI, i o njemu će biti reči kasnije u tekstu.

Najviši sloj Androida predstavljaju korisničke aplikacije. Android aplikacije se tipično razvijaju u programskom jeziku Java, oslanjajući se na Java programsku spregu koju obezbeđuje sloj radnog okvira. Postoji mogućnost da se Android aplikacije razvijaju i u drugim programskim jezicima, ali većinom nisu toliko dobro podržani kao Java. Na slici je prikazana arhitektura Android operativnog sistema.



Sl. 1. Arhitektura Android operativnog sistema

III. OPIS REŠENJA

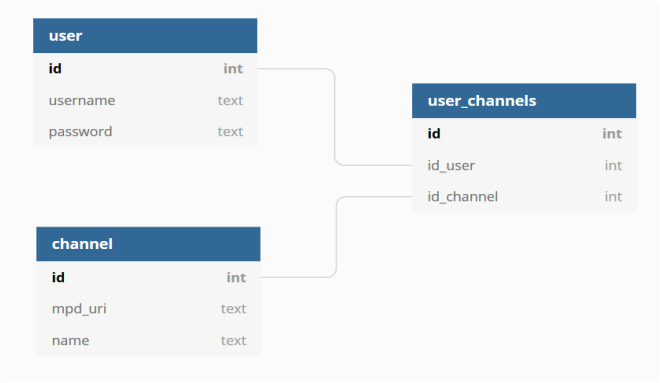
Sistem se sastoji iz dve aplikacije: Java aplikacija koja sadrži bazu podataka i grafički interfejs i native aplikacija pomoću koje se video skida sa MPD linka i emituje na odgovarajući način. Za uređaj na kojoj se razvija aplikacija izabran je Raspbery Pi 3 Model B. Koristi 64-bitni procesor sa četiri jezgra i poseduje memoriju od 1 GB. Celokupna instalacija Android operativnog sistema na uređaj je trivijalan posao koji se sastoji iz par koraka. Na slici je prikazan korišćeni uređaj.

Sl. 2. Raspberry Pi 3 Model B

Java aplikacija je realizovana kroz okruženje Android Studio. Jedna od tih prednosti jeste SQLite, kao podrška za rad sa bazom podataka. SQLite je ugrađeni DBMS u svaki Android uređaj.

Korišćenje SQLite baze podataka na Androidu ne zahteva

nikakve dodatne instalacije, te nije potrebno pokretanje pozadinskih procesa. Takođe, kako on nema klijent-server arhitekturu i to omogućava višestruki i istovremeni pristup podacima, jer se kod izvršava pri pozivu. Konkretno, u ovom radu klase za rad sa bazom podataka smeštene su u paket *database* i to su klase koje predstavljaju projekciju tabela kanal, korisnik i kanal-korisnik iz baze podataka. Respektivno, to su klase koje sadrže podatke o korisnicima, kanalima i vezi između korisnika i kanala kojima oni imaju pristup.



Sl. 3 Dijagram klasa u bazi podataka

Aplikacija sadrži nekoliko scena: scena za prijavljivanje na sistem, scene sa porukama o grešci prilikom unosa netačne lozinke ili korisničkog imena i scena sa grafičkim elementom za pregled video sadržaja i izborom kanala. Korisnik će pri instalaciji aplikacije već posedovati korisničko ime i lozinku koju je dobio od operatera kod kojeg je pretplaćen na paket kanala, što će iskoristiti prilikom prijavljivanja na sistem, prikazano na slici 3.

Sl. 4. Prijavljivanje na sistem

U zavisnosti od uspešnog ili neuspešnog scenarija prilikom logovanja korisnik se upućuje respektivno na stranicu sa kanalima koje može da prati ili na stranicu sa porukom o grešci.

Nakon uspešne prijave na sistem korisniku se pojavljuje početni ekran koji sadrži listu kanala na koje je korisnik pretplaćen kao i ravan, grafički element koji omogućava pregled video sadržaja u Android aplikacijama. Na ovom ekranu korisnik ima mogućnost da klikne na sliku koja predstavlja određeni kanal i u tom trenutku počinje reprodukcija sadržaja koja se odnosi na taj kanal, prikazano na slici broj pet.

Sl. 5. Biranje kanala

Native aplikacija je realizovana koristeći teoriju iz adaptivnih audio i video tokova. Adaptivni bitrejt prenos je tehnika koja se koristi za prenos multimedijalnog sadržaja putem kompjuterske mreže. Ranije se za ovaj prenos koristio protokol RTP, dok se danas koristi HTTP. Kako bi korisnicima pružili sadržaj koji je adaptivan u odnosu na internet konekciju velike kompanije su predložile svoja rešenja. Tako je Apple predložio HTTP Live Streaming (HLS) koji je podržan na njegovoj platformi IOS, kao i radni okvir za reprodukciju QuickTime X. Microsoft pokušava sa Smooth Streaming koji predstavlja još jedan od servisa u ISS paketu. Ipak među svima njima izdvojio se DASH [5], kao jedino rešenje koje je internacionalni standard za obezbeđivnje neprekidnog videa, ali na uštrb kvaliteta slike.

DASH jeste tehnologija koja omogućava da se multimedijalna datoteka deli na više delova koji se zatim koristeći HTTP protokol dostavljaju klijentu. Pored samih segmenata video datoteke prenose se i podaci koji nose informacije o samim segmentima. Segmenti mogu da sadrže bilo koju vrstu multimedijalnih podataka u bilo kom formatu. Preporuka je da se koristi MPEG4 datoteka, odnosno MPEG2 transportni tok za prenos tih datoteka. DASH koristi HTTP infrastrukturu koja se zapravo koristi za dostavljanje svakog WWW sadržaja što dozvoljava uređajima kao što su televizor konektovan na internet, digitalni TV prijemnici ili čak Raspberry Pi da koriste ovu tehnologiju na lak način.

Ideja native dela aplikacije jeste da se multimedijalni sadržaj, koji predstavlja prenos uživo, skida sa jedne od MPD veza, zatim skladišti i na kraju prikazuje korisniku. Za tu potrebu ključnu ulogu igraju dve biblioteke libplayer i libdash. Njihov način dodavanja u Android Studio i prevođenje u sistemu je opisan kroz makefile-ove. Paralelno se pomoću libdash biblioteke dohvata sadržaj MPD veza, a nakon preuzimanja segmenta on se šalje odgovarajućem objektu koji je zadužen za njegovu reprodukciju.

Najpre, objekat se kreira kao instance klase PlayerHandle, koja je jedna od klasa biblioteke libplayer. Zatim sledi postavljanje u početno stanje tog objekta odgovarajućim parametrima. Nakon ovoga, u klasi Stream koja proširuje klasu IAdaptionStreamDownloaderListener se implementiraju određene funkcionalnosti u odgovarajućim nasleđenim metodama za primanje sadržaja, početak skidanja i neuspeh pri skidanju. U metodi za primanje skinutog sadržaja između ostalog se obavlja i dohvatanje informacije o skinutom sadržaju odnosno da li se radi o audio ili video tipu segmenta i na osnovu tog podatka šalje se podatak funckiji feedPlayer iz

čijeg imena možemo zaključiti da se radi o funkciji koja služi za prikaz skinutog sadržaja.

IV. TESTIRANJE

Da bi se proverila ispravnost aplikacije izvršena je dinamička provera programa izvođenjem konačnog broja testova i upoređivanjem sa očekivanim ponašanjem programa. Osim same provere cilj testiranja je i stvaranje određenog nivoa sigurnosti da program radi ono što je opisano u zahtevima. U ovom radu kod testiranja fokus je na delu aplikacije pisanom na programskom jeziku Java. Sama aplikacija je testirana na dva načina: automatski i manuelno. Korišćeni su JUnit i PowerMockito [6] radni okvir koji pruža usluge za testiranje Android aplikacija. Od metoda za testiranje koje pruža ovaj radni okvir korišćen je metod koji je generalno najviše u upotrebi – assertion sistem. Testovi koji su pisani za ovaj sistem su koristeći ovaj metod poredili vrednosti testiranih metoda sa očekivanim. Rezultati testova su prikazani u tabeli.

Tabela I Rezultati testova

Opis grupe testova	Rezultat
Prijava i odjava korisnika	Prošli
Reprodukcija MP4 sadržaja	Prošli
Prikazivanje sadržaja različitog kvaliteta	Prošli

Što se tiče automatskog testiranja napisane su između ostalih i metode koje testiraju prijavljivanje korisnika na sistem sa različitim parametrima.

Manuelno testiranje je, pored standardnog načina uz pomoć testera, izvršeno i korišćenjem programa Monkey. On predstavlja alatku koja se pokreće na uređaju i koja generiše pseudo-slučajne ulaze i tako simulira manuelno testiranje aplikacije. Testiranje je dodatno proverilo ispravnost aplikacije i demonstriralo kako aplikacija radi u različitim slučajevima.

V. ZAKLJUČAK

Današnji korisnici uređaja potrošačke elektronike su sofisticiraniji od svojih prethodnika u kontekstu da više obraćaju pažnju na detalje vezane za doživljaj korišćenja datih uređaja. Proizvod s toga treba prilagoditi korisnicima, ali i potrebama na tržištu.

Ostvarena je ideja da se podigne poslednja verzija Android Things sistema na pomenutoj ploči i napiše aplikacija za reprodukciju sadržaja, koja će se sastojati iz native aplikacije koja se oslanja na postojeći DTV sprežni sloj, koja reprodukuje TV sadržaj preko mreže i nakon toga se integriše preko JNI sloja sa Java aplikacijom. Sve ovo je na kraju potvrđeno rezultatima testiranja.

U odnosu na referentni rad [3], može se složiti sa tim da Raspberry Pi nudi bolje mogućnosti od platformi LG i Panasonic, zato što Raspberry Pi nema striktno ograničeni API koji ne bi dozvoljavao pokretanje aplikacija u pozadini, čime bi se smanjila mogućnost korišćenja ovih platformi u veoma važnoj ulozi – obrada podataka.

Planovi za dalji razvoj rada su usmereni ka snimanju Widevine [7] zaštićenog sadržaja, kao i dodavanje podrške za snimanje drugih vrsta dinamičkih adaptihivnih tokova podataka, prvenstveno HLS [8].

ZAHVALNICA

Ovaj rad je delimično finansiran od strane Ministarstva za prosvetu, nauku i tehnološki razvoj Republike Srbije, na projektu broj: TR32029.

LITERATURA

[1] „StatCounter'', http://gs.statcounter.com/os-market-share/mobile/worldwide poslednji put uspešno pristupljeno 23. April 2019

[2] N.Kuzmanovic, T. Maruna, M. Savic, G. Miljkovic, Google's Android as an application environment for DTV decoder system, *14th International Symposium on Consumer Electronics*, 2010

[3] M. Yusufov, I. Kornilov, Roles of Smart TV in IoT-environments: a Survey, 13th FRUCT conference, 2012

[4] I. Pan, N. Lukić, „Projektovanje i arhitekture softverskih sistema: Sistemi zasnovani na Androidu'', FTN izdavaštvo, Novi Sad, 2015

[5] „DASH'', https://en.wikipedia.org/wiki/Dynamic_Adaptive_Streaming_over_HTTP, poslednji put uspešno pristupljeno 22. April 2019

[6] „PowerMockito'', https://github.com/powermock/powermock/wiki/mockito, poslednji put uspešno pristupljeno 22. April 2019

[7] B. Lazarević, M. Jovanović, D. Živković, Đ. Glišić „Realizacija aplikacije za snimanje MPEG-DASH toka podataka i zaštitu snimljenog sadržaja na uređajima sa Android operativnim sistemom'', 62. Konferencija za elektroniku, telekomunikacije, računarstvo, automatiku i nuklearnu tehniku (ETRAN), 11-14 Jun 2018, Palić, Srbija

[8] D. Jugović, M. Lutovac Banduka „Proširenje i integracija HLS programske podrške u Android bazirane sisteme'' 2017 Telecommunications Forum Telfor (TELFOR), Beograd, 2017

ABSTRACT

In the last ten years, the use of digital television has been expanding, technologies have been developed and new devices have emerged. There are also requests for new functionalities by users. Developing an application for watching TV on a set top box is something we have already seen before. In addition, this causes some complications if it is a standalone engineer because of the inaccessibility of the device in which applications or the lack of root rights will be developed on the same. As an excellent innovative solution in this situation, a platform for developing the Internet of Things application, offered by Google itself: Android Things. This paper presents a concrete solution for the reproduction of television content using this platform. For the implementation of the solution, the Rasberry Pi 3 Model B.

ONE SOLUTION OF REPRODUCTION MULTIMEDIA CONTENT ON INTERNET OF THINGS DEVICE

Marijana Gligorić, Vladimir Nešić, Nikola Vranić, Miloš Subotić, Đorđe Glišić

Jedno rešenje zaštite podataka na Linux Set Top Box uređajima

Aleksandra Keča Despotović, Boris Mlikota, Mario Radonjić, Miroslav Bako, *Member, IEEE*

***Apstrakt*— U ovom radu je prikazano jedno rešenje implementacije zaštitnog sistema za Linux Set Top Box (STB) uređaje. Cilj ovog rada je da prikaže kako je moguće, korišćenjem Linux kontejnera (eng. Sandboxing), postići veći nivo zaštite uređaja, i na koji način je moguće izolovati kritične procese od neželjenog pristupa podacima. Dat je kratak osvrt na teorijske osnove korišćenja Linux kontejnera, predlog implementacije ove metode na STB-u, kao i prikaz uticaja korišćenja ove metode na ukupne performance sistema.**

***Ključne reči*— STB; Set Top Box; TV; Linux kontejneri; LXC; Kernel namespace; CAS;**

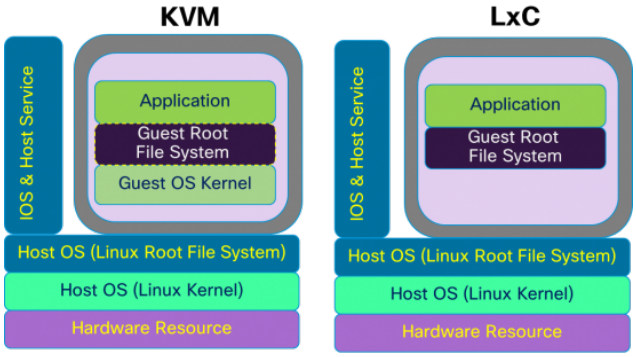
I. Uvod

Razvojem digitalne televizije sve više se javlja potreba za zaštitom tv sadržaja od neovlašćenog pristupa. Tokom vremena razvijeni su brojni metodi zaštite, u vidu kodovanja signala, ali većina tih metoda nije pružala dovoljan nivo zaštite u slučaju poverljivih podataka. Kako je rasla potreba za sve većom zaštitom podataka, tako se povećavao broj zloupotreba, u vidu zaobilaženja zaštitnih sistema i dobavljanja podataka od strane lica koja tim podacima nemaju dozvoljen pristup. Razvojem i komercijalizacijom digitalne televizije razvijali su se i različiti tipovi zaštite digitalnog televizijskog signala. U digitalnoj televiziji, kodovanje je uspešan način zaštite video i audio sadržaja, međutim postoji potreba da se procesi, u kojima se obavlja dekodovanje, zaštite od neželjenog pristupa. Kako bi dekodovanje signala unutar procesa ostalo zaštićeno, bilo je potrebno pronaći način kako bi se i okruženje, u kome se vrše procesi dekodovanja signala, učinilo bezbednim i zaštićenim. Da bi bio sprečen neovlašćen pristup bilo kom delu sistema, neophodno je obezbediti maksimalan nivo zaštite, i tako na uspešan način izolovati procese koji imaju pristup podacima koje je potrebno zaštititi. Jedan od mehanizama izolacije delova sistema, koji su razvijeni na Linux platformi, je korišćenje linux kontejnera.

II. Teorijske osnove

Linux kontejneri pružaju mogućnost izolovanja željenih procesa u logičke celine, takozvane "kontejnere", kako bi obezbedili veću izolaciju između procesa i podataka koji se nalaze u različitim kontejnerima. Cilj je postići veći nivo zaštite osetljivih podataka, bez nepotrebnog usporavanja sistema.

Linux kontejneri omogućavaju stvaranje okruženja koje je veoma slično virtualnoj mašini, ali bez opterećenja sistema koje unosi vitrualna mašina, u vidu pokretanja sopstvenog kernela i simulacije hardwera. Poređenje arhitekture Linux kontejnera i virtualne mašine prikazan je na slici 1.



Sl. 1. Poređenje arhitekture virtualne mašine i Linux kontejnera

Da bi se iskoristile prednosti linux kontejnera, koriste se već postojeće osobine linux kernela, kao što su *namespace*-ovi, kontrola pristupa i kontrolne grupe.

A. Kernel namespace

Namespace je način na koji linux kernel izoluje ili virtualizuje sistemske resurse od jednog ili više procesa. Primeri resursa koji mogu biti virtualizovani su proces ID, hostnames, user ID, pristup mreži, međuprocesna komunikacija i filesystem.

Namespace-ovi su mehanizmi kernela koji obezbeđuju izolaciju procesa (ili grupe procesa) i vizuelizaciju sistemskih resursa za te procese. Postoji sedam vrsta *namespace*-ova:

- MNT (Mount *namespace*) - kontroliše mount point
- PID (Proces ID *namespace*) - kreira PID-ove za svaki proces, koji ne zavise od drugih *namespace*-ova, pa će proces nakon kreiranja imati poseban PID za svaki *namespace*
- NET (Mrežni *namespace*) - virtualizuje mrežu na način da je svaki mrežni interfejs, virtualni ili fizički, u jednom trenutku dostupan samo u jednom *namespace*-u, a svaki *namespace* ima jedinstvene IP adrese, routing tabele, liste socketa, firewall...
- IPC (Međuprocesni komunikacioni *namespace*) - Interprocess communication sprečava procese u

Aleksandra Keča Despotović – RT-RK.doo, Narodnog Fronta 23a, 21000 Novi Sad, Srbija (e-mail: Aleksandra.Keca-Despotovic@iwedia.com)
Boris Mlikota - RT-RK.doo, Narodnog Fronta 23a, 21000 Novi Sad, Srbija (e-mail: Boris.Mlikota@rt-rk.com)
Mario Radonjić - RT-RK.doo, Narodnog Fronta 23a, 21000 Novi Sad, Srbija (e-mail: Mario.Radonjic@iwedia.com)
Miroslav Bako - RT-RK.doo, Narodnog Fronta 23a, 21000 Novi Sad, Srbija (e-mail: Miroslav.Bako@rt-rk.com)

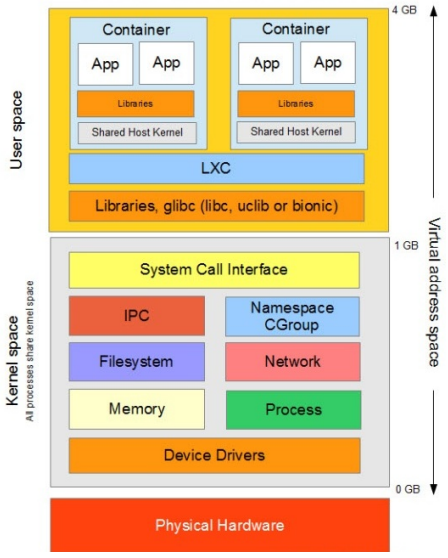
različitim IPC *namespace*-ovima da koriste iste načine komunikacije u isto vreme
- UTS (UNIX timesharing system *namespace*) - omogućava da sistem deluje kao da ima različita host i domain imena za različite procese.
- USER (Korisnički *namespace*) – pruža mogućnost da svaki kontejner ima sopstvenog korisnika (User ID) ili grupu (Group ID), kako bi se omogućila root-prava pristupa unutar kontejnera
- CGROUP (Control group *namespace*) - sprečava da drugim *namespace*-ovima bude vidljivo kojoj kontrolnoj grupi neki proces pripada

Namespace-ovi predstavljaju osnovni koncept linux kontejnera. Linux sistem inicijalizuje po jednu instancu svakog tipa *namespace*-a, a nakon inicijalizacije moguće je dodatno kreiranje željenih *namespace*-ova.

Kako bi korišćenje ovih osobina kernela, u cilju kreiranja linux kontejnera, bilo brže i jednostavnije za korisnika, ove osobine i funkcionalnosti su grupisane u jedinstven paket, koji se naziva LXC (LinuX Container).

B. LXC (Linux Container)

LXC je interfejs koji je kreiran radi jednostavnijeg i bržeg korišćenja linux kernel funkcionalnosti. Pomoću jednostavnog interfejsa i alata, omogućava linux korisnicima lak i brz način kreiranja i upravljanja linux kontejnerima. LXC se sastoji od alata, template-a i biblioteke, koji čine sloj softvera koji je prilagodljiv i podržava sve funkcionalnosti kernela, na koji se oslanja. Slikoviti prikaz LXC arhitekture prikazan je na slici 2.



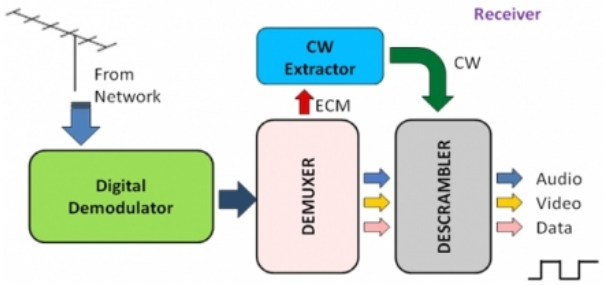
Sl. 2. LXC arhitektura

Uz pomoć LXC-a omogućena je izolacija procesa od neželjenog pristupa, od strane neovlašćenih korisnika ili drugih procesa. Linux kernel obezbeđuje CGROUP funkcionalnost koja omogućava ograničenje i određuje nivoe prioriteta pojedinih resursa (CPU, memorija, I/O blok, mreža...), bez potrebe za startovanjem virtualne mašine, kao i *namespace* funkcionalnost koja omogućava totalnu izolaciju operativnog okruženja, sa tačke gledišta aplikacije. Korišćenjem LXC-a moguće je postići nivo virtuelizacije operativnog sistema na osnovu virtualnog okruženja, koje se sastoji iz sopstvenih procesa i mrežnog prostora, umesto korišćenjem standardnih virtualnih mašina. Prednost korišćenja LXC-a, u poređenju sa sličnim rešenjima, kao što su Linux-VServer ili OpenVZ, je to što korišćenje LXC-a ne zahteva nikakve izmene u kernelu, već koristi mehanizme koji su već ugrađeni u kernel.

III. METODE ZAŠTITE PODATAKA NA STB UREĐAJIMA

U sklopu DVB (Digital Video Broadcasting) standarda definisan je standard za kontrolu pristupa, tzv. CAS (Conditional Access System), koji definiše metode zaštite digitalnog televizijskog signala. Na osnovu standarda, definisani su određeni uslovi koje je potrebno ispuniti, kako bi bio omogućen pristup digitalnom sadržaju od strane sistema koji vrši kontrolu pristupa. Zaštita podataka je ostvarena kombinacijom skremblovanja i kodovanja. Podaci se skrembluju pomoću 48-bitnog zaštitnog ključa, zvanog control word, koji se automatski generiše i koji u trenutku deskremblovanja mora biti poznat, a može se menjati i više puta u minuti. Da bi se deskremblovanje odvijalo pravovremeno i bez zastoja, zaštitni ključ mora biti dostupan u trenutku kada je potreban. Kako bi ključ bio zaštićen prilikom prenosa do prijemnika, potrebno je da se koduje kao ECM poruka (Entitlement Control Message). Nakon prijema, ključ je potrebno dešifrovati, što se postiže pomoću EMM poruke (Entitlement Management Message). Sadržaj ECM i EMM poruka nije standardizovan, već zavisi od sistema zaštite koji se koristi. Trenutno, na trzistu postoji više CAS-ova, kao sto su VideoGuard, Irdeto, Nagravision, Conax, Viaccess, Cisco, Mediaguard...

Na slici 3 prikazana je blok šema korišćenja CAS-a na STB uređaju.



Sl. 2. CAS na STB uređaju

Ovi sistemi se zasnivaju na prijemu ECM i EMM poruka, koji se dopremaju na STB preko mreže, pa samim tim zahtevaju određeni nivo sigurnosti i izolacije procesa u kojima se vrši prijem ovih važnih podataka.

IV. KORIŠĆENJE LXC-A NA STB UREĐAJIMA

U cilju bolje zaštite podataka koji se smatraju osetljivim, na STB uređajima, koji koriste linux platformu, poželjno je koristiti neku metodu izolacije delova sistema. Jedna od metoda zaštite koja je realizovana u ovom radu je korišćenje

LinuX Container-a (LXC-a). Podaci, koji se smatraju osetljivim, moraju biti izolovani u poseban kontejner, koji ne sme da ima pristup mreži (osim interprocesne komunikacije), niti da prima podatke iz izvora koji se smatraju nepouzdanim.

U slučaju korišćenja kontejnera ne smeju postojati task-ovi koji imaju root/administrativne privilegije. Takođe, kontejneri ne smeju imati direktan pristup informacijama o drugim kontejnerima, kako bi se obezbedila izolacija. Ukoliko postoji potreba za razmenom informacija između dva kontejnera, mora se obezbediti poseban kanal za razmenu podataka (zauzeti memorijski prostor – buffer), a podaci koji se razmenjuju moraju biti formatirani na unapred definisan način, tj da ispoštuju unapred definisan komunikacioni protokol. Ovi kanali za razmenu podataka između izolovanih kontejnera su najslabija tačka sistema, pa ih je potrebno obezbediti.

V. Primer korišćenja LXC-a

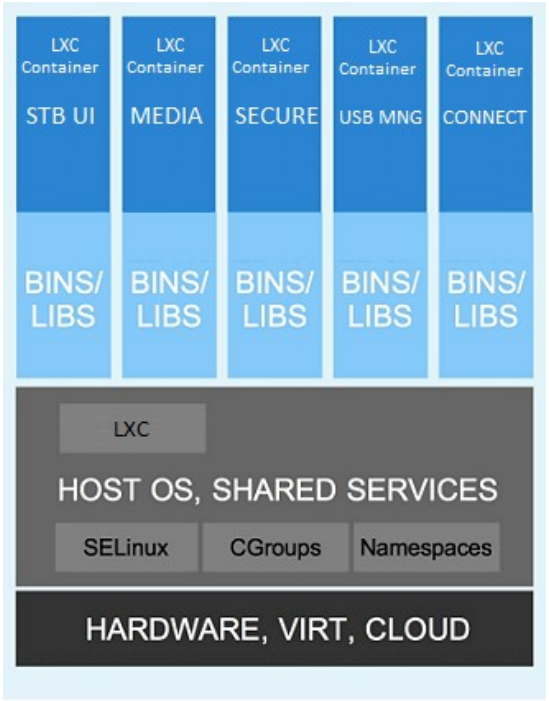
Kako se ne bi degradirali robusnost i pouzdanost sistema, program je potrebno podeliti na manje celine, tako što će se u kontejnere grupisati komponente koje imaju sličnu funkcionalnost i sličan nivo potrebne zaštite. Prilikom definisanja nivoa zaštite određenog kontejnera, treba voditi računa o sledećem:

- Server/klijent – Serverske funkcije su vidljive i dostupne spolja, pa postoji veća potreba za zaštitom u odnosu na klijentske funcije. Poželjno je odvojiti ove funkcionalnosti u razdvojene kontejnere.
- Pristup periferijama – poželjno je da se funcionalnosti koje pristupaju istim periferijama, ili grupi periferija, odvoje u zasebne kontejnere, kako bi delovi sistema koji pristupaju periferijama bili međusobno izolovani.
- Privilegije – ukoliko neke funkcionalnosti zahtevaju korišćenje operacija sa administrativnim pravima, poželjno je odvojiti ih u zaseban kontejner, i time obezbediti da budu izolovane od ostatka sistema.
- Mrežna komunikacija – funkcionalnosti koje koriste pouzdanu komunikaciju treba da se izoluju od onih koje koriste nepouzdanu komunikaciju.

Ne postoji jedinstven način razdvajanja funkcionalnosti, ali svakako treba voditi računa da se što više poštuju gore navedeni predlozi. Ukoliko određeni kontejner ima veća prava pristupa, to se javlja veća potreba da se taj kontejner zaštiti i izoluje. Delovi koda koji imaju manje prava poželjno je izdvojiti u zasebne kontejnere, kako bi se u slučaju neovlašćenog pristupa tom delu koda obezbedilo da se ne može uticati na druge delove koda, za koje je potrebno imati veća prava pristupa, i time ugrozila zaštita podataka koji moraju ostati zaštićeni.

Na osnovu navedenih preporuka i razmatranja, na STB-u se funkcionalnosti mogu izolovati u zasebne kontejnere, na sledeći način:

- STB UI – Kontejner koji izoluje browser za renderovanje main STB UI. Browser menager je odgovoran za upravljanje ulaznim parametrima RC (remote controler) dobijenim iz DirectFB (Direct frame buffer), koje zatim prosleđuje HBBTV-u, glavnom korisničkom intefejsu ili drugim kontejnerima, ukoliko ima potrebe za tim. Takođe treba da poseduje LIRC deamon koji se koristi za dekodovanje IR poruka prosleđenih ili pristiglih iz drajvera.
- MEDIA - Unutar ovog kontejnera se nalaze procesi koji su zaduženi za skladištenje Audio/Video podataka, komunikacije sa eksternim serverom, pristup spoljnoj memoriji prilikom procesa snimanja sadržaja...
- SECURE - Unutar secure kontejnera se nalaze osetljivi podaci, potrebni za dekodovanje i deskremblovanje, kojima je potrebna najveća zaštita (security biblioteke, sertifikati, licence...)
- USBMNG - Unutar ovog kontejnera nalazi se proces koji upravlja USB spoljnom memorijom. Ovaj proces je zadužen, između ostalog, za pokretanje i formatiranje USB uređaja (samo osnovne operacije, ne upravlja upisom ili čitanjem medija podataka). Ovaj kontejner mora imati obezbeđena administrativna prava pristupa.
- CONNECT - Unutar ovog kontejnera se nalazi proces koji je zadužen za upravljanje internet komunikacijom.



Sl. 4. Podela STB funkcionalnosti u izolovane LXC kontejnere

Na slici 4 se može videti grafički prikaz podele funkcionalnosti u LXC kontejnere.
Primena linux kontejnera, ili neke metode izolacije sistema, često je obavezan mehanizam, koji linux STB mora da poseduje, kako bi mogao uspesno biti ugrađen neki sistem zaštite u STB.

VI. Uticaj korišćenja LXC-a na performanse sistema

Kako bi se sagledao uticaj korišćenja LXC-a na performanse sistema, obavljena su određena merenja, čiji su rezultati zatim upoređeni sa rezultatima dobijenim korišćenjem sistema koji ne koristi mehanizam LXC-a. U cilju evaluacije korišćenjen je isti STB uređaj i u oba slučaja je bila startovana aplikacija i svi programi učitani u memoriju.

U tabeli 1 su prikazani rezultati merenja broja izvršenih operacija sa celim brojevim i operacija sa brojevima sa pokretnim zarezom, tokom vremena od 1 s, sa brojem ponavljanja 10000000. Prikazane su prosečne vrednosti.

Tabela 1

Prosečan broj izvršenih operacija tokom 1s

	Sa LXC	Bez LXC
Broj operacija sa celim brojevima	4739336.5	4672897.0
Broj operacija sa brojevima sa pokretnim zarezom	1180.55	1179.21

U tabeli 2 i 3 su redom prikazani i upoređeni rezultati merenja brzine upisa i čitanja celobrojnih podataka u RAM memoriju.

Tabela 2

Brzina upisa celobrojnih podataka u RAM memoriju

	Sa LXC	Bez LXC
1 Kb blok	5393.89 MB/s	5470.93 MB/s
4 Kb blok	5833.71 MB/s	6667.90 MB/s
16 Kb blok	6704.24 MB/s	6474.12 MB/s
64 Kb blok	6650.04 MB/s	5793.39 MB/s
256 Kb blok	4529.99 MB/s	4271.54 MB/s
1024 Kb blok	1258.92 MB/s	1800.65 MB/s
4096 Kb blok	1167.96 MB/s	1079.12 MB/s
16384 Kb blok	1083.21 MB/s	967.47 MB/s

Tabela 2

Brzina čitanja celobrojnih podataka iz RAM memorije

	Sa LXC	Bez LXC
1 Kb blok	5982.03 MB/s	5930.57 MB/s
4 Kb blok	6629.49 MB/s	6631.23 MB/s
16 Kb blok	6639.82 MB/s	5960.21 MB/s
64 Kb blok	3327.35 MB/s	2979.99 MB/s
256 Kb blok	2931.33 MB/s	2949.76 MB/s
1024 Kb blok	1333.88 MB/s	1204.70 MB/s
4096 Kb blok	1517.67 MB/s	1468.60 MB/s
16384 Kb blok	1501.93 MB/s	1521.87 MB/s

U tabeli 4 upoređena su vremena pokretanja sistema sa i bez korišćenja LXC-a. Vreme pokretanja, u ovom slučaju, podrazumeva period od trenutka pokretanja sistema, do trenutka kada su svi programi ucitani u memoriju i aplikacija pokrenuta.

Tabela 4

Prosečno vreme pokretanja sistema

	Sa LXC	Bez LXC
Vreme pokretanja sistema	101 s	101 s

Na osnovu prikazanih rezultata zaključeno je da korišćenje LXC-a značajno ne utiče na performance sistema, uz postignut viši nivo zaštite STB uređaja, korišćenjem LXC-a.

VII. Zaključak

Na osnovu praktične realizacije gore opisanih metoda zaštite, uspešno je postignut viši nivo zaštite STB uređaja, korišćenjem LXC paketa. Ispoštovane su navedene preporuke prilikom grupisanja funkcionalnosti, i postignuta je bolja zaštita bez značajnog uticaja na performanse sistema. Korišćenjem LXC-a kreiran je potreban nivo sigurnosti okruženja, kako bi uspešno mogao biti integrisan bilo koji CAS. STB uređaji, koji se zasnivaju na LXC izolaciji procesa, su uspešno praktično realizovani, i već se nalaze u komercijalnoj upotrebi, donoseći brojne pogodnosti, kako operaterima, tako i korisnicima. Zbog povećanog nivoa zaštite, predstavljaju najsavremeniji i najsigurniji uređaj za prijem i obradu digitalnog televizijskog signala.

Zahvalnica

Hvala kolegama Mlikota Borisu, Radonjić Mariu i Bako Miroslavu na njihovoj stručnoj pomoći, znanju i kolegijalnosti, bez koje ovaj rad ne bi bio uspešno realizovan.

Literatura

[1] https://www.linkedin.com/pulse/conditional-access-system-ganga-r-jaiswal
[2] http://www.infoworld.com/article/3072929/linux/containers-101-linux-containers-and-docker-explained.html
[3] http://blogs.cisco.com/enterprise/what-the-heck-is-a-service-container
[4] https://4.bp.blogspot.com/-_A_bhJee3bc/VsxT7aloFjI/AAAAAAAAArc/TqkqhPlekEo/s1600/lxc-architecture-www.hackthesec.co.in.jpg
[5] https://my.oschina.net/jxcdwangtao/blog/828651
[6] https://www.toptal.com/linux/separation-anxiety-isolating-your-system-with-linux-namespaces
[7] https://en.wikipedia.org/wiki/LXC
[8] https://en.wikipedia.org/wiki/Linux_namespaces

Abstract

This paper presents implementation of security protection system on Linux Set Top Box (STB) devices. Goal of this paper is to represents how to achieve a higher level of protection on STB devices, using Linux container methods, and how critical processes and data can be isolated from unauthorized access. In this paper was given a short overview of Linux container methods and suggested one solution for using this methods on STB device. It is shown how using of this methods have an effect on system performances.

One solution of data protection in Linux Set Top Box

Aleksandra Keča Despotović, Boris Mlikota, Mario Radonjić, Miroslav Bako

Daljinska obrada mamografskih slika korišćenjem Matlab Web Servisa

Marina Milošević, Dejan Vujičić, Željko Jovanović, Đorđe Damnjanović i Maja Radović

***Apstrakt*— U ovom radu predložena je metoda za daljinsku obradu mamografskih slika (mamograma) bazirana na primeni Matlab Web Servisa. Kreiran je udaljeni sistem za obradu slike u vidu Matlab aplikacije koja se izvršava na serveru. Grafički korisnički interfejs omogućava korisniku da učita mamografsku sliku koja se šalje na server, unese ulazne parametre neophodne za obradu slike i nakon obrade prikaže dobijene rezultate. Predstavljeni računarski sistem za obradu mamografskih slika baziran je na primeni metoda za segmentaciju mamograma kojima je moguće povećati vidljivost najranijih pokazatelja tumora - mikrokalcifikacija. Vidljivost mikrokalcifikacija značajno je poboljšana primenom Sobelove metode za izdvajanje ivica i metode za povećavanje kontrasta slike.**

***Ključne reči*—Mamografija; Mikrokalcifikacije; Segmentacija slike; Matlab; Web servis.**

I. Uvod

Kancer dojke je najčešći oblik maligniteta kod žena, koji karakteriše neprekidan porast obolelih [1]. Dijagnostička metoda koja omogućava precizno otkrivanje promena na tkivu dojke u najranijem stadijumu razvoja je mamografija [2,3].

Mikrokalcifikacije su veoma važan i ponekad jedini znak prisutnosti kancera dojke u početnom stadijumu razvoja, zbog čega je detektovanje mikrokalcifikacija veoma važan deo dijagnoze. U pitanju su mala ležišta kalcijuma u tkivu dojke, koja su na mamografskom snimku (mamogramu) vidljiva kao vrlo mali objekti visokog intenziteta u odnosu na okolno tkivo [4]. Prisutnost šuma na slici, nehomogena pozadina mikrokalcifikacija na mamogramu, kao i pojava svetlijeg tkiva dojke u odnosu na mikrokalcifikacije, u velikoj meri otežavaju njihovo detektovanje i mogu dovesti do toga da radiolog previdi postojeće mikrokalcifikacije. Iz tog razloga, razvoj CAD (engl. *Computer Aided Detection - CAD*) algoritama koji omogućavaju upotrebu računara u cilju potvrđivanja mišljenja radiologa ili u cilju donošenja uporedne dijagnoze kada nije moguće konsultovati drugog radiologa, je od neprocenjivog značaja. U osnovi CAD sistema opisanog u ovom radu stoji metoda za izdvajanje ivica primenom Sobelovog operatora i metoda za povećavanje kontrasta mamograma.

Značaj realizovanog CAD sistema za donošenje uporedne dijagnoze se povećava ukoliko je on dostupan većem broju korisnika u bilo kojem trenutku. Primenom Matlab web servisa omogućen je pristup opisanom CAD sistemu za detekciju mikrokalcifikacija na mamogramu sa udaljenih lokacija u svakom trenutku.

Matlab web servis ima ulogu da omogući klijentima (korisnicima) da sa udaljenih lokacija posredstvom internet pregledača pristupaju Matlab aplikacijama, koji se nalaze na serverskom računaru.

Opisani CAD sistem za detekciju mikrokalcifikacija na mamogramu napisan je u Matlabu i smešten je na lokalnom računaru koji obavlja ulogu servera. Posredstvom internet pregledača korisnik učitava mamografsku sliku i unosi potrebne vrednosti koje se šalju na server kao ulazni parametri CAD sistema za detekciju mikrokalcifikacija. Nakon izvršavanja Matlab programa na serverskom računaru, rezultat obrade unešenih podataka, tj. segmentirani mamogram postaje dostupan korisniku putem internet pregledača.

II. Primenjene metode

A. *Detektovanje mikrokalcifikacija na mamogramu*

Pre same analize mamograma sprovedena je predobrada snimaka koja obuhvata eliminisanje šuma i izdvajanje područja od interesa, kako bi se obezbedili optimalni uslovi za dalju obradu. U cilju povećanja vidljivosti mikrokalcifikacija, na mamograme su primenjene dve segmentacione metode: metoda za izdvajanje ivica primenom Sobelovog operatora i metoda za povećavanje kontrasta.

1) *Predobrada mamografskih snimaka*

Predobrada mamograma ima za cilj da obezbedi bolje uslove za izdvajanje željenih objekata. Faza predobrade predstavljenog CAD sistema sastoji se od dva segmenta: uklanjanja šuma sa slike i izdvajanja područja od interesa.

Efikasno uklanjanje šuma sa slike bez narušavanja kvaliteta obrađene slike dovodi do poboljšanja rezultata kasnije obrade.

Marina Milošević – Fakultet tehničkih nauka Univerziteta u Kragujevcu, Svetog Save 65, 32000 Čačak, Srbija (e-mail:marina.milosevic@ftn.kg.ac.rs).

Dejan Vujičić – Fakultet tehničkih nauka Univerziteta u Kragujevcu, Svetog Save 65, 32000 Čačak, Srbija (e-mail: dejan.vujicic@ftn.kg.ac.rs).

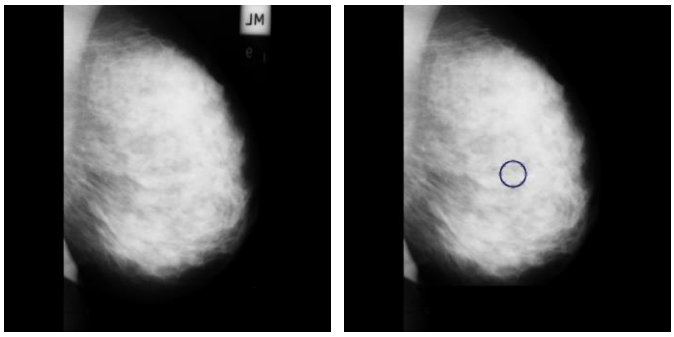
Željko Jovanović – Fakultet tehničkih nauka Univerziteta u Kragujevcu, Svetog Save 65, 32000 Čačak, Srbija (e-mail: zeljko.jovanovic@ftn.kg.ac.rs).

Đorđe Damnjanović– Fakultet tehničkih nauka Univerziteta u Kragujevcu, Svetog Save 65, 32000 Čačak, Srbija (e-mail: djordje.damnjanovic@ftn.kg.ac.rs).

Maja Radović – Fakultet tehničkih nauka Univerziteta u Kragujevcu, Svetog Save 65, 32000 Čačak, Srbija (e-mail: maja.radovic@ftn.kg.ac.rs).

Uklanjanje šuma sa mamograma izvršeno je primenom medijan filtra [5].

Izdvajanje područja od interesa koje podrazumeva dojku izdvojenu iz pozadine, je neophodan segment faze predobrade, posebno u slučaju skeniranih analognih mamograma koji imaju problem nejednako osvetljene pozadine i često sadrže nepoželjne objekte koje je potrebno ukloniti. Za izdvajanje područja od interesa primenjena je procedura predstavljena u [6]. Rezultat primene ove procedure na mamogram iz Mini-MIAS baze podataka [7] sa kojeg je uklonjen šum (Sl. 1(a)), prikazan je na Sl. 1 (b). Na Sl. 1 (b) zaokruženo je područje na kojem su prisutne mikrokalcifikacije.



(a) (b)

Sl. 1. Predobrada mamograma: a) Mamogram mdb266 iz Mini-MIAS baze podataka sa kojeg je uklonjen šum, b) Dojka izdvojena iz pozadine sa zaokruženim klasterom mikrokalcifikacija.

2) Segmentacija mamografskih snimaka

U cilju povećanja verovatnoće otkrivanja klastera mikrokalcifikacija i pojedinačnih mikrokalcifikacija, primenjena procedura za segmentiranje mamograma obuhvata dve faze.

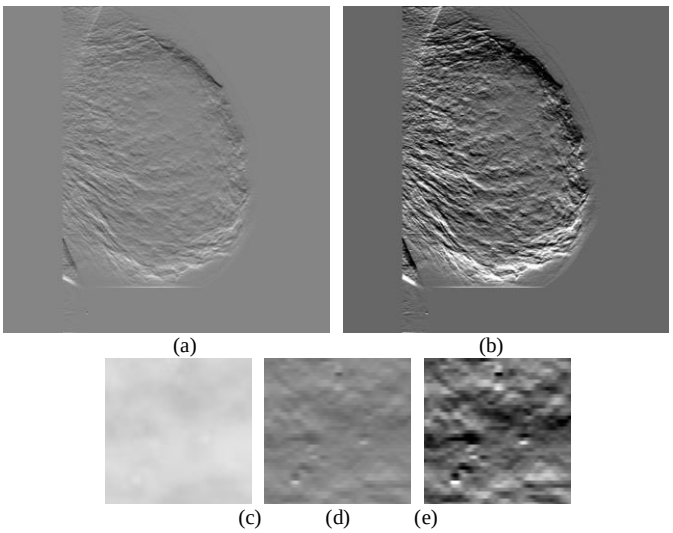
Prva faza segmentacije mamograma podrazumeva proces kojim se slika deli na dva dela, pozadinu i ivice mikrokalcifikacija. Ivice mikrokalcifikacija izdvojene su primenom Sobelovog operatora [8]. Sobelov detektor ivica je najpoznatiji i najpopularniji među klasičnim metodama za detekciju ivica zbog svoje jednostavne implementacije i dobrih performansi. Rezultat primene ove metode je dvodimenzionalna mapa gradijenata izračunatih u svakoj tački slike, pri čemu je amplituda gradijenta velika na ivicama objekata, dok je unutar i izvan objekata uglavnom niska [9]. Dvodimenzionalna mapa y-gradijenata (projekcija gradijenta slike na y-osu), dobijenih primenom Sobelovog operatora, prikazana je na Sl. 2 (a). Na ovoj slici, vidljivost mikrokalcifikacija je malo bolja u poređenju sa originalnim mamogramom.

U drugoj fazi segmentacije mamograma, na sliku dobijenu primenom Sobelovog operatora primenjena je operacija povećavanja kontrasta. Povećavanje kontrasta, kojim se postiže bolja vidljivost objekata na slici, ostvareno je primenom operacije za podešavanje vrednosti intenziteta piksela slike. Transformacija intenziteta piksela izvršena je preslikavanjem vrednosti ulazne slike u nove vrednosti na izlaznoj slici tako što se vrednosti intenziteta piksela koje se nalaze između zadatih graničnih vrednosti u ulaznoj slici (*low_in*, *high_in*) preslikaju u vrednosti između zadatih graničnih vrednosti u izlaznoj slici (*low_out*, *high_out*). Sve vrednosti manje od *low_in* se preslikavaju u vrednost *low_out*, a vrednosti veće od *high_in* se preslikavaju u vrednost *high_out*.

Primena opisane procedure za povećavanje kontrasta rezultirala je značajnim poboljšanjem vidljivosti mikrokalcifikacija (Sl. 2 (b)). Mikrokalcifikacije su postale lako uočljive, jasno se vidi gde se nalaze i kakvog su oblika, što je veoma bitno zbog njihove klasifikacije na maligne i benigne.

Na kraju, lokalizacija klastera i pojedinačnih mikrokalcifikacija može se izvršiti upoređivanjem segmentiranih i originalnih mamograma.

Sl. 2 (c) prikazuje klaster mikrokalcifikacija izdvojen iz originalnog mamograma. Rezultati prve i druge faze segmentacije primenjene na izdvojene delove mamograma prikazani su na Sl. 2 (d) i Sl. 2 (e), respektivno.



(a) (b)

(c) (d) (e)

Sl. 2. Segmentacija mikrokalcifikacija: (a) Dvodimenzionalna mapa y-gradijenata dobijenih primenom Sobelovog operatora, (b) Rezultat primene operacije povećavanja kontrasta, (c) Klaster mikrokalcifikacija izdvojen iz originalnog mamograma, (d) Rezultat prve faze segmentacije izdvojenih mikrokalcifikacija, (e) Rezultat druge faze segmentacije izdvojenih mikrokalcifikacija.

Vizuelnom procenom utvrđeno je da primenjena metoda segmentacije daje zadovoljavajuće rezultate kod detektovanja mikrokalcifikacija. Efikasnost ove metode potvrđena je klasifikacijom originalnih i segmentiranih mamograma u dve kategorije, kategoriju mamograma koji sadrže mikrokalcifikacije i kategoriju mamograma koji ih ne sadrže, primenom klasifikacione metode opisane u [6].

Mini-MIAS baza podataka sadrži 23 mamograma na kojima su prisutne mikrokalcifikacije, pri čemu je sa tih mamograma moguće izdvojiti 28 klastera mikrokalcifikacija. U klasifikacionom testu korišćeno je ukupno 60 uzoraka i to 28 slika koje sadrže klastere mikrokalcifikacija i 32 slike koje predstavljaju delove mamograma na kojima nema mikrokalcifikacija. U tri klasifikaciona testa primenjene su tri različite klasifikacione šeme: klasifikator zasnovan na podržavajućim vektorima (engl. Support Vector Machine - SVM) [10], klasifikator koji koristi metodu k-najbližih suseda (engl. K-Nearest Neighbor - k-NN) [11] i naivni Bajesov klasifikator [12].

Rezultati predstavljeni u [13] pokazali su da je u sva tri klasifikaciona testa najuspešnija bila klasifikacija segmentiranih delova mamograma sa povećanim kontrastom. Zatim sledi klasifikacija segmentiranih delova bez povećanja kontrasta, dok je najmanje uspešna klasifikacija delova originalnih mamograma. Deo pomenutih rezultata prikazan je u Tabeli 1.

Tačnost klasifikacije nešto veća od 50%, čini klasifikaciju delova originalnih mamograma neuspešnom jer je 50% ispravno prepoznatih uzoraka jednako slučajnom pogađanju. Nešto veća tačnost klasifikacije postignuta je klasifikovanjem delova mamograma segmentiranih primenom Sobelovog operatora. Tačnost klasifikacije ostvarena testiranjem segmentiranih uzoraka sa povećanim kontrastom, opravdava primenu predložene metode za detektovanje mikrokalcifikacija.

TABELA I
TAČNOST KLASIFIKACIJE ORIGINALNIH I SEGMENTIRANIH MAMOGRAMA OSTVARENA PRIMENOM SVM KLASIFIKATORA, K-NN KLASIFIKATORA I NAIVNOG BAJESOVOG KLASIFIKATORA.

	Tačnost klasifikacije	
SVM klasifikator	Originalni mamogrami	56.7 %
	Sobel-segmentirani	65 %
	Nakon povećanja kontrasta	88.3 %
K-NN klasifikator	Originalni mamogrami	58.3 %
	Sobel-segmentirani	60 %
	Nakon povećanja kontrasta	73.3 %
Naivni Bajesov klasifikator	Originalni mamogrami	53.3 %
	Sobel-segmentirani	65 %
	Nakon povećanja kontrasta	78.3 %

Na osnovu rezultata prikazanih u Tabeli 1, koji su detaljno izloženi i analizirani u [13], može se zaključiti da je svaka faza segmentacije mamograma doprinela poboljšanju vidljivosti mikrokalcifikacija.

B. Matlab web servis

Internet je danas najefikasniji način za komunikaciju i razmenu informacija. Takođe, omogućava korisnicima obradu podataka pomoću web baziranih aplikacija, pri tom nezahtevajući dodatni softver za pokretanje tih aplikacija. Matlab web servis (MWS) je programski dodatak MATLAB programskom paketu koji omogućava korisnicima da pristupaju aplikacijama napisanim u ovom programu preko interneta koristeći standardne web tehnologije. MWS nije deo standardne Matlab instalacije, već se instalira kao poseban dodatak. U ovom radu korišćen je programski dodatak Modelit Matlab Webservice Toolbox [14].

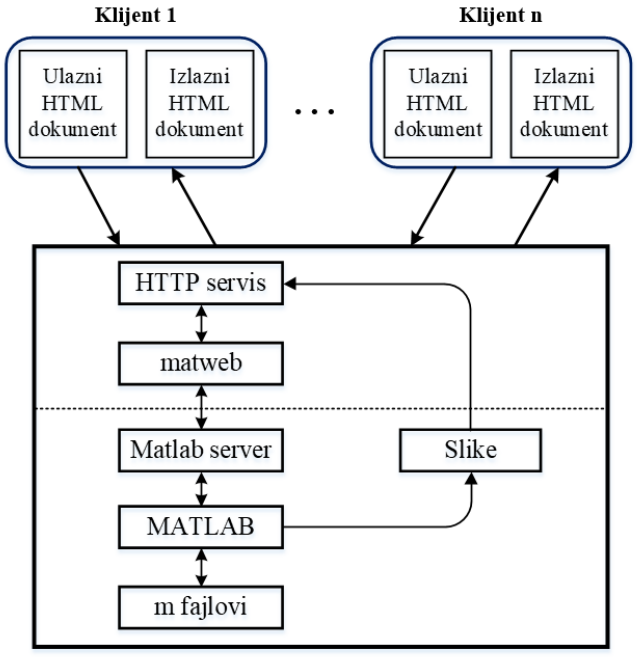
Najjednostavnija konfiguracija MWS podrazumeva da korisnik sa klijentskog računara, pomoću internet pregledača zadaje vrednosti ulaznih parametara i pristupa aplikacijama napisanim u Matlab-u, koje se nalaze na serverskom računaru. Nakon izvršavanja Matlab programa na serveru, dobijeni rezultati se prosleđuju korisniku preko internet pregledača. Na ovaj način, korisnici mogu da pristupaju aplikacijama napisanim u Matlab-u preko interneta sa bilo koje udaljene lokacije na jednostavan način, bez dodatnih troškova i potrebe da instaliraju ovaj programski paket.

MWS aplikacija predstavlja kombinaciju Matlab izvršnih datoteka (m fajlova), HTML (HyperText Markup Language) i grafičkih fajlova. Realizacija MWS aplikacije se može podeliti na sledeće korake:

- Kreiranje korisničkog interfejsa u vidu HTML dokumenta (web strane) za unos ulaznih podataka od strane klijenta i za prikaz rezultata dobijenih nakon izvršavanja Matlab programa,
- Pisanje Matlab programa u vidu izvršnog m fajla koji:
  - preuzima podatke koje je klijent uneo u ulaznom HTML dokumentu,
  - obrađuje unete podatke i generiše zahtevane izlazne podatke (numeričke vrednosti, tekst, slike, itd.),
  - Prosleđuje dobijene rezultate u izlazni HTML dokument.
- Navođenje naziva glavnog izvršnog m fajla i pripadajućih konfiguracionih podataka u konfiguracionom fajlu.

Na Sl. 3 prikazan je blok dijagram koji opisuje komunikaciju između klijentske aplikacije (klijenta) i Matlab-a preko interneta. Klijentska aplikacija (najčešće je to internet pregledač) otvara web stranu na kojoj je omogućen unos vrednosti ulaznih parametara, kako numeričkih, tako i grafičkih (slike).

Nakon unosa potrebnih ulaznih vrednosti od strane klijenta u ulaznom HTML dokumentu, klijentski zahtev se prosleđuje *matweb* komponenti preko HTTP servisa. Komponenta dijagrama *matweb* je TCP/IP (Transmission Control Protocol / Internet Protocol) klijent koji komunicira sa Matlab serverom. Njegova uloga je da korišćenjem HTTP servisa (Common Gateway Interface - CGI) preuzme podatke unete na web strani i da ih prosledi Matlab serveru.



Sl. 3. Komunikacija između klijentske aplikacije i Matlab-a preko interneta.

Matlab server predstavlja višenitni TCP/IP server koji upravlja komunikacijom između klijentske aplikacije i Matlab-a. Konfigurisan je tako da osluškuje zahteve na TCP/IP portovima koji su definisani u *matlabserver.conf* datoteci. Matlab server učitava traženi m fajl u Matlab i po završetku njegovog izvršavanja prosleđuje dobijene izlazne podatke *matweb* komponenti. Nakon toga, *matweb* ih posredstvom HTTP servisa predaje klijentu u izlaznom HTML dokumentu.

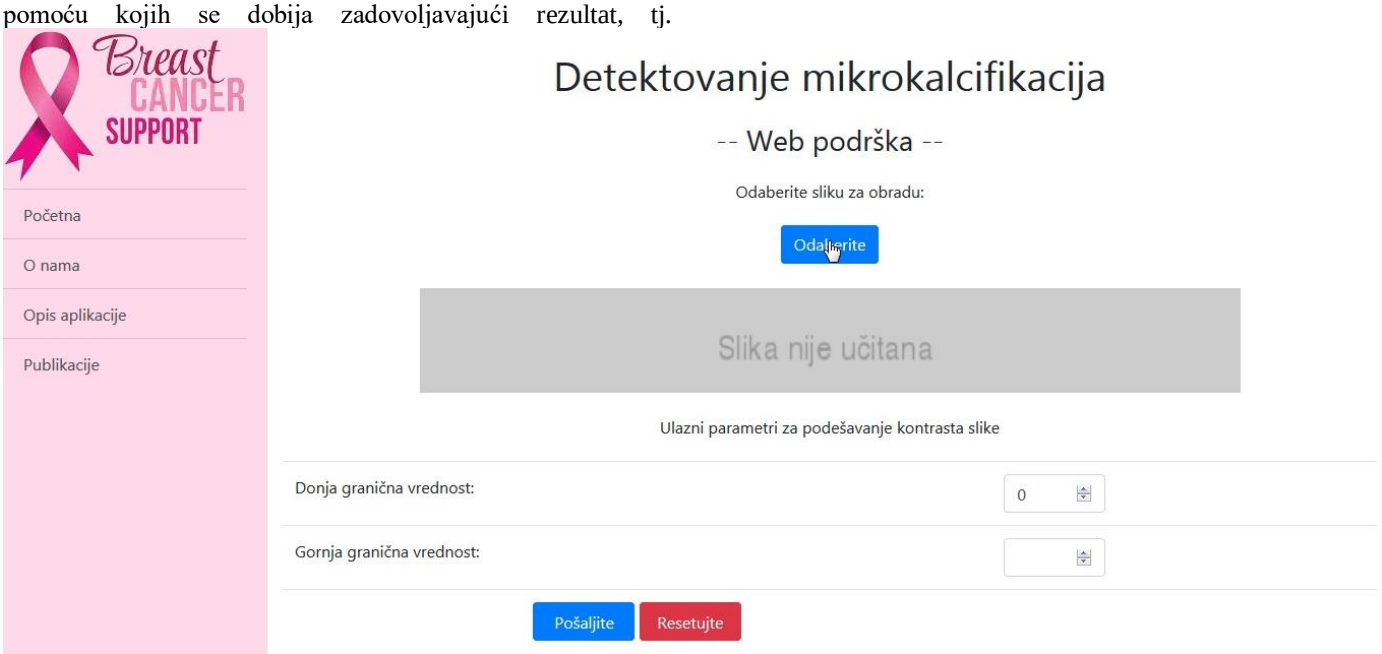
III. Rezultati

Korisnički interfejs MWS aplikacije sastoji se od dva osnovna HTML dokumenta, jednog za učitavanje mamograma koji se obrađuje i unos vrednosti ulaznih parametara i drugog za prikaz rezultata dobijenih nakon izvršavanja Matlab koda, kao i nekoliko pomoćnih HTML dokumenata čija je uloga da korisnicima olakšaju rad sa aplikacijom. Aplikaciju može da koristi svaki korisnik interneta, na svom računaru sa odgovarajućim internet pregledačem. MWS aplikacija testirana je pomoću Mozilla Firefox pregledača korišćenjem lokalnog računara koji je obavljao ulogu servera.

Korisnički interfejs za učitavanje mamograma i unos vrednosti ulaznih parametara prikazan je na Sl. 4. Klikom na dugme "*Odaberite*" korisnik bira mamogram koji će se obrađivati. Nakon učitavanja željenog mamograma i njegovog prikaza koji potvrđuje da je mamogram ispravno učitan, potrebno je zadati donju i gornju graničnu vrednost za podešavanje kontrasta slike dobijene primenom Sobelovog operatora. Polja predviđena za zadavanje graničnih vrednosti omogućavaju korisniku da izabere jednu od ponuđenih vrednosti iz dozvoljenog opsega (opseg vrednosti od 0 do 1, sa korakom 0,1). Korisnik može da zadaje različite vrednosti ulaznih parametara sve dok ne pronađe optimalne vrednosti pomoću kojih se dobija zadovoljavajući rezultat, tj. mamogram na kojem se jasno vidi da li su mikrokalcifikacije prisutne. Klikom na dugme "*Pošaljite*" učitana slika i unete vrednosti se prosleđuju aplikaciji na obradu. Dugme "*Resetujte*" omogućava brisanje svih zadatih vrednosti, uključujući i odabrani mamogram.

Nakon obrade unešenih podataka, otvara se nova web strana na kojoj se prikazuju originalni mamogram i mamogram dobijen nakon primene Sobelovog operatora i operacije za povećavanje kontrasta slike (Sl. 5). Vizuelnim upoređivanjem segmentiranog i originalnog mamograma moguće je utvrditi lokacije na kojima su prisutne mikrokalcifikacije ili zaključiti da mikrokalcifikacije nisu prisutne na datom mamogramu. Klikom na dugme "*Sačuvajte*" korisnicima aplikacije je data mogućnost čuvanja obrađenog mamograma na sopstvenom računaru, radi dalje analize i obrade.

Na levoj strani HTML dokumenata prikazanim na Sl. 4 i Sl. 5 nalaze se linkovi sa informacijama koje korisnicima olakšavaju primenu aplikacije, posebno prilikom prvog susreta sa aplikacijom. Na linku *O nama* date su osnovne informacije o autorima aplikacije. Kratak opis Matlab aplikacije za detektovanje mikrokalcifikacija dat je na web strani na koju ukazuje link *Opis aplikacije*. Ovde su prikazane informacije o mikrokalcifikacijama, predobradi mamograma, Sobelovom detektoru ivica i povećavanju kontrasta mamograma. Prilikom prvog pristupa aplikaciji neophodno je da korisnik pročita ovaj informativni tekst, kako bi mogao da zada ispravne vrednosti ulaznih parametara. Izborom linka *Publikacije* otvara se web strana na kojoj je dat spisak publikovanih naučnih radova autora aplikacije, čija je tema detektovanje mikrokalcifikacija.



Sl. 4. Web strana za učitavanje mamograma i unos vrednosti ulaznih parametara.

Sl. 5. Web strana za prikaz rezultata obrade mamograma.

IV. Zaključak

U radu je predstavljen sistem za obradu mamograma koji je zahvaljujući primeni Matlab Web Servisa dostupan preko interneta. Korisnički interfejs čine web stane za unos podataka i prikaz rezultata. Prema tome, korisnicima nije potreban dodatni softver, pored internet pregledača, da bi koristili ovu web baziranu aplikaciju.

Korisnici MWS aplikacije za otkrivanje mikrokalcifikacija na mamogramu mogu da učitaju sopstveni mamogram i da unose različite vrednosti ulaznih parametara sve dok ne dobiju zadovoljavajući rezultat, tj. mamogram na kojem se jasno vidi da li su mikrokalcifikacije prisutne.

Najvažnija prednost primene MWS-a je što korisnici mogu koristiti aplikacije napisane u Matlab-u bez instaliranja Matlab softvera, čija je cena licence poprilično visoka. Korisnici takođe ne moraju da imaju neko posebno znanje o ovom programskom paketu i o programiranju u njemu da bi pokrenuli aplikacije. Za izvršavanje MWS aplikacije dovoljno je imati odgovarajući web pregledač i biti upoznat sa namenom Matlab aplikacije, kako bi se na ispravan način zadale vrednosti ulaznih parametara aplikacije.

MWS aplikacije su dostupne preko standardnog HTTP protokola. Ovo omogućava povezivanje sa Matlabom, ne samo preko internet pregledača, već preko bilo koje aplikacije koja implementira ovaj protokol, kao što su na primer Java apleti.

Prednosti same Matlab aplikacije su velika tačnost detektovanja mikrokalcifikacija čak i kod mamograma sa vrlo gustim tkivom dojke i veoma kratko vreme obrade mamograma.

Pored brojnih prednosti, MWS ima i neke nedostatke. Jedan od nedostataka je fiksna struktura aplikacija napisanih u Matlab-u. U nekim slučajevima, promena strukture aplikacije bi bila korisna. Dalji rad biće usmeren ka nadograđivanju opisane aplikacije za otkrivanje mikrokalcifikacija, koje se ogleda u pružanju mogućnosti izbora jednog od nekoliko ponuđenih detektora ivica.

Zahvalnica

Ovaj rad podržan je projektima Ministarstva prosvete, nauke i tehnološkog razvoja Republike Srbije, broj: TR32043, III-47003 i III-41007.

Literatura

[1] A. Papadopoulos, D.I. Fotiadis, A. Likas, "An Automatic Microcalcification Detection System Based on A Hybrid Neural Network Classifier," *Artificial Intelligence in Medicine*, vol. 25, pp. 149–167, 2002.

[2] S.W. Fletcher, J.G. Elmore, "Mammographic screening for breast cancer," *The New England Journal of Medicine*, vol. 348, pp. 1672-1680, 2003.

[3] N.M. Hambly, M.M. McNicholas, N. Phelan, G.C. Hargaden, A. O'Doherty, F.L. Flanagan, "Comparison of Digital Mammography and Screen-Film Mammography in Breast cancer Screening: A review in the Irish Breast Screening Program," *American journal of roentgenology*, vol. 193, no. 4, pp. 1010-1018, 2009.

[4] M.N. Gürcan, Y. Yardımcı, A.E. Çetin, "Microcalcification segmentation and mammogram image enhancement using nonlinear filtering", Bilkent University, Department of Electrical and Electronics Engineering Bilkent, Ankara, Turkey.

[5] S.E. Umbaugh, "Computer Vision and Image Processing: A Practical Approach Using CVIPtools," New York: Prentice-Hall PTR, 1998, ISBN: 0132645998.

[6] M. Milosevic, D. Jankovic, A. Peulic, "Comparative Analysis of Breast Cancer Detection in Mammograms and Thermograms," *Biomed Eng/Biomed Tech*, vol. 60, no. 1, pp. 49-56, 2015, DOI: 10.1515/bmt-2014-0047.

[7] J. Suckling, J. Parker, D.R. Dance, S. Astley, I. Hutt, C.R.M. Boggis, I. Ricketts, E. Stamatakis, N. Cerneaz, S-I. Kok, P. Taylor, D. Betal, J. Savage, "The mammographic images analysis society digital mammogram database," *Exerpta Medica*, vol. *1069*, pp. *375–378*, 1994.

[8] P.E. Danielsson, O. Seger, "Generalized and Separable Sobel Operators," *Machine vision for three-dimensional scenes*, Academic Press, 1990.

[9] O. Castillo, P. Melin, "Type-2 Fuzzy Logic: Theory and Applications," Berlin: Springer-Verlag, 2008.

[10] B.E. Boser, I.M. Guyon, V.N. Vapnik, “A training algorithm for optimal margin classifiers,” 5th Annual ACM Workshop on Computational Learning Theory, Pittsburgh, Pennsylvania, pp. 144-152, 1992.
[11] N.S. Altman, “An introduction to kernel and nearest-neighbor nonparametric regression,” *Am Stat*, vol. 46, no. 3, pp. 175-185, 1992.
[12] I. Rish, “An Empirical Study of the Naive Bayes Classifier,” Workshop on Empirical Methods in Artificial Intelligence (IJCAI-01), 2001.
[13] M. Milošević, “Unapređenje procesa detekcije raka dojke primenom računarskog sistema za dijagnostiku integrisanog u medicinski informacioni sistem,” doktorska disertacija, Elektronski fakultet, Univerzitet u Nišu, Niš, Srbija, 2016.
[14] Modelit Webserver Toolbox for Matlab, available at: https://www.modelit.nl/index.php/webserver-toolbox-for-matlab (accessed on: October 2018)

ABSTRACT

In this paper, we propose a method for remote processing of mammographic images (mammograms) based on Matlab Web Service. We created a remote image processing system in the form of a Matlab application running on the server. The graphical user interface allows the user to upload the mammographic image that is sent to the server, enter the input parameters necessary for image processing and shows results obtained after processing. The presented computer system for mammograms processing is based on mammograms segmentation methods which can increase the visibility of the earliest tumor indicators - microcalcifications. The visibility of microcalcifications is significantly improved using the Sobel edge detection method and image contrast enhancement method.

Remote processing of mammographic images using Matlab Web Service

Marina Milošević, Dejan Vujičić, Željko Jovanović, Đorđe Damnjanović, Maja Radović

How to Build Internet Exchange Point from scratch

Nenad Krajnović, *Senior Member, IEEE*

***Abstract*—This Internet Exchange Points (IXPs) play a major role at the core of the Internet peering ecosystem. Because of that, IXPs functioning is of crucial importance for the Internet economy. IXPs should provide high availability in operation, high throughput and minimal latency. Typical IXP is built on layer 2 Ethernet switches. Using Ethernet switches has many challenges, such as how to implement redundant links in the network topology without using Spanning-tree protocol, or how to protect the network from broadcast storm, or how to prevent traffic leaking from IXP members. Choice of datacenters to be present in, and their number and connection topology should be coupled not only with technological, but also financial issues, yet with network neutrality approach always coming first. Besides Ethernet drawbacks, additional problem is how to achieve proper propagation of all routing prefixes toward IXP members. For solving this issue, all IXPs are using route-servers for simplification of route prefixes announcements among IXP members. And route-servers are based on using open-source software, such as BIRD, Quagga or OpenBGPD, with all pros and cons of using open-source software. Since the Internet stability is of major importance, operator of an IXP is responsible for controlling route prefixes announcement of every IXP member. This is related with Internet Routing Registries which should have updated information about route prefixes that every Autonomous System (AS) is announcing. Unfortunately, IRR databases are not completely accurate which means the operator of IXP should find the approach to filter announcement in proper way. All those issues should be overcome to end up with fully functional IXP which is important milestone in Internet architecture. Besides global Internet, IXPs are also of major importance for every country. IXP provides that all local content stay local instead to go around the globe in case of IXP missing, which significantly reduce the expenses for ISPs, lower latency, improve stability, and introduce scalable capacity. This is very important taking into account the importance of Internet for economy of every modern country. This paper presents experience and best practice implemented in building an Internet Exchange Point in the case of Serbian Open Exchange.**

***Index Terms*— Internet; IXP; BIRD; evolution of IXP.**

I. Introduction

IXPs became critical infrastructure for every country and their functioning is of crucial importance for the Internet economy. In this article it will be covered main technical, operational, and technological aspects, but will reflect on equally important issues of building IXP community, building trust in superb operational performance, and sharing know-how, and best practices around the technological platform. IXPs should provide high availability in operation, high throughput and minimal latency. Typical IXP is built on layer 2 Ethernet switches. Choice of datacenters to be present in, and their number and connection topology should be coupled not only with technological, but also financial issues, yet with network neutrality approach always coming first. Additional problem is how to achieve proper propagation of all IP prefixes toward IXP members. For solving this issue, vast majority of IXPs are using route-servers for simplification of IP prefixes announcements among IXP members. Route-servers are very often based on open-source software, such as BIRD, Quagga or OpenBGPd [1], with all pros and cons of using open-source software. Since the Internet stability is of major importance, operator of an IXP is responsible for controlling IP prefixes announcement of every IXP member. This is related with Internet Routing Registries (IRR) which should have updated information about IP prefixes that every Autonomous System (AS) is announcing. Unfortunately, IRR databases are not always completely accurate, due to fast changes in ISP business, and lack of time, and genuine interest of ISPs to keep the IRR database fully up-to-date, which means the operator of IXP should find the approach to filter announcement in proper way. In addition to that, there is global initiative to improve Internet routing stability named MANRS (Mutually Agreed Norms for Routing Security) [2].

Importance of open communication, sharing of best practices, and protecting other market players from security challenges coming from your own network, could not be stressed enough. Building superb technological platform without having a proper set of rules of engagement is dead end. All those issues should be overcome to end up with fully functional IXP which is important milestone in overall Internet architecture. Besides other advantages, IXP provides that all local content stay local instead to traverse around the globe in case of IXP missing, which significantly reduce the expenses for ISPs, keep the latency low, improve stability, and introduce scalable capacity. IXPs are also important milestone in national Internet infrastructure security. Operating IXP inside the country, ISPs can count on stable work even in case of huge DDoS (Distributed Denial of Service) attack which can come from the Internet. This article presents experience and best practice implemented in building an Internet Exchange Point in the case of Serbian Open Exchange - SOX [3].

Nenad Krajnović is with SOX – Serbian Open Exchange, 78 Todora Dukina, 11000 Belgrade, Serbia (e-mail: krajko@sox.rs).

II. Layer 2 IXP Realization

Today, IXPs are based on Layer 2 Ethernet technology. Ethernet technology offers very good scalability of access link capacity (starting from gigabit Ethernet, over 10G Ethernet to 100G and 400G Ethernet with link aggregation possibility) and the network itself (it is easy task to add one additional Ethernet switch in the network). With Ethernet switches, it is very easy to establish point-to-point connection between users by using virtual LAN (VLAN) and QinQ functions, if it is necessary. Because of this service, support of jumbo Ethernet frames is a must. Since IXP can be treated as telecommunication services exchange, this possibility is very important. At SOX, besides common Internet Exchange services, service of point-to-point virtual links based on VLAN technology are offered.

Ethernet has some functionality that could cause problems in everyday operation of an IXP. One of the most dangerous is broadcast storm. To protect the network, switches must have possibility to rate-limit broadcast traffic. The limit for broadcast traffic should be very small, since in standard working conditions it almost does not exist as a regular traffic. The same goes for unsolicited multicast traffic, in the event it exists in the network. It is important to remember that IPv6 is using multicast instead of broadcast. So, if IXP support IPv6 protocol, multicast traffic will be present. Some IXP members utilize Private VLANs to exchange multicast streams between networks, such as TV channels, and in that case multicast per specific VLAN should be fully enabled. Traffic filtering on Layer 2 level is also very important. It is not unusual situation that IXP customers physical link from IXP terminate on Ethernet switch and forward traffic via VLAN to the router. In such situation, it is very possible, almost imminent, that IXP customer would either send or receive some unsolicited traffic such as BPDU (Bridge Protocol Data Unit) frames from Spanning Tree Protocol, CDP (Cisco Discovery Protocol) frames, LLDP (Link Layer Discovery Protocol) frames, etc. Forwarding of such frames can cause blocking of interfaces on IXP's customers equipment. To prevent this predicament, incoming traffic should be filtered by MAC addresses and ethertype field in packet. Only 3 ethertypes should be accepted: IPv4 (0x0800), ARP (0x0806) and IPv6 (0x86dd). If the Ethernet port on IXP's switch is in access mode than "one MAC per port" feature should be used. Only frames with acceptable source and destination MAC addresses should be accepted and all other traffic should be blocked. In case the port is in trunk mode, traffic filtering should be reduced because in some VLANs can exists regular STP, CDP, LLDP frames. For those situations, per VLAN Layer 2 filtering feature would be preferable. One additional measure to put under control broadcast traffic is to establish *honey-pot* server for all IPv4 addresses that belongs to address block used for peering but that are not allocated to IXP users. Very useful tool for solving this issue is *ARP sponge* [4] developed by AMS-IX. This tool, written in Perl, listens on peering LAN for ARP traffic and when the number of ARP requests for certain IP address exceeds a threshold, it sends out an ARP Reply using its own MAC address. Final measure for controlling broadcast traffic is to setup ARP timeout timer to at least 4 hours. Changes of connected devices on IXP infrastructure are very rare, especially changing of port for the connection. Because of that, ARP timeout timer can be set to such a high value.

For the physical connectivity, optical cables are preferred media. Since the Ethernet ports should be 1G/10G, the best way (cheapest, smallest space is required) is to use SFP/SFP+ (Small Form-factor Pluggable Module) ports. And to be in position to monitor optical links on physical level, all ports and SFP/SFP+ modules should have DDMI (Digital Diagnostics Monitoring Interface) functionality which allows reading all technical parameters from optical interfaces. It is good practice to permanently monitor optical level on all optical ports via SNMP polling (not many vendors offer this feature). Those data are very useful for troubleshooting when the errors occurred. As example, Fig 1. represents degradation of attenuation of passive CWDM module.

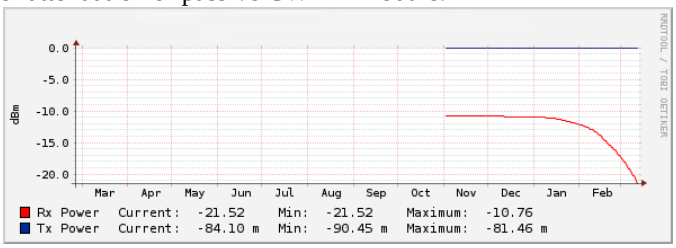


Fig. 1. CWDM passive module degradation over time

Regarding required throughput of the switch, market research should be done. Some statistics shows that 54% of IXPs in the world have traffic which is up to 20Gbps [5]. So, it is unwise to spend a lot of money for very powerful switch whose potential will not be used. SOX started with few Gbps of traffic in total, and today, after 6 years of growth, it is at 200Gbps of total traffic. SOX is not completely in accordance with statistics and that is the reason for recommendation for market research before starting this business.

Traffic analysis is one of very important function on IXP. The operator of IXP should have the proper tool to identify irregular traffic. To achieve that task, IXPs are using sflow analysis [6]. Sflow is abbreviation of "statistical flow" and represents one way for analyzing traffic that is passing the network devices. It analyze statistical sample of the traffic (e.g. every 8192nd packets). It is less demanding for the equipment than netflow but still generate good statistics and information about traffic. This is especial true in case of IXP with multi Gbps traffic. As a part of subsystem for traffic analysis, sflow collector should exist. There is a lot of commercial software for this purpose. SOX developed its own software based on pmacct [7] open source project. Advantage of in house software development is that the software can be tailored to fulfill all specific requests of operating IXP network.

Besides already mentioned features, few more are important for proper functioning of IXP. Ethernet switches should support SNMPv3 for network management, SSHv2 for

remote access to the switch, remote logging for sending all logs to central syslog server where logs should be analyzed and saved for the future use, port mirroring with preferably remote span extension for the purpose of traffic sniffing during the troubleshooting. Dual power supply is obligatory request for all equipment that should be part of critical Internet infrastructure.

III. IP Prefixes Exchange and BGP

As it is well known, BGP protocol is the major protocol for network layer reachability information exchange between ISPs denoted with Autonomous System (AS) numbers. When ISPs come to IXP, they might establish one BGP session with every member of IXP, which is huge administrative and technical work. Instead of this approach, the majority of IXPs are using route-servers for distribution of information about IP prefixes. Basic principle of work of IXP is presented on Fig. 2.

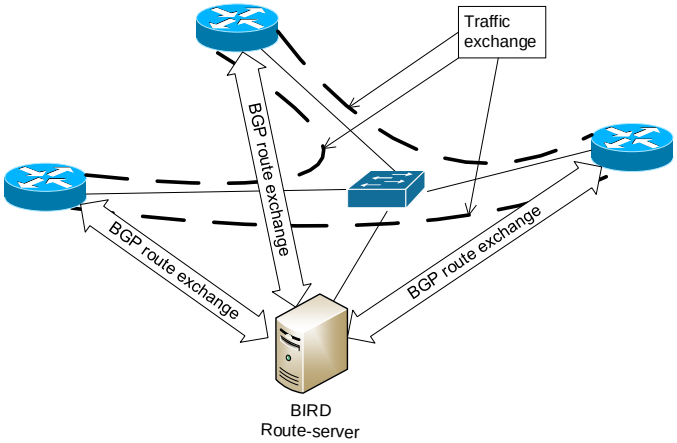


Fig. 2. Route-server in IXP network

Route-server is a simple Intel-based server connected to the same LAN as all other member's routers of IXP. Route-server is running BGP process in route-server mode and it has established BGP session with every router on LAN. In route-server mode, when BGP protocol receive IP prefix announcement from one BGP peer, it forwards it keeping the same next_hop attribute and without adding its own AS number to the AS_path attribute. Final result is that the traffic is going directly between routers and IP prefixes information are exchanged via route-server. In this setup every router has to establish only one BGP session with route-server. In reality, and for the purpose of redundancy, IXPs maintain two route-servers with the same routing policy. BGP protocol chooses the best route based on few attributes and one of the most important is AS_path attribute. AS_path contains all AS number through which this announcement passed. And the most preferred route is the one that has the shortest AS_path attribute. Since the traffic is forwarded directly between routers, this is the same situation as if they were directly connected. Because of that, best practice for the route-server is to remove IXP's AS number from AS_path attribute which is announced to the routers of IXP customers. The drawback of removing AS number is that standard BGP mechanism for detecting routing loops based on detecting AS number in AS_path attribute does not work. IXPs should implement different methods for detection of routing loops. Those methods are prefix tagging with specific community number and detection of it on incoming routing policies and exact route filtering on all incoming prefix announcements.

Additional feature of route-server is reflected in the fact that it has to maintain separate BGP table for every peering partner. Basic principle of route-server functioning is presented on Fig. 3.

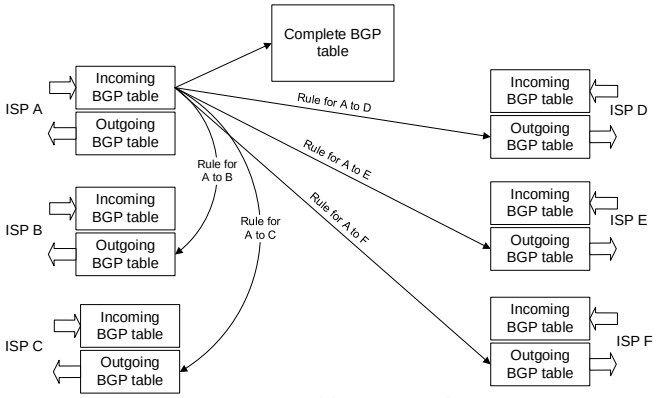


Fig. 3. Basic principle of functioning of route-server

Main task to achieve on route-server is to provide possibility for every ISP to independently exchange routing information with all IXP members. For that purpose, route-server has 2 BGP table for every peering neighbor, one is for incoming announcements and the other is for outgoing announcements. There is separate rule for distribution of IP prefix information between every two BGP tables on route-server (for clarity Fig. 1 shows only rules for the distribution information from incoming BGP table for ISP A toward all other outgoing tables on route-server). Route-server also has one common BGP table which is used for monitoring. There are tree well know open-source software for route-server: Quagga [8], BIRD [9] and OpenBGPd [10]. SOX is using BIRD software for its route-servers. BIRD software is dominant at IXPs around the world [1] and is optimize for integration with scripts.

Previously described scenario is minimal scenario for proper functioning of IXP. The operator of IXP has full control of all IP prefix information forwarding. But, this is not what the ISPs want. They want to have possibility to control announcement of their IP prefixes by themselves. That's the reason for implementation BGP community based control system. One of the BGP protocol attributes is community attribute which is well-known transitive attribute. Every IP prefix can be tagged with community attribute of some value and this can be done by ISPs. IXP operator should define the values of community attributes that will result with some specific action on route processing on route-server. The best current practice is that first two bytes of community attribute is AS number of IXP and second two bytes is specific value which is equivalent of required operation. For example, SOX

has AS number 13004, code 100 represents the action "do not announce route to ISP 10", code 101 represents the action "announce route with 1 SOX AS number prepend to ISP 10", etc. For blocking announcement of the route to ISP 10, it should have community 13004:100. Each route can have multiple community attributes which represents required handling policy of the routes on route-server. With community attribute system implemented on route-server every IXP member have full control of the announcement of their IP prefixes without direct access to the configuration of BIRD software on route-server. Novel development in this area is acceptance of Large BGP community [11] which allows more flexible mechanism for controlling BGP prefixes propagation. Problem with it is that still not all major vendors support it.

Nevertheless the foregoing description of the community based announcement control mechanism, some ISPs prefer to establish direct BGP session with IXP members. In principle, this does not interfere with normal operation of IXP, but many IXP operators insist that every IXP member must have BGP session with route-server. IP prefix filtering is an additional function of route-server which is very important for Internet stability. BGP protocol does not have any kind of route authentication which means that anyone can announce anything. There have been problems in the past that error in the configuration of the BGP protocol caused temporary loss of Internet access of some autonomous systems. In order to prevent similar events in the future, IXP has a responsibility to carefully control all the announcements of its members. The main method of control announcements is to generate IP prefix filters based on the information in Internet Routing Registry. Since ISPs do not keep IRR database in fully up-to-date state, manual tuning of route filters is a necessity. In addition to route filters, limiting the number of announced IP prefixes is another method of controlling the process of IP prefixes exchange. There is one tool specifically designed for operation of IXP by name IXP manager [12]. At the time being, SOX didn’t install it in the network management subsystem, but it is very useful software which helps the IXP operator to manage the network. The last measure of protection the LAN for peering (better known as peering LAN) between IXP members is not to announce peering LAN IP prefix to the Internet. IXP members have duty not to announce that IP prefix inside their network and outside to the Internet. This measure is protecting peering LAN from malicious activities coming without possibility of control/remedy, from the global Internet, thus reducing the possibility of various attacks, and improving troubleshooting of a problem, should it occur, obviously originating from the networks of IXP members. Problem of route-server operation are further addressed in RFC 7947 [13] and RFC 7948 [14], and SOX route-servers configuration is fully compliant with these recommendations.

IV. IXP Network Design

Equal importance of financial cost and redundancy in network topology in order to achieve high availability of IXP services is never ending story. At the beginning, IXP would typically start with one site and one Ethernet switch. As the time go by, traffic is growing, number of connection is increasing and, sooner or later, IXP have to expand to additional sites, which incurs both technological and financial questions. Opening a new POP (Point Of Presence), will enable more customer networks to connect, incurring both new revenue and new cost. As the number of POPs increase, topology should reflect market and traffic demands, retaining high availability, as there is just one switch. Biggest world IXPs like AMS-IX and DE-CIX become global IXPs with PoPs all around the world [15],[16].

Choice of fast re-route protocols is quite extensive, ranging from simple LACP (Link Aggregation Control Protocol), STP (Spanning Tree Protocol), RSTP (Rapid Spanning Tree Protocol), all the way to different MPLS options. After testing many options, we adopted well known KISS (Keep It Simple and Stupid) approach, over dimensioning backbone link and bonded multiple 10G links via LACP. If the new site is in the same city, we are talking about Metro Ethernet Network or, in case of another city, this becomes WAN network. In both cases, network topology should be carefully designed. First of all, internal interconnection, IXP backbone, links should have enough capacity to support growth of IXP traffic and to have requested availability. When we consider traffic, we are talking about tens of gigabits per second, so, dark fiber links or wavelengths in DWDM systems are the only viable choice. Fiber cut in today word is not rare situation. SOX's experience is that we have one fiber cut per week on average for international links. That imposes requirements for careful planning of network topology. Connection between sites should go over at least two physically different paths. Capacity of every physical link should be high enough to be in position to support total traffic in case of break of another link. It looks like network is over dimensioned but, since the traffic is constantly growing, capacity is not exaggerated. And availability of the service is one of the major requests for every IXP. Important question is which level of over dimensioning should be used when we are talking about link capacity. Choice of new IXP POP locations is often derived from suggestions of existing and new potential IXP members/customers, stressing once more importance of communication, collaboration, and cooperation. Adopting this open, and sometimes difficult and lengthy process, IXP upgrades its customers into members/partners in joint activity, tailored toward their exact needs. Very difficult for engineering part of the company, but very important for the whole IXP existence. Practice shows that operators are using management systems that measure 5 minute average traffic. Problem with this approach in case of Internet traffic is its burstiness. Traffic measurements carried out in SOX network shows that intermittent increase of the traffic can go up to 3 times of 5 minute average. Based on all this results, SOX has

adopted the rule that the link should be upgraded when the maximal traffic reach 50% to 70% of link capacity. Based on these recommendations, SOX internal network design is presented on Fig. 4.

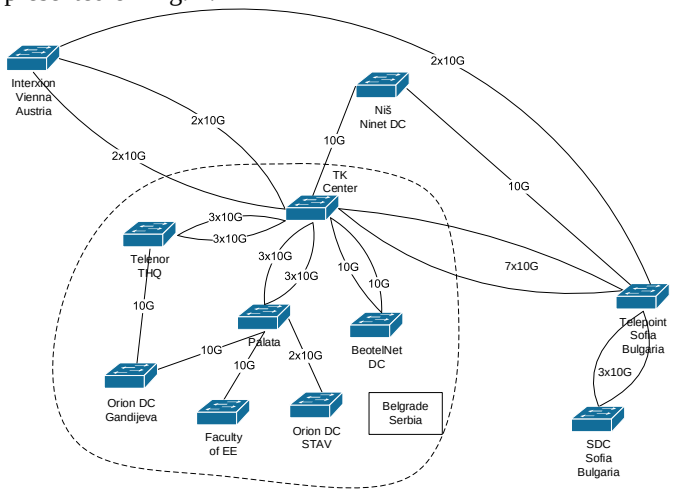


Fig. 4. SOX network topology

V. IXP Services

IXP exists primary because of customers or members (name depends on formal organization of IXP). And the primary service is peering between ISPs connected to the IXP. Everyday development of the Internet pushes IXPs to expand its service portfolio. With the raise of popularity of multimedia content, CDNs became crucial. CDNs main task is to be as close as possible to as much as possible end users, and the right place to achieve that task is to connect CDN to IXP. Besides ISPs, CDNs are the second most important element of IXP. More CDNs are connected to IXP, more ISPs will come to connect to that IXP. Two approaches can be used for connecting CDNs to IXP. On large IXPs, CDNs are using their own routers for connection. On smaller IXPs, CDNs install only caching servers and IXP is responsible for CDN traffic routing. This requests one new element in IXP architecture - router. Standard approach is to buy brand name router (such as Cisco, Juniper...) which is expensive solution acceptable for big ISPs, yet not so affordable for an small, just established IXP. In case of small IXP, price is very important element in decision process. Alternative solution for router is software router based on Linux and BIRD [9]. Limitation of software based router is 3Mpps (Mega packets per second) of throughput with equivalent 16Gbps throughput of real traffic on IXP [17]. The price of software based router is significantly lower than the price of brand name device and the performances are good enough for small IXP. SOX is using software based routers for some of the services with very good results.

Very important service for IXP is hosting anycast root DNS servers. In the process of end user accessing to any content on the Internet, first step is to contact local resolving DNS server which contact root DNS server. If the root DNS server is far from local DNS, significant latency is introduced in Internet access, and diminishes end user experience. Positioning root DNS server on local IXP decrease latency to only few milliseconds which increase the quality of experience for end users. SOX is hosting 4 anycast root DNS servers on its network (I-, J-, K- and L- root DNS servers) and it resulted with significantly faster Internet response in the region. Besides that, SOX's customers are hosting few more root DNS servers which make SOX infrastructure very well covered with this type of service.

Further expansion of SOX IXP was in the direction of interconnecting with other IXPs. Interconnection of two IXPs that are Layer 2 based is not a simple task as establishing the cooperation, high performance router is required. This router must have high throughput, support of BGP protocol and powerful mechanism for route announce/acceptance. If we want to have full IXPs integration, this router should have possibility to delete its own AS number from AS_path attribute. By removing AS number while announcing prefixes to other IXP, who is also removing its AS number when announce something to the IXP customers, IXP customers would see like they have direct peering session with customer of neighboring IXP. Here stay the same remarks as previously - since the AS number is removed, another method of detecting routing loops should be implemented. SOX implemented direct connection with few regional IXPs which increase the traffic volume on all IXPs in cooperation and advance regional traffic exchange.

As more and more traffic is passing IXP network, it becomes natural place for DDoS mitigation and protection systems. DDoS mitigation system can be installed anywhere in the network. But, if it is installed at end user network, it would be useless to protect network from saturation of Internet links. Because of that, DDoS mitigation system should be installed closer to Internet backbone where the links have enough capacity to overcome DDoS attack. And natural place is IXP network. Generally speaking, IXP links have high capacity to survive DDoS attack. When the DDoS mitigation system filters malicious traffic, regular traffic can be forwarded to the end user without saturation of the user's link. As the IXP grow, DDoS mitigation system become a must have service.

VI. Lessons Learned

In previous paragraphs, best practice knowledge is presented. Besides that, there are some lessons that we learned and that cannot be found in any best practice document. L2 filtering functionality is part of almost all brand name Ethernet switches. Unfortunately, many implementations of this function do not work properly. In worst case, it does not work at all, and in some of cases, it works partially. Since this is very import function for IXP, it should be tested before starting with operational work. And switch throughput should be checked also. We tested some Ethernet switches with specified throughput of hundreds of Gbps which was overloaded with 8Gbps. This is especially problem if we are dealing with bare-metal switches, which are very popular these days because of very good price/performance ratio.

One important lesson was learned on using DAC (Direct Attach Copper) cables for 10G Ethernet. This cable is very cheap and viable solution for interconnection of the devices inside the rack cabinet. But, this cable is consuming more energy than the fiber optic interface. We connected two Ethernet switches with DAC cable for 10Gbps. When the switches were connected, everything was working correctly. Traffic increased till 4,5Gbps when the problem started. One of the switches (we'll call it A) started detecting line down, while the other one (switch B) keep the line active all the time. What was the problem? Line driver on 10G SFP+ port on switch B did not have enough power to feed DAC cable when the traffic increase (higher traffic consume more power) and the voltage on the cable was lower than it should be. Switch A detected that as the line down and stop sending traffic. As soon as the traffic stop, switch B got enough power to feed the DAC cable and the traffic could continue normally. This resulted with link flapping when the traffic increases. Changing DAC cable with the new one didn't solve the problem since it was not the problem at all. Instead of DAC cable, we put multimode 10G SFP+ module and multimode optical patch cable and the problem was solved permanently (multimode SFP+ module consume less power than the DAC cable). Since the price of multimode SFP+ module decreased significantly and it is comparable with the price of DAC cables, our recommendation is not to use DAC cable at all, use multimode cables.

Sflow statistics are already mentioned as important information for operation of IXP. Our testing shows that not all switches generate valid sflow statistics of the traffic. Comparing with number of bytes counted on interface, the difference goes up to 60%. And this was especial case with bare-metal switches. Since gathering of sflow statistics is resource intensive task, it looks like bare-metal switches are not optimized for this function. We even noticed that some switches are sending sflow packets without all necessary data, like VLAN identification of the traffic was missing. We learned that this feature should be deeply tested before use of the switch.

At last but not the least, close cooperation with IXP members is crucial, network neutrality, and open procedures, as well as collaboration with major Internet companies, like ICANN, RIPE, Verisign, ISOC, Netnod, and respective CDNs, are proving to be very beneficial. Regular meetings with IXP members, empowering members with traffic control features, but also including members into decision making process building community is equally beneficial. Presenting latest technical and technological solutions, as well as introducing best practices, equally in operational excellence, and in educational events constitute second pillar in building successful IXP, while first pillar being superb highly available IXP service described in detail in this article.

VII. Conclusion

This article presents all the best practices that we accepted and implemented during the operation of the SOX IXP. Besides all technical issues previously mentioned, one very important thing for every IXP is open connection policy which is publicly announced. Good relationship with local community is necessity to achieve their confidence in your work. Without good communication, collaboration and cooperation with local community no hardware neither software can help in establishing and successfully operating IXP.

Acknowledgment

The author wish to thanks Jane Coffin from Internet Society, Martin Levy from CloudFlare, Kurt-Erik "Kurtis" Lindqvist from LINX, Veni Markovski from ICANN, and Arnold Nipper from DE-CIX for significant help and support at the beginning of SOX.

References

[1] "Euro-IX Internet Exchange Points 2016 Reports", https://www.euro-ix.net/media/filer_public/a7/e0/a7e09822-ded4-4b0d-8051-6a8bc2db798a/euro-ix-ixp-report-2016-final.pdf
[2] https://www.manrs.org/
[3] http://www.sox.rs/
[4] "Controlling ARP Traffic on AMS-IX platform", https://www.ams-ix.net/ams/documentation/more
[5] Remco van Mook, "The $ 1,000 Internet Exchange", RIPE 71 meeting, https://ripe71.ripe.net/presentations/30-1000-dollar-exchange-ripe71.pdf
[6] RFC 3176, "InMon Corporation's sFlow: A Method for Monitoring Traffic in Switched and Routed Networks", Sept. 2001.
[7] http://www.pmacct.net/
[8] http://www.nongnu.org/quagga/
[9] http://bird.network.cz/
[10] http://www.openbgpd.org/
[11] RFC 8195, "Use of BGP Large Communities", June 2017.
[12] http://www.ixpmanager.org/
[13] RFC 7947, "Internet Exchange BGP Route Server", Sept. 2016.
[14] RFC 7948, "Internet Exchange BGP Route Server Operations", Sept. 2016.
[15] https://www.ams-ix.net/ams
[16] https://www.de-cix.net/
[17] N. Krajnovic: "Characteristics of the Traffic on Serbian Open Exchange", 3rd International Conference on Electrical, Electronic and Computing Engineering IcETRAN 2016, Zlatibor, Serbia, june 13 -16, 2016.

One Solution for Fast Reroute in OpenFlow Networks

Nataša Maksić and Aleksandra Smiljanić, *Member, IEEE*

***Abstract*—This paper presents a solution for fast reroute in the implementation of Ethernet switching fabric based on OpenFlow. This solution adds fast traffic protection by using OpenFlow Fast Failover groups. The paper presents our design for improved reliability in software defined networks (SDN), and describes its implementation.**

***Index Terms*—OpenFlow; Fast Reroute; Ethernet.**

I. Introduction

TELECOMMUNICATION devices have grown in complexity over the past decades and introduction of new ideas has grown harder in the process. The packet switching devices became complex products manufactured by large companies. They became inaccessible for researches whose goal is to implement and evaluate some of the emerging ideas in packet forwarding. OpenFlow was introduced as an answer to this problem: it is a protocol which would enable configuration of packet matching and processing from a remote controller. This would eliminate need to write the software executing on the packet switching device, and enable simpler implementation of new algorithms which would control packet forwarding.

We have proposed implementation of the Ethernet switching fabric using OpenFlow ([1]-[3]). The standard network of Ethernet switches would be based on spanning tree protocol and learning switches. Due to the spanning tree protocol, it would be able to use only selected subset of links which do not form loops. Hence, there would not be more than one path between any two switches. Such architecture would not be suitable for achieving high performance. Our proposed solution is based on using Proxy ARP and aggregate flows, which provides the same service as standard Ethernet switches to the attached users. At the same time, it has the ability to direct packets according to the selected algorithm, in order to achieve high performance. In [1] and [2], we describe flow establishment based on LB-ECR algorithm for data center networks. In [3], we describe an algorithm for adaptive flow routing.

The important property of high performance networks is fast reaction to network failures. In standard Ethernet networks, link failure requires the recalculation of the spanning tree, which has the potential to disrupt traffic across the network for seconds. When transmitting time critical flows we wish to reduce downtime to less than 50ms. In packet networks, such traffic protection can be accomplished by fast reroute techniques. Fast reroute techniques enable redirection of the traffic in the packet switching device which detected failure.

Principle of OpenFlow is that the calculation of forwarding paths is performed by the controller. However, if a controller would perform traffic redirection, the message containing information about the failure would need to be carried to the controller and processed by the controller. Then, the controller would need to send messages to the device in order to redirect traffic. Such procedure could not be performed in short time period. Hence, the OpenFlow standard provided ability to perform fast traffic redirection based on link failure. This feature is called OpenFlow Fast Failover group and it was introduced to the version 1.1 of the OpenFlow standard [4].

Fast reroute is not a novel technique. It is well established and deployed in MPLS networks. RFC 4090 describes varieties of MPLS fast reroute and elements of RSVP-TE protocol for the configuration of MPLS fast reroute. Also, IP fast reroute is an active area of research, with the goal to provide fast traffic protection within IP packet routing.

The manufacturers of routers have implemented MPLS fast reroute and, to some extent, IP fast reroute techniques. The algorithms for finding protection paths for MPLS are implemented in routers, and proven in practice. Hence, the established algorithms for fast reroute can be applied to OpenFlow networks in order to provide fast traffic protection. These algorithms can provide OpenFlow networks with fast traffic protection needed by high performance applications. Flexibility of OpenFlow will enable further development of these algorithms, and possibly further increase in packet forwarding performance. In this paper, we implement our proposed solution in Ryu OpenFlow controller [5], and evaluate it using Mininet [6].

Section II presents an overview of the related work in the area of OpenFlow traffic protection. III introduces OpenFlow Fast Failover groups. Section IV presents the solution proposed in this paper. This section firstly describes OpenFlow network for which the fast traffic protection is implemented. Then, it describes configuration steps for protection in OpenFlow network, and illustrates the protection flows on an example. Section V describes evaluation of the protection configuration. Chapter VI contains conclusion.

Nataša Maksić is with the School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail: maksicn@etf.rs).

Aleksandra Smiljanić is with the School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail: aleksandra@etf.rs).

II. Related work

Paper [7] proposes a solution which incorporates establishment of secondary route by the controller. This solution uses BFD [8] for failure detection. It uses OpenFlow Fast Failover groups and precalculated secondary paths to provide fast protection. The controller would perform recalculation of new paths after it receives information about the failure.

Paper [9] proposes introduction of aggregate flow which would carry traffic of all affected flows around a failure. In this case, authors propose reconfiguration instead of protection, i.e. they do not use Open Flow Fast Failover groups. Instead, they rely on the controller to reroute the flow to the new path, and redirect traffic from failed link to this path.

In paper [10], authors use OpenFlow Fast Failover groups to redirect traffic. The controller periodically recalculates path of protection flows. Additionally, authors propose periodical rerouting of traffic flows in order to avoid congestion.

Paper [11] proposes an algorithm for calculation of backup paths which incorporates current queue lengths. These queue lengths are converted to link distances and used in Dijkstra algorithm. Authors do not use OpenFlow Fast Failover groups, and redirection of the flow affected by failure to backup path is performed by controller.

Paper [12] describes OpenFlow solution which uses OpenFlow Fast Failover groups for fast flow protection followed by the flow recalculation and reconfiguration performed by the OpenFlow controller.

Some papers propose the modification of the OpenFlow protocol in order to improve protection capabilities. In paper [13], authors argue that OpenFlow Fast Failover mechanism is not adequate since local detour paths may not be available or may be inefficient. They propose OpenState, an extension to OpenFlow in which internal switch state is added to flow match criteria, and this state can be changed as a part of flow actions. By using such state machines, authors propose a scheme for redirecting packets from nodes that precede the node that detected failures on the OpenFlow path. In paper [14], authors propose OpenFlow protocol extension in order to provide fast monitoring, and end-to-end protection of the path instead of fast reroute protection.

So, the idea of fast reroute protection is well known and established. In OpenFlow, its implementation needs to be adapted to the particular OpenFlow design.

Since fast reroute is well established in MPLS, many of the papers in area of OpenFlow protection are based on the concepts known from MPLS fast reroute techniques. RFC 4090 defines two varieties of fast reroute traffic protection: Facility Backup and 1-to-1 backup. Facility Backup is further split into link protection and node protection. The main difference between the two varieties is that in Facility Backup, packets from multiple LSP affected by failure are transmitted through one protection LSP. In 1-to-1 backup, each protection LSP carries traffic of one LSP affected by failure. Reconfiguration after failure plays important role in MPLS. Reconfiguration can be performed after fast reroute in order to find new paths for the traffic, or it can be used instead of protection if the duration of the downtime is not critical. Ideas from MPLS fast reroute are used in OpenFlow solutions. OpenFlow solution proposed in [9] is similar to MPLS Facility Backup, and OpenFlow solutions proposed in [7], [10]-[13] are similar to 1-to-1 MPLS protection.

III. OpenFlow Fast Failover

OpenFlow standard has enabled traffic redirection in case of failure without the need for controller actions. This redirection is performed using OpenFlow Fast Failover Groups. Fast Failover Groups are introduced in the 1.1 version of the OpenFlow standard [4].

Fast Failover Group contains list of buckets. Each bucket contains specification of port which is monitored, and set of actions to be performed if the port is in the operational state. At each moment in time, one of the buckets with the port that is available is used. Buckets are sorted, and the system chooses the first bucket in the list with the port that is in the operational state. Typically, two buckets will be configured, one for protected path, and one for protection path.

The controller needs to configure groups, primary and secondary path in advance, and the switch will redirect traffic to secondary path after the detection of the failure. This will protect the traffic in the case of failure. In the meantime, the controller will receive information about the failure; and, depending on the algorithm, it may perform reconfiguration of the affected flows.

The failure detection mechanism is not defined within the OpenFlow group interface. Open vSwitch has the ability to use BFD [8] to detect failure. The advantage of using BFD is that it monitors not only the physical state of the link, but also the functionality of the neighboring switch. Packet processing in Open vSwitch is described in [15]. Open vSwitch can process certain packet types independently of OpenFlow tables. These packet types include BFD. In this way, we can use BFD for fast detection of the state of link, and neighboring switch. The detection speed is determined by the interval between BFD messages, and the timeout interval after which the port is transferred to the failed state, and the traffic is redirected to the protection path.

IV. Elements of the OpenFlow implementation

A. Proxy ARP and Aggregate flows

Our OpenFlow implementation performs the functions of Ethernet switching fabric. Such network can be deployed in high performance computer networks, such as data center networks. In comparison with the standard Ethernet networks, OpenFlow network can apply various flow forwarding algorithms in order to improve communication performance.

The computers attached to the switching fabric need to observe this switching fabric as a network of standard Ethernet switches. Proxy ARP function enables computers to learn each other's Ethernet addresses, and to start IP communication. This function is implemented through the controller. The switch forwards an ARP request to the controller, which, then, sends the requests from other switches

to the computers attached to those switches. When the ARP reply is received on one of the switches, the controller instructs the switch attached to the computer that issued the request to send the ARP reply to this computer. When the controller receives ARP request or ARP response it stores information about server IP and Ethernet address as well as to which switch and which port of that switch the server is connected. This information speeds up Proxy ARP procedure and expires after predefined time period. The details of implementation of Proxy ARP functionality using Ryu controller are presented in [1].

First IP packet of the application flow which arrives to OpenFlow switch is forwarded to the controller. In our implementation, the controller assigns application flows to aggregate flows. One aggregate flow can carry more application flows. There may be more than one aggregate flow between two computers. Aggregate flows are established in advance. Application flows are assigned to aggregate flows according to selected flow routing algorithm. Our implementation of the creation of aggregate flows using Ryu controller is described in [1].

The aggregate flows enable faster creation of new application flows. Controller needs to modify flow tables only in switches attached to communicating computers. In the switch attached to the computer that sent first IP packet of the flow, the controller creates OpenFlow rule to include outgoing packets of the application flow to selected aggregate flow. In the destination computer of the IP packet, the controller creates OpenFlow rules to extract packets of the application flow from the aggregate flow and to forward them to the port where the destination computer is attached. It should be noted that behind OpenFlow switch port, we can have Ethernet network with more than one computer.

B. Fast Reroute

Design of fast reroute includes creation of flows that will carry packets of the flows affected by failure. In this proposal, we carry traffic of one aggregate flow with one protection flow. In fast reroute, the node that detects the failure also redirects packets to protection flows. Hence, we need to create protection flow from each of the switches on the aggregate flow path in order to provide protection from any failure on the path. We can separate three cases for this configuration: first switch of the aggregate flow path, last switch of the aggregate flow path and intermediate switches. In each of the switches with the exception of the last switch on the path, we create OpenFlow Fast Failover groups.

Fig. 1 shows the Ryu code for creation of the OpenFlow Fast Failover group. This group contains two buckets. It selects one of two ports. Ports are represented by integer identifiers stored in variables port1 and port2. References of objects and methods used in Figure 1 are taken from the Ryu documentation [5]. Object parser used in Fig. 1 represents the OpenFlow protocol.

In Fig.1, we first define the Python list named *buckets* that will contain instances of object OFPBucket. The constructor for this object takes parameters *weight*, *watch_port*, *watch_group*, *actions*. We use value 0 for the parameter *weight* since this parameter is standardized for OpenFlow Load Balancing groups, and it is not important for Fast Failover groups. The second parameter, *watch_port* is the identifier of the port which is monitored for failure. This parameter receives the port identifier. The third parameter, *watch_group*, defines the group that is observed for failure. The group is considered failed if all of its buckets are failed [16]. Since we are not interested in monitoring groups, we use the value of proto_v1_3.OFPGT_ALL for this parameter. The last parameter in the constructor for object OFPBucket is *actions*. This parameter contains the list of actions that will be performed if the bucket is selected based on the states of the ports in the Fast Failover group.

```
buckets = []
b= parser. OFPBucket (0, port1,
        ofproto_v1_3.OFPP_ANY, actions1)
buckets.append(b)
b= parser. OFPBucket (0, port2,
        ofproto_v1_3.OFPP_ANY, actions2)
buckets.append(b)
group=parser.OFPGroupMod(datapath,
        ofproto_v1_3.OFPGC_ADD,
        ofproto_v1_3.OFPGT_FF,
        group_id,
        buckets)
datapath.send_msg(group)
```

Fig. 1. Creation of OpenFlow Fast Failover group.

In Fig 1, we create two buckets, one for the port that belongs to the protected aggregate flow, and the other for the port belonging to the protection flow.

After the buckets are created, we can create Fast Failover group. For this purpose we create object of the class OFPGroupMod. The constructor for this object has the following parameters: *datapath*, *command*, *type*, *group_id* and *buckets*. Parameter *datapath* identifies the switch to which the command will be sent. Parameter *command* defines operation which should be executed on the group. Possible values of this parameter are OFPGC_ADD, OFPGC_MODIFY and OFPGC_DELETE [16]. Since we are creating new group in Fig. 1, we use the value OFPGC_ADD. The parameter *type* defines type of the group. Possible values are OFPGT_ALL, OFPGT_SELECT, OFPGT_INDIRECT and OFPGT_FF [16]. Since we are creating the Fast Failover group, we use value OFPGT_FF. Parameter *group_id* contains the integer identifier of the group. Finally, parameter *buckets* contains the list of buckets in the group that was previously created.

After the object group is created, it can be sent to the switch using method *send_msg*.

When creating buckets, we obtain two ports from the calculated paths of aggregate flow and protection flow. The OpenFlow actions assigned to buckets will differ in the output ports to which packet should be forwarded, and in the actions regarding addition of VLAN tags.

First, we will observe the case of the switch connected to server, and we consider configuration for the reception of packets from the server. Configuration for the case without protection is described in [1]. For the case with protection, the match rules stay the same. For TCP and UDP traffic, they

specify input port, protocol type, source and destination IP address, as well as the source and destination port. For the case with protection, the action for this match is a reference to the OpenFlow Fast Failover group. This group monitors two ports, and has one set of actions for each port. These sets of actions are based on the set of actions for unprotected case, described in [1]. According to the Proxy ARP procedure [1], they set the MAC address to the MAC address of the destination server, add VLAN tag of the aggregate flow and forward packets to the switch output port. In the Fast Failover group, these two sets use different values for the VLAN tag and output port corresponding to the protected path and the protection path. So, we observe paths as unidirectional, and in the first switch on the path, the Fast Failover group directs traffic either to the protected path or protection path.

In the intermediate switches along the protection path, between the first switch and last switch, the Fast Failover groups have also two entries. One entry directs traffic along the protection path. In the case of failure of the next hop on the protected path, the second entry directs packets on the protection path. Configuration for intermediate switches without protection is presented in [1]. Match is performed based on the input port and tag value, and the action is forwarding packets to the appropriate ports. With protection, match remains the same. Action for the protected path is to forward the packet to the port on the protected path. However, actions for the protection path include two steps: changing the tag of the packet and forwarding the packet to the port. In Fig. 2, we can observe creation of the actions list for the protection path in Ryu controller.

```
actions2 = [parser.OFPActionSetField
(vlan_vid=(vlan2 | ofproto_v1_3.OFPVID_PRESENT)),
parser.OFPActionOutput(out_port2)]
```

Fig. 2. Actions for protection flow

Finally, in the last switch of the unidirectional path, Fast Failover groups are not used. The switch only needs to extract packets from the protected flow, and from each of the arriving protection flows in the same way as in [15].

In the intermediate switches of the protection path, the match and actions are defined in the same way as for the intermediate switches of the protected path.

C. Traffic restoration

OpenFlow specification states that in Fast Failover groups, the switch will execute actions from first bucket with the port that is operational [16]. The first bucket is defined as the highest priority bucket. The first bucket contains actions for the protection path. After the failed link or switch is repaired, the traffic will automatically be returned to the primary path.

Transfer of traffic to the protection path may lead to the increased traffic on some link. In some solutions, for example [3], the controller continuously monitors link utilizations and performs necessary actions to avoid congestions. Such algorithms may also be applied to the paths using the protection. If the controller detects the possibility of congestion it can calculate new path for protected flow or protection flow and configure this path in the OpenFlow network.

D. Example protected flow

For illustration of the flow protection, we will observe network shown in Fig. 3.

In Fig. 3 we have six OpenFlow switches. We observe two servers connected to switches S1 and S4. These two servers need to communicate. According to the algorithm, the controller will assign communication between these two servers to aggregate flow with the path S1-S2-S3-S4. Both directions of this flow use the same path.

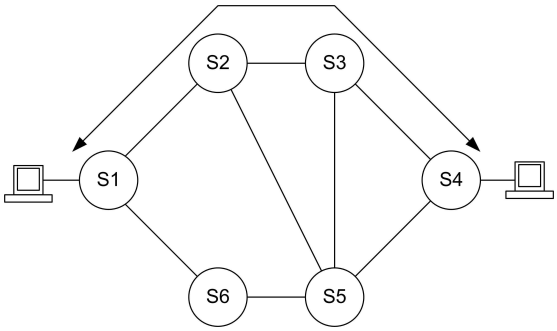


Fig. 3. Example flow which is required to be protected

In our implementations, two directions of the flow will be protected independently. Fig. 4 shows one direction of the flow and protection flows for that direction. Protected flow is drawn with a full line, and protection flows are drawn with the dashed line. Each of the protection flows provides protection for failure detected in one of the switches along the protected path. In Fig. 4, values of the flow tags are shown. The protected flow in direction from S1 to S4 has tag 1. In our implementation, tag is carried in the VLAN identification field of the Ethernet header. The protected flow that starts from switch S1 has tag 3. This flow will be used if switch S1 detects the failure. Hence, the flow with tag 3 protects from failure of the link between switches S1 and S2, or from failure of the switch S2.

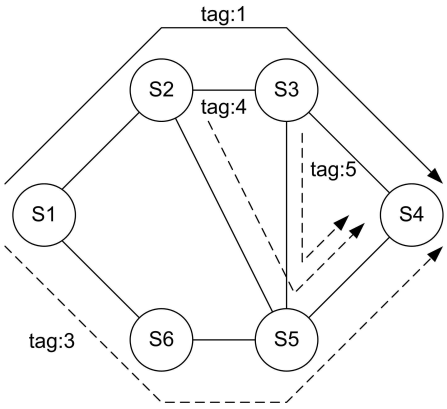


Fig. 4. Protection paths in one direction of the flow

The flow with tag 4 is a protection flow for failure of the link S2 to S3 or failure of the switch S3. Finally, the flow with tag 5 is a protection flow for failure of the link S3 to S4. We

cannot create protection from the failure of switch S4, because the flow has to use this switch.

In Fig. 5, we can observe the protected path from switch S4 to S1 with tag 2, and protection paths with tags 6, 7 and 8.

V. Evaluation using Mininet

We have validated operation of the described algorithm using Mininet [6]. Mininet setup which we use is described in [3]. Based on this setup we can evaluate the configuration generated by the controller in different topologies. For the topology presented in Fig. 3, we can generate traffic in Mininet virtual hosts, activate and deactivate links in the evaluation network using the Mininet terminal. For example, we can simulate failure of the link between switches S2 and S3 using command "link S2 S3 down", and confirm that the protection works.

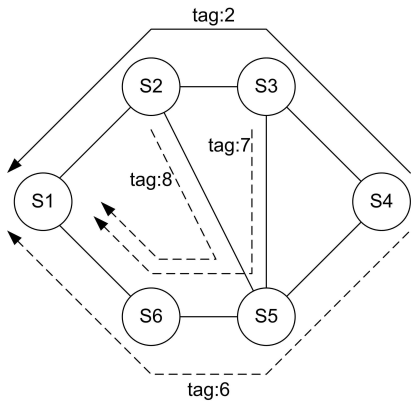


Fig. 5. Protection paths in the second direction of the flow

In this development environment, we can observe configuration of switches and transmitted traffic in flows. For example, by issuing command "sudo ovs-ofctl -O OpenFlow13 dump-groups S1", we can observe the configuration of the OpenFlow group in the first switch of the path. This command is issued in the Mininet virtual machine. Fig. 6 shows the output of this command.

```
group_id=1,type=ff,bucket=watch_port:"S1-
eth1",actions=set_field:00:00:00:00:00:02-
>eth_dst,push_vlan:0x8100,set_field:4097-
>vlan_vid,output:"S1-eth1",bucket=watch_port:"S1-
eth2",actions=set_field:00:00:00:00:00:02-
>eth_dst,push_vlan:0x8100,set_field:4099-
>vlan_vid,output:"S1-eth2"
```

Fig. 6. Configuration of the OpenFlow Fast Failover group in the first switch of the path

In Fig. 6, we can observe configuration of the OpenFlow group created at switch S1 by implementation of the proposed algorithm at the controller. We can observe that the group type is ff (Fast Failover), and we can observe the configuration of two buckets. The first bucket monitors the state of port eth1, which is the port toward switch S2 and belongs to protection path. The second bucket monitors the port eth2 which connects the switch S1 with the switch S6, and belongs to protection path. According to the Proxy ARP algorithm [1], both sets of actions set the Ethernet destination address to the address of the destination server. Also, both sets of actions add tags to the packet. Values of the tags differ, one tag is assigned to the protected path and one is assigned to the protection path. Finally, both sets of actions forward packets to the output ports. Higher priority bucket belongs to the protected path and forwards packet towards switch S1. The second bucket belongs to the protection path, and forwards packet to switch S6.

An example of configuration of Fast Failover groups in the intermediate switch along the path is shown in Fig. 7.

```
group_id=0,type=ff,bucket=watch_port:"S2-
eth2",actions=output:"S2-eth2",bucket=watch_port:"S2-
eth3",actions=set_field:4100->vlan_vid,output:"S2-eth3"

group_id=1,type=ff,bucket=watch_port:"S2-
eth1",actions=output:"S2-eth1",bucket=watch_port:"S2-
eth3",actions=set_field:4104->vlan_vid,output:"S2-eth3"
```

Fig. 7. Configuration of Fast Failover groups in switch S2

If Fig. 7, we can observe two groups at switch S2. One group protects the flow in one direction, and the other in another direction. In Fig. 4, it is shown that the protection path with tag 4 starts at switch S2. The protected path and protection path in switch S2 are configured by the first group in Fig. 7. In Fig. 7, the action rule for the second bucket contains action set_field:4100->vlan_vid which changes the value of the packet VLAN tag. By setting the value 4100, the rule set the following values inside the VLAN tag: it sets the DEI (Drop Eligible Indicator) bit to 1 and the VID (VLAN Identifier) field to 4. Since bits in the VLAN tag are processed by the OpenFlow standard, the DEI bit has not the meaning it had in standard Ethernet. In the OpenFlow specification [16], this bit is called OFPVID_PRESENT, and it indicates that the VLAN identifier is set.

Similarly, the second group in Fig. 7 corresponds to the flows in Fig. 5. The first bucket in this group corresponds to the protected flow with tag 2 and the second bucket corresponds to the protection flow with tag 8. The action for the protected path is to forward unmodified packet towards switch S1. The action for the protection path is to change value of VLAN identifier from 2 to 8 and to forward packet toward switch S5.

VI. Conclusion

This paper presents one solution for traffic protection in OpenFlow networks. Presented solution uses OpenFlow Fast Failover groups and it is based on Ryu controller. The proposed solution is built on the previous work by the authors of this paper on the creation of the OpenFlow Ethernet switching fabric [1]-[3].

The target application of the proposed OpenFlow switching fabric are data center networks [1]-[3]. Application of OpenFlow in data centers enables customization of the packet forwarding algorithms based on specific needs of the data center. Such customizations could increase the performance

and utilization of the communication network in data center. The introduction of fast traffic protection adds important new component which could be used in a design of such new algorithms. Using OpenFlow Fast Failover groups, the traffic can be redirected around the failed link or switch much faster than in standard Ethernet networks.

ACKNOWLEDGMENT

This work was supported within the project TR-32022 by the Serbian Ministry of Science and Education, and by companies Telekom Srbija and Informatika.

REFERENCES

[1] N. Maksić, "Two-Phase Load Balancing for Data Center Networks using OpenFlow," Telfor 2017.
[2] N. Maksić, "Two-Phase Load Balancing for Data Center Networks using OpenFlow," Telfor Journal, Vol. 10, No. 1, 2018.
[3] N. Maksić, Aleksandra Smiljanić, "A Scheme for Congestion Avoidance using OpenFlow," IcETRAN 2018.
[4] "OpenFlow Switch Specification, Version 1.1.0," Open Networking Foundation, February 2011.
[5] "Ryu, Release 4.30, "User Documentation, https://osrg.github.io/ryu/ resources.html#documentation, November 2018
[6] B. Lantz, B. Heller, N. McKeown, "Network in a Laptop: Rapid Prototyping for Software-Defined Networks", Hotnets 2010, Moterey, CA, USA, October 2010.
[7] N. L. M. van Adrichem, B. J. van Asten and F. A. Kuipers, "Fast Recovery in Software-Defined Networks," 2014 Third European Workshop on Software Defined Networks, IEEE, Sept. 2014.
[8] D. Katz, D. Ward, "RFC5880: Bidirectional Forwarding Detection (BFD) ," Internet Engineering Task Force (IETF), June 2010
[9] X. Zhang, Z. Cheng, R.P. Lin, L. He, S. Yu, H. Luo, "Local Fast Reroute with Flow Aggregation in Software Defined Networks," IEEE Communications Letters, Volume: 21, Issue: 4, April 2017.
[10] Y.-D. Lin, H.-Y. Teng , C.-R. Hsu , C.-C. Liao and Y.-C. Lai, "Fast Failover and Switchover for Link Failures and Congestion in Software Defined Networks," 2016 IEEE International Conference on Communications (ICC), May 2016
[11] N. M. Sahri and K. Okamura, "Openflow Path Fast Failover Fast Convergence Mechanism," Proceedings of the Asia-Pacific Advanced Network 2014 v. 38, p. 23-28.
[12] W. J. A. Silva, "Make Flows Great Again: A Hybrid Resilience Mechanism for OpenFlow Networks," Information 2018, 9, 146; doi:10.3390/info9060146
[13] A. Capone, C. Cascone, A.Q.T. Nguyen, B. Sanso, "Detour Planning for Fast and Reliable Failure Recovery in SDN with OpenState," Design of Reliable Communication Networks (DRCN), At Kansas City, MO, USA, March 2015.
[14] J. Kempf, E. Bellagamba. A. Kern, D. Jocha, A. Takacs, P. Skoldstrom, "Scalable Fault Management for OpenFlow," 2012 IEEE International Conference on Communications (ICC), Ottawa, ON, Canada, June 2012.
[15] "Open vSwitch, Release 2.10.90," User Documentation, https://docs.openvswitch.org/en/latest/, August 2018
[16] "OpenFlow Switch Specification, Version 1.3.1," Open Networking Foundation, September 2012.

Implementation of the MPLS Label Switching Procedure for the High-Speed Software Routers

Mihailo Vesović, *Graduate Student Member, IEEE,* Hasan Redžović, *Graduate Student Member, IEEE,* and Aleksandra Smiljanić, *Member, IEEE*

***Abstract*— Software routers are preferred over the classical routers when high flexibility and low costs are the requirements. Software routers are usually developed for the general purpose computers, which run standard operating systems not optimized for high-speed packet processing. For this reason, the fast I/O frameworks are used to speed up the data plane. In this paper, we have implemented Multiprotocol Label Switching (MPLS) procedure using the Data Plane Development Kit (DPDK) fast I/O. MPLS has a good synergy with the software routers as it is lightweight, scalable and protocol-independent. Our goal is to show its maximal forwarding speeds in such environment.**

***Index Terms*—Data Plane Development Kit, Fast I/O, MPLS Protocol, Software Routers**

I. Introduction

ROUTERS can be defined as a dedicated hardware devices used to perform the routing functionality. Logically, the router device is composed of the three main parts – control, data and management plane. The data plane is used to forward the traffic, while the control plane is used to decide on the routing paths. The management plane is focused on the device configuration and monitoring. Software routers use general purpose computer instead, and implement functionalities of all three planes in software. Single server (or workstation) with multiple network cards, operating system and routing software can be considered as a fully-fledged software router.

There are many candidates for the software control plane. Protocol suites like Quagga [1], BIRD [2] or Xorp [3] are all promising candidates, as they are open source projects. Management plane can be completely custom program. Unlike the control and management planes, the design of the data plane is more problematic, as it requires high packet processing speeds. Even though the standard operating systems have their own network protocol stacks, their performance is not adequate. The maximum achievable IP packet forwarding rate of the Linux protocol stack is around 4 Mpps (Mega packets per second) [4], which can be enough if 1 Gbit/s Network Interface Cards (NIC) are used. However, currently deployed servers usually have multiple 10 Gbit/s NICs installed, each of which requires the processing speeds of at least 14.88 Mpps per port.

The reasons why the Linux protocol stack is slow are multiple: expensive system calls, packet replications, additional metadata, etc [4]. In the recent years, fast I/O frameworks were developed in order to speed-up the data plane processing. Commonly used fast I/O frameworks are netmap [4], Data Plane Development Kit (DPDK) [5] and PF_RING [6]. The main idea is to pass the slow network protocol stack, and deliver the packets directly to the user application. This allows users to write their own packet processing software, tailored to their specific needs. The fast I/O frameworks process packets in batches, which significantly reduces the system call overhead. Additionally, packet replications are avoided, as well as the unnecessary memory allocations [4]. We can conclude that the data plane can be implemented as the combination of the high performance I/O framework and the user-space application, while the control plane can be any of the standard control plane suites, e.g. Quagga.

Multiprotocol Label Switching (MPLS) protocol inherently has some noteworthy properties which are of interest for software routers. The main advantage is that MPLS tunnels allow the packets to be forwarded on the additional paths, not only the shortest ones. As a consequence, it can be used to implement traffic engineering, which can improve link utilization and resiliency of the network. MPLS is classified as the layer 2.5 protocol and it can be used to encapsulate any other layer 3 protocol. Thus, backbone routers in the MPLS network (LSRs – label switched routers) are not required to know how to process any underlying protocol. This allows the backbone hardware infrastructure to be unified. Additionally, backbone routers do not need to have the complete Internet Protocol (IP) routing tables, reducing the memory consumption. MPLS also helps in the creation of peer-to-peer Virtual Private Networks (VPN) [7].

IP based routing is the standard in the today's networks, but the IPv4 longest prefix matching lookup can be expensive. It is necessary to use specific data structures, like multibit tries, in order to simplify the lookup process. Lookup is performed by traversing the tree from the root to the leaves until the most specific route is found. The lookup operation has a complexity of $O(W/k)$, where W is the IP address length, and k is the stride size. However, the insertion and deletion of routes are complex procedures which require the reorganization of the

Mihailo Vesović is with the School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11120 Belgrade, Serbia (e-mail: mikives@etf.rs).

Hasan Redžović is with the Innovation Center of School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11120 Belgrade, Serbia (e-mail: hasan.redzovic@ic.etf.bg.ac.rs).

Aleksandra Smiljanić is with the School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11120 Belgrade, Serbia (e-mail: aleksandra@etf.rs).

tree, with complexities of $O(W/k + 2^k)$. Memory consumption, on the other hand, can be as high as $O(2^k NW/k)$ for a set of N prefixes [8].

On the other hand, MPLS lookup is much simpler. MPLS labels have a local meaning in the network. Inside the MPLS network, the routing is performed solely on the MPLS label, rather than the IPv4 destination address. The packets are routed via a Label-Switched Path (LSP), which is set up by some control protocols, such as Label Distribution Protocol (LDP) [9] or RSVP-TE [10]. MPLS lookup and table updates can be performed in the $O(1)$ time [11]. It is best to use direct lookup, as there could be not more than 2^{20} possible labels. In this way, the memory consumption is limited as well. Direct lookup can be implemented as the data read from the memory address which corresponds to the MPLS label, i.e. the MPLS lookup is a simple array indexing.

In this paper, we have implemented the MPLS label switching procedure for the backbone software routers using the DPDK. Our implementation is a natural extension of the DPDK *l3fwd* application, and as such, it is parallelized. In our previous work [12], we have implemented IPv4 lookup procedure using the netmap framework. This application was parallelized using the OpenMP, where we have leveraged the advantages of the Receive Side Scaling (RSS) to disperse the packets across the multiple receive (RX) packet queues. We have also implemented MPLS forwarding for the netmap platform [13], but this implementation was only single-threaded with the limited performance. In this paper, we will show that our MPLS DPDK label switching application outperforms both of our previous solutions even without the RSS capability, for a single RX ring per input interface.

II. Fast I/O With the DPDK

DPDK (Data Plane Development Kit) is the set of libraries and drivers for the high-speed packet processing. DPDK allows the creation of network applications which require fast packet handling, for example high-throughput software switches or routers. It can be used with Linux and FreeBSD operating systems and it supports most common computer architectures [5].

The unique programming interface is available to the user via the EAL layer (Environment Abstraction Layer). The EAL hides the specific details from the users, thus allowing them to easily develop network applications. DPDK uses run-to-completion model. In this model, the core is polling for the packets on the input through an API, which are, afterwards, processed by the same core and forwarded to the specific output. Using the DPDK structures, pipelined processing can also be implemented.

DPDK allows the efficient utilization of CPU cores through the creation of multithreaded applications. The main thread is designed for the resource allocation (e.g. ring or mempool structures), which the other threads will use. DPDK application has at least one logical core thread (lcore), which is the instance of the application that is running on one core.

Interfaces, i.e. ports can have multiple receive and transmit (TX) queues associated with them, which allows parallel processing. During the configuration phase, the number of RX/TX queues is specified for each interface, as well as the number of descriptors per queue. Queues are implemented in the form of descriptor rings, and they allow FIFO operations with bulk/burst queueing/dequeueing. The descriptors are the objects that hold the pointers to the *rte_mbuf* structures where the packets are stored.

The main problem with the existing fast I/O platforms is that the packet processing functionality needs to be written from the scratch, which is time-consuming and error-prone. Luckily, the DPDK framework offers many libraries that can be used to quickly implement desired network applications. For example, the DPDK packet framework allows definition of completely custom packet processing, which is implemented by connecting standard blocks such as ports (hardware rings, software rings, etc.) and lookup tables (Exact Matching (EM), Longest Prefix Matching (LPM), Access-Control Lists (ACL) etc.). The *rte_tables* library helps users to create efficient lookup tables for batch packet processing. The library allows the user to specify the addition or removal of the MPLS header to the packet, which can be used to implement the ingress and egress MPLS functionalities.

The additional advantage of the DPDK framework is the richness of its sample applications which provide insight about the various functionalities that can be implemented. For example, application *l3fwd* is used to perform the high-speed IPv4/IPv6 packet forwarding. The IP processing is parallelized through the usage of multiple lcores. Each lcore has its own set of input and output queues assigned to it. Lcore fetches the packets in the batch from its RX queues, process them, and sends them to the appropriate TX queues. The forwarding decision is based on the packet content. For the IPv4 packets, it is the destination IPv4 address, upon which the LPM is applied. The LPM is performed via the tables designed for fast LPM lookup (*rte_tables* library). Result of the lookup process is the output interface where the packet should be sent.

III. Implementation Details

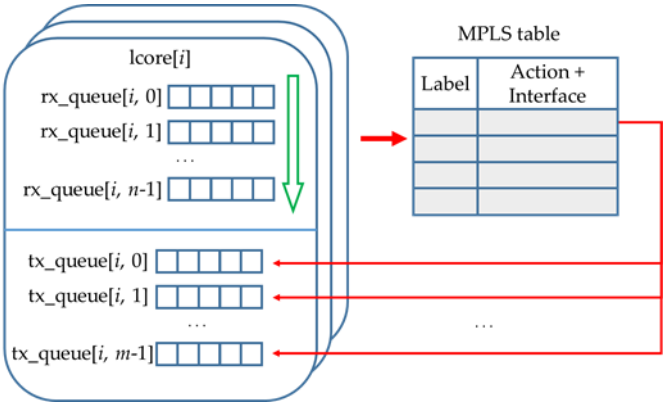
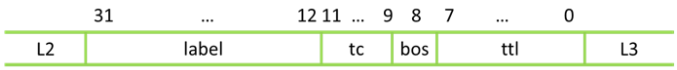


Fig. 1. Lcore RX/TX queues assignment and MPLS processing diagram

The MPLS forwarding application is partially based on the *l3fwd* application program code. Function of the main thread is similar. It parses the user arguments, allocates the resources, configures the ports, queues and lcores and checks whether the configuration is valid. At the end, the main thread launches all the lcores, which are performing the MPLS specific processing.

In the MPLS forwarding application, the number of RX queues per port is defined by the user. Users can specify the arbitrary number of queues per port, and to assign these queues to the specific lcore for processing. The number of TX queues at each port is equal to the number of active lcores, one TX queue per each lcore. In Fig. 1, we can see that each $lcore[i]$ $(i = 0,1,...,num\_lcores-1)$, has multiple RX/TX queues associated with it. The number of RX queues, n, is given via the user configuration, while the number of TX queues is $m = num\_lcores$. The packets are read from all the RX queues in a round-robin fashion. Afterwards, the packet processing is applied to all of the received packets. In this application, only the MPLS packets are processed, while the other packets are discarded. MPLS packets are modified and forwarded to the output TX queues.

Packet reads and packet writes are all performed in a burst mode. The burst size can be configurable, with the default value of 32. The packets are stored in the intermediate buffers until the specified number of packets is available for one port. If the number of packets available after some fixed period of time is less than this specified value, the packets are sent to their output. This is done in order to have latency limited.



MPLS lookup table		
Label	Action	Data
0x00000 ... 0xFFFFF	PUSH | POP | Label Switch + TTL Dec	New Label(s) + Output Interface

Fig. 2. MPLS header and MPLS lookup table

MPLS protocol has a 32-bit long header, inserted between L2 and L3 headers. Its fields are MPLS label, Traffic Class (TC), Bottom of Stack (BOS) and Time to Live (TTL). MPLS label is used to forward the traffic in the MPLS network instead of IPv4/IPv6. Its main advantages is that it is only 20 bits long, allowing simple lookup procedure. TC is used for fulfilling of the QoS requirements. The TTL field is similar to the TTL from IPv4 protocol. In fact, it inherits the TTL value from the IPv4 at the ingress router, and it is used to modify the TTL value of the IPv4 at the egress router. The BOS field indicates the last MPLS header in the case where multiple MPLS headers may exist.

The MPLS table is indexed by the MPLS label. The lookup result is the action and the output interface where the packet should be sent. The most common action is the MPLS label switching, where the old label is replaced with the new label. In this case, the lookup result also provides the new MPLS label. Other possible actions are addition/deletion of the MPLS header (push/pop), as well as the packet dropping.

In our implementation, we have created the MPLS table as an array of 2^{20} entries. For the LSRs, we have implemented only the label switching action. The lookup result has additional information about the neighbor destination MAC address, as we do not have Address Resolution Protocol (ARP) implemented. The label switching functionality could have been implemented using direct access *rte_tables* also. However, this is impractical since it imposes additional processing overhead compared to the simple array indexing.

The main MPLS function was optimized to achieve as high processing speed as possible. Some of the techniques, like function inlining, were necessary to avoid the costly function calls. Parsing of the header requires usage of the standard logical operations, like shifting, logical AND and logical OR operations. However, we have noticed some issues with the code where too many logical operations are called at the same time. If we were about to split the header into the 4 different fields, as in Fig. 2, and later reassemble it, the lower performance would be achieved, probably due to the processor resource starvation. However, we have split the MPLS header into only two distinct parts, the MPLS label part and the remaining part, and the performance was couple of Mpps better.

IV. Testing Environment

Testing environment setup is shown in Fig. 3. We have two workstations that are connected via 10G Ethernet copper cables. The first workstation is used as a packet generator. It generates packets on two available ports, with the speeds of 10 Gbit/s per port. This machine also receives and measures the forwarded traffic. The packets are generated using the DPDK Pktgen application [14], version 3.5.9.

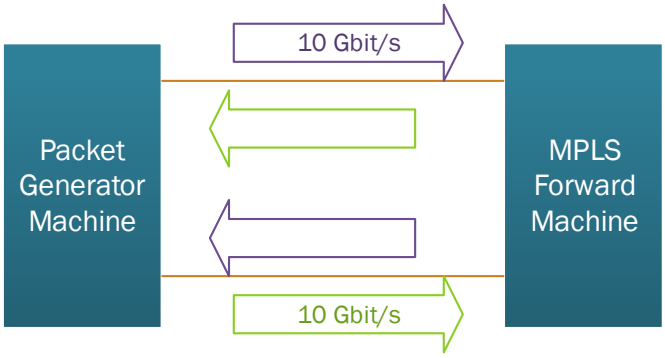


Fig. 3. Testing setup

The second workstation is used as a test machine where the MPLS forwarding application is installed. The application receives the packets, performs the MPLS lookup, and forwards the traffic back to the packet generator. The MPLS lookup table is set to forward the traffic from the port 0 to the port 1 and vice-versa. The application is based on the DPDK version 18.05.1. Characteristics of the MPLS forwarding

machine are summarized in Table I.

TABLE I
TESTING MACHINE SPECIFICATION

Processor	Intel Core i7-3770K CPU L1d/i cache: 32K L2 cache: 256K L3 cache: 8192K Frequency: 3.50GHz
RAM Memory	2 x 8 GB DDR3 Frequency: 1600 MHz
Network Interface Card	Intel Ethernet X710-DA4 Quad Port, Driver: i40e
Operating System	CentOS Linux 7, x86_64 Kernel: 3.10.0-862

V. RESULTS

We have measured the packet forwarding rates of the MPLS application for different packet sizes in the range of 64 B to 1522 B, and 1024 descriptors per queue by default. We have generated the 20 Gbit/s traffic using the DPDK pktgen. Packet rates are dependent on the packet size and can be calculated by the formula:

$$pkt\_rate = bit\_rate/(8\cdot(pkt\_size+20)) \qquad (1)$$

MPLS packet processing rates are shown in Fig. 4. From the graph, we can see that the forwarded traffic is the same as the generated traffic, except for the shortest packets. We have achieved the forwarding rates of 26.22 Mpps (17.76 Gbit/s) for 64 B packets, which is 3.54 Mpps lower than the maximal possible value of 29.76 Mpps.

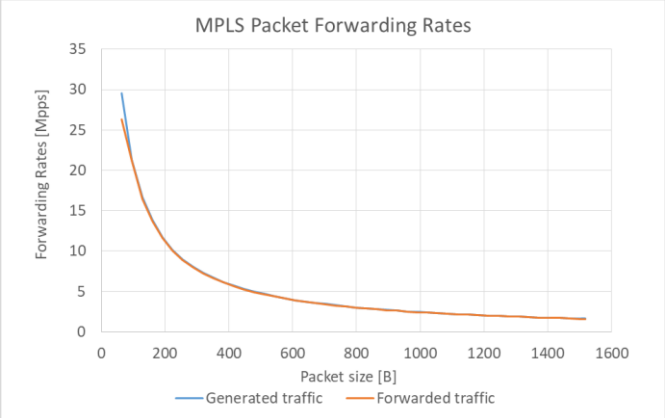


Fig. 4. MPLS packet forwarding rates for different packet sizes

We have also wanted to measure the influence of the number of RX/TX descriptors on the forwarding rates. For that purpose, we have measured the forwarding rates for the 64 B packets, and different number of descriptors per RX/TX queue. The measurements have been performed for the case where two ports are used, with one RX queue and one TX queue per each (port, lcore) pair. The results are shown in Fig. 5. The peak performance is around 27.45 Mpps for 256 RX/TX descriptors per queue. When the number of descriptors is small, fewer packets are available for processing which is the limiting factor. On the contrary, if the number of descriptors is too high, we will experience another performance drop. The reason is the suboptimal L1 cache usage.

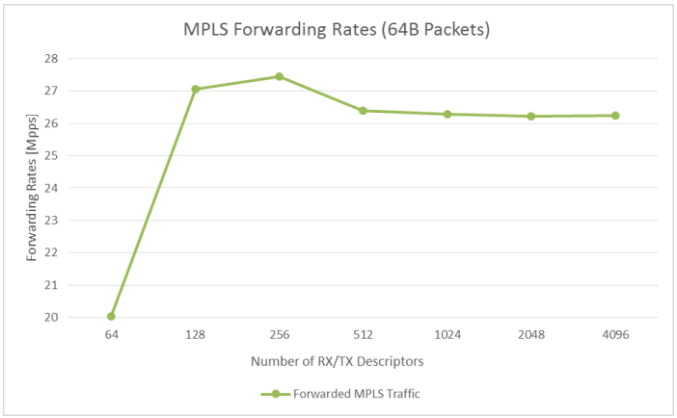


Fig. 5. MPLS packet forwarding rates for different number of descriptors per RX/TX queue and 64 B size packets

In the previous test, we have seen that the best performance is achieved for 256 descriptors per queue. However, this value can vary depending on the number of queues used and the cache memory size. Thus, it is best to start from the default value of 1024 first, and, afterwards, to fine tune the performance. It is necessary to pinpoint that we have used the same value for the number of RX and TX descriptors. However, in reality, these values be independently set.

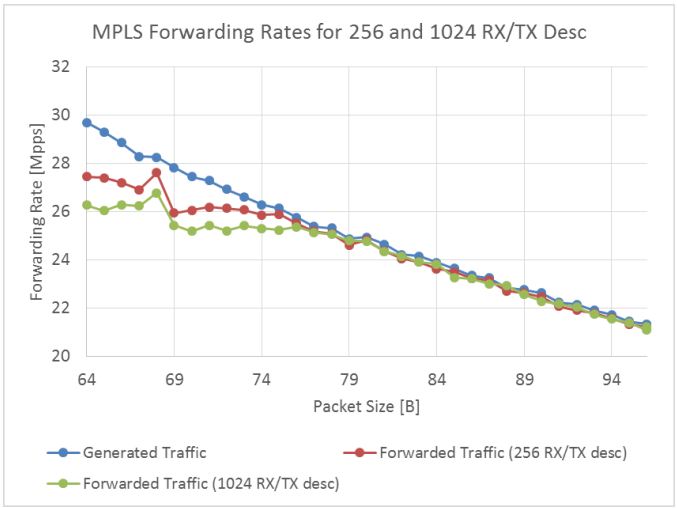


Fig. 6. MPLS packet forwarding rates for packets sizes in range [64B, 96B] and different number of descriptors (256 and 1024)

In Fig. 6, we can see the difference in the forwarding rates for the case where we have used optimal (256 RX/TX

descriptors) and the default value (1024 RX/TX descriptors) for different packets sizes. Due to clarity, we have only performed measurements for the packets sizes between 64 B and 96 B. Indeed, as we can see from the Fig. 6, for the packets longer than 77 B there is no huge difference in the performance between generated and forwarded traffic. The performance difference can vary up to 1 Mpps, depending on the packet size. We can also see the performance drop when increasing the packet size from 68 B to 69 B. This is because the packets become longer than the cache line size of 64 B (the last 4 bytes of the packet belong to the Ethernet trailer CRC, and these are stripped at the NIC level).

Currently obtained results are much better in comparison to our previous work. In [13], we have measured the performance of single-threaded MPLS application implemented using the netmap framework. For that application we have achieved the performance of around 8 Mpps on FreeBSD operating system using a single port. Our current implementation achieves nearly the maximal packet rates when only one port is used, which is 1.85 times better. In [12], we have parallelized the IPv4 lookup procedure using OpenMP and netmap. In that paper, we have reached the speeds of 25.26 Mpps for two ports and 64 flows per port. Our current MPLS implementation is 2.19 Mpps better even though only one flow per port is used. Unfortunately, in comparison to the *l3fwd* DPDK application, MPLS forward performance is slightly lower. The *l3fwd* application forwards nearly all of the generated traffic, approaching the maximal value of 29.76 Mpps.

VI. Conclusion

The MPLS protocol has important properties that allow better throughput, resiliency and security in the networks. Due to the fact that the lookup procedure is simplified, and that MPLS can support traffic engineering, it is essential protocol for the implementation of software routers which have limited packet processing rates. Fast I/O frameworks allow higher packet processing rates compared to the standard Linux kernel protocol stack. DPDK framework is a good fast I/O candidate, not only because it allows high throughputs, but because it is open source project with lots of sample applications and libraries to help with the implementation of network applications.

In this paper, we have implemented the MPLS label switching functionality for the MPLS LSR software routers using DPDK framework. Our application achieves the packet rates of 27.45 Mpps for the shortest 64 B packets when two ports are used. We have examined the performance variation when different number of RX/TX descriptors is used, and concluded that there is a tradeoff between optimal cache usage and available number of processing slots.

In the future work, we plan to implement and measure the performance of ingress and egress MPLS routers. We will also try to optimize the MPLS forward performance to be comparable with the *l3fwd* application. Additionally, we will try larger testing setup with more NICs and workstations.

Acknowledgment

This work was supported by the Serbian Ministry of Science and Education (project TR-32022), and by companies Telekom Srbija and Informatika.

References

[1] "Quagga Routing Suite," [Online]. Available: https://www.nongnu.org/quagga/. [Accessed 24. April 2019].

[2] "The BIRD Internet Routing Daemon," [Online]. Available: https://bird.network.cz/. [Accessed 24. April 2019.].

[3] "XORP," [Online]. Available: http://www.xorp.org/. [Accessed 24. April 2019.].

[4] L. Rizzo, "Netmap: a novel framework for fast packet I/O," in *21st USENIX Security Symposium (UENIX Security 12)*, 2012.

[5] "DPDK Data Plane Development Kit," [Online]. Available: https://www.dpdk.org/. [Accessed 24. April 2019.].

[6] "PF_RING High-speed packet capture, filtering and analysis," ntop, [Online]. Available: https://www.ntop.org/products/packet-capture/pf_ring/. [Accessed 24. April 2019.].

[7] L. De Ghein, MPLS Fundamentals - A Comprehensive Introduction to MPLS Theory and Practice, Indianapolis, USA: Cisco Press, 2016.

[8] M. Á. Ruiz-Sánchez, E. W. Biersack and W. Dabbous, "Survey and Taxonomy of IP Address Lookup Algorithms," *IEEE Network*, vol. 15, no. 2, pp. 8-23, 2001.

[9] L. Andersson, I. Minei and B. Thomas, RFC 5036: LDP Specification, IETF, 2007.

[10] D. Awduche, L. Berger, D. Gan, T. Li, V. Srinivasan and G. Swallow, RFC 3209: RSVP-TE: Extensions to RSVP for LSP Tunnels, IETF, 2001.

[11] S. Khanvilkar, F. Bashir, D. Schonfeld and A. Khokhar, "Multimedia Networks and Communication," in *The Electrical Engineering Handbook*, Academic Press, 2005, pp. 408-416.

[12] M. Vesović, A. Smiljanić and M. Tomašević, "Speeding up IP Lookup Procedure in Software Routers by Means of Parallelization," *Telfor Journal*, vol. 9, no. 1, pp. 2-7, 2017.

[13] M. Vesović, H. Redžović, A. Smiljanić and S. Gajin, "Evaluation of Netmap Framework for MPLS Protocol Implementation," in *Proceedings of 3rd International Conference on Electrical, Electronic and Computing Engineering IcETRAN 2016*, Zlatibor, Serbia, 2016.

[14] "The Pktgen Application," [Online]. Available: https://pktgen-dpdk.readthedocs.io/en/latest/. [Accessed 25. April 2019.].

Integration of the NETCONF Protocol in the Internet of Things by means of RESTful Web Services

Dalibor Đumić, Sretenka Došlić, Marija Antić and Boško Milić

***Abstract*— Herein a new network management protocol called Network Configuration Protocol (NETCONF) will be introduced through an overview and empirical study. The features and capabilities of the NETCONF are a document-oriented approach based on Extensible Markup Language (XML) and the challenge will be an integration of the NETCONF capabilities in the Internet of Things (IoT) using REpresentational State Transfer web services (RESTful). In the end, the goal is to implement this integration through building a home automation server based on any single board computer supporting Linux operating system (OS) and RESTful web application as the client. The web application shall communicate with the server through the NETCONF protocol and allow a user to control devices in his home. In the end, the benefits of applying such these technologies in the home automation or domotics will be discussed and concluded.**

***Index Terms*—NETCONF, Internet of Things, REST, IHGNM, home automation, domotics**

I. Introduction

There is a number of heterogeneous devices present in-home network which is expected to be connected. Here, devices are based on different hardware platforms, controller services are also of a different nature, the software components that enable network access on them are also of different nature for each device. For example, health and lifestyle wearable Internet of Things (IoT) devices like a smartwatch, wristband have different capability in terms of memory usage, power consumption, processing speed as compared to smart home appliances like a washing machine.

IoT devices can be generally classified based on their key characteristics like communication pattern, memory usage, data processing capability, and power consumption. For example, devices like smart keys or a smart washing machine don't need to be connected always and are switched on to perform certain tasks when required; these devices consume less power for communication. This paper is focused on such low power or normally off devices in a home network as they need a gateway for effectively managing operations and communication with the Internet.

The paper is organized as follows. Section 2 addresses important background information on device management in the home automation system. In Section 3, the NETCONF protocol and its features are introduced. The proposed integration of the NETCONF protocol in the IoT is detailed in Section 4. Possible use cases are discussed in the Section 5. Finally, the lessons learned, and the main conclusions are addressed in Section 6.

II. Background

Particularly, there are two common approaches for managing devices in home automation system: simple network management and intelligent home gateway network management (IHGNM). Each of these approaches is as per the application. [1]

If the devices have enough resources that it supports a direct connection to the internet in a secure manner and does not support multiple device classes, the implementation is possible with a lightweight machine to machine (LWM2M). LWM2M is a remote device management standard. In this, IoT devices can be directly managed. This is an example of simple network management. [2]

In IHGNM architecture, multiple device classes from low resource constrained to high resource-constrained devices are considered. The proposed architecture monitors many heterogeneous devices in a home network. Here, IHGNM architecture for a high resource-constrained device in a home network is discussed.

III. Empirical Study of the NETCONF

A. NETCONF

The Network Configuration Protocol (NETCONF) is a network management protocol that can install, manipulate, and delete the configuration of the devices in the network. Its purpose is managing network devices, retrieving its configuration data, and uploading or manipulating new configuration data of the network devices. That means devices

Dalibor Đumić is with the RT-RK Institute for Computer Based Systems, Patre 5, 78000 Banja Luka, Bosnia and Herzegovina (e-mail: Dalibor.Djumic@ rt-rk.com).

Sretenka Došlić is with the RT-RK Institute for Computer Based Systems, Patre 5, 78000 Banja Luka, Bosnia and Herzegovina (e-mail: Sretenka.Doslic@rt-rk.com).

Marija Antić is with the RT-RK Institute for Computer Based Systems, Narodnog Fronta 23a, 21000 Novi Sad, Serbia (e-mail: Marija.Antic@rt-rk.com).

Boško Milić is with the RT-RK Institute for Computer Based Systems, Patre 5, 78000 Banja Luka, Bosnia and Herzegovina (e-mail: Bosko.Milic@rt-rk.com).

on the network can take different states according to their configuration. [3]

Configuration datastores are used for switching between these states. A configuration datastore is the complete set of configuration information which is needed to change a device's state from its initial default state into a desired operational state. NETCONF supports multiple configuration datastores (candidate, running, startup) as well as event notification feature. [4]

- "running" - the currently active configuration of a device and it is always present.
- "startup" - the configuration that will be used during the next startup.
- "candidate" - a configuration that may become a "running" configuration through an explicit commit.

Manipulating device configuration is possible using NETCONF operations. Operations are invoked as Remote Procedure Calls (RPCs) from the client to the server. Some minor operations are:

- "commit" - commits the "candidate" configuration to "running",
- "copy-config" - copy one configuration datastore to another,
- "edit-config" - changes the contents of a configuration database,
- "get-config" - retrieves configuration datastore,
- "lock" - prevent changes to a datastore from another party, and
- "unlock" - releases lock on a datastore. [5]

Configuration data stored on devices and the protocol messages between devices are encoded in Extensible Markup Language (XML) on both client and server side. The client can be a script or application typically running as part of a network manager. The server is typically a network device.

There is a rule that a device on the network must support at least one NETCONF session. The main NETCONF message exchange between client and server in a single NETCONF session is illustrated in Fig. 1.

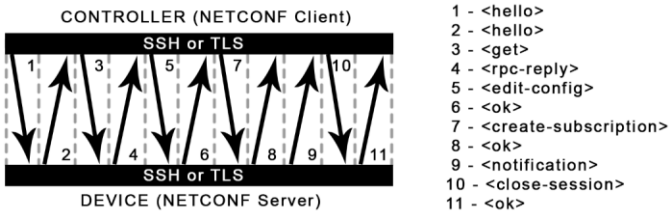


Fig. 1. NETCONF messages

The information that can be retrieved from the server is separated into two classes: configuration data and state data. Configuration data is the set of writable data that is needed to transform a system from its initial default state into its current state. State data is the additional data on a system that is not configuration data such as read-only status information and collected statistics. For specifying NETCONF data models and operations, the YANG data modeling language is used.

B. YANG

A YANG module defines a hierarchy of data that can be used for NETCONF-based operations, including configuration data, state data, RPCs, and notifications. This allows a complete description of all data sent between both NETCONF client and server side. Some of YANG statements are previewed in Table I. [5][6]

TABLE I
YANG STATEMENTS

Statements	Description
augment	Extends existing data hierarchies
choice	Defines mutually exclusive alternatives
container	Defines a layer of the data hierarchy
extension	Allows new statements to be added to YANG
feature	Indicates parts of the model are optional
grouping	Groups data definitions into reusable sets
key	Defines the key leafs for lists
leaf	Defines a leaf node in the data hierarchy
leaf-list	A leaf node that can appear multiple times
list	A hierarchy that can appear multiple times
notification	Defines notification
rpc	Defines input and output parameters for an RPC
typedef	Defines a new type
uses	Incorporates the contents of a "grouping"

YANG modules can be translated into an equivalent XML syntax called YANG Independent Notation (YIN), allowing applications using XML parsers and Extensible Stylesheet Language Transformations (XSLT) scripts to operate on the models. YANG defines a set of built-in types and has a type mechanism through which additional types may be defined. Derived types can restrict their base type's set of valid values using mechanisms like range or pattern restrictions that can be enforced by clients or servers. That means that the modeler can add constraints to the data the model to prevent impossible or illogical data. [6]

These constraints give clients information about the data being sent from the device and also allow the client to know as much as possible about the data the device will accept so

the client can send correct data. Table II briefly describes some other common YANG constraints:

TABLE II
YANG STATEMENTS

Statements	Description
length	Limits the length of string
mandatory	Requires the node appear
max-elements	Limits the number of instances in list
min-element	Limits the number of instances in list
must	XPath expression must be true
pattern	Regular expression must be satisfied
range	Value must appear in range
reference	Value must appear elsewhere in the data
unique	Value must be unique within the data
when	Node is only present when XPath expression is true

The module, which is the base unit of definition in YANG, defines a single data model. It contains three types of statements: module-header statements, revision statements, and definition statements. The module header statements describe the module and give information about the module itself. The revision statements give information about the history of the module and the definition statements are the body of the module where the data model is defined. [5] To use YANG, YANG modules must be defined to model the specific problem domain. These modules are then loaded, compiled, or coded into the server. [4] Finally, a NETCONF server may implement any number of modules. [5]

IV. PROPOSED METHODOLOGY

After the empirical study of the NETCONF protocol and retrieving its features, the proposed integration of the NETCONF protocol in the Internet of Things should be divided into two parts: server side and client side, as it is shown in following Fig. 2.

Fig. 2. An overview of the integration of the NETCONF protocol in the Internet of Things

A. Server Side

Since the main goal is to implement this integration in home automation, an ideal server for this case is shown in Fig. 3.

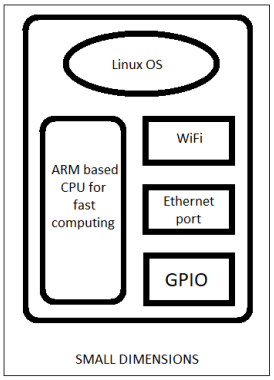


Fig. 3. An overview of ideal single board computer

It should be a single board computer which possesses the following important characteristics:

- small physical dimensions,
- able to boot Linux Operating System,
- has General Purpose Input Output (GPIO) pins for interfacing with the sensors and devices,
- has Ethernet port and/or WiFi module, and
- CPU based on ARM for fast computing.

Great match for the single board with these following characteristics is Raspberry Pi 3 B+, which is based on 1.4GHz 64-bit quad-core ARM Cortex-A53 processor. [7]

The good thing about Raspberry Pi is that it has the GPIO module which can be used through several programming languages such as C, C#, Python, Java, etc. The fact is that the integration will be implemented by using Python programming language and it makes Raspberry Pi a perfect match. To build a server, it is necessary to install Netopeer2, a set of tools implementing network configuration based on the NETCONF protocol. [8] Each room in a home would have installed sensors and relays for controlling devices. A server would be connected via appropriate connection lines to these rooms as it is shown in following Fig. 4.

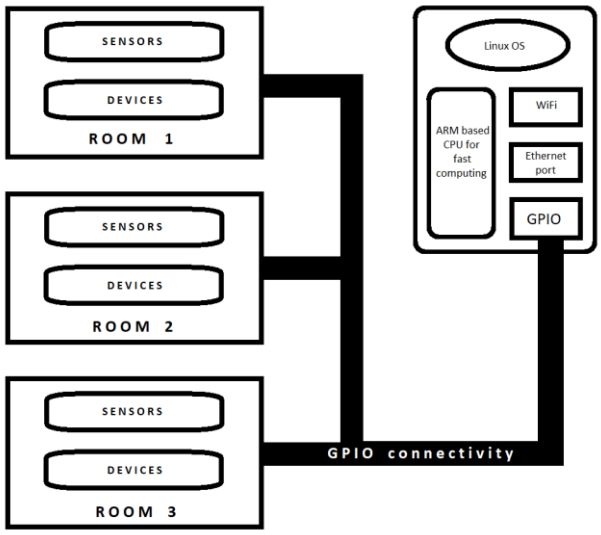


Fig. 4. GPIO connectivity of home rooms and server side

For each room, a custom YANG module should be created, and each custom YANG module should have data such as temperature, humidity, open or closed status, turned off or turned on status, etc. The structure of the simplest custom YANG module for a room is shown in the section "Appendix". On that way, the server can easily manage the information related to the sensors and relays in the home through these custom YANG modules. [9]

B. Client Side

On the client side, any device which supports the NETCONF protocol can communicate with the server. However, the challenge is to develop an application by means of RESTful services. It should send the RPC commands such as "edit-config" or "get-config" directly to the NETCONF server in order to retrieve information about rooms in the user's home. Finally, its interface must be user-friendly and rich with data charts, data graphs, toggle buttons, etc.

The very first step is to develop a script which shall "talk" with the NETCONF server. Thanks to enormous possibilities of the Python programming language, it is possible to communicate with the server via the NETCONF protocol by using ncclient library. The ncclient library facilitates client-side scripting and application development around the NETCONF protocol. [10]

The next step is to develop a web application and merge it with the script based on ncclient library. There are many high-level Python web frameworks and one of them is Django. Django is specific because it encourages rapid development and clean, pragmatic design. [11] By combining Django and ncclient, a powerful user-friendly web application is created, and it will fulfill its main purpose – to collect all information about the conditions such as temperature and humidity in the rooms of the user's home and to control devices in the rooms of the user's home, all of it over the NETCONF protocol.

V. Potential Use Cases

Some of the application of the domotics based on the technologies in the proposed methodology will be discussed in the potential use cases.

A. Case I: Smart House

The domotics in smart houses would be focused on saving energy by turning the lights on to illuminate dark stairs and turning them off when nobody is around. It would also turn the lights on and off even when there is nobody at home, to simulate occupation, or automatically turning down air conditioning systems when the external temperature drops, controlling cameras and security devices, etc.

B. Case II: Smart Hospital

The automation of the healthcare sphere is one of the most urgent and, perhaps, the most difficult tasks in the world. The concept of smart hospitals will let people save time and take better care of them. Moreover, it will help doctors organize their schedules more effectively, making it possible to avoid long queues in hospitals while medical staff can manage their time easily.

C. Case III: Smart Hotels

In the hotel industry, the priority should be turning the hotels into a smart hotel, because it can significantly improve the customer experience, make life easier for staff, and save owners money. Using a smart room, guests are able to control the various components and get their room exactly how they like it. They also find it both faster and easier to obtain important information. In the end, creating a smart hotel can also reduce the number of operational costs.

VI. Conclusion

Through the empirical study of the NETCONF protocol, great capabilities of the NETCONF protocol are discovered. It allows us to have an unlimited number of YANG modules with different structures of the data. This characteristic of the NETCONF protocol is the main reason for using it in domotics.

The integration of the NETCONF protocol in the Internet of Things is not a challenge anymore. Thanks to the powerful Python Web framework and ncclient Python library, it is possible to develop a rich web application which can be outperformed on many devices such as single board computers, desktop computers, notebooks, and even the tablets.

Appendix

```
module room1 {
    namespace "urn:sysrepo:room1";
    prefix room1;
    revision 2019-03-19 {
        description "Initial revision.";
    }
    typedef room-temperature {
        description "Temperature of room";
        type uint8 {
            range "-20..1000";
        }
    }
    container ac{
        description "Configuration container of the AC.";
        leaf ac-switch {
            description "Main switch for switching ON or OFF";
            type boolean;
            default false;
        }
        leaf temperature {
            description "Slider for adjusting the desired
temperature";
            type desired-temperature;
            default 25;
        }
    }
    container room-state {
        description "State data container of the room.";
        config false;
        leaf temperature {
            description "Actual temperature inside the room.";
            type room-temperature;
        }
    }
}
```

ACKNOWLEDGMENT

This work was partially supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia, under grant number III44009-2.

REFERENCES

[1] C. Pham, Y. Lim, Y. Tan. "Management architecture for heterogeneous IoT devices in home network", *IEEE 5th Global Conference on Consumer Electronics*, pp.1-6, IEEE, October 2016

[2] S. Rao, D. Chendanda, C. Deshpande, V. Lakkund, "Implementing LWM2M in constrained IoT devices", *IEEE Conference on Wireless Sensors (ICWiSe)*, pp. 52-57, IEEE, August 2015

[3] R. Enns, M. Brojklund, J. Schoenwaelder and A. Bierman, "Network Configuration Protocol (NETCONF)", Internet Engineering Task Force (IETF), ISSN: 2070-1721, June 2011. [Online]. Available: https://tools.ietf.org/html/rfc6241

[4] M. Dallaglio, N. Sambo, F. Cugini, P. Castoldi, "Management of sliceable transponder with NETCONF and YANG", *International Conference on Optical Network Design and Modeling*, pp. 1 – 6, IEEE, May 2016

[5] P. Shafer, "An Architecture for Network Management using NETCONF and YANG", Internet Engineering Task Force (IETF), ISSN: 2070-1721, June 2011, [Online]. Available: https://tools.ietf.org/id/draft-ietf-netmod-arch-07.html

[6] M. Brojklund, "YANG – A Data Modeling Language for the Network Configuration Protocol (NETCONF)", *Internet Engineering Task Force (IETF)*, ISSN: 2070-1721, October 2010. [Online]. Available: https://tools.ietf.org/html/rfc6020

[7] The Raspberry Pi Foundation. "Raspberry Pi 3 Model B+", [Online], Available: https://www.raspberrypi.org/products/raspberry-pi-3-model-b-plus/

[8] Czech Educational and Research Network (CESNET), "Netopeer2 – The NETCONF Toolset", [Online], Available: https://github.com/CESNET/Netopeer2

[9] sysrepo - YANG-based datastore for Unix/Linux application, [Online], Available: http://www.sysrepo.org/static/doc/html/start_page.html

[10] S. Bhushan, L. Poulopouls, Python library for NETCONF clients, [Online], Available: http://ncclient.readthedocs.org/

[11] Django Software Foundation, [Online], Available: https://www.djangoproject.com/

Comparison of RCIED Activation Responsive and Active Jamming Reliability

Mladen Mileusnić, Predrag Petrović, Aleksandar Lebl and Branislav Pavić

***Abstract*— In this paper we compared the time required for the successful jamming of remote controlled improvised explosive devices activation using active and responsive jamming methods. As a representative of active jamming method we analyzed jamming signal generation using frequency sweep, whereas for the analysis of the possible activating signal presence based on responsive jamming procedures we supposed Fast Fourier Transform (FFT) implementation. Taking into account the current technology state, it is proved that the time required to achieve the successful jamming relied on FFT analysis may be less than in the case of active sweep jamming. When very fast specialized processors are applied to FFT estimation with the highest clock speed, the required time to achieve effective jamming may be up to several tens of times less based on FFT detection for responsive jamming than in the case of active sweep jamming.**

***Index Terms*— active and responsive jamming; RCIED - remote controlled improvised explosive devices; frequency sweep; Fast Fourier Transform; jamming reliability.**

I. Introduction

THE common characteristic of all remote controlled improvised explosive devices (RCIED) is that they are activated by wirelessly transmitted messages. The results of RCIED activation message could be disasterous regarding people lives (VIP persons) and the equipments damages. All elements related to activation signal characteristics (signal power, frequency, implemented modulation method, message duration) are completely unknown. This fact produces great problems in the realization of RCIED activation jammers. Contributions [1] and [2] provide a general overview of jammers types, communications jamming requirements and their efficiency analysis. Modern communications jamming principles and techniques may be found in [3].

There are two basic approaches to the jammer implementation. The first one is active jamming, which consisted of continuous predefined jamming signals sending independently of the RCIED activating message characteristics. In this concept there are no „look through“ phases to detect the activation message existence and the jamming signal characteristics are selected in general using previous experience and expectations. The most important freely selected jamming signal parameter is the RF signal level. This level has to be as high as possible to successfully prevent activating message reception. Two key features which are not optimally chosen relate to continuous jamming regardless of RCIED activation message existence and the RF jamming signal level necessary for jamming successfulness due to the fact that the activation signal level is unknown.

The alternative approach to jammer implementation is responsive jamming concept. In this case the jamming signal characteristics can be optimized using look through intervals to detect the activation message existence and its level. That's why it is possible to send the jamming signal only during activation message presence and jamming signal level can be adjusted to the activation message level in order to successfully deny the threat. A wide range of active and responsive jammers may be found in [4]-[13].

It may be concluded from this short presentation of active and responsive jamming characteristics that active jamming is always successful, while responsive jamming efficiency depends on activation message detection reliability. The question is whether responsive jamming reliability may be greater than for active jamming. In this paper we compare the reliability of mostly implemented active jamming method – frequency sweep [14]-[18] to the reliability of a representative method for activating signal eventual presence detection in order to generate jamming signal according to the activation signal characteristics by implementation of Fast Fourier Transform (FFT) in the analysis [13].

The method for RCIED activation signal frequency spectrum estimation based on FFT analysis is presented in Section 2 with the emphasis on the required time for calculation. Section 3 is devoted to frequency sweep jamming and to determination of required time to realize one complete jamming cycle. In Section 4 jamming reliability on the basis of FFT analysis is compared to the frequency sweep jamming reliability, whereby two special purpose processors are considered for FFT calculation. Reliability estimation is based on the required time to allow successful jamming. Section 5 explains the influence of pipeline signal processing on the performances of responsive jammer. At the end, the paper conclusion is given in Section 6.

II. Signal Spectrum Estimation on the FFT Basis

FFT is the calculation procedure, which allows relatively fast estimation of discretized signal frequency spectrum. Starting from n time samples of analyzed signal, this procedure gives a snapshot of signal frequency spectrum also in n points, i.e. n spectrum lines are obtained. FFT is the optimum method taking into account the required number of

Mladen Mileusnić is with the Institute IRITEL, Batajnički put 23, 11080 Belgrade, Serbia (e-mail: mladenmi@iritel.com).

Predrag Petrović is with the Institute IRITEL, Batajnički put 23, 11080 Belgrade, Serbia (e-mail: presa@iritel.com).

Aleksandar Lebl is with the Institute IRITEL, Batajnički put 23, 11080 Belgrade, Serbia (e-mail: lebl@iritel.com).

Branislav Pavić is with the Institute IRITEL, Batajnički put 23, 11080 Belgrade, Serbia (e-mail: bane@iritel.com).

mathematical operations for signal spectrum determination. There are $(n/2)\cdot log_2(n)$ complex multiplications and $n\cdot log_2(n)$ complex additions [19]. The limitation for n is that the condition $n=2^a$ must be satisfied, where a is the positive integer number. This is a significant saving in the number of mathematical operations and in the required calculation time comparing to the classical method of frequency spectrum estimation by Discrete Fourier Transform (DFT). Namely, it is necessary to perform n^2 complex multiplications and n^2-n complex additions to obtain n frequency spectrum components by DFT on the base of n time samples.

Let us suppose that f_s is the frequency of analyzed signal sampling. The sample acquisition time is then:

$$T = \frac{n}{f_s} \tag{1}$$

The frequency resolution on the base of sample acquisition time may be determined as:

$$df = \frac{1}{T} = \frac{f_s}{n} \tag{2}$$

Therefore, frequency resolution is improved when acquisition time is increased, i.e. the space between frequency spectral components of the analyzed signal is lower.

Constant advancements in processor realization technology and mathematical algorithm improvements are visible in two aspects of FFT calculation progress. On the one side the number of points in which frequency spectrum is determined is constantly increased, and on the other side the required FFT calculation time for some exactly determined number of frequency spectrum components is constantly decreased, chronologically, successively according to presentations in [20], [21], [19], [22]. We selected two approaches referenced in [19] and [22] due to very fast processing algorithms.

Data presented in [22] is related to the FFT calculation time as a function of the number of signal time samples implemented for FFT calculation, i.e. as a function of the obtained frequency components number in the analyzed spectrum. The presented data is for processor clock of 1GHz. It is further emphasized in [22] that improvement may be achieved by processor clock speed increase to 1.25GHz. Besides, it is stated in [23] that maximum processor clock frequency may be even 1.4GHz. On the base of these data, the FFT calculation time (T_{cal} in ms) is presented in Table I as a function of the number of points used in a calculation, for a processor clock of 1.25GHz and for 8 processor cores. The value of the constant K is 1024 in the first column of the Table I.

The time of FFT calculation (T_{cal} in μs) according to the data emphasized in [19] is presented in the Table II. The processor clock in this case may be in the range between 60MHz and 150MHz [24]. That's why data are presented for the mean processor frequency of 100MHz. FFT hardware accelerator (HWAFFT) is one of the parts in the processor implemented according to [19]. HWAFFT is intended for faster FFT calculation. Data in Table 2 are related to the case when HWAFFT is implemented. The number of points is relatively small (till 1024) where FFT is calculated comparing to the number of points, where FFT results are presented in Table I.

The total time, which is needed for signal analysis in a jammer (T_{an}) before (eventually) starting RCIED activation jamming signal emission, consists of three components: sample acquisition time (T), FFT calculation time (T_{cal}) and the time, which is necessary to compare obtained signal frequency components after FFT calculation (T_{comp}) in order to determine whether it is necessary to start jamming. When considering the last component (T_{comp}), there is not such a data in a literature, because calculation is very specific. For our analysis, we supposed that taking equal values of T_{cal} and T_{comp} is a quite good approximation, i.e.

$$T_{an} = T + T_{cal} + T_{comp} = \frac{n}{f_s} + 2\cdot T_{cal} \tag{3}$$

TABLE I
THE TIME OF FFT CALCULATION AS A FUNCTION OF THE NUMBER OF CALCULATION POINTS FOR THE PROCESSOR PRESENTED IN [22]

Number of points for FFT calculation	Calculation time T_{cal} [ms] (8 cores, 1.25GHz)
16K	0.1051
32K	0.1584
64K	0.2517
128K	0.5128
256K	0.9488
512K	2.4824
1024K	5.1226

TABLE II
THE TIME OF FFT CALCULATION AS A FUNCTION OF THE NUMBER OF CALCULATION POINTS FOR THE PROCESSOR PRESENTED IN [19]

Number of points for FFT calculation	Calculation time T_{cal} [μs] (with HWAFFT, 100MHz)
8	1.3
16	1.7
32	3.21
64	4.36
128	9.12
256	16.68
512	37.4
1024	73.15

III. ACTIVE JAMMING USING FREQUENCY SWEEP

Frequency sweep is often used method of active jamming. It is necessary to linearly change signal frequency step by step from its minimum value (*f1*) to the maximum one (*f2*) in order to realize a sweep. It is a readily implemented jamming method, because a significantly smaller power is necessary in relation to wideband jamming based on Additive White Gaussian Noise (*AWGN*) [16] - [18]. Linear frequency change of jamming signal frequency is practically approximated by a stepwise change, as it is presented in Fig. 1. There are two parameters, besides outmost sweep signal frequencies *f1* and

f2, which model signal frequency change: frequency change step (f_Δ) and each step time interval duration while the same signal frequency is generated (T_Δ).

The total step number in jamming realization may be represented by an equation:

$$N = \frac{f2 - f1}{f_\Delta} \tag{4}$$

while the duration of one total sweep cycle may be represented as:

$$T_{sw} = N \cdot T_\Delta = \frac{f2 - f1}{f_\Delta} \cdot T_\Delta \tag{5}$$

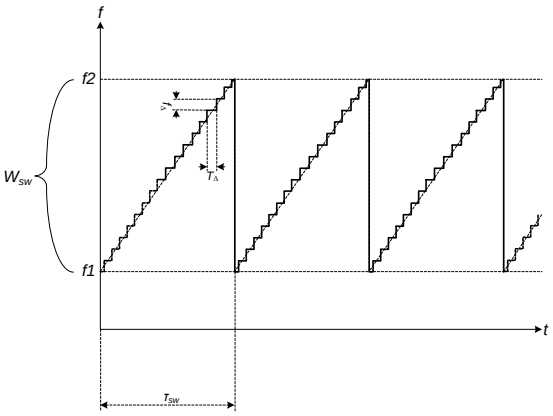


Fig. 1. Practical implementation of frequency sweep in the RCIED activation active jamming.

IV. Comparison of the Active and Responsive Jamming Reliability

The starting data in our analysis will be the required time to realize sweep jamming of RCIED activation signal under the condition that at least once jamming signal frequency and RCIED activation message frequency are approximately equal. After that we shall determine the total time from the beginning of the analyzed signal sample acquisition including FFT calculation time until the start of jamming signal emission on the detected RCIED activation signal frequency. In order to achieve comparison requirements of these two results, we shall suppose that the number of steps in sweep procedure (N) is equal to the number of points where the analyzed signal spectrum is estimated (n).

There is a number of D/A converters, which may be implemented for jamming signal generation. One typical example may be found in [25]. This D/A converter is used in our jammer solution [15]. It generates analog signal from the samples, whose maximum frequency is pretty high (f_s=3.5GHz), thus enabling maximum generated signal frequency f_g=1.4GHz.

Let us suppose that RCIED activation jamming is realized by jamming signal generation in a frequency band between *f1*=20MHz and *f2*≈1.33GHz. We shall further define the frequency change step f_Δ=20kHz, and the frequency sweep step duration T_Δ=200ns (very short time, a greater time is used in a practical realization - T_Δ=1μs or more [15]). On the base of (4) we obtain the number of steps in a sweep procedure realization N=65536=64K, and on the base of (5) sweep procedure duration is T_{sw}≈13.1ms.

Let us now have a device for responsive jamming, which collects the analyzed signal samples (performs A/D conversion) at the same frequency f_s=3.5GHz as the frequency of samples for D/A converter [26]. Number of samples that need to be collected, and consequently obtained number of the analyzed signal discrete frequency components is the same as the number of sweep steps (n=65536=64K). In this way it is achieved that the accuracy of jamming signal frequency *df* according to (2) is the same as the step of sweep signal frequency change f_Δ according to (4). In the case of responsive jamming, the signal sample acquisition time from (1) will be T≈18.7μs, while on the basis of Table 1 the FFT calculation time is T_{cal}=251.7μs. According to our adopted approximation it is also T_{comp}=251.7μs. Taking into account these three values, the total analysis time on the basis of (3) is T_{an}=522.1μs.

Comparing two time intervals which are the main indicators of active and responsive jamming reliability (T_{sw} and T_{an}), it is concluded that responsive jamming on the basis of FFT implementation is significantly more reliable (in our example ≈25 times) than active jamming by frequency sweep. One additional element which is an advantage of responsive jamming is the fact that after the analysis process (T_{an}) jamming may be performed only on the detected activation signal frequency with no time limit. In the case of active jamming, signal frequency is only during a short time period T_Δ approximately equal to the activation signal frequency. Therefore, due to greater jamming time it is also higher the probability that jamming is successful when responsive jamming is implemented.

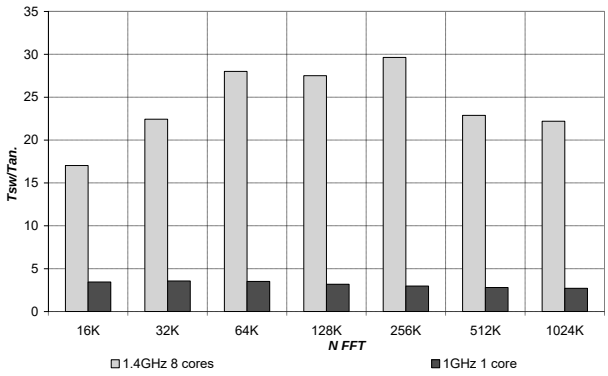


Fig. 2. The reliability of responsive jamming on the FFT basis according to [22] in relation to active frequency sweep jamming.

Fig. 2 presents the results of reliability comparison of responsive jamming based on FFT analysis according to data from [22] in relation to active jamming based on frequency sweep. When active jamming is considered, jamming time on each frequency is adopted to be 200ns. Two extreme cases are chosen for responsive jamming: a) the first one, which allows maximum analysis speed (maximum specified processor clock frequency 1.4GHz and maximum number of processor cores – 8) and b) the second one, which corresponds to the minimum

analysis speed (minimum processor clock frequency 1GHz and minimum number of processor cores – 1). The required number of processor cycles for FFT calculation as a function of the number of FFT points is taken according to Table 1 from [22]. After that, it is determined the necessary time for FFT analysis in some cases. The graph in Fig. 2 indicates greater reliability of responsive jamming on the base of FFT in relation to active jamming using frequency sweep in all jamming conditions, because the calculated relation of jamming reliability is always between two extreme cases presented in Fig. 2 (i.e. this relation is always greater than 1).

Fig. 3 presents the results of reliability comparison of responsive jamming based on FFT analysis according to data from [19] in relation to active jamming based on frequency sweep. The processor implemented for FFT calculation operates on lower frequencies (between 60MHz and 150MHz) [24] than in the example from Fig. 2. That's why it is adopted that analyzed signal sampling is realized on a significantly lower frequency (800MHz) than in the example in Fig. 2. This further means that jamming is realized in this case for lower frequencies (till 320MHz).

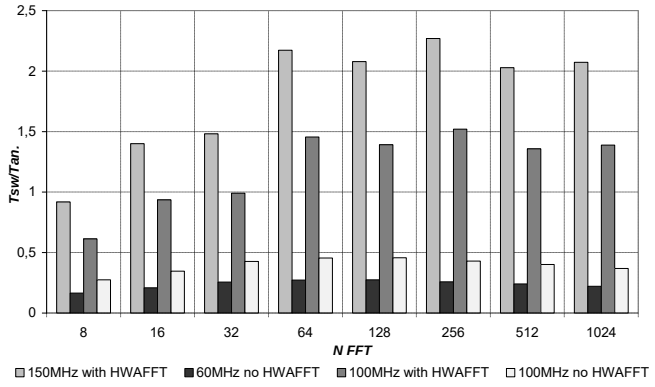


Fig. 3. The reliability of responsive jamming on the FFT basis according to [19] in relation to active frequency sweep jamming.

The results in Fig. 3 are presented separately in the case that HWAFFT is used for an analysis and when it is avoided. The maximum processing rate in this case is achieved if HWAFFT is used together with maximum processor clock frequency (150MHz), while the minimum processing rate is if HWAFFT is not used and the processor clock frequency is minimum (60MHz). These results are presented by first two vertical graphs for each number of frequencies in FFT analysis. Besides these two graphs, the results for mean processor clock frequency (100MHz) are presented when HWAFFT is used and when it is not used. The required number of processor cycles to calculate FFT for some number of points in FFT analysis is determined on the base of Table 3, i.e. Table 4 from [19].

Using the analysis the results in Fig. 3 it can be concluded that the application of HWAFFT allows also in this case that responsive jamming using FFT may be more reliable than the active jamming by frequency sweep (except for the smallest number of analyzed frequencies – 8, which is unlikely to occur in practice). On the contrary, if HWAFFT is not used, responsive jamming reliability using FFT becomes lower than the reliability of frequency sweep jamming (because the relation $T_{sw}/T_{an}<1$). For the mean processor clock frequency (100MHz) and with the HWAFFT implementation for the smaller number of points in the analysis, frequency sweep implementation is more reliable, while for the greater number of points the analysis on the base of FFT is better.

V. Influence of Pipeline Signal Processing

There is one additional possibility to improve performance of responsive jammer, which is realized by FFT implementation [27]. This possibility is related to pipeline processing. The features of such processing may be explained by the illustration shown in Fig. 4.

The time intervals reserved for signal samples collection, signal processing using FFT and decision making on the base of calculated FFT values are presented by different hatching in Fig. 4. In "normal" processing mode (row N_1) the result of decision making is available during time intervals T_3 and T_6. Hardware components used for sample collection are only active during time intervals T_1 and T_4. FFT is calculated only during time intervals T_2 and T_5, while decision is made only during T_3 and T_6. Roughly speaking, all dedicated hardware parts are active approximately one third of time.

Pipeline signal processing contributes to better hardware utilization and in the same time to achieve more reliable results, because they are more often available. These can be explained by rows N_2, N_3 and N_4 in Fig. 4.

The characteristic of pipeline processing is that sample collection, FFT analysis and decision making are realized in each time interval. For the samples collected during T_1 decision is available in time interval T_3 (row N_2). Further, for the samples collected during T_2 decision is available in time interval T_4 (row N_3), and so on. In this way the results of an analysis are available three times more often, while using approximately the same hardware.

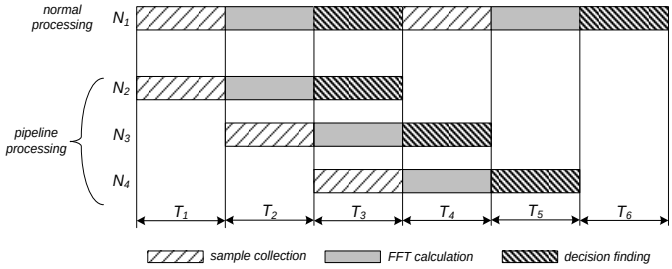


Fig. 4. Illustration of pipeline signal processing for responsive jamming realization.

VI. Conclusions

The results of the calculation presented in this paper have proved that responsive jamming may be more reliable than active jamming. The required time for secure jamming signal generation on the frequency of RCIED activation message is analyzed as a criterion of jamming reliability. If active jamming is realized by generation of frequency sweep signal, jamming is considered to be successful in the case that

complete cycle of frequency change from the lowest frequency towards the highest one is passed. Process of responsive jamming by FFT implementation includes period of signal sample acquisition for the future analysis, the time required for FFT calculation and the time of calculated frequencies comparison in order to make a decision about (eventual) jamming signal generation. As the result of complete analysis based on FFT, the frequency of RCIED activation signal is obtained, and jamming on the exactly determined frequency may be initiated. The results are compared for two processors, which are specialized for FFT calculation. One of these two processors provide that responsive jamming based on FFT implementation is more reliable than active jamming using frequency sweep. In this case analysis rate is several times, and even up to several tens of times greater when FFT is implemented in the analysis in relation to the frequency sweep rate. For the second analyzed processor reliability of responsive jamming depends on processor hardware characteristics such as the processor clock frequency and whether hardware accelerator (HWAFFT) is applied. If HWAFFT is included in the analysis with the higher processor clock frequency, jamming based on FFT analysis is certainly more reliable. On the contrary, if HWAFFT is not used with the lower processor clock frequency, the speed of FFT analysis may not approach jamming speed realized by frequency sweep.

It can be summarized that the results of comparative analysis presented in this paper prove that at up-to-date technological development level, RCIED activation responsive jamming may be very reliable and very often even more reliable than active jamming.

ACKNOWLEDGMENT

The paper is written in the framework of project TR 32051, which is cofinanced by Ministry of Education, Science and Technological Development of Republic of Serbia, 2011-2019.

REFERENCES

[1] P. Petrović, M. Šunjevarić, „Radio Surveillance and Jamming Systems and Techniques“, Proceeding of EEE Conference: Trends in telecommunications development (organized by Faculty of Electrical Engineering and the company Electronics Industry), Belgrade, November 1988., pp. 17.1-17.22., in Serbian.

[2] P. Petrović, „Syllabus EW Course: Introductory Course of Electronic Warfare Systems, Techniques and Technologies“, Preliminary material for lectures at the Royal Academy in Shrivenham, Research Gate, January 2011.

[3] R. Poisel, „Modern Communications Jamming Principles and Techniques“, Second Edition, Artech House, Boston/London, 2011.

[4] Security and Defense Technologies (Secintel): „Portable RF Jammer“, http://www.secintel.com/ecom-prodshow/portable_rf_jammer.html.

[5] Homeland Security Strategies GB LTD, „VIP-300T Covert IED Jammer – Trunk Mounted Version, http://www.secintel.com/media/pdf/vip300T_Covert_IED_Jammer.pdf.

[6] Security and Defense Technologies (Secintel), „Reseller Demonstration Kits“,http://www.secintel.com/ecom-prodshow/stationary_jammer.html.

[7] Security and Defense Technologies (Secintel): „Backpack Jammer“, http://www.secintel.com/ecom-prodshow/backpack_jammer.html.

[8] HSS Development, „Technologies + Solutions“, http://www.secintel.com/media/pdf/hsscatalog.pdf.

[9] Phantom Technologies LTD., „Our Security Solutions“, http://www.phantom.co.il/ufiles/attaches/RCJ1390LT-H_new.pdf.

[10] Elisra Electronic Systems ltd, „EJAB Family Electronic Jammers Against Bombs“, http://www.mw-elisra.com/pdf/EJAB-BrochureLIGHT001.pdf.

[11] Aselsan, „GERGEDAN – Portable RCIED Jammer System (Vehicle Type)“, http://www.aselsan.com.tr/en-us/capabilities/electronic-warfare-systems/electronik-support-and-electronic-attack-systems/gergedan-portable-rcied-jammer-system-%28vehicle-type%29.

[12] J. Mietzner, P. Nickel, A. Meusling, P. Loos, G. Bauch, „Responsive communications jamming against radio-controlled improvised explosive devices“, *IEEE Communications Magazine*, Vol. 50, Issue 10, pp. 38-46, October 2012.

[13] K. Wilgucki. R. Urban, G.Baranowski, P. Grądzki, P. Grądzki, P. Skarżyński, „Automated protection system against RCIED“, In Proc. Military Communication Institute (MCI), 2012.

[14] P. Petrović, N. Remenski, P. Jovanović, V. Tadić, B. Pavić, M. Mileusnić, B. Mišković, „WRJ 2004 Wideband Radio Jammer against RCIEDs“, tehničko rešenje – novi proizvod na projektu tehnološkog razvoja TR32051 pod nazivom „Razvoj i realizacija naredne generacije sistema, uređaja i softvera na bazi softverskog radija za radio i radarske mreže“, 2011., http://www.iritel.com/images/pdf/wrj2004-e.pdf.

[15] M. Mileusnić, P. Petrović, B. Pavić, V. Marinković-Nedelicki, J. Glišović, A. Lebl, I. Marjanović, „The Radio Jammer Against Remote Controlled Improvised Explosive Devices“, 25th Telecommunications Forum (TELFOR), Belgrade, November 21-22th 2017, pp. 151-154, https://ieeexplore.ieee.org/document/8249309, ISBN 978-1-5386-3072-3.

[16] M. Mileusnić, B. Pavić, V. Marinković-Nedelicki, P. Petrović, D. Mitić, A. Lebl, „Analysis of Jamming Successfulness against RCIED Activation“, 5th International Conference IcETRAN 2018, Palić, June 11-14th 2018., Proceedings of Papers, pp. 1206-1211, ISBN 978-86-7466-752-1, paper awarded as the best one in the section Telecommunications.

[17] M. Mileusnić, P. Petrović, B. Pavić, V. Marinković-Nedelicki, V. Matić, A. Lebl, „Jamming of MPSK Modulated Messages for RCIED Activation“, 8th International Scientific Conference on Defensive Technologies OTEH 2018, Belgrade, 11-12th October 2018, pp. 380-385, ISBN 978-8681123-88-1.

[18] M. Mileusnić, B. Pavić, V. Marinković-Nedelicki, P. Petrović, D. Mitić, A. Lebl, “Analysis of jamming successfulness against RCIED activation with the emphasis on sweep jamming,“ *Facta Universitatis, Series Electronics and Energetics*, Vol. 32, No. 2, June 2019, https://doi.org/10.2298/FUEE1902211M, ISSN: 0353-3670, pp. 211-229., the extended and revised version of the paper from the IcETRAN 2018.

[19] M. McKeown, „Texas Instruments: FFT Implementation on the TMS320VC5505, TMS320C5505, and TMS320C5515 DSPs“, Application Report SPRABB, June 2010 – Revised January 2013, pp. 1-28.

[20] E. Çetin, R. C. S. Morling, I. Kale, „An Integrated 256-point Complex FFT Processor for Real-time Spectrum Analysis and Measurement“, IEEE Proceedings of Instrumentation and Measurement Technology Conference, Vol. 1, May 1997., pp. 96-101.

[21] Pipelined FFT/IFFT 256 points (Fast Fourier Transform) IP Core User Manual, Unicore Systems Ltd., pp. 1-17., https://opencores.org/websvn/filedetails?repname=pipelined_fft_256&path=%2Fpipelined_fft_256%2Ftrunk%2FDOC%2Ffft256_um.pdf.

[22] X. Li, E. Blinka, „Very large FFT for TMS320C6678 processors“, Texas Instruments, 2015., pp. 1-6.

[23] Texas Instruments, „Multicore Fixed and Floating-Point Digital Signal Processor“, SPRS691 – November 2010 – Revised March 2014., pp. 1-242.

[24] Texas Instruments, „TMS320C5505 Fixed-Point Digital Signal Processor“, SPRS660F – August 2010 – Revised September 2013., pp. 1-157.

[25] Analog Devices, „AD9914: 3.5 GSPS Direct Digital Synthesizer with 12-Bit DAC“, Rev. F, February 2017., pp. 1-46.

[26] Texas Instruments, „ADC12J4000, 12-Bit, 4GSPS ADC With Integrated DDC“, SLAS989D – January 2014 – Revised October 2017., pp. 1-96.

[27] L. Karlsson, „Method, System and Apparatus for Maximizing a Jammer's Time-on-Target and Power-on-Target", United States Patent Application, Publication No. US 2006/0164283 A1, 27. July 2006.

Direct Ranging and Direction of Arrival Estimation of Non-cooperative Radio Transmitters

Dragan D. Golubović, Nenad J. Vukmirović and Miljko M. Erić

***Abstract*— A MUSIC type algorithm for direct one-step ranging and direction of arrival estimation of non-cooperative narrowband radio transmitters with antenna array is proposed. The method is applicable for far field or near field, single or multi-user non-cooperative signal scenario with planar or spherical waves on antenna array. Antenna array can be with distributed antennas but also it can be classical antenna array with co-located antennas on small distance. The proposed algorithm is suitable for application in massive MIMO systems for 5G. Properties and performance of the proposed algorithm are illustrated and evaluated by simulations. To reduce the computational cost of the algorithm, and provide valid results of simulation, adaptive searching grid in calculation of cost function is applied.**

***Index Terms*— direct localization, direct direction-of-arrival and range estimation, source localization, adaptive searching grid, massive MIMO, 5G.**

I. Introduction

The focus of this paper is joint one-step ranging and Direction of Arrival (DOA) estimation of non-cooperative narrowband radio transmitters with an antenna array. This joint estimation problem is equivalent to location estimation.

Localization in cellular wireless systems, such as in massive MIMO for 5G, is up-to-date research topic. There are two main groups of localization methods: classical widely investigated two-step localization methods (such as RSS, TDOA, DOA) and one-step (direct) position determination (position estimation methods). A method for direct position determination (direct position estimation) proposed in [1,2] was developed for a system with distributed antenna subarrays applied in a spatial semi-coherent signal scenario. A method for direct localization in massive MIMO system with distributed antennas is proposed in [3]. The method for direct position estimation with a millimeter-wave massive MIMO system based on distributed steerable phased antenna arrays is proposed in [4]. The method for joint range/DOA estimation of radio transmitters was a subject of papers [5,6]. In the paper [7] the authors proposed a method for joint range and azimuth estimation of acoustic signals with microphone arrays. In this paper, the system and signal model proposed in [7] for acoustic signals on microphone array are generalized to the radio signals on an antenna array. An antenna array can be classical with co-located antennas on small distance (around half of carrier wavelength), or with distributed antennas. The direction of arrival of the signal at the antenna array is defined in relation to the reference antenna, so signal model can be applied both for planar and spherical waves. There is no limitation of the antenna geometry (it can be with collocated or distributed antennas). Some of antennas can be grouped in distributed antenna subarrays. Not every antenna array geometry is equally good for joint ranging and DOA estimation. It is also assumed that signal scenario is fully spatially coherent (Line of Sight –LOS).

Based on the signal model at the antenna array a MUSIC type algorithm for 2D joint (direct) ranging and direction of arrival estimation of non-cooperative narrowband radio transmitters is proposed.

The proposed algorithm is numerically complex and demanding. We made some modifications and developed an estimation algorithm with less computation than MUSIC width a fixed searching grid. This algorithm is with a variable grid depending on the distance. This searching grid is adaptive and the main goal of this paper is to show and explain the performance of the proposed algorithm analyzing the RMSE of the range and azimuth depending on the distance from the reference origin. The performance depends on the signal-to-noise ratio (SNR) and the comparison for a different SNR values at the antenna array was shown. Using proposed algorithm it is possible to achieve the accuracy of 10^{-2}m for distance estimation and the accuracy of 10^{-3} degrees for azimuth estimation. We showed also the RMSE of the range and azimuth for different scenarios, where a signal source is inside the antenna array, where it is close to the antenna array and where it is far away from the antenna array. The first scenario illustrates the problem of ambiguity and is interesting for application in distributed massive MIMO systems, others are interesting from the aspect of the massive MIMO system. In the last section, we present the simulation results which illustrate properties and performance of the proposed algorithm.

II. Problem Formulation

In this part, a mathematical model and problem formulation are discussed. It will be assumed that the signal arriving at antenna array is narrowband.

The antenna array receives the radio signals (in the

Dragan D. Golubović is with University of Belgrade, School of Electrical Engineering, Bulevar Kralja Aleksandra 73, 11020, Belgrade, Serbia and Information Technlogy School, Cara Dušana, 11080, Belgrade, Serbia (e-mail: dragan.golubovic@gmail.com).

Nenad J. Vukmirović is with University of Belgrade, School of Electrical Engineering and Innovation Centre of School of Electrical Engineering, Bulevar Kralja Aleksandra 73, 11020, Belgrade, Serbia (e-mail: vn135023p@student.etf.bg.ac.rs).

Miljko. M. Erić is with University of Belgrade, School of Electrical Engineering, Bulevar Kralja Aleksandra 73, 11020, Belgrade, Serbia (e-mail: miljko.eric@etf.rs).

mathematical model their number is denoted by K), and the spatial sampling of the radio signal is performed. The number of spatial samples is determined by the number of antennas ($L > K$). It is assumed that the antenna array consists of L antennas arbitrarily located in space at the locations defined by vector $\mathbf{p}_i$, where i=1, 2…L. For simplicity, we suppose that antennas have identical omnidirectional characteristics. The origin of the coordinate system is in the center of the array. Fig. 1 shows the geometrical model of the system.

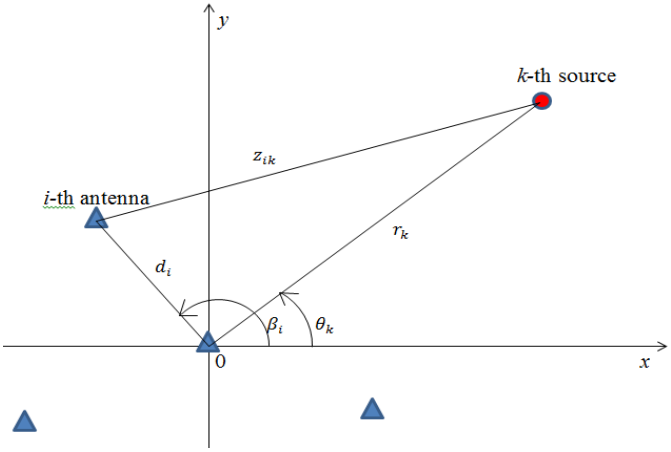


Fig. 1. Proposed geometrical 2D model of the system.

In this model, d_i indicates the distance of the i-th antenna from the reference origin, r_k is the distance of the k-th source from the reference origin, and z_{ik} is the distance between the i-th antenna and the k-th source. θ_k is azimuth of arrival of the k-th signal and β_i is the angle between x-axis and i-th antenna direction. The position of the i-th antenna in the antenna array can be expressed as:

$$\mathbf{p}_i = [p_{ix} \quad p_{iy} \quad p_{iz}]^T . \tag{1}$$

In this relation, p_{lx} , p_{ly} and p_{lz} are x, y, and z coordinates of the i-th antenna respectively. Because only a 2D case was analyzed, for the purpose of the analysis in this paper, z-coordinates will always be equal to zero, both for antennas and signal sources locations. Since the antenna array consists of L antennas, the matrix $\mathbf{p}$ with antenna positions has the form:

$$\mathbf{p} = \begin{bmatrix} p_{1x} & p_{1y} & p_{1z} \\ \vdots & \dots & \vdots \\ p_{Lx} & p_{Ly} & p_{Lz} \end{bmatrix} . \tag{2}$$

d_i is calculated as:

$$d_i = \sqrt{p_{ix}^2 + p_{iy}^2} . \tag{3}$$

We also have:

$$z_{ik}^2 = r_k^2 + d_i^2 - 2 r_k d_i \cos\theta_{ik} . \tag{4}$$

Determining the location of the signal source is possible if we estimate the distance of the signal source from the reference origin, as well as the angle between this direction and the x-axis. Therefore, further analysis will be carried out using radial coordinates (r_k, θ_k) for the signal source location, and (d_i, β_i) for the location of i-th antenna in the antenna array.

After defining all the necessary geometric parameters, a mathematical model of the summary signal can be formed at the i-th antenna:

$$x_i(m) = \sum_{k=1}^{K} v_i(r_k, \theta_k) s_k(m) + n_i(m) . \tag{5}$$

Where i is the antenna index (i=1,2,…L), M is the number of signal samples based on which the location will be estimated (m=0,1,…M-1), K is the number of signals, $s_k(m)$ is the m-th sample of the k-th signal, $n_i(m)$ is the m-th noise sample at the i-th antenna in the antenna array and $v_i(r_k, \theta_k)$ are responses of antenna elements to a signal from the desired source position. We assumed that only the additive white Gaussian noise is present [1].

Equation (5) can be formulated in matrix form using the following equations:

$$\mathbf{x}(\mathbf{r},\boldsymbol{\theta};m) = \sum_{k=1}^{K} \mathbf{v}(r_k, \theta_k) s_k(m) + \mathbf{n}(m) \tag{6}$$

$$\mathbf{x}(\mathbf{r},\boldsymbol{\theta};m) = \mathbf{V}(\mathbf{r},\boldsymbol{\theta})\mathbf{s}(m) + \mathbf{n}(m) \tag{7}$$

$$\mathbf{V}(\mathbf{r},\boldsymbol{\theta}) = \left[\mathbf{v}(r_1,\theta_1) \quad \mathbf{v}(r_2,\theta_2) \quad \cdots \quad \mathbf{v}(r_K,\theta_K)\right]. \tag{8}$$

In the above relations, $\mathbf{v}(r_k, \theta_k)$ represents the steering vector of the k-th source signal, where r_k and θ_k are the unknown parameters that will be estimated. By determining these two parameters, localization has been successfully performed for the k-th source. The dimension of the vector $\mathbf{v}(r_k, \theta_k)$ is Lx1, and the individual responses of the antenna elements to the signal from the desired direction are determined by the relation:

$$v_i(\theta_k, r_k) = \frac{c_{ik}}{z_{ik}} \exp(-j \frac{2\pi}{\lambda_c} z_{ik}) . \tag{9}$$

As can be seen from the observed relation, the distance z_{ik} enters the formulation of the steering vector, and therefore represents a very important parameter. λ_c is the carrier wavelength of the signal coming to the antenna array, and c_{ik} are the coefficients associated with the antenna gain. If it is assumed that omnidirectional antennas were used, these coefficients become independent of the antenna and the source, and it can be assumed that their value is equal to 1 for simpler analysis [4] [5].

Of particular importance is the process of normalization, because then the signal model can be generalized. If λ_c and λ_g are the carrier wavelength of the incoming signal and the wavelength of the upper limit frequency in the signal spectrum, respectively, then we can define a new parameter:

$$q_2 = \frac{f_c}{f_g} = \frac{\lambda_g}{\lambda_c} \tag{10}$$

This parameter also enters the formulation of the propagation vector:

$$v_i(\theta_k, r_k) = \frac{1}{z_{ik}} \exp(-j2\pi q_2 \cdot \frac{z_{ik}}{\lambda_g}). \tag{11}$$

This is a narrowband model but we can generalize it to also hold for broadband signals:

$$v_i(\theta_k, r_k, h) = \frac{1}{z_{ik}} \exp(-j2\pi(q_2 + q_1 \cdot \Omega_h) \cdot \frac{z_{ik}}{\lambda_g}). \tag{12}$$

Where $\Omega_h \in [-0.5, 0.5]$ and $q_1 = \Delta f_{BW} / f_g$, where Δf_{BW} denotes spectral bandwidth of the signal. It is very important to note that the mapping of radio signals from a reference point (reference antenna) to the antenna array will be done in the frequency domain, even in the case of a narrowband signal. By choosing the appropriate parameters, this model is generally applicable for both narrowband and broadband signals.

Based on this mathematical model, it is necessary to estimate unknown parameters, which are the range and the azimuth of the incoming signal source, respectively.

III. MUSIC TYPE ALGORITHM FOR DIRECT AZIMUTH AND RANGE ESTIMATION

The formulation of the MUSIC method is based on the specific properties of the correlation matrix of the signal incoming at the antenna array. Common to all high resolution methods for estimating the direction of arrival of the radio signal is the estimation of the covariance matrix using spatial samples of signals at the antennas and its interpretation on the subspace of the signal and the subspace of the noise. We are looking for eigenvalues and corresponding eigenvectors of this matrix. In the MUSIC method, the information about azimuth and elevation of incoming radio signals is obtained from the fact that the vectors propagating the current directions of arrival, are orthogonal to the eigenvectors, which correspond to the minimum eigenvalues of the covariance matrix. Appropriate eigenvectors borders with the subspace of the noise. Other eigenvectors, which correspond to the highest eigenvalues borders with the subspace of the signal.

In our model, it is necessary to modify the algorithm by not looking for azimuth and elevation, but azimuth and the range of the transmitter from the reference origin. When forming a criterion function, it is necessary to define the the noise subspace matrix as:

$$\mathbf{U}_N = [\mathbf{u}_{K+1} \quad \mathbf{u}_{K+2} \quad \cdots \quad \mathbf{u}_L] \tag{13}$$

Columns of this matrix are noise eigenvectors that correspond to the lowest eigenvalues. To determine the location of the radio signal source, a 2D criterion function of the MUSIC method is formed and defined as follows:

$$P(r, \theta) = \frac{1}{\mathbf{v}^H(r, \theta)\mathbf{U}_N\mathbf{U}_N^H\mathbf{v}(r, \theta)} \tag{14}$$

The arguments of the maximum of this criterion function (r_k, θ_k) are in fact estimations of unknown parameters and which will be determined (range and azimuth) using our approach.

This method is very accurate in determining locations, but it has great numerical complexity. It is based on a fixed searching grid, and the number of points where we calculate the criterion function rapidly grows when the resolution of the grid is increased. It is important to say that we choose azimuth and range resolutions separately. If the resolution is better, we need more time to estimate the location of the signal source. When the distance of the signal source is lower, estimation error will be smaller, and we can choose better resolution to locate the source without any problem. But for larger distances, it will be very inefficient to estimate the location using great resolution, because it takes a lot of time.

Therefore, we made some modifications end developed an adaptive estimation algorithm with less computation than MUSIC with a fixed searching grid.

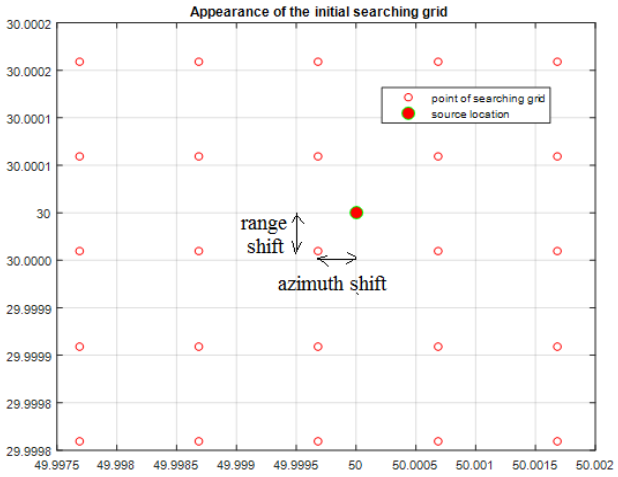


Fig. 2. Initial searching grid around the location of radio signal source (azimuth is 50º, range is 30 meters, `az_rez` is 0.001º and `range_rez` is 0.0001 m)

Fig. 2 shows an example of using the initial searching 5×5 point grid around the location of the radio signal source.

First of all, we must change dimensions of the observation window to reduce the complexity of the algorithm. The basic idea is to select a window with 5×5 points around the source location and to search for the maximum of the criterion function only in that window. Azimuth resolution is defined as `az_rez`, and range resolution is `range_rez`. In order to ensure that no location is favoured during the estimation, we will move the initial window by some random values for both, azimuth and range. These random values are in [-`range_rez`/2, `range_rez`/2] for range shift and in [-`az_rez`/2, `az_rez`/2] for azimuth shift of the window.

Our approach is to estimate the location using the arguments of the maximum of the criterion function. But it

is clear that the real maximum is not in the initial searching grid window. Because of that, we have developed the method with a moving searching grid window. If the maximum of the criterion function is on a border of the window, we must move the window toward to the left or to the right border (for azimuth) and up and down border (for range) depending on where the maximum is located. Also we must ensure the overlapping of two windows during the moving of the window. The process is over when the maximum of the criterion function is not on any border of the searching grid window. Fig. 3 shows the moving of the initial searching grid window along the azimuth and range dimension, respectively.

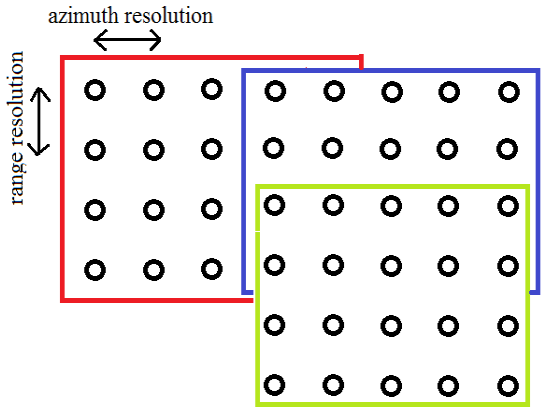


Fig. 3. The moving of the initial searching grid window (red) along the azimuth axis (blue) and along the range axis (green).

This method has a greatly reduced numerical complexity. But when we reduce the resolution cell (better resolution), it can be very hard to find the maximum of the criterion function because it will be on some border of the window for a long time (many iterations). Because of that, it again has great numerical complexity.

If the range of the source increases, it requires more time for estimation of the location, because estimation error is larger, and we must solve this problem. The solution is to use an adaptive searching grid, where resolution is better for close locations, and rougher for distant locations. This dependence is approximately linear, and will be analyzed in the next chapter of this paper.

The searching grid is adaptive and the main goal of this paper is to show and explain the performance of the proposed algorithm analyzing RMSE of the range and azimuth depending on the distance. This is the case where we change the size of the window and move the window at the same time.

IV. Simulation Results

The simulation of the hypothetical system is based on the mathematical model presented in the previous chapters of this paper and implemented with the MATLAB software package. The goal is to determine the location of the radio signal source using antenna arrays at the reception side and the adaptive MUSIC algorithm. Of course, it is first necessary to define the parameters of the hypothetical system to be simulated.

Radio signals were 16-QAM. Also, the model is also applicable to other signal classes. We used the signal model defined in (12) where q_1=0.001 and q_2=0.9. f_c is 1GHz. At the reception side, there is a circular antenna array with 5 antennas arranged in the (x,y)-plane [6]. In the first scenario, a signal source is inside the antenna array, the distance between antenna elements is 25m, and the signal source is 10m from the reference origin. Other scenarios are where it is close to the antenna array (30m) and where it is far away from the antenna array (500m). The distance between antenna elements is 1 meter for both scenarios. The vectors of the antenna positions are expressed in meters (Fig. 4.).

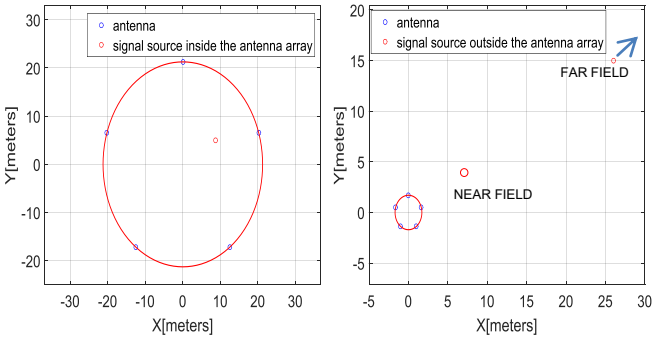


Fig. 4. Circular antenna array with 5 antennas located in the (x,y)-plane where source signal is inside the antenna array (left), and where it is outside the antenna array (right)

The model, which is presented in this paper, is generally applicable to all geometries of antenna arrays. For simplicity, it will be assumed that the antennas have identical omnidirectional characteristics. It is assumed that the additive white Gaussian noise is present. The sources are in 2D space, and they have only x and y coordinates. The signal covariance matrix is estimated based on 200 signal samples at the reception side. We also assume that the signal at the antennas have approximately equal SNR values of 20dB. Because we are analyzing RMSE of range and azimuth depending of the distance, it is important to keep this value constant. Therefore, we can follow RMSE depending only on the distance. This is ensured by increasing the power of the source signal at the distance of 1 meter from the source, bearing in mind that the power decreases with the square of the distance [7]. The azimuth angle in these simulations is 30º and does not change.

Mesh and contour plots of the modulus of the criterion function are shown in Fig. 5 for different scenarios.

The first scenario illustrates the problem of ambiguity and is interesting for application in distributed massive MIMO systems, others are interesting from the aspect of the massive MIMO system. First two scenarios show us that we can determine range and azimuth, but in the third scenario the criterion function is constant across the r-dimension, and the accuracy of the algorithm is reduced for distance estimation.

As we can see, we can estimate the location, but the searching window is too large and it has great computational complexity. The number of points where we calculate the criterion function rapidly grows when we reduce the size of the resolution cell. However, we can get localization results with similar accuracy using our approach at reduced numerical cost.

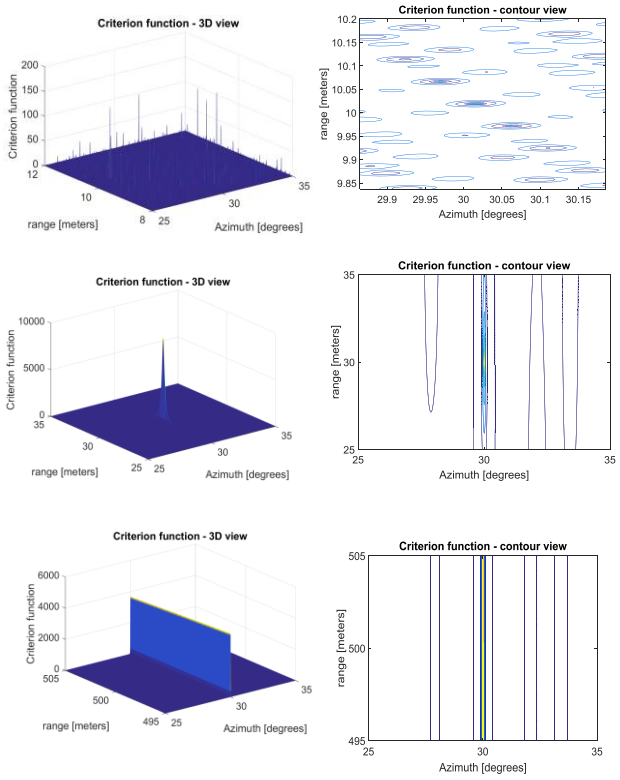


Fig. 5. Mesh and contour plots of the modulus of the criterion function for different scenarios: a) signal source is inside the antenna array and range is 10m b) signal source is close to the antenna array and range is 30m c) signal source is far away from the antenna array and range is 500m (azimuth is 30º, `az_rez` is 0.001º and `range_rez` is 0.001 m for all scenarios).

5×5 points searching grid is used, with adaptive resolution changing depending of the distance of the radio signal source from the reference origin. We propose a solution where azimuth and range resolution are different for smaller distances and for greater distances (from 50 to 500 meters): `range_rez=distance*0.00008-0.000079` meters (from 5 to 50 meters), `range_rez=distance*0.0003-0.006` meters (from 50 to 500 meters), `az_rez= 0.001/(distance+5)+0.00001` (from 5 to 50 meters) and `az_rez=0.001/distance+0.00001` (from 50 to 500 meters). 8192 simulations were made, and the first 100 estimates were shown in Fig. 6.

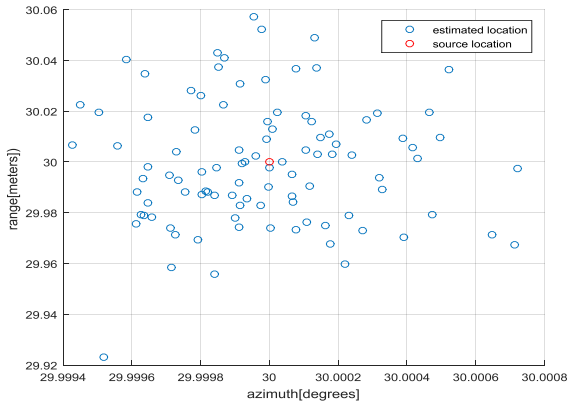


Fig. 6. Source Scatter plot of azimuth/range error (the first 100 estimates of 8192)

The proposed algorithm calculates RMSE (Root Mean Square Error) of range and azimuth depending on the distance. We use different numbers of simulations to see how the simulated results converge. Fig.7 shows the dependence of $RMSE_{distance}$ by choosing several distances where sources were located. We choose different values of SNR at antenna array. In this case, the distance axis is in logarithmic scale.

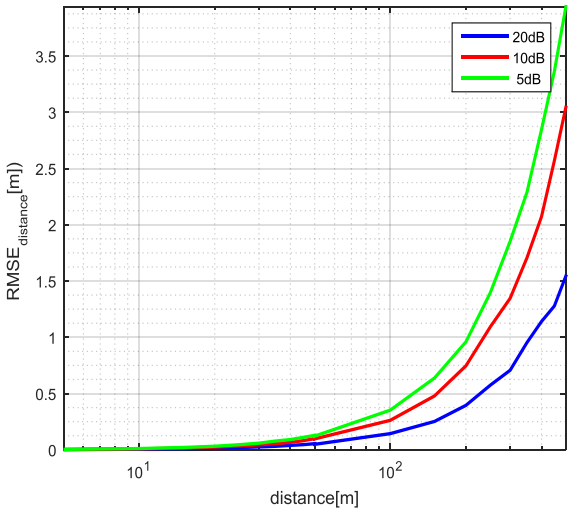


Fig. 7. $RMSE_{distance}$ for different distances of radio signal source and for different values of SNR at antenna array (from 5 to 500 meters) where signal source is outside the antenna array

As we can see from the plot, RMSE increased with the distance, especially for greater distances. This is definitely an expected result. But for smaller distances it is very hard to show the $RMSE_{distance}$ because the error is too small. Because of that, we have to show a plot for smaller distances from 0 to 50m (Fig. 8) and for different values of SNR at antenna array. These values directly affect the performance of the proposed algorithm.

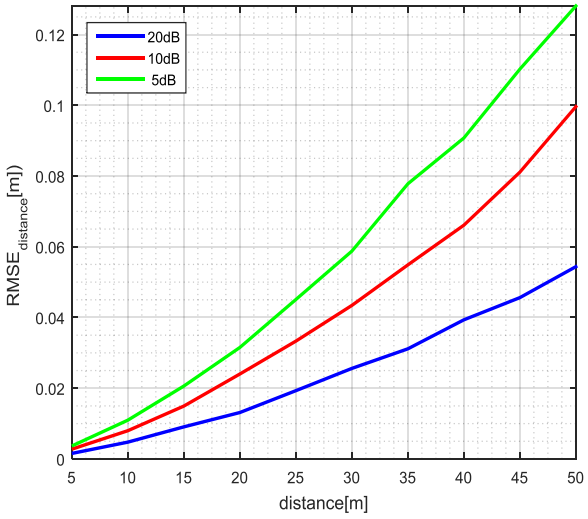


Fig. 8. Plot of $RMSE_{distance}$ for different distances of radio signal sources (5 to 50 meters) and for different values of SNR at the antenna array where signal source is outside the antenna array

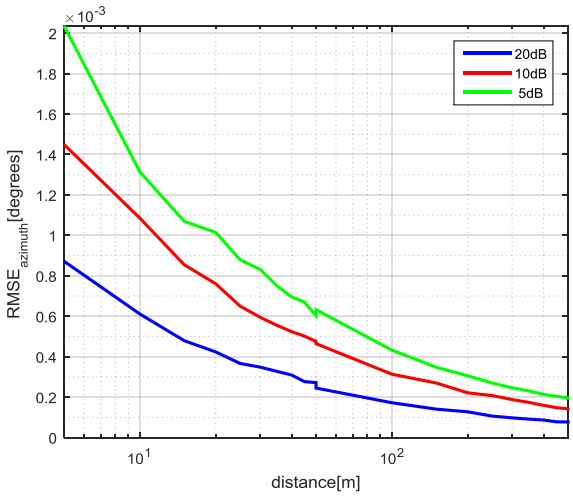


Fig. 9. $RMSE_{azimuth}$ for different distances of radio signal sources (from 5 to 500 meters) and for different values of SNR at antenna array where signal source is outside the antenna array

$RMSE_{azimuth}$ was shown in Fig. 9 and it decreases as the distance increases. The estimation error is smaller for larger values of SNR at the antenna array. If the number of simulations is greater, the RMSE estimates are more accurate and converge to the real value.

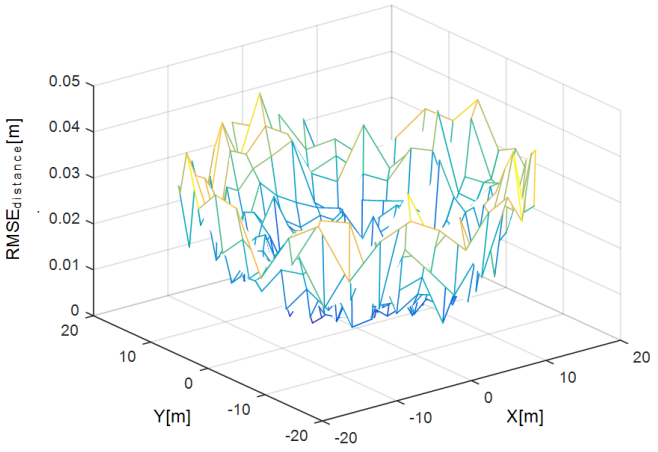


Fig. 10. $RMSE_{distance}$ for different distances of radio signal sources (from 1 to 20 meters) and for all azimuths where signal source is inside the antenna array

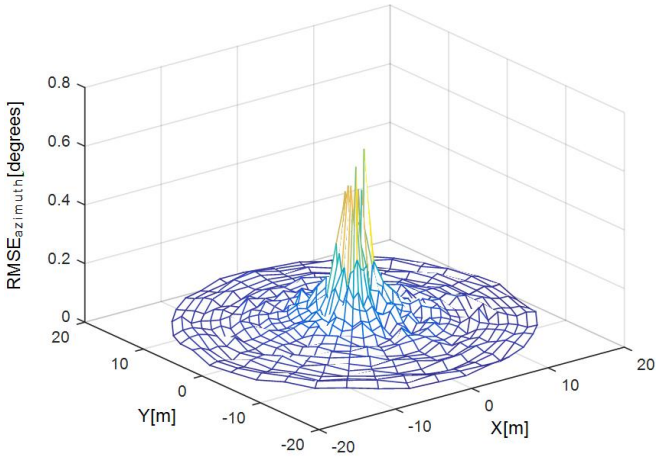


Fig. 11. $RMSE_{azimuth}$ for different distances of radio signal sources (from 1 to 20 meters) and for all azimuths where signal source is inside the antenna array

Fig. 10 and Fig. 11 show the $RMSE_{azimuth}$ and $RMSE_{distance}$ for all azimuths and for different distances of radio signal sources where radio signal source is inside the antenna array. As we can see, symmetry can be noticed and azimuth estimation is less accurate then it was in previous cases where signal source was outside the antenna array.

V. Conclusion

In this paper, we presented a localization algorithm where we simultaneously estimated the range and azimuth of the radio signal sources. A modified-version of the 2D MUSIC algorithm gives good results but the drawback is a heavy computational load. To solve this problem, we used an adaptive searching grid, where resolution is better for close locations, and rougher for distant locations. We have changed the size of the searching grid window and moved the window at the same time. It was shown that different SNR values at the antenna array directly affect the performance of the proposed algorithm. We also showed the RMSE of the range and azimuth for different scenarios, where signal source is inside the antenna array, where it is close to the antenna array and where it is far away from the antenna array. We showed and explained the results.

References

[1] A. Weiss, "Direct position determination of narrowband radio frequency transmitters," IEEE Signal Process. Lett, vol. 11, no. 5, pp. 513–516, May 2004.

[2] A. Weiss, A. Amar, "Direct position determination of multiple radio signals," EURASIP J. Applied Signal Process., no. 1, pp. 37–49, Jan. 2005.

[3] N. Garcia, H. Wymeersch, E. Larsson, A. Haimovich, and M. Coulon, "Direct Localization for Massive MIMO,", IEEE Trans. Signal Process., vol. 65, no. 10, pp. 2475–2487, May 2017.

[4] Nenad Vukmirović, Miloš Janjić, Petar M. Djurić and Miljko Erić "Position estimation with a millimeter wave massive MIMO system based on distributed steerable phased antenna arrays", . EURASIP Journal on Advances in Signal Processing (2018) 2018:33 https://doi.org/10.1186/s13634-018-0553-9, special issue Network localization.

[5] A.M. Elbir, T. E. Tuncer, "Angle and position estimation for far-field and near-field multipath signals'', 22nd Signal Processing and Communications Applications Conference (SIU), 2014.

[6] J.F. C. Xiao, L. Z. Xian, D.Zhang, "A New Algorithm for Joint Range-DOA-Frequency Estimation of Near-Field Sources", EURASIP Journal on Applied Signal Processing, 386–392, 2004.

[7] Y.D. Huang, M.Barkat, "Near-Field Multiple Source Localization by Passive Sensor Array", IEEE Transactions on Antennas and Propagations, Vol.39, No.7, JULY 1991.

[8] J. Chen, L. Yip, J. Elson, H. Wang, D. Maniezzo, R. Hudson, K. Yao, and D. Estrin, "Coherent Acoustic Array Processing and Localization on Wireless Sensor Networks", Proceedings of the IEEE, vol. 91, no. 8, pp. 1154–1162, 2003.

[9] J. C. Chen, K. Yao, and R. E. Hudson, "Acoustic Source Localization and Beamforming: Theory and Practice", EURASIP Journal on Applied Signal Processing, pp. 359–370, 2003.

[10] Y. Tamai, S. Kagami, H. Mizoguchi, K. Sakaya, K. Nagashima, T. Takano, "Circular Microphone Array for Meeting System", Sensors, Vol. 2, pp. 1100 _ 1105, 2003.

[11] H. V. Trees, Optimum Array Processing – part IV Detection, Estimation and Modulation Theory, Wiley Interscience, 2002.

Implementation of algorithm for excision of point targets from distributed radar detections

Pavle Petrović, Nemanja Grbić, Nikola Stojković, *Member, IEEE*, Dejan Nikolić, *Member, IEEE*, Nikola Lekić, *Member, IEEE*

***Abstract*—With maximal range of 200 nmi HFSWR (High Frequency Surface Wave Radar) represents a very effective solution for remote sensing of wide maritime areas. Although HFSWR may be used for many purposes, in this paper the focus is on vessel detection. Since any vessel fits into one HFSWR resolution cell all of them are considered point-like targets. In some circumstances it was noticed that reflections from vessels are occupying multiple consecutive cells, thus forming falsely distributed targets. The paper proposes an algorithm which solves this anomaly by pinpointing vessels exact location within distributed area. The algorithm is tested on real data collected from HFSWR systems located in Gulf of Guinea.**

***Index Terms*—radar, HFSW radar, CFAR detector, Ship detection, marine systems.**

I. Introduction

IN recent years a demand for monitoring of the deep seas is rising. One of the solutions lies in the usage of HFSWR (High Frequency Surface Wave Radar). A general overview of HFSWR principles can be found in [1-3].

HFSWR being discussed in this paper is described in detail in [4] and only a brief overview of it will be presented here. Radar site geometry is presented in Fig 1.

HFSWR transmitter and receiver are separated, each with its own antenna array. Transmitters role is to generate a signal of desired strength and direct it towards a region the HFSWR is designed to monitor.

To accomplish this it is equipped with power amplifiers and planar antenna array consisting of 4 quarter wave monopole antennae [5]. Radiation pattern of such array is presented in Fig. 1. HFSWR receiver array contains 16 quarter-wave monopole linear array, each connected to a dedicated RF receiver. Signal received by each antenna is RF processed and digitalized separately. Further processing of digitalized signal is done on data from all 16 channels conjointly.

The rest of the paper is organized as follows: in section II a short description of signal processing techniques and challenges in which advances are made is presented. Section III presents the details of Excision Algorithm. Section IV consists of data collected from an HFWSR site and presents the advances achieved by the implementation of said algorithm. Conclusions and possible future research topics are given in section V.

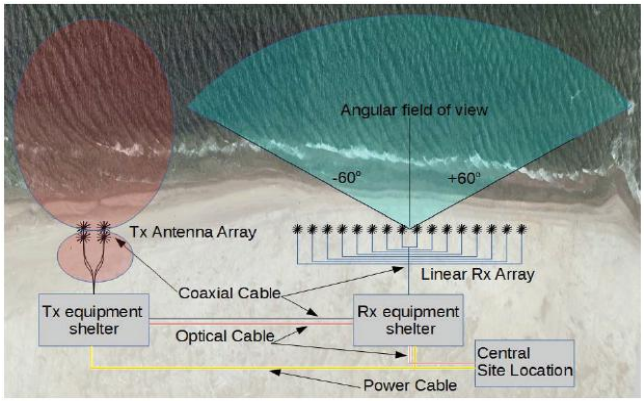


Fig. 1. HFSWR deployment zone.

II. Signal processing in HFSWR and challenges addressed in this paper

Block schematic of the digital signal processing (DSP) chain is presented in Fig 2. and its detailed description is given in [1], [6]. The focus of this paper is on the way how to improve design of CFAR (Constant False Alarm Rate) detection algorithms in a non-homogeneous clutter environment. Since CFAR algorithms perform the best when the environment contains only homogeneous noise any source of non-homogeneous clutter will degrade their performance. In HF band Bragg lines are a dominant source of non-homogeneous clutter. They form when wavelength of a reflected electromagnetic wave is equal to double the distance between peaks of sea waves [7]. Additionally, HF electromagnetic waves can interact strongly with ionosphere, causing hardly predictable behaviours resulting with localized, but non-homogeneous clutter. Both of these effects lead to increase of false detection, which requires specific algorithms in order to maintain reasonable false alarm rate. Some previous work regarding these problems may be found in [8], [9].

Pavle Petrović is with the School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia, VLATACOM Institute, Bulevar Milutina Milankovica 5, 11070 Belgrade, Serbia(e-mail: pavle.petrovic@vlatacom.com).

Nemanja Grbić is with the School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia, VLATACOM Institute, Bulevar Milutina Milankovica 5, 11070 Belgrade, Serbia(e-mail: nemanja.grbic@vlatacom.com).

Nikola Stojković is with the School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia, VLATACOM Institute, Bulevar Milutina Milankovica 5, 11070 Belgrade, Serbia(e-mail: nikola.stojkovic@vlatacom.com).

Dejan Nikolić is with VLATACOM Institute, Bulevar Milutina Milankovica 5, 11070 Belgrade, Serbia(e-mail: dejan.nikolic@vlatacom.com).

Nikola Lekić is with VLATACOM Institute, Bulevar Milutina Milankovica 5, 11070 Belgrade, Serbia(e-mail: nikola.lekic@vlatacom.com).

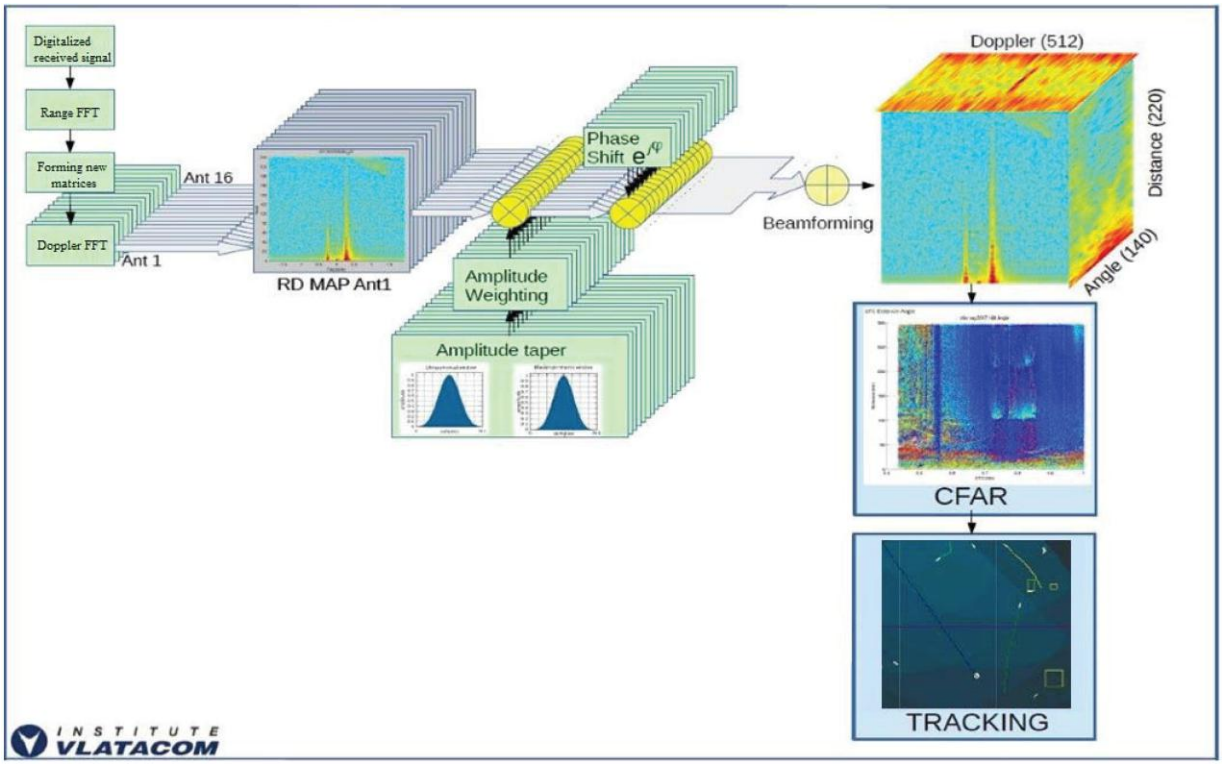


Fig. 2. HFSWR signal processing chain (taken from [4])

During regular operation of HFSWR a number of distributed targets were observed, which may be consequence of non-homogeneous clutter environment. A distributed target is characterized by a number of detections from CFAR that have one or two equal coordinates (coordinates being range from radar, azimuth angle relative to true north and radial speed relative to receiver array) while the rest of coordinates differ by a small amount, as little as one resolution cell. In other words, those targets occupy multiple adjacent resolution cells. When the positions of these distributed targets are compared with data received from AIS (Automatic identification system) [10] it becomes clear that this distribution originates from single vessel, not a group of vessels. An example is presented in Fig. 3.

Fig. 3. Falsely Distributed detection. Four red markers represent detections found by CFAR while a white marker under them represents position received from AIS.,

To the best of our knowledge little attention is paid to this occurrence and no adequate solution could be found. These detections will move with the ship so tracker will rarely be able to eliminate them, since they will form tracks parallel to the track of the true ship detection [11]. In this way a group of vessels are created in Command & Control system in place of a single vessel. For that reason an algorithm that can properly eliminate false targets, while simultaneously maintaining true targets is needed. Additionally, it should, as much as possible, be able to recognize and avoid discarding detections that originate from multiple vessels sailing closely to each other.

III. Description of the Proposed Excision Algorithm

Excision algorithm implemented here operates on 3-D data structures called RDA Cubes (Range Doppler Azimuth Cubes). They consist of 200 × 140 × 512 fields, each of which can be declared a detection by CFAR algorithm. Each of these fields is a result of previous signal processing, and contains information about signal level received from a particular resolution cell. Dimensions of the RDA cube are related to the resolution HFSWR. In this implementation they contain information about 200 range resolution cells, 140 azimuth resolution cells and 512 Doppler shift resolution cells. These parameters dictate the dimensions of cube axes [4]. Excision consists of four steps:

1. Chose one of CFAR detections.
2. Form a sub-window around picked detection such that none of its outer cells are designed as detections by CFAR.
3. Using cells contained inside a sub-window calculate the most probable position of the target.

4. Keep the most probable detection while discarding the rest.

These steps are repeated until there are no more detections to be processed. In this implementation detections are listed by azimuth angle and they are processed in that order. The list of detections needs to be refreshed after each iteration because excision can (and often will) discard some detections. Detections that are once declared excised are not processed again.

It is important to ensure that sub-window is large enough to encompass whole falsely distributed target, but small enough so it doesn't encompass multiple closely spaced point-like targets, whenever it is possible. It was concluded empirically that most of falsely distributed targets are multiplied along only one detection parameter (range, Doppler shift or azimuth). Because of that, and the fact that 3-D data structures that are structured like a cuboid (for example three dimensional matrices in Matlab) are convenient to work with, sub-window extending is done in six directions: in a positive and a negative way along each axis. If there exists a detection in one of six faces of a sub-window in the current iteration, sub-window is extended in those directions. A simplified, two dimensional sub-window forming is presented in Fig. 4.

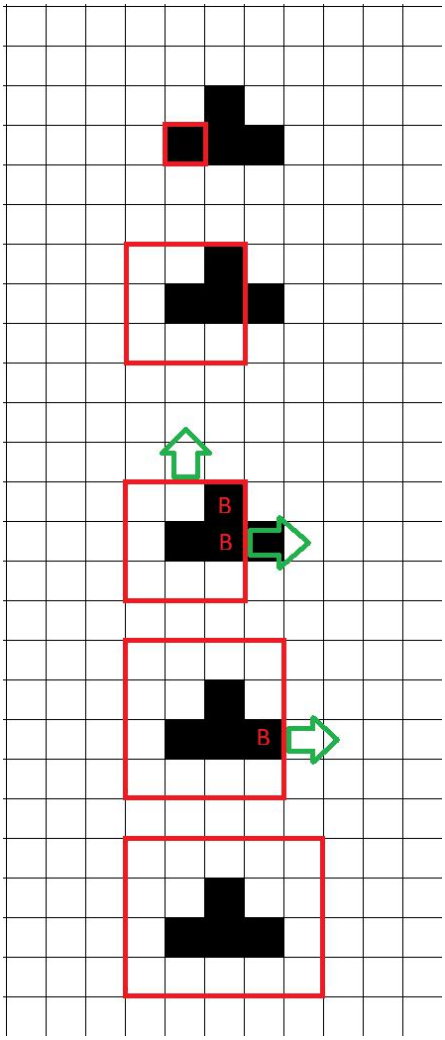


Fig. 4. Steps of sub-window forming for four connected detections in two dimensions, shown as black fields. Sub-window is presented as a red rectangle, directions of expansions as green arrows and detections found in outermost layer of a sub-window are marked with a red letter "B".

A single detection is chosen (marked with a red square, first from top in Fig. 4.) and first iteration of sub-window is formed from all the cells bordering it. Then all the cells contained in bordering layer of a sub-window are checked for detections. In the example from Fig. 4. two detections are found, existing in top and right face of a sub-window. Sub-window is expanded in those directions. These steps are repeated until there are no more detections in the outermost layer of a sub-windows (Fig. 4. bottom).

In the next step, a detection with the highest SNR (Signal to noise ratio) is picked. This will be used as a reference point. From that point all the other detections included in a sub-window are listed. Starting from them, position shift vector is calculated relative to reference cell using the weighting function. Cumulative position adjustment vector $\mathbf{P_{shift}}$ is calculated as shown in (1), while weighting function w in this implementation is shown in (2).

$$\vec{P}_{shift} = \sum_{i=0}^{n_{det}} \frac{\vec{P}_i - \vec{P}_{ref}}{\left\| \vec{P}_i - \vec{P}_{ref} \right\|} w_i; \tag{1}$$

$$w_i = \frac{10^{\frac{-(SNR_i^{dB} - SNR_{ref}^{dB})}{20}}}{2}; \tag{2}$$

Where n_{det} is the number of detections found inside the sub-window, $\mathbf{P_{ref}}$ represents the coordinates of a detection with the highest SNR, while $\mathbf{P_i}$ are coordinates of i-th detection listed in a sub-window. The unit of measurement of $\mathbf{P_{shift}}$ is resolution cell: a position adjustment vector of amplitude 1 will move the calculated position of the target in the given direction by one resolution cell. Separate coordinates of the target are calculated from amplitude and angle of $\mathbf{P_{shift}}$ and are converted to real units (km for range, degree for azimuth and Hz for Doppler shift).

The result of such a weighted method is that detections that are closer to the reference and have a SNR closer to it are going to have a greater influence on its position. This characteristic allows interpolation of a targets position based on several discreet measurements. A simple example would be two detections of very similar SNRs, detected at the same range and with a very similar radial speed, only differing by one azimuth resolution cell. To decide that the ship is exactly in one or the other resolution cell would not be accurate enough. Considering that the signal level in both cells is similar it is safe to assume that the target is somewhere between them. In this particular case excision algorithm is going to result in a position shift along the line connecting two detections, while the amplitude of this shift will be inversely proportional with the difference in SNR difference: theoretically, if the two detections have exactly equal SNRs final position of detection will be calculated as the middle of the line segment between the two detections.

Considering that this algorithm needs to perform in real time remote sensing system it is important to evaluate its complexity. For that purpose the worst case scenario will be discussed. A distributed target composed of n individual targets is used for that analysis. From the point of steps required for sub-window forming the worst case scenario is having all the distributed detections aligned in a single line. In that case full sub-window forming will take n-1 steps.

Every other configuration of bordering targets will be completely enclosed at least as fast as that because at least one step will have a chance of enclosing more than one detection at the same time. Complexity of this step is evaluated as O(n). Finding a detection with maximal SNR is a well known problem which has a complexity of O(n). Lastly, position shift vector for each detection except the reference one is calculated as shown in Eq. 1. Calculation time grows linearly with the number of detections inside sub-window, giving this step the complexity of O(n). These three steps combined give the entire algorithm a linear complexity, or O(n).

IV. Field tests

This algorithm was implemented in C++ programming language where it is applied on a preliminary detection list generated by CFAR. Test samples are samples of real data collected from two HFSWRs operating in Gulf of Guinea, Africa. A distributed target shown in Fig. 3. is the result of CFAR algorithm without excision algorithm applied. The same CFAR detection file is processed by excision algorithm and the same region shown in Fig. 3. is shown in Fig. 5.

Fig. 5. Detection excised from a distributed target (red marker). Informations about ship from AIS (white marker) are showed on the right.

From Fig. 5. it can be seen that the position of radar target is now unambiguous. Exact positions of falsely distributed targets (FDT rows) and resulting target after excision, along with distances from AIS position are given in Table I.

Table I
Positions of targets before and after excision

Marker	Latitude	Longitude	Distance from AIS
AIS	N 6.10009°	E 3.57250°	/
FDT1	N 6.09575°	E 3.58583°	1.55 km
FDT2	N 6.10222°	E 3.58000°	0.86 km
FDT3	N 6.11424°	E 3.56889°	1.64 km
FDT4	N 6.10816°	E 3.57389°	0.91 km
Excised	N 6.10863°	E 3.57232°	0.94 km

It is important to note that AIS is only used for confirming already detected targets, not for detection, but here it will be used as a reference for accurate position of the target. Even though excised detection is not the most accurate one, in comparison with position reported by AIS, mean value of distances from AIS of all falsely distributed targets is almost 35% higher than the distance of excised target form AIS. Additionally, total number of detections is decreased when excision is applied, thus false alarms are significantly reduced. Considering the whole system has to process all the data from one integration cycle before the next set of data arrives this reduction in number of targets that need processing is of great importance.

The time needed for the excision algorithm alone was not measured separately for it is eclipsed by the time needed to perform the CFAR algorithm. The time needed for the entire step of detection does not seem to vary more than several tens of milliseconds depending on whether excision is turned on or off. Considering that the duration of entire CFAR step of HFSWR is on the order of several seconds, a delay like that is negligible. Fig. 6. shows times needed to perform entire CFAR algorithm with and without excision algorithm applied.

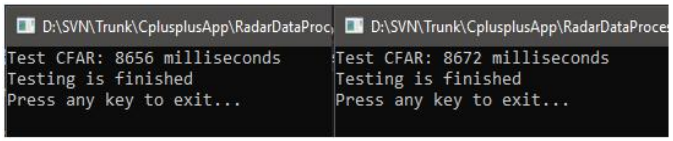


Fig. 6. Execution times for entire CFAR step without excision (left) and with excision enabled (right).

It is clearly shown that execution time difference is negligible, since it adds only around 16ms to the total CFAR execution time.

Significant reduction to the number of detections of CFAR output files allows for parameters of CFAR to be set so that the threshold will be lower, thus letting more detections, false and true, to be found. Most of the false detections resulting from waves, which are distributed by nature, will then largely be annulled by excision algorithm, leading to a smaller number of detections overall, while a lower threshold will allow for vessels that are further away to be detected. Tracker will then be responsible for discerning detections stemming from vessels from leftover detections stemming from waves. Fig. 7. shows one such situation, where light blue cone describes the zone in which HFSWR is projected to detect and track ships. Fig. 7. shows that in the main coverage area of the radar excision algorithm successfully filtered out most of the detections that were falsely distributed. It can be seen that calculated detection positions are closely related to positions received from AIS. Additionally, it can be seen that ships that are more than 200km away from HFSWR are successfully detected. Positions of those distant targets can be estimated with more precision by applying multi radar multi target fusion algorithms to data originating from multiple HFSWR systems covering the same area [12]. Targets detected outside its cone (left and right) are generally less accurate because of beam widening when receiver beam is steered so far off of bore-sight (more than 50° in either side).

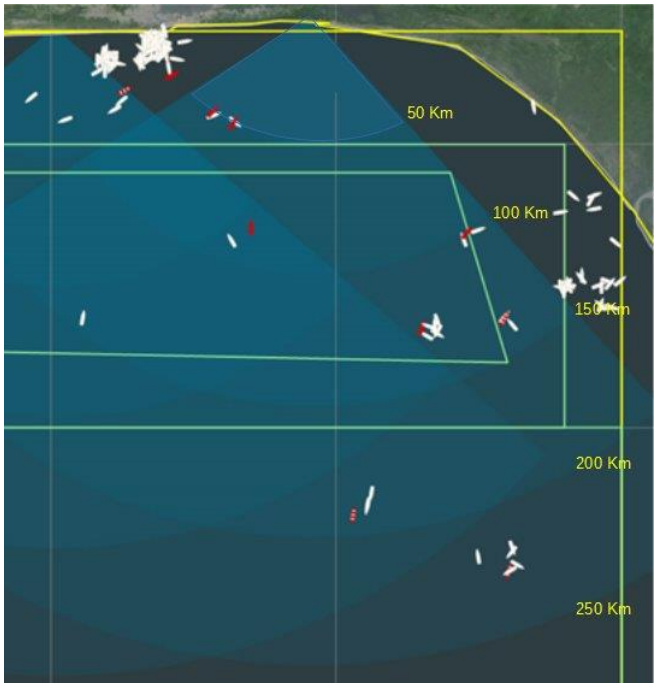


Fig. 7. Detections (red markers) correlated with AIS (white markers) after performing excision algorithm. Each arc represents 50km range away from radar.

Please note, two targets residing in the same range-angle cell, but having different Doppler shift can easily be differentiated by their radial speed. However, two targets that are in the same range and azimuth resolution cell can be merged into one in they have a very similar radial speed, if any of them is distributed. In practice such situation may be resolved by further signal processing, i.e. multi-radar fusion algorithm [12], [13]. However, a potentially better solution could be found during future research.

V. Conclusion

In this paper an example of how post-processing of CFAR detection data can be used to eliminate uncertainties related to target position. Algorithm for excising distributed targets is implemented and tested on real data collected from operating HFSWR systems located in the Gulf of Guinea. It was shown that, when applied, excision algorithm reduces the number of false targets while performing a type of interpolation to deduce where a real radar target is most probably located.

For the future work authors intend to continue optimisation of the weighting function in order to improve algorithm effectiveness.

References

[1] G. A. Fabrizio, *High Frequency Over-The-Horizon Radar,* 1st ed. New York City, USA: McGraw-Hill Education, 2013.

[2] L. Sevgi, A. Ponsford, H.C. Chan, "An integrated maritime surveillance system based on high-frequency surface-wave radars. Part 1. Theoretical background and numerical simulations", *IEEE Antennas and Propagation Magazine*, Volume: 43 Issue: 4, Aug 2001, Page(s): 28 -43

[3] A. Ponsford, A. Ponsford, H.C. Chan, "An integrated maritime surveillance system based on high-frequency surface-wave radars. Part 2. Operational status and system performance", *IEEE Antennas and Propagation Magazine*, Volume: 43 Issue: 5, Oct 2001, Page(s): 52 -63

[4] D. Nikolic, B. Dzolic, N. Tošic, N. Lekic, V. D. Orlic, B. M. Todorovic: "HFSW Radar Design: Tactical, Technological and Enviromental Challenges," Proc. OTEH 2016, Belgrade, Serbia, pp. 349-354, 6.-7. October 2016.

[5] C. A. Balanis, "Linear Wire Antennas" in *Antenna Theory: Analysis and Design,* 3rd ed. Hoboken, New Jersey, USA: Wiley-Interscience, 2005, ch. 4, sec. 7, pp. 191-193.

[6] N. Stojkovic, D. Nikolic, P. Petrovic, N. Tosic, I. Gluvacevic, N. Stojiljkovic, N. Lekic: "An Implementation of DBF and CFAR Models in OTHR Signal Processing," Proc. CSPA 2019, Penang, Malaysia, pp. 7-11, 8.-9. March 2019.

[7] O. M. Phillips, "Radar returns from the sea surface—Bragg scattering and breaking waves," *J. Phys. Oceanogr.*, 18, 1065–1074, DOI: 10.1175/1520-0485(1988)018<1065:RRFTSS>2.0.CO;2

[8] X. Lu, "Enhanced Detection of Small Targets in Ocean Clutter for High Frequency Surface Wave Radar," Ph. D. dissertation, Dept. of Electrical and Computer Engineering, University of Victoria, Victoria, British Columbia, Canada, 2009.

[9] D. Ivkovic, M. Andric, B. Zrnic, P. Okiljevic, N. Kozic: "CATM-CFAR Detector in the Receiver of the Software Defined Radar," *Scientific Technical Review*, Volume 64, Issue 4, pp. 27-38, 2014.

[10] IMO. *Resolution MSC.74 Annex 3 Recommendation of Performance Standards AIS;* IMO: London, UK, 1998.

[11] N. Stojkovic, D. Nikolic, B. Dzolic, N. Tosic, V. Orlic, N. Lekic, B. M. Todorovic: "An Implementation of Tracking algorithm for Over-The-Horizon Surface Wave Radar," Proc. TELFOR 2016, Belgrade, Serbia, 22.-23. November 2016.

[12] D. Nikolic, N. Stojkovic, Z. Popovic, N. Tosic, N. Lekic, Z. Stankovic, N. Doncov: "Maritime Over the Horizon Sensor Integration: HFSWR Data Fusion Algorithm," *Remote Sensing*, Volume 11, Issue 7, article no. 852, Apr. 2019, DOI:10.3390/rs11070852

[13] D. Nikolic, N. Stojkovic, N. Lekic: "Maritime over the Horizon Sensor Integration: High Frequency Surface-Wave-Radar and Automatic Identification System Data Integration Agorithm," *Sensors*, Volume 18, Issue 4, article no. 1147, Apr. 2018, DOI:10.3390s18041147

Analysis of different window function effects on DBF in HFSWR signal processing

Nemanja Grbić, Pavle Petrović, Dejan Nikolić, *Member, IEEE*, Nikola Stojković, *Member, IEEE,* Vladimir Orlić, *Member, IEEE*

***Abstract*—In this paper a few implementations of Digital Beam-forming (DBF) technique used in High Frequency Surface Wave Radar (HFSWR) signal processing are examined. All of them are based on phase shifting principle, but differ in so called "window" functions used for preparation of the data during the beam-forming process. Three different sets of magnitudes of weights used to form Hamming, Blackman and modified Blackman, window functions will be evaluated in this paper. Although choosing an optimal window function is situational, in this paper it is shown that for HFSWR DBF algorithm, modified Blackman can be regarded as the window of choice. Data used for this evaluation is obtained from HFSWR sites located in the Gulf of Guinea.**

***Index Terms*—HFSWR, DBF, Window functions, main lobe, side lobes.**

I. Introduction

MARITIME criminal activities and political disputes of areas away from territorial waters are an increasing occurrence, dealing significant damage to economies and endangering lives in the process. In order to take any action at all, consistent surveillance of the maritime area is mandatory. Not only does this apply to territorial waters, but for the entirety of the Exclusive economic zone (EEZ) [1] as well. Since EEZ consists of a 200 nautical mile (370 km) strip of water from the shoreline, some nations have hundreds of thousands of square miles of area to cover. Adding further to difficulties is the curvature of the earth preventing direct line of sight. Even though this all seems quite difficult, there are ways to monitor the EEZ.

One approach utilizes optical and microwaves sensors. To bypass their limitation, which is their range and direct line of sight requirement, these sensors are mounted on mobile platforms, such as satellites, air planes, ships, etc. However, this introduces limitations of the mobile platforms. The need to resupply these platforms, them being manned or unmanned, presents a drawback to EEZ monitoring which needs to be done constantly. Poor weather conditions can completely prevent deployment of mobile platforms.

Another approach is to use a network of long range stationary sensors, specifically high frequency surface wave radars. These provide detection well beyond the horizon due to surface waves following the curvature of the earth, therefore, eliminating direct line of sight requirements [3]. Since these radars are stationary, data availability is much more consistent compared to mobile platforms approach. Another important advantage these radars have is their much lower annual upkeep cost.

However, HFSWR has its disadvantages. Most notable one being its size, specifically, the area of land needed to deploy an HFSWR site [3]. In some areas of the world, finding a suitable location to fit both tactical and technical needs can be quite difficult. In some cases, these are remote locations far from settled areas. This creates logistical challenges such us communication, power routing and travelling conditions for maintenance crews. In order to ensure both unmanned and constant operation, various control algorithms are mandatory [4], [5]. Finally, there are different environmental [6] and man-made [7] factors that need to be accounted for, further increasing the system's complexity.

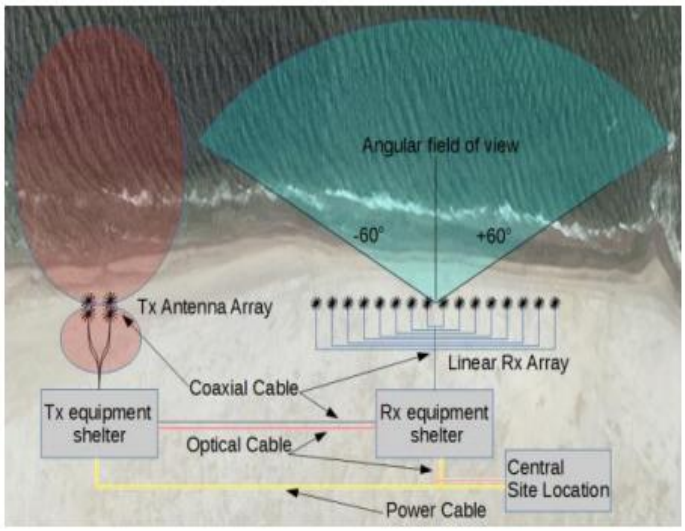


Fig. 1. Usual OTHR site deployment [3].

Nemanja Grbić is with the School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia, VLATACOM Institute, Bulevar Milutina Milankovica 5, 11070 Belgrade, Serbia(e-mail: nemanja.grbic@vlatacom.com).

Pavle Petrović is with the School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia, VLATACOM Institute, Bulevar Milutina Milankovica 5, 11070 Belgrade, Serbia(e-mail: pavle.petrovic@vlatacom.com).

Dejan Nikolić is with VLATACOM Institute, Bulevar Milutina Milankovica 5, 11070 Belgrade, Serbia(e-mail: dejan.nikolic@vlatacom.com).

Nikola Stojković is with the School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia, VLATACOM Institute, Bulevar Milutina Milankovica 5, 11070 Belgrade, Serbia(e-mail: nikola.stojkovic@vlatacom.com).

Vladimir Orlić is with VLATACOM Institute, Bulevar Milutina Milankovica 5, 11070 Belgrade, Serbia(e-mail: vladimir.orlic@vlatacom.com).

HFSWR consists of 2 antenna arrays, transmitter and receiver array. Additionally, some HFSWR sites are equipped with additional support equipment, power suppliers, communication equipment, and armed security. The Rx array consists of 16 quarter wave monopole antennae [8] with spacing between each element being 0.5 wavelengths. To ensure multiple HFSWR sites do not interfere with each other different operating frequencies are used.

II. DBF AND EXAMINED WINDOW FUNCTIONS

Complete signal processing is described in detail in [9]. In this paper main focus is on DBF. Input for DBF are so called Range Doppler (RD) maps. RD maps are matrices created using range and Doppler processing [9], done for each angle of incidence within the angular field of view. This is done for each antenna individually.

Each element of a single RD map for one antenna is multiplied by a complex weight coefficient, phase-shifting beam-forming [10]. A set of complex weight coefficients is formed by multiplying the steering vector for the angle of incidence of the current RD map by a set of real numbers.

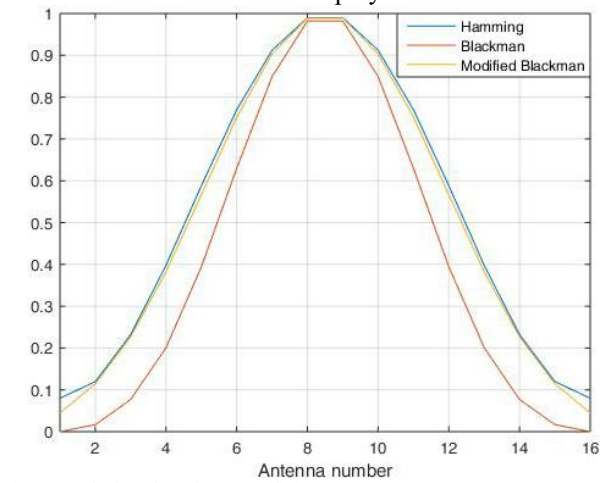


Fig. 3. Window functions coefficients.

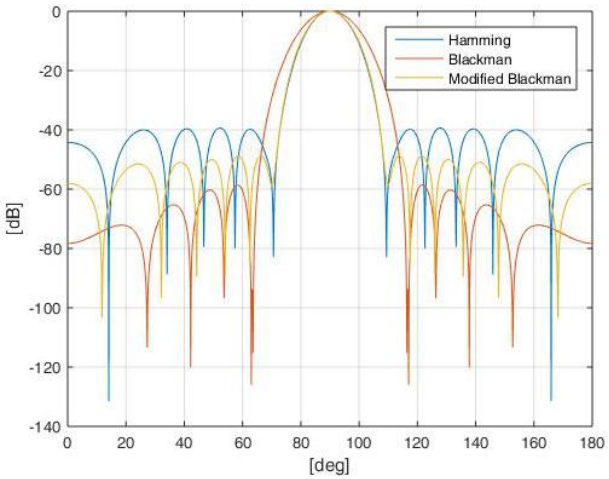


Fig. 4. AF generated using mentioned window functions.

This set of real numbers is the window function mentioned, and it affects array factor (AF), most notably main lobe width and side lobes intensity. Three window functions will be analyzed, Hamming [11], Blackman [11], and modified Blackman which will be explained in this paragraph. Figure 3. shows coefficients for all three window functions. Figure 4. shows array factors for those three windows. Both figures are generated using matlab. Equation (1) is the analytical form for Hamming window function, where $0 \le n \le M$ with M+1 being the window length,

$$w(n) = 0.54 - 0.46\cos(\frac{2\pi n}{M}) . \quad (1)$$

Equation (2) is the analytical form of Blackman window function, where $0 \le n \le M-1$, with N being the window length, while M is N/2 when N is even and (N+1)/2 when N is odd,

$$w(n) = 0.42 - 0.5\cos(\frac{2\pi n}{L-1}) + 0.08\cos(\frac{4\pi n}{L-1}) . \quad (2)$$

The other half of the window is obtained by flipping the first half around mid-point.

By comparing AF for Hamming and Blackman window functions it can be seen that Blackman window function gives a wider main lobe while drastically reducing side lobe levels. However, it can be seen from fig 3. that the weight coefficient amplitudes on the ends are practically zero. This means that antennae 1 and 16 have little influence on AF making them redundant. To avoid this, Blackman window was modified and tested. Modification is done by generating a set of Blackman coefficients for 20 members, but leaving the first two, and last two out of the window function. This solution is a compromise between Hamming window and default Blackman window function. The reason why rectangular window, which has the narrowest main lobe, is not mentioned is because of its pronounced side lobes. It was estimated that lower side lobe intensity is needed for HFSWR. Characteristics of all three AF are displayed in the following table.

TABLE I
ARRAY FACTOR CHARACTERISTICS

Window type	Main lobe 3 dB width [deg]	The first side lobe compression [dB]
Hamming	10.8939	39.2
Blackman	13.8923	58.62
Mod. Blackman	11.0983	48.83

III. TEST RESULTS

Range Doppler Angle (RDA) cubes, contain information about target radial speed and position (angle and distance). Using RDA cubes sampled from HFSWR in the gulf of Guinea, three beam-formers were tested

From an RDA cube itself, it is difficult to see the influence of beam-forming. However, taking an RA (Range Angle) map, from the RDA cube, for one specific Doppler frequency, allows for clear perception of beam-former influence. To confirm that results from RA maps are in fact targets, synchronization with Automatic identification system (AIS) was used. AIS and its synchronization with HFSWR are covered in detail in [12-14].

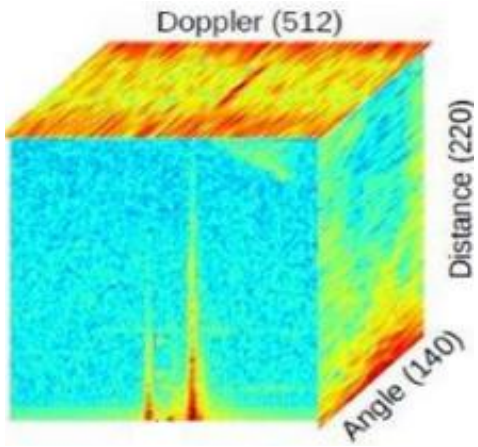


Fig. 5. RDA cube structure.

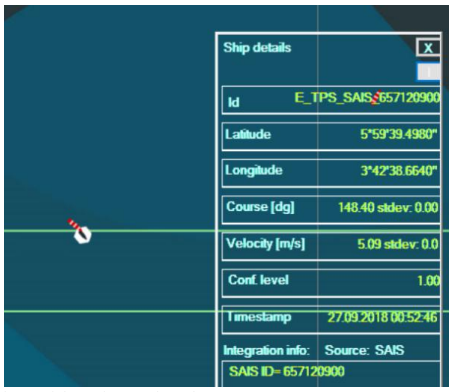


Fig. 6. Tracker with estimated and AIS targets.

The white ship symbol in the picture represents an AIS target. The red ship symbol represents HFSWR estimates after beam-forming and Constant False Alarm Rate (CFAR). From AIS, the target's exact position and velocity can be seen. This means that all RDA parameters for the target are known, and allows for an appropriate RA map to be chosen. The target is about 60 km from the site near the edge of HFSWR field of view.

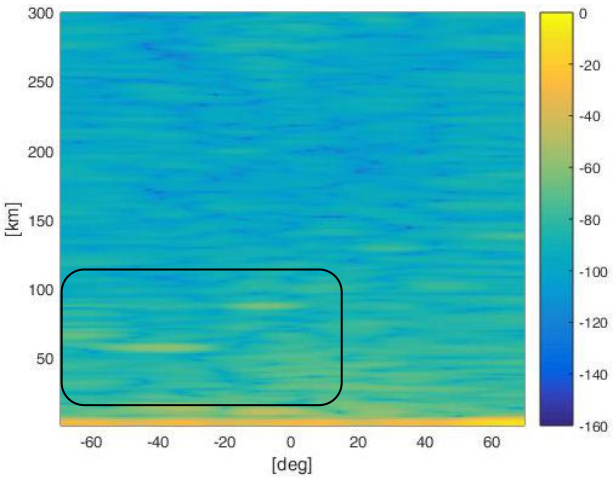


Fig. 7. Chosen RA map with highlighted target area.

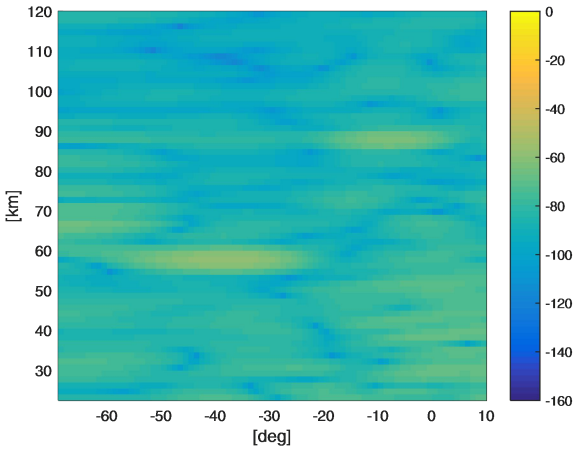


Fig. 8. Highlighted area displaying the target at 57 km range.

The RA maps on Fig 7. and 8. are made using Hamming window and their purpose is to demonstrate real data based on relevant parameters. Since "naked eye" cannot easily notice what difference different window functions have on target distribution in RA map, the maps are reduced to logarithmic amplitude versus angle plot at specific range. Since target is located at 57 km range, normalized logarithmic amplitude versus angle plot is given for that distance (Fig. 9).

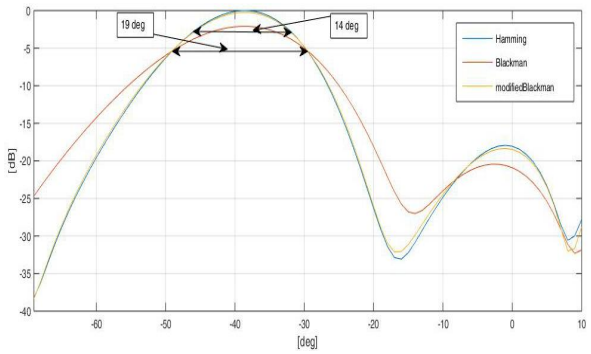


Fig. 9. Normalized logarithmic amplitude versus angle at 57 km range, with marked 3-dB main lobe width.

From Fig. 9. it can be seen that the target is distributed over arc nearly 20 degree wide. The reason for this is AF main lobe width. From table I it can be seen that all analysed window functions give at least 10 degrees of 3 dB main lobe width, which in turn gives a wide-angle displayed target. As mentioned, there are more steps in the signal processing after beam-forming that address this issue. However, a narrower main lobe is always desirable. From the presented plot it can be seen that Blackman window function displays the target over a wider area compared to the other two due to its AF having a wider main lobe. Although theoretical side lobe levels are at least 40 dB less than signal level (see Fig. 4.), in figure 9. it can be seen that are approx. 25 dB less than signal level. In RA the side lobes are manifested as constant range lines, which are interrupted due to AF shaping function influences. Their signal level is considerably lower than that of the real target and in most cases they will be eliminated during consecutive signal processing steps. However, sometimes this effect can cause appearance of false alarms.

IV. Conclusion

The advantage of DBF is that it is entirely done with software, so any changes to the beam-former can be immediately implemented. Henceforth, 3 window functions are described and evaluated here. It is worth noting that choosing an appropriate window function can be difficult and entirely situational, since it comes down to a compromise between main lobe width and side lobe intensity. In the case presented in this paper the modified Blackman window function shows the best performance since it provided the best of trade of between main-lobe width and side-lobes compression.

For the future work, more complex window functions will be implemented and tested at the various data sets in order to find the optimal function for the targeted application.

References

[1] United Nations, Law of the Sea, Part V – Exclusive Economic Zone. August 2011.

[2] G. A. Fabrizio, *High Frequency Over-The-Horizon Radar,* 1st ed. New York City, USA: McGraw-Hill Education, 2013.

[3] D. Nikolic, N. Tosic, B. Dzolic, N. Grbic, P. Petrovic, A. Djurdjevic, N. Lekic: "Tailoring OTHR deployment in order to meet conditions in remote Equatorial areas," Proc. CSPA 2019, Penang, Malaysia, pp. 7-11, 8.-9. March 2019.

[4] B. Dzolic, D. Nikolic, N. Tosic, N. Lekic, V. Orlic, B. Todorovic, "System for Remote Monitoring And Control of HF–OTHR Radar," Proceedings of 7th International Scientific Conference on defensive technologies (OTEH) 2016, Belgrade, Serbia, October 2016.

[5] P. Petrovic, N. Grbic, B. Dzolic, N. Lekic, M. Peric, "Software for Monitoring of Direct Path Test Data for HFSW Over the Horizon Radar," Proc. of IcETRAN 2018, Palic, SR, June, 2018.

[6] M. C. Peel, B. L. Finlayson, T. A. McMahon, "Updated world map of the Koppen – Geiger climate classification,"*Hydrol. Earth Syst. Sci.*, vol. 11, pp 1633 – 1644, Oct. 2007, DOI: 10.5194/hess-11-1633-2007

[7] N. Tošic, A. Samčovic, N. Lekic, B. Todorovic, S. Jankovic, S. Mladenovic, "Analiza interferencije u HF opsegu uzrokovane LED reflektorom koriscenjem slike spektrograma", In proc. Of Telecommunications forum TELFOR, November 2018, Belgrade, Serbia – published in Serbian

[8] C. A. Balanis, "Linear Wire Antennas" in *Antenna Theory: Analysis and Design,* 3rd ed. Hoboken, New Jersey, USA: Wiley-Interscience, 2005, ch. 4, sec. 7, pp. 191-193.

[9] N. Stojkovic, D. Nikolic, P. Petrovic, N. Tosic, I. Gluvacevic, N. Stojiljkovic, N. Lekic: "An Implementation of DBF and CFAR Models in OTHR Signal Processing," Proc. CSPA 2019, Penang, Malaysia, pp. 7-11, 8.-9. March 2019.

[10] H. L. Van Trees, "Optimum Array Processing," John Wiley & Sons, Inc., 2002. ISBN 9780471093909

[11] Oppenheim, Alan V., Ronald W. Schafer, and John R. Buck. Discrete-Time Signal Processing. Upper Saddle River, NJ: Prentice Hall, 1999.

[12] IMO. *Resolution MSC.74 Annex 3 Recommendation of Performance Standards AIS;* IMO: London, UK, 1998.

[13] D. Nikolic, N. Stojkovic, Z. Popovic, N. Tosic, N. Lekic, Z. Stankovic, N. Doncov: "Maritime Over the Horizon Sensor Integration: HFSWR Data Fusion Algorithm," *Remote Sensing*, Volume 11, Issue 7, article no. 852, Apr. 2019, DOI:10.3390/rs11070852

[14] D. Nikolic, N. Stojkovic, N. Lekic: "Maritime over the Horizon Sensor Integration: High Frequency Surface-Wave-Radar and Automatic Identification System Data Integration Algorithm," *Sensors*, Volume 18, Issue 4, article no. 1147, Apr. 2018, DOI:10.3390s1804114

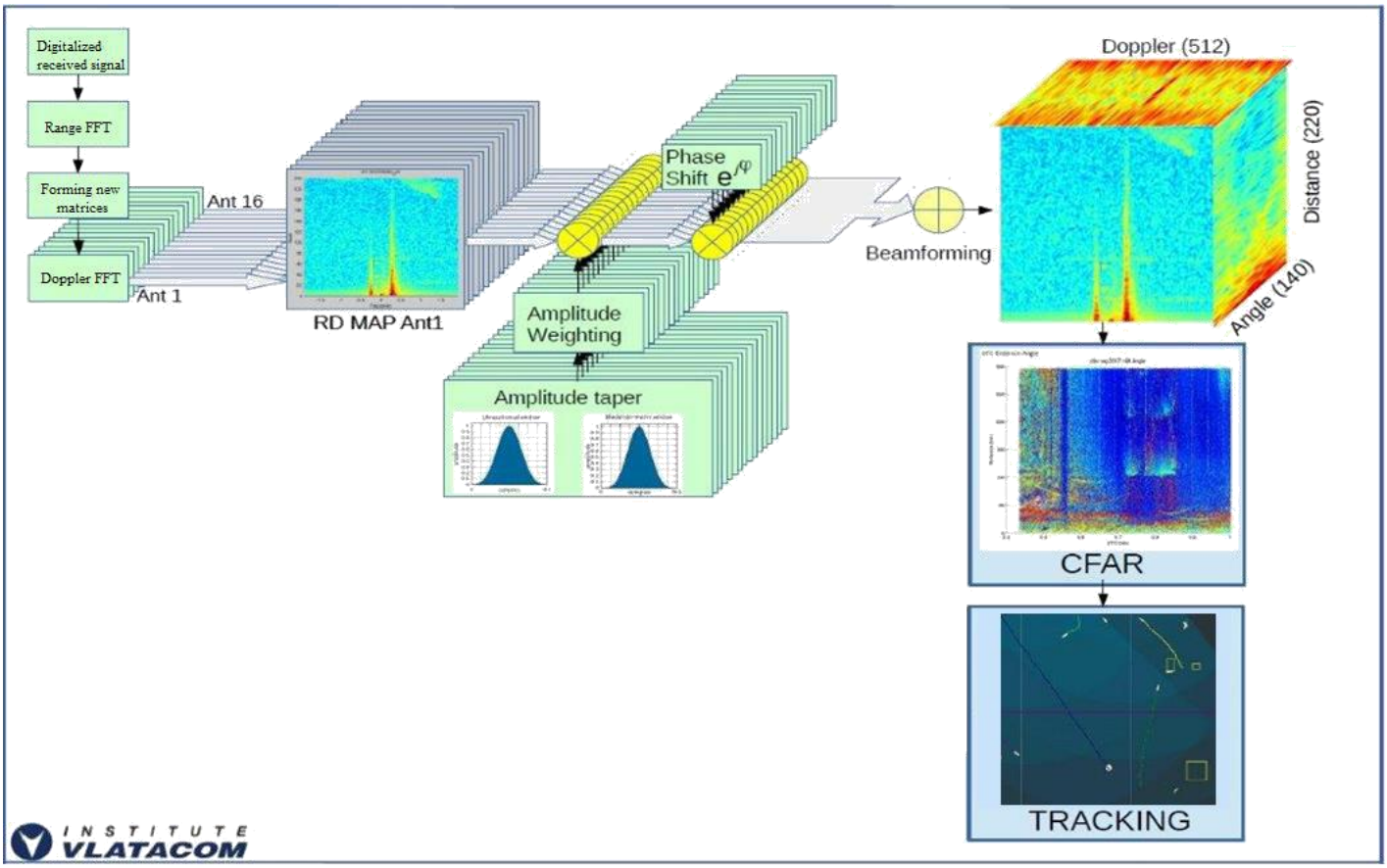


Figure 2. Signal processing steps [9].

Implementacija tunelovanja Q-SIG preko SIP u privatnoj telefonskoj mreži sa integrisanim uslugama funkcionalnog korisnika

Slađan Svrzić, Zoran Čiča, Zoran Miličević i Zoran Perišić

***Apstrakt*— U radu se daje objašnjenje za novi način povezivanja ISDN PABX i IP orijentisanih PABX (PINX) u privatnoj automatskoj telefonskoj mreži integrisanih usluga funkcionalnog korisnika, primenom postupka tunelovanja mrežne Q signalizacije (Q-SIG) kroz privatnu IP mrežu (Intranet) sa SIP. Dat je kratak prikaz Standarda ECMA-355 za postupak tunelovanja Q-SIG kroz IP/SIP mreže i opisana je njegova praktična primena na delu automatske telefonske mreže integrisanih usluga, za međusobno povezivanje učesnika sa različitih krajnjih ISDN/Q-SIG PABX, preko tranzitnih IP/Q-SIG PINX, a pri čemu IP/SIP mreža ima ulogu Interventne mreže (*Intervening Network*, IVN).**

***Ključne reči*— Privatna automatska telefonska mreža integrisanih usluga; Q signalizacija; tunelovanje Q-SIG; IP orijentisane PABX.**

I. UVOD

Q-SIG je signalizacioni sistem koji je dizajniran za upotrebu u korporativnim telefonskim mrežama (*Corporate Telephony Network*, CTN). Standardizovan je na globalnom nivou i podržan od strane vodećih svetskih proizvođača Privatnih automatskih telefonskih centrala (*Private Automatic Branch Exchange*, PABX) [1]. Koristeći proverenu ISDN i IP tehnologiju, Q-SIG obezbeđuje opseg osnovnih i dopunskih usluga za poboljšanje poslovnih rezultata korisnika privatnih telekomunikaciono-informacionih sistema. Q-SIG je jedini sistem signalizacije za korporativne mreže koji obezbeđuje nezavisnost od proizvođača i garantovanu interoperabilnost na globalnom nivou [1].

Q-SIG protokol mrežne signalizacije funkcioniše između PINX (*Private Integrated services Network eXchange*) u okviru privatne mreže sa integrisanim uslugama - PISN (*Private Integrated Services Network*) [2]. U okviru PISN se putem režima komutacije kola korisnicima obezbeđuju kako osnovne, tako i dopunske usluge [3]. U okviru Standarda ECMA-143 (*Upravljanje pozivima u podršci osnovnim uslugama*) [4], zatim u Standardu ECMA-165 (*Generički funkcionalni protokol za podršku dopunskim uslugama*) [5], kao i u nizu drugih standarda koji određuju pojedinačne dopunske usluge, specificirana je uloga Q/SIG za podršku tim uslugama. Ime Q-SIG je izvedeno iz činjenice da se koristi za signalizaciju u "Q" referentnoj tački PINX. "Q" referentna tačka je logička tačka razgraničenja, tj. nalazi se na krajevima logičke veze dve PINX, koja se naziva inter-PINX linkom (IPL). Fizička veza sa svakom od PINX vrši se na referentnoj tački „C". U tom slučaju, ulogu IVN (*Intervening Network*) mogu preuzeti namenski kanali za komunikaciju (analogni ili digitalni) ili komutirane komunikacije (za virtuelne privatne mreže) [1].

U ostatku rada je u okviru drugog poglavlja predstavljena Q-SIG u okviru SIP IP. U trećem poglavlju opisana je primena tunelovanja Q-SIG preko SIP IP u okviru PISN funkcionalnog korisnika, kroz predloženi i testirani model mreže. Na kraju su data zaključna razmatranja.

II. Q-SIG U OKVIRU SIP IP

ECMA-355 je standard koji veoma uspešno produžava upotrebu Q-SIG i u IP orijentisanim privatnim telefonskim mrežama (*Private Telephony Network*, PTN) [6], tako što specificira tunelovanje Q-SIG preko *Session Initiation Protocol* (SIP) [7]. Tunelovanje Q-SIG preko SIP omogućava međusobno pozivanje između PABX, "ostrva" mreža sa komutacijom kola koje koriste Q-SIG, i u slučaju da su one međusobno povezane IP mrežom (koja koristi SIP) [8], i to bez gubitka Q-SIG funkcionalnosti. Ovaj standard olakšava uvođenje novih poboljšanih SIP i *Session Description Protocol* (SDP) funkcionalnosti [9], a koje su specificirane nakon objavljivanja ranijih izdanja ovog standarda. Ta poboljšanja uključuju mogućnost primene zaštite korisnog signala i mehanizme za funkcionalniju razmenu informacija preko SDP, kao i primenu novouvedenog indikatora za uočavanje promena u proceduri signalizacije.

Standard prvenstveno specificira tunelovanje Q-SIG preko SIP protokola u okviru korporativnih telefonskih mreža, a koji se obično odvija preko IP mreže, pri čemu se telefonski pozivi smatraju vrstom multimedijalne sesije u okviru koje se razmenjuju samo audio signali [8].

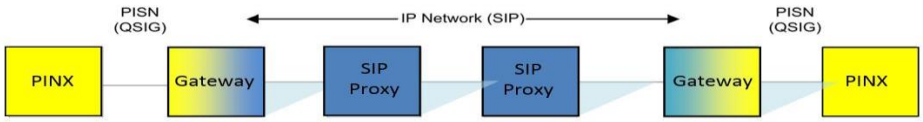
Često su velike CTN takve arhitekture da sadrže granične PISN koje koriste Q-SIG, kao i sopstvene centralne IP mreže koje koriste SIP. Tada postoje dva slučaja u scenariju pozivanja između učesnika. U prvom slučaju, Q-SIG poziv ili Signalna konekcija nezavisna od poziva (*Call Independent Signalling Connection*, CISC) mogu poticati od "A" učesnika

Slađan Svrzić – Towersnet, dr Agostina Neta 16/24, 11070 Novi Beograd, Srbija (e-mail: s.svrzic@towersnet.rs).

Zoran Čiča – Elektrotehnički fakultet, Univerzitet u Beogradu, Bulevar Kralja Aleksandra 73, 11020 Beograd, Srbija (e-mail:zoran.cica@ etf.rs).

Zoran Miličević – Uprava za informatiku i telekomunikacije, GŠ VS, Raška 2, 11000Beograd, Srbija (e-mail: zoranmilicko@gmail.com).

Zoran Perišić – Uprava za informatiku i telekomunikacije, GŠ VS, Raška 2, 11000Beograd, Srbija (e-mail: zperisic@gmail.com).



Sl. 1. Raspored entiteta kod realizacije poziva od Q-SIG, preko SIP, na Q-SIG [6].

spojenog na PISN i završiti kod "B" učesnika povezanog na IP mrežu, ili obrnuto. U takvim situacijama, gejtvej obezbeđuje međusobno povezivanje Q-SIG i SIP na granici između PISN i IP mreže. Realizacija osnovnog interaktivnog poziva preko gejtveja specificirana je u Standardu ECMA-339 [10].

Drugi slučaj je kada Q-SIG poziv ili CISC, koji potiču od "A" učesnika povezanog na PISN, u okviru koje se koristi Q-SIG, prolaze preko IP mreže koristeći SIP, i završavaju kod "B" učesnika povezanog sa drugom PISN (ili drugim delom iste PISN), u okviru koje se koristi Q-SIG.

Standard ECMA-355 tretira navedeni slučaj, jer se prilikom takvog načina povezivanja sačuvaju sve mogućnosti Q-SIG u toku prolaza kroz IP mrežu [6]. To se postiže tunelovanjem, tj. enkapsuliranjem Q-SIG poruka unutar SIP zahteva i SIP odgovora, koji se razmenjuju u kontekstu propisanog SIP dijaloga. Ovakva se veza smatra "*end-to-end*" povezivanjem, a arhitektura za to se može postići korišćenjem gejtveja na svakom prelazu između PISN, koja koristi Q-SIG, i IP mreže, koja koristi SIP [6], kao što je to prikazano na slici 1.

Svaki gejtvej može da obezbedi međusobno povezivanje PISN i IP mreže, na način kako je to opisano u ECMA-339, tj. samo za uslugu osnovnog poziva-BC (*Basic Call*), kako je to specificirano u Standardu ECMA-143 [4]. Mnoge druge mogućnosti Q-SIG (podrška za dopunske usluge i dodatne mrežne usluge) su, kao vlasničke, specificirane u drugim standardima i specifikacijama karakterističnim za konkretnog proizvođača. Neke od tih dodatnih mogućnosti Q-SIG su pogodne za međusobno povezivanje sa SIP, dok druge nisu, pošto u SIP za to ne postoje odgovarajući elementi ili se pak te karakteristike u okviru SIP ostvaruju na način, koji nije kompatibilan sa Q-SIG.

Rešenje za takve situacije nađeno je u tunelovanju Q-SIG poruka kroz IP mrežu, tj. u njihovoj enkapsulaciji u okviru SIP poruka, tako da ne postoji mogućnost gubljenja delova Q-SIG kod pomenutog "*end-to-end*" povezivanja. U tom slučaju, jedan od dva gejtveja kreira SIP dijalog sa drugim gejtvejom, a SIP poruke u okviru tog dijaloga koriste se za enkapsulaciju Q-SIG poruka. Kroz upotrebu SDP opisanu u RFC 3264 [9], dijalog takođe uspostavlja sesiju u kojoj medijski tokovi, nose korisničke informacije (npr. govor) između dva Q-SIG gejtveja. U stvari, tada ta dva gejtveja funkcionišu kao Q-SIG *Transit PINX*, prenoseći Q-SIG poruke sa malom ili nikakvom modifikacijom.

Sa razmatranim rešenjem za tunelovanje, IP mreža obezbeđuje inter-PINX konekciju (IPC) između dva gejtveja, koji onda funkcionišu kao tranzitne PINX. Tunel koji obezbeđuje SIP za Q-SIG poruke deluje kao kanal za signalizaciju D_Q, a medijski tokovi funkcionišu kao kanali za korisničke informacije U_Q [11]. Slika 2 ilustruje ovaj koncept.

Svaka PINX se fizički povezuje sa interventnom mrežom preko interfejsa na referentnoj tački "C". Interventna mreža obezbeđuje fizičku IPC između referentnih tačaka "C" krajnjih PINX. Funkcije mapiranja, unutar svake od PINX, vrše mapiranje D_Q-kanala i U_Q-kanala u "Q" referentnoj tački preko jedne, ili više IPC ostvarenih u referentnoj tački "C".

Standard ECMA-336 specificira upotrebu funkcija mapiranja za slučaj kada je interventna mreža zasnovana na IP [11] i kada se koristi za uspostavu sledećih tipova IPC:

- za TCP konekciju, za prenos signalnih informacija i informacija o kontroli resursa [12] i
- za par UDP tokova, po jedan tok u svakom smeru, za prenos korisničkih informacija preko RTP [13].

Jedan inter-PINX link zahteva jednu TCP konekciju, za podršku D_Q-kanalu, i jedan par UDP tokova po U_Q-kanalu. Pored prenosa Q-SIG protokola, TCP veza je takođe potrebna za prenos informacija o kontroli resursa za uspostavljanje UDP tokova.

Ovaj standard podržava dva načina interkonekcije između krajnjih PINX: "Na zahtev" (gde se jedna TCP veza i par UDP tokova uspostavljaju na početku svakog poziva i brišu na kraju tog poziva) i "Polu-stalnu" (gde jedna TCP veza sa neograničenim trajanjem prenosi Q-SIG u ime više poziva). U slučaju takve interkonekcije, TCP veza može podržati nula, jedan ili više poziva u isto vreme.

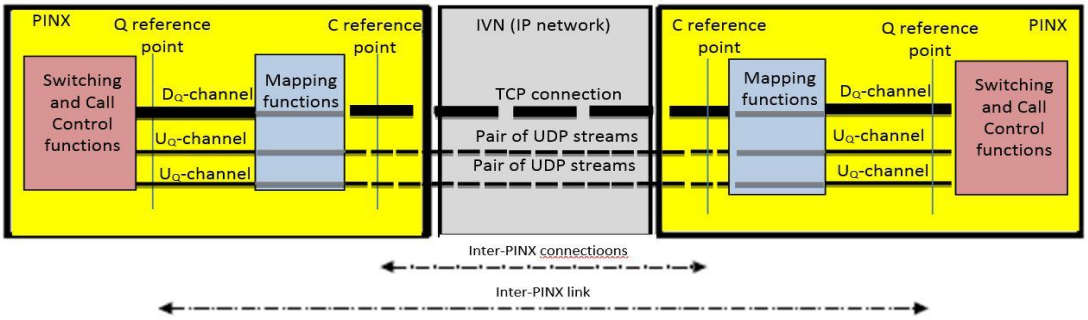
Par UDP tokova za korisničke informacije se uspostavlja na početku svakog poziva i briše se na kraju tog poziva [11].

III. PRIMENA TUNELOVANJA Q-SIG/SIP IP U OKVIRU PISN FUNKCIONALNOG KORISNIKA

Pod pojmom funkcionalni korisnik (FKo) podrazumeva se specifična organizacija koja, zbog svoje namene, ustrojstva, zadataka i teritorijalne razuđenosti, ima izuzetno velike zahteve u pogledu savremenih telekomunikacija, a pre svega za posedovanjem integrisanog sistema telekomunikacija za dovoljno brzu obradu i prenos tačnih i zaštićenih informacija.

Pri tome, odavno je postalo veoma važno da se kroz prenos i komutaciju informacija (govornih i negovornih) korisnicima u okviru takve organizacije ponudi široki spektar, najsavremenijih telekomunikacionih korisničkih usluga i mrežnih servisa, čiju tehničku podršku treba da pruži, ne samo na TDM već i na IP platformi, savremeno organizovana i funkcionalno orijentisana privatna telefonska mreža (PTN).

Kao takva, PTN razmatranog FKo treba da predstavlja jedinstvenu telekomunikaciono-informatičku platformu, te da po svojoj kompleksnosti, organizaciji, dostignutom stepenu tehničke integracije i geografskoj rasprostranjenosti koincidira ka performansama, odrednicama i standardima savremene



Sl. 2. IPC koncept (polu-stalna konekcija) [11].

korporativne telefonske mreže.

To onda znači, da nije potpuno zatvorena i okrenuta samoj sebi, već da se putem, različitih i geografski razuđenih, interkonekcija povezuje sa javnom fiksnom i mobilnim telekomunikacionim mrežama, kao i sa postojećom, funcionalnom, digitalnom, tranking, mobilnom radio-mrežom TETRA (*Terrestrial Trunked Radio*).

Treba reći da je Automatska telefonska mreža (AtlMr) razmatranog funkcionalnog korisnika kompleksna privatna mreža automatske telefonije, sa učešćem više desetina ISDN i IP orijentisanih PABX, povezanih po mešovitoj arhitekturi „zvezde“ i „petlje“ i višenivovski strukturiranih kao krajnje, čvorne, tranzitne i glavna tranzitna. Primenom Q-SIG, kao sistema CCS (*Common channel signalling*) orijentisane signalizacije, u toj AtlMr obezbeđena je interoperatibilnost između digitalnih automatskih telefonskih centrala (DATC) različitih proizvođača. To je postignuto u samoj fiksnoj telekomunikacionoj mreži funkcionalnog korisnika, zatim sa DATC iz drugih privatnih telekomunikacionih mreža (sa kojima je u interkonekciji), te i sa DATC javnih operatora fiksne i mobilne telefonije. Na taj način, obezbeđena je potpuna nezavisnost od različitih proizvođača ISDN DATC, kao i predviđeni obim osnovnih i dopunskih korisničkih usluga i mrežnih servisa, bez obzira na koji su komutacioni čvor priključeni. Pri tome, u vezi sa daljom primenom Q-SIG u ATlMr, za dotičnog funkcionalnog korisnika se nesumnjivo nameću prioriteti da se iste usluge i servisi realizuju, ne samo na njenim homogenim ISDN delovima, već i na delovima sa IP/SIP interventnim mrežama iz okvira Intraneta.

Odgovarajući na te obaveze, u poslednjih nekoliko godina osavremenjen je, ili potpuno zamenjen, određen broj čvornih i tranzitnih DATC, koje su postale interoperatibilne, na nivou IP/SIP, i sa izgrađenom IP mrežom za komutaciju paketa. Tu se prvenstveno misli na savremene komutacione sisteme, koji su posebno dizajnirani za rad u PISN uz interoperatibilnost sa IP/SIP okruženjem (PINX).

Praktično rešenje tunelovanja Q-SIG kroz IP proxy, primenom SIP/UDP, ostvareno je na delu ATlMr FKo na kome je povezivanje između učestvujućih tranzitnih PINX, realizovano snopom od 30 IP trankova preko LAN 10/100 Mb/s Etherneta. Učestvujuće krajnje ISDN PABX (DATC) povezane su na pripadajuće tranzitne PINX putem prenosničkih snopova od po 30 TDM trankova (sa E1/ISDN PRI) i sa primenom sistema mrežne signalizacije tipa Q-SIG.

Tranzitne PINX su izgrađene u arhitekturi Linijskih modula interfeja (LIM-ova) sa 7U magacinima, takozvanih *Media Gateway* LIM-ova, koji se sastoje od *Telephony Servera* i *Media Gateway*-a, bitnih za rad kako sa IP/SIP, tako i sa ISDN/Q-SIG okruženjem [15]. Ta verzija sistema komutacije pruža SIP ekstenzijama i SIP trankovima visok nivo funkcionalnosti, sa podrškom koja je u skladu sa više od 45 RFC standarda, kao osnovu za integraciju sa aplikacijama objedinjenih komunikacija trećih strana i organizaciju modernijih i efikasnijih poslovnih procesa, kao i za vezu sa spoljnim svetom putem SIP trankova (kao alternacija tradicionalnim PSTN mrežama) [15].

Primenjeni sistem je sertifikovan za veliki broj SIP mrežnih provajdera i odgovara SIP *Connect* 1.1 standardu. Njegov SIP tranking interfejs je takođe osnova za integraciju sa sistemima treće strane, kao što su *Microsoft Linc 2013/Skipe-for-Business, IBM Sametime SUT* i druge platforme za omogućavanje interoperatibilnosti između različitih komunikacionih platformi. Pored toga, sa ovim sistemom podržan je tip SIP tranka putem koga se vrši umrežvanje sa prenosom svih standardizovanih učesničkih i mrežnih usluga. To znači, da je na taj način omogućena i transparentnost funkcija između, njime povezanih, čvorova [15].

LIM se sastoji od jednog *Server*-a na koji, po potrebi, može biti povezano od jednog do petnaest *Media Gateway*-a, ali se u konkretnom slučaju radi o samo jednom *Media Gateway*-u po LIM-u. *Telephony Server*-i, kao osnovni delovi servera na obe tranzitne PINX, poseduju operativni sistem *Novells SUSE® Linux Enterprise Server (SLES)* verzije 11, sa 64-bitnom arhitekturom. Implementiran je server u obliku serverske ploče ASU-E (*Aastra Server Unit Ebedded*) sa poboljšanim performansama, koja je bazirana na *COMXpress* standardu u konfiguraciji sa: procesorom 2.26GHz Com2Duo, RAM 4GB (proširiva do 8GB), SATA HDD kapaciteta 160GB, podržanim SW RAID1 i 2 Eternet porta [16].

Media Gateway Unit je implementirana MGU2 ploča, koja u osnovnoj varijanti podržava do 4 interfejsa ISDN E1 (ili T1), mobilne lokale, IP lokale (H.323, SIP – uključujući IP DECT i WiFi), IP prenosnike i IP umrežavanje putem H.323 i SIP, te Q-SIG umrežavanje. Na svakoj učestvujućoj tranzitnoj PINX u LIM-u je implementirana po jedna MGU2 ploča, koja u

LIM kabinetima ima tačno određenu fizičku poziciju [17].

Osnovne komponente instaliranog *Media Gateway*-a su: TDM *Switch*, *Ethernet* (Layer 2) i *Media Stream Processor* (MSP).

TDM *Switch* na MGU2 ima centralnu funkciju za komutaciju svih medijskih veza između trankova, ekstenzija i pomoćnih funkcija, koje se prenose kroz njega. TDM *Switch* je neblokirajuće vremensko komutaciono polje (tipa: *Time-Space-Time*) za 2048 x 64kb/s vremenskih pozicija.

Na MGU2 ploči postoje dve standardne 10/100/1000 *Base-T* LAN veze označene sa LAN 0 i LAN 1. *Ethernet* paketi tipa "*Non-RTP*" za signalizaciju, se preko sviča Sloja 2 usmeravaju na procesor uređaja-DP (*Device Processor*), dok se "RTP" paketi koji prenose medije (VoIP), preusmeravaju na DSP u MSP.

MSP je sistem procesor za VoIP telefonske aplikacije i ujedno zajednička baza resursa koju dele sve DSP orijentisane funkcije vezane za MGU2, kao što su VoIP, T.38 i DTMF prijemnici [17].

RTP paketi [14], koji se šalju preko IP mreže, podložni su slučajnim varijacijama kašnjenja, dolasku van redosleda i riziku da budu odbačeni. Ovi negativni efekti smanjuju kvalitet prenošenog audio signala, pa se za ublažavanje tog efekta na MGU2 koristi kolo *Jitter Buffer*-a, koje tu aktivnost realizuje na račun povećanih kašnjenja kod prenosa govora. Duga kašnjenja kod prenosa govora, posebno u kombinaciji sa ehom na daljem kraju, čine da eho postane primetniji i uznemiravajući. Iako se za pokušaj minimizacije tog kašnjenja koristi ugrađeni poništavač eha, mora se imati u vidu, da na njegov rad može imati negativan uticaj situacija kada dođe do smanjenja kvaliteta govora (npr. zbog gubitka paketa) [17].

Ugrađeni poništavač eha EC (*Echo Canceler*) uklanja eventualno postojeći eho predajnog signala (povratni signal), iz prijemnog signala. Eho je obično uzrokovan refleksijama koje se javljaju kod 2/4-žičnog prenosa, ali i kao akustični eho u telefonskim aparatima. EC se koristi samo za pozive preko IP mreže, koja radi u režimu sa komutacijom paketa (VoIP).

Važno je napomenuti da je konfiguracija EC takva, da postoji samo jedan na MGU2 ploči, pa tako utiče na sve VoIP pozive u njoj, uključujući i inter-Media GW pozive preko IP-a. Prema tome, prilikom testiranja izbor podešavanja EC morali smo realizovati kao kompromis, u smislu izrečenog [17].

Za sve odlazne RTP (audio) media tokove, pri VoIP pozivu, MGU2 kreira slučajnu 32-bitnu vrednost SSRC (*Sinchronization Source*). Ova vrednost se koristi za sve RTP pakete u tom toku, sve vreme dok je on aktivan. Ako se uspostavljeni poziv stavi na čekanje, ili se izvrši bilo koja promena trenutno aktuelnog RTP media toka od strane sistema komutacije (npr. promena DTMF relejnog režima), aktuelni media tok se zatvara i zamenjuje novim, sa novom SSRC vrednošću.

Na odgovarajućem dolaznom RTP media toku, MGU2 potvrđuje sve primljene RTP pakete. Paketi sa bilo kojom SSRC vrednošću će biti prihvatani sve dok se primaju parovi paketa sa uzastopnim brojevima i istim SSRC [17].

Glavne funkcije implementirane MGU2 ploče su sledeće: posreduje u svim komunikacijama ka komutacionom sistemu; poseduje interfejse za digitalne trakove E1/T1; obezbeđuje RTP/SRTP, DTMF detekciju, DTMF i faksimil tonove preko RTP; otklanja eho i smanjuje džiter u VoIP komunikaciji [17].

Prilikom implementacije MGU2 ploče u sistem provodi se procedura inicijalizacije njegovih funkcija kao Media Interfejsa (važne su za RTP podatke u medija tokovima) i kao Signalizacionog GW-a (važan za manipulaciju sa Q-SIG porukama). U tom smislu, prijavljuju se njegova IP adresa *Default* GW-a, zatim IP adresa Media Interfejsa i IP adresa Signalizacionog GW-a [17].

Primenjeni Sistem komutacije podržava, kako IPv6, tako i IPv4 adresiranje, pa u konkretnom rešenju PINX rade kao izvorna IPv4 mreža za Menadžere poziva i komponente GW-a. Pri tome, platforma HV i OS softver rade koristeći IPv4 "*dual stack*" interfejse za IP mrežu. Softverske komponente elemenata (Menadžer poziva i GW) konfigurisani su za obavljanje i međusobno korišćenje standardnog IPv4 adresiranja, kako iz perspektive signalizacije, tako i iz perspektive medija. Instalirana je inter-Server komunikacija IPv4 "*dual stack*", tako da ceo sistem (svi serveri) koristi istu IP verziju. IPv4 adrese, iz 32-bitnog opsega izvorne IP mreže, podržane su za SIP terminale, SIP klijente i SIP trankove. *Media Server* podržava i IPv4 i IPv6 za medijske tokove (RTP i RTCP) i upravljanje odlaznim pozivima (SCTP). Pošto MGU2 i *Media Server* podržavaju IPv4, to onda ne ometa primenu mera bezbednosti (zaštitu signalizacije) u sistemu [17].

Interventna mreža je realizovana putem sopstvenog IP proxy, koji je izveden na bazi 26-to portnih 10/100 *Base-T Ethernet CISCO* svičeva tipa SQ200-26pp, na oba kraja, i sa spojnim putem kroz 10/100 *Base-T Ethernet* kapacitet Optičkog digitalnog sistema prenosa brzine 625 Mb/s, a po kojima se takođe koristi SIP.

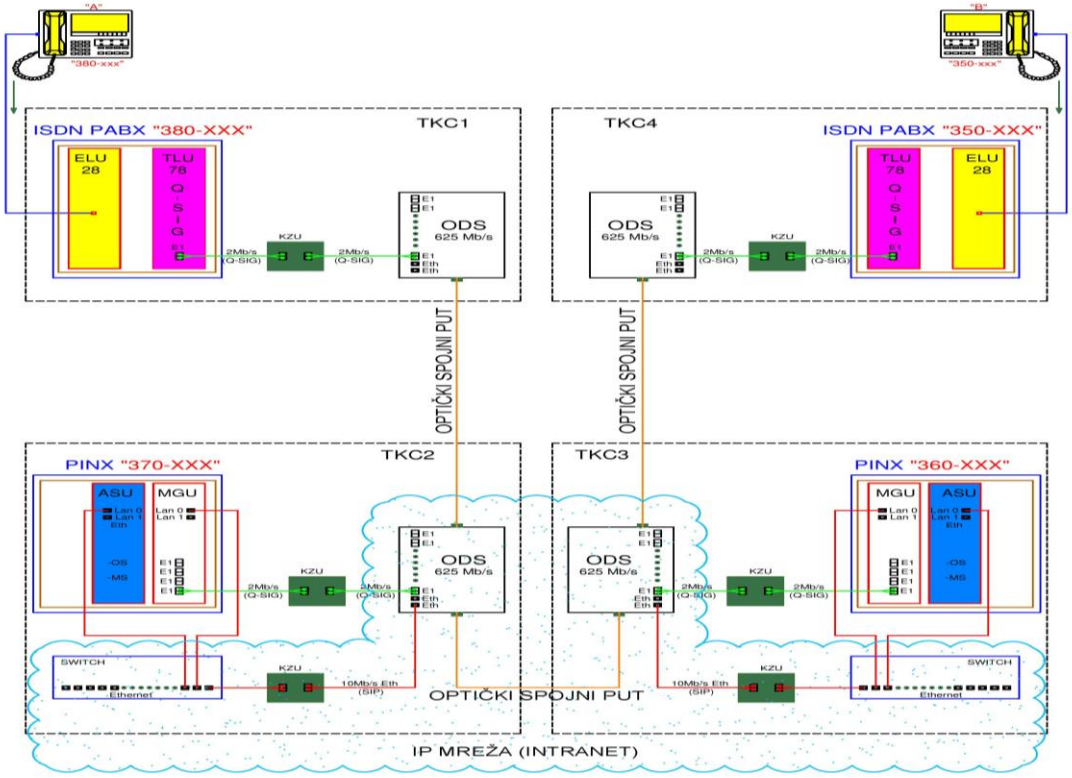
Tako izgrađena arhitekura dotičnog dela AtlMr FKo dala je hardversku i softversku mogućnost za realizaciju neophodne TCP konekcije i tunelovanje Q-SIG kroz IP proxy, između tranzitnih PINX (sa TKC2 i TKC3) u kojima su integrisani *MediaServeri* i *Media GW*-i, a u svrhu odvijanja mrežnog telefonskog saobraćaja kroz novoformirani SIP prenosnički snop od 30 trankova (vidi sliku 3).

Prvi korak, posle nabavki i instaliranja neophodnih softverskih licenci za SIP rutu i proračunati broj SIP trankova u njoj, manjih intervencija u Planu analize cifara i biranja-NADAP (*Number Analysis Dialing Plane*), te fizičkog ostvarenja svih potrebnih konekcija na testiranoj opremi, bio je da se na propisani način (na obe tranzitne PINX) definišu i iniciraju **nove SIP rute**. To je ostvareno preko postojećeg *Sistem Manager*-a i aplikativnog softverskog paketa PUTTY tako, što je softverski formiran i iniciran novi prenosnički snop, kapaciteta 30 SIP trankova (sa protokolima TCP za signalizaciju i UDP za media tokove), te na isti usmeren relevantni deo mrežnog telefonskog saobraćaja (odlazni pozivi iz lokalnog okruženja ili iz tranzita).

Za administriranje specifičnih podataka o **novoj ruti sa SIP trankovima** korištena je komanda "**sip_route**". Novi podaci korišćeni su, kao dopunski, uz već ranije korišćene podatke za

definisanje tradicionalnih ruta (ISDN, digitalnih, analognih...).

Kada se definiše **nova SIP ruta** i u okviru nje formira prenosnički snop potrebnog kapaciteta **SIP trankova**, prvo se koriste naredbe "**sip_route-set**", pa onda "**ROCAI**", "**RODAI**" i "**ROEQI**", te na kraju "**RODDI**". Način realizacije definisanja **nove rute** i podizanja **SIP trankova** u okviru nje, obavlja se u 5 osnovnih koraka i to:



Sl. 3. Realizacija tunelovanja Q-SIG/SIP IP na delu ATIMr funkcionalnog korisnika

1. sip_route -set [-**profile** <ime profila trunk-a] -**route I** -**uristring0** "**sip:? @** <SIPrekURI>", [ostali parametri sip_ route potrebni ili zahtevani profilom - Parametri koje uključuju -**Profile** -].

2. ROCAI: ROU = I, SEL= SIG = {D11 = **A** za SIP rutu}, ostali servisni parametri.

3. RODAI: ROU = I, TIPE = TL66, VARI = 00000000, VARC = 00000000, VARO = 00000000; Ako je **sip_route -profile** postavljen, onda VARI, VARC, VARO moraju imati nule. Ako nije tako, onda se ova konfiguracija postavlja iz profila navedenog kao -Profil parametri linijskog protokola-.

4. ROEQI: ROU = I, TRU = <lim> - <prvi redni broj> && <lim> - <poslednji redni broj>. TRU parametar definiše LIM-ove i kapacitet koji se koriste za SIP signaliziranje na ovoj ruti. (Napomena: Komanda "**sip_ rute**" mora biti izvršena pre "**ROEQI**" koja povezuje opremu sa rutom).

5. RODDI: ROU = I, DEST = <dest-number>. Definišu se pristupni, jednoznačni i dvoznačni kodovi, tzv. destinacije, koje se, prilikom odlaznih poziva, usmeravaju na ovu rutu.

Znači, uz administriranje podataka specifičnih za SIP trankove, tradicionalno se administriraju i "RO" podaci o ruti. U okviru naredbe "**sip_route-set**", na obe tranzitne PINX, definisane su i inicirane **nove rute sa SIP**, uz uvodenje sledećih parametara: **route**=25 (ruta je dobila svoj broj); **protocol** = udp (uveden je *Default*-ni UDP protokol); **profile** =*Default* (postavljen je difoltni profil, pošto se ruta uspostavlja između PINX istog tipa i proizvođača, a radi što bolje saradnje između njih); **service**=*PRIVATE* (parametar je uslovljen postojećom licencom za rad centrale u privatnoj mreži); **uristring0**=sip:?@192.168.X.X (definisane su IP adrese PINX na oba kraja veze komunikacije, a **uristring0** se koristi kod kreiranja poruke "*request URI*" u okviru SIP zahteva (recimo *INVITE*) po tipu: "nepoznati javni broj"); **remoteport**=*default* (radi se o "*Remote host port*-u" čiji se rang proteže od 0 do 65535; vrednost *default*=5060 definiše da je paketski protokol TCP/UDP); **accept**=ALL (Tip podudaranja koji treba da se izvrši prilikom upravljanja pozivima. Vrednost je definisana kao ALL= "prihvati sve pakete"); **priority**=255 (Prioritet prilikom uparivanja podataka dolaznih poziva rute je 0-255, najniži prioritet = 255. Obzirom da je parameter **accept**=ALL, to onda uslovljava da se prioritet mora postaviti na *Default*-nu vrednost 255); **register**=NO_REG (postavljena je takva vrednost, pošto relevantna PINX nema udaljene LIM-ove); **trusted**=*NO_TRUSTED* (Parametar je vezan sa parametrom **profile**=*Default*. SIP ruta prema pouzdanoj mreži, poverava odredište rute uz ograničenje strane porekla informacije.); **challenge**=no (Izazov za dolazni *INVITE* zahtev na ovom prenosničkom snopu. Definisano je "ne", pošto prenosnički snop na ovoj ruti ima prirodu *EMERGENCY*.);**supervise** =

NO_SUPERVISION (Definisano je da se ne vrši supervizija SIP poruka, t.j da opcije slanja paketa signalizacije ne čekaju na poruku "200ok", kao odgovor druge strane); **codecs**=G723,G729A,G729AB,PCMA,PCMU;**handleasexn** = no (Ova postavka se koristi da omogući korišćenje trank podataka za ekstenzije dolaznih poziva. Od broja polja pa na dalje treba da se podudara sa (pre) registrovanom ekstenzijom. Format je "da" ili "ne". Definisano je"ne".).

Testiranim rešenjem se kao prelazi, između delova PINS koje koriste ISDN/Q-SIG (krajnje PINX na TKC1 i TKC4), i IP mreže, koja radi po SIP (Intranet kao IVN), koriste GW-i (*Media GW*-i iz sastava tranzitnih PINX na TKC2 i TKC3).

Scenario realizovanog povezivanja krajnjih ISDN PINX, putem tunelovanja Q-SIG preko IP/SIP u privatnoj ATlMr funkcionalnog korisnika, prikazan je na slici 3. To je situacija, nastala na delu ATlMr FKo, koja se u potpunosti poklapa sa standardnom situacijom prikazanom na slici 1.

Na taj način, Q-SIG poziv koji je iniciran od "A" učesnika povezanog na PINX numeracije "380-xxx" sa TKC1, koja koristi Q-SIG na prenosničkom snopu prema tranzitnoj PINX sa TKC2, prolazi kroz tranzitnu PINX i njen *Media GW* (igra ulogu Ulaznog GW), koji po redu enkapsulira Q-SIG poruke u SIP (kroz TCP za signalizaciju i UDP za medija tokove), a zatim primenom tunelovanja Q-SIG prelazi preko IP mreže, koja koristi SIP, do druge tranzitne PINX sa TKC3, gde se na njenom *Media GW* vrši prevođenje SIP u Q-SIG (igra ulogu Izlaznog GW), te završava na učesniku "B", povezanom na krajnju PINX numeracije "350-xxx" sa TKC4, koja koristi ISDN/Q-SIG (na prenosničkom snopu za povezivanje sa tranzitnom PINX na TKC3).

Testiranim rešenjem je postignuto da, po pitanju prenosa osnovnih i dopunskih korisničkih usluga i svih mrežnih usluga (kao i posebnih *Mitel "proprietary"* usluga), ne postoje ograničenja kako u smislu kompatibilnosti Ulaznog i Izlaznog GW, tako i u smislu Q-SIG mogućnosti i ekvivalenata SIP IP mreže, te nema gubljenja delova Q-SIG na prikazanom "*end-to-end*" povezivanju.

Putem tunelovanja Q-SIG, IP mreža obezbeđuje inter-PINX TCP konekciju između dva *Media GW* iz sastava tranzitnih PINX, kada tunnel koji SIP obezbeđuje za prenos Q-SIG poruka deluje kao DQ-kanal za signalizaciju, dok uspostavljeni UDP media tokovi funkcionišu kao UQ-kanali za prenos korisničkih informacija (govora, modemskih informacija, podataka i sistemskih poruka).

Na uspostavljenom SIP prenosničkom snopu realizovana je i uspešna spoljna-vansistemska zaštita informacija, kako svih prenešenih signalizacionih kriterijuma iz okvira Q-SIG, tako i svih korisniških informacija koje se prenose kroz medija tokove. Za to su upotrebljeni uređaji za grupnu zaštitu IP/SIP paketskog prenosa, na nivou 10/100Mb/s *Base-T Etherneta*, koji su montirani sa obe strane prenosničkog SIP snopa, tj. na TKC2 i TKC3 (slika 3). Tako je postignuta neophodna neprekidnost zaštite informacija na celom spojnom putu od ISDN PABX "380-xxx" sa TKC1, preko tranzitnih PINX "370-xxx" i PINX "360-xxx" sa TKC2 i TKC3, do ISDN PABX "350-xxx" sa TKC4.

IV. ZAKLJUČAK

Realizovano rešenje za organizaciju telefonskog saobraćaja na delu ATlMr FKo povezivanjem dve tranzitne PINX putem organizacije **SIP prenosničkog snopa**, kroz **vlastitu IP mrežu** (*Intranet*), predstavlja novinu u organizaciji PTN razmatranog funkcionalnog korisnika. Korišćenje, od ranije primenjenog i u praksi dokazanog, sistema mrežne signalizacije tipa Q-SIG, putem njenog tunelovanja i prenosa bez degradacije kroz IP mrežu sa SIP protokolom, doprinelo je da se u ATlMr FKo ne naruši od ranije **izgrađeni status PINS**, i na delovima gde se njeni TDM/ISDN delovi povezuju sa IP proxy. Ovakvo rešenje je **uslovljeno** veličinom ATlMr i potrebnim finansijskim sredstvima, tj. nije bilo moguće realizovati jednovremenu tranziciju kompletne postojeće privatne telefonske mreže na bazi komutacije kola u mrežu sa komutacijom paketa, pri čemu su **zadržane** sve njene postojeće performanse.

LITERATURA

[1] InterConnect Communications Ltd (ICC) Merlin House, Station Road, Chpstow, Gwent NPG 5BP, United Kingdom, "Q-SIG, The Handbook for Communications Managers", August 1995 (ISBN 1 870935 09 8).

[2] Standard ECMA-133, "Private Integrated Services Network (PISN)-Reference Configuration for PISN Exchages (PINX)".

[3] Standard ECMA-142, "Private Integrated Services Network (PISN)-Circuit Mode 64 kbit/s Bearer Services-Service Description, Functional Capabilities".

[4] Standard ECMA-143, "Private Integrated Services Network (PISN)-Circuit mode bearer services- Inter –Exchange Signalling Procedures and Protocol".

[5] Standard ECMA-165, "Private Integrated Services Network (PISN)-Generic Funtcional Protocol for the Support of Supplementary Services - Inter –Exchange Signalling Procedures and Protocol".

[6] Standard ECMA-355, "Corporate Telecommunication Networks-Tunneling Q-SIG over IP", Ecma International, Rue du Rhone 114, CH-1204 Geneva, third Edition/June 2008.

[7] J. Rosenberg, H.Schulzrinne. el al, "SIP session initiation protocol," RFC 3261.

[8] J.Postel, "Internet Protocol", *IE RFC 791*.

[9] J.Rosenberg, H.Schulzrinne, et al., " An Offer/Answer Model with the Session Description Protocol (SDP)", *IE RFC 3264*.

[10] *Standard ECMA-339*, "Corporate Telecommunication Networks-Signalling interworking between Q-SIG and SIP- Basic services", Ecma International, Rue du Rhone 114, CH-1204 Geneva.

[11] *Standard ECMA-336*, "Private Integrated Services Network (PISN)-Mapping Functions for the Tunneling of Q-SIG trough IP Networks", Ecma International, Rue du Rhone 114, CH-1204 Geneva, June 2002.

[12] *IE TFC 761*, "Transmission Control Protocol".

[13] *IE TFC 768*, "User Datagram Protocol".

[14] *IE RFC 1889*, "RTP: a transport protocol for real-time applications".

[15] *Mitel Networks Corporation*, "MiVoice MX-ONE System Description", 21/1551-ASP 113 01 Uen AA, 28.08.2018.

[16] *Mitel Networks Corporation*, "MiVoice MX-ONE Media Server Description", 70/1551-ANF 901 14 Uen F, 05.09.2018.

[17] *Mitel Networks Corporation*, "MiVoice MX-ONE Media Gateway Unit MGU2 Description", 21/1551-ANF 901 36 Uen D, 26.04.2017.

Abstract

This paper explains a new way of connecting ISD and IP oriented PABX in a Private automatic telephone network (ATN) of integrated services for private users (PINX), by performing network Q Signalization (Q-SIG) tunneling through their private SIP oriented IP network (Intranet). It also shows as to how an ECMA-355 standard utilized Q-SIG tunneling through IP/SIP and describes a practical solution, realized within the ATN integrated services of the private integrated services network (PINS SM) for interconnecting users from differing end ISDN/Q-SIG PABX, through transit IP/Q-SIG PINX (the IP/SIP network performs the role of an Intervening network-IVN).

Q-SIG over SIP Tunneling in PISN with Integrated Services of Functional User

Slađan Svrzić, Zoran Čiča, Zoran Miličević and Zoran Perišić

Роминг у 802.11 мрежама и његова експериментална карактеризација

Данило Лазовић, Зоран Станковић, Јован Бајчетић,

***Апстракт*— Главни циљ овог рада је приказ основних принципа експерименталне карактеризације бежичних рачунарских мрежа. Развој метода и поступака мерења и анализе мерених резултата који се односе на праћење роминг догађаја клијента мреже и његовог утицаја на пакетски пренос података у IEEE 802.11 мрежном окружењу. У складу са тим, полазећи од неких постојећих решења за мерење handoff интервала, комбинацијом, надоградњом и прилагођењем реалним условима у мрежи, развијен је и примењен поступак мерења и анализе резултата који се односи на идентификацију, основну карактеризацију роминг догађаја и праћење његовог утицаја на перформансе пакетског преноса у реалном окружењу IEEE 802.11 мреже што је и главни допринос овог рада.**

***Кључне речи*— 802.11 мреже, роминг процес, handoff интервал, мерење пакетског протока мреже.**

I. УВОД

Поступци за мерење handoff интервала су до сада представљени у разним саопштењима – у лабораторијским условима [1, 2], као и у реалним условима [3]. У експерименталној поставци [1, 2] су коришћена два AP (Access Point) уређаја, станица за надгледање саобраћаја и референтна станица која је вршила роминг (прелажење из одговорности једног у одгорворност другог AP-а) између та два уређаја, док се код реалне архитектуре једне 802.11 мреже [3] користило више AP уређаја, станица за надгледање саобраћаја и референтна станица која је вршила роминг између AP уређаја. Мерени резултати који су добијени у тим радовима су били у складу са очекиваним резултатима који су доступни у литератури.

Полазећи од метода мерења handoff интервала изложених у [1-3], њиховом комбинацијом, надоградњом и прилагођењем у односу на расположиву опрему и услове и ограничења која су евидентирана у једном реалном роминг окружењу попут 802.11n мреже Електронског факултета у Нишу, развијен је поступак намењен експерименталној карактеризацији роминг догађаја у Wi-Fi мрежи са већим бројем AP уређаја.

Предложени поступак има следеће особине:

1. Поступак је развијен са намером да се омогуће мерења у реалним (не-лабораторијским) условима Wi-Fi мреже са већим бројем AP уређаја;
2. Поступак омогућава добијање информација о времену и месту настанка роминг догађаја, трајању handoff интервала и нивоима сигнала AP уређаја у току роминг догађаја, затим праћење на ком каналу у 2.4 GHz ISM опсегу клијент комуницира пре, током и после роминг догађаја, идентификацију области у сервисној зони мреже у којој се приликом задате трасе кретања дешавају роминг догађаји (роминг области), као и мерење параметара пакетског преноса података пре, за време и након роминг догађаја.
3. Предложене мерне процедуре могу да се релативно лако прилагоде професионалној опреми као што су специјализовани мерачи нивоа EM поља, Wi-Fi тестери, сензори за позиционирање унутар просторије и слични уређаји, са циљем да се добију резултати високе тачности који би били од великој значаја за ефикасну анализу рада Wi-Fi мреже.

Рад је реализован кроз шест поглавља. Након увода, у другом поглављу је представљен handoff интервал. Треће поглавље даје начин на који су се реализовала мерења, четврто поглавље представља поступак експерименталне карактеризације роминга у IEEE 802.11 мрежама. У петом поглављу су представљене особине саобраћаја које су мерене на „Траси 1“. Закључак је дат у шестом поглављу.

II. HANDOFF ИНТЕРВАЛ

Ради реализације процедуре роминга (handoff процес), мобилна станица размењује са приступном тачком (AP) уређене секвенце порука. Физички слој два AP уређаја омогућава повезивање овог логичког механизма и шаље информације о стању пареметара уређаја са једне инстанце на другу. Пренос на нивоу битова који обављају најмање три учесника – мобилна станица, претходни AP уређај и наредни AP уређај у потпуности се реализује коришћењем функција физичког слоја (Сл. 1). Претходни AP уређај је AP уређај са којим је мобилна станица имала везу пре извршења handoff-a, а AP уређај са којим је мобилна станица остварила комуникацију након handoff-a назива се наредни AP уређај.

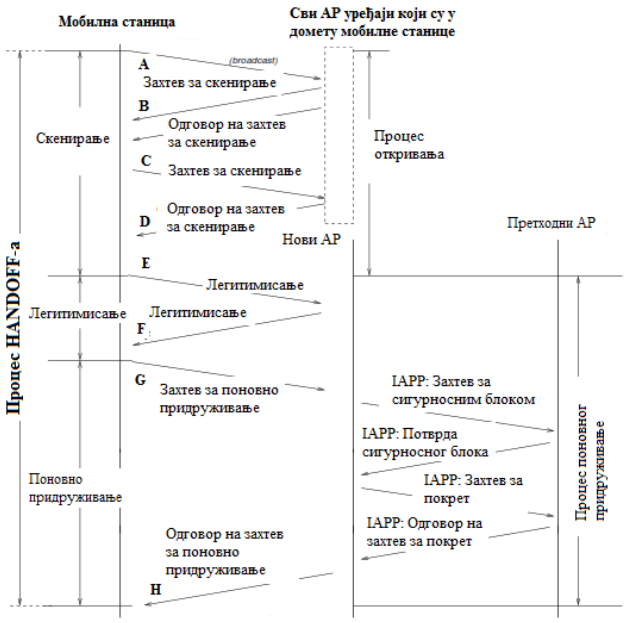
Процес handoff-a се спрводи у две фазе [3]**:**

1. Откривање AP уређаја у окружењу и
2. Поновно придруживање.

Данило Лазовић, Војна академија, Универзитет одбране у Београду, Павла Јуришића Штурма 33, 11000 Београд, Србија (e-mail: danilo.lazovic136@gmail.com).

Зоран Станковић, Електронски факултет, Универзитет у Нишу, Александра Медведева 14, 18000 Ниш, (e-mail: zoran.stankovic@elfak.ni.ac.rs).

Јован Бајчетић, Војна академија, Универзитет одбране у Београду, Павла Јуришића Штурма 33, 11000 Београд, Србија (e-mail: bajce05@gmail.com).



Сл. 1. Процес HANDOFF-а

Од самог произвођача Wi-Fi опреме зависе разлози због којих мобилна станица мења АР уређај на који је тренутно повезана. Ови разлози могу да буду изазвани различитим променама каракетристика преноса и стања пријемног сигнала које открива мобилна-пријемна станица. Најчешћи разлози за handoff одлуку су следећи [4]:

1. Низак ниво сигнала са АР уређаја на који је повезана мобилна станица (најчешћи ниво сигнала после којег мобилна станица мења АР уређај и који може да се у литератури нађе за поједине мрежне уређаје износи око -70 dBm);
2. Мобилна станица детектује веома лош однос сигнал-шум што се тиче сигнала који се прима са АР уређаја на који је повезана мобилна станица;
3. Велики број изгубљених рамова података у комуникацији мобилне станице са АР уређајем. Рамови се шаљу, али за њих не стиже потврда о успешном пријему путем АСК рама;
4. Мобилна станица изгуби конекцију са тренутним АР уређајем због неког техничког проблема.

III. Експериментална карактеризација роминга у IEEE 802.11n мрежи Електронског факултета у Нишу

Експериментална карактеризација роминга у IEEE 802.11n мрежи Електронског факултета у Нишу обухватила је мерење особина пакетског саобраћаја између мобилне станице која врши роминг и АР уређаја који учествују у роминг процесу. У оквиру ове активности вршена су мерења брзине преноса сегмената транспортног слоја мреже (протока TCP сегмената) током кретања мобилне станице, као и мерење броја изгубљених сегмената које шаље мобилна станица.

У оквиру локације дефинисана је једна траса кретања мобилног корисника са таквим простирањем које обезбеђује појаву бар једног роминга. У нивоу приземља факултета дефинисана је траса под називом “Траса 1 - приземље.

A. IEEE 802.11n мрежа Електронског факултета у Нишу

У згради Електронског факултета у Нишу где је постављена архитектура IEEE 802.11 мреже која омогућава извршење роминга извршена су сва мерења која се односе на експерименталну карактеризацију роминга у IEEE 802.11n мрежи. Мрежа која се налази у згради назива се ELFAK-INTERNET и састоји се од више АР уређаја који су распоређени на тачно одређеним локацијама у згради (приземље, М1 и по спратовима). Распоред АР уређаја дат је на Сл. 2. Четири АР уређаја се налазе у приземљу факултета у холу, још четири АР уређаја се налазе у три амфитеатра (Сл. 2). Уређаји произвођача Ubiquity Networks модел UniFi Access Point (2.4 GHz) се користе у 802.11 мрежи у приземљу, осим АР уређаја у амфитеатру А1. У амфитеатру А1 су инсталирана 2 АР уређаја истог произвођача, али напреднијег модела – UniFi Access Point - AC (2.4 GHz и 5 GHz).

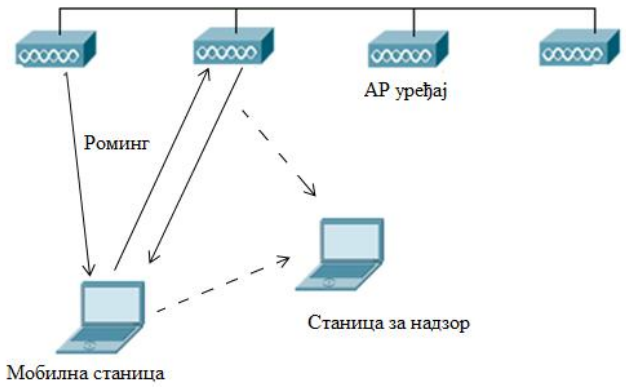
SSID који је идентичан називу целе мреже (ELFAK-INTERNET) имају сви АР уређаји. Приликом повезивања на факултетску мрежу клијент добија IP адресу коју задржава приликом вршења роминга, тако да се сви роминг процеси врше искључиво на слоју везе података (слој 2). Рамови из слоја везе података и информација о MAC (Media Access Control) адресама клијента и АР уређаја са којим клијент комуницира користили су се за идентификацију роминг догађаја приликом мерења која су се односила на експерименталну карактеризацију роминга.

Сл. 2. Распоред АР уређаја на приземљу факултета

IV. Мерни систем за експериментаљну карактеризацију роминга у IEEE 802.11n мрежи и поступак мерења

У циљу експерименталне карактеризације роминга у IEEE 802.11n мрежи Електронског факултета у Нишу користио се мерни систем који се састојао од два мобилна лаптоп рачунара који су у себи имали уграђене IEEE 802.11n мрежне адаптере и на којима су били инсталирани одговарајући софтвери за надгледање саобраћаја на слоју везе података и транспортном слоју (Сл. 3). Први рачунар је представљао клијента – мобилну

станицу, док је други коришћен за надгледање саобраћаја на мрежи и прикупљање података са исте (Станица за надзор, Сл. 3).



Сл. 3. Архитектура мерног система

A. *Мобилна станица*

Мобилна станица се након везивања на ELFAK-INTERNET мрежу кретала изабраним трасама, генерисала пакетски сабраћај ка Интернету и при томе улазила у процес роминга у одређеним областима трасе. Саобраћај мобилне станице се у току кретања надгледао и прикупљао од стране рачунара за надзор у циљу мерења handoff интервала и особина пакетског сабраћаја мреже приликом роминг догађаја. Као мобилна станица је коришћен лаптоп рачунар марке FUJITSU на којем је инсталиран Windows 8.1 Pro, 2,3 GHz са 6 GB RAM који у себи има интегрисану мрежну картицу.

Прикупљање и снимање размене пакета (пакетски саобраћај) између мобилне станице и AP уређаја је вршено уз помоћ програмског пакета WireShark, програмског пакета Acrylic WiFi Professional, и програмског пакета Matlab.

B. *Рачунар за надзор саобраћаја на мрежи*

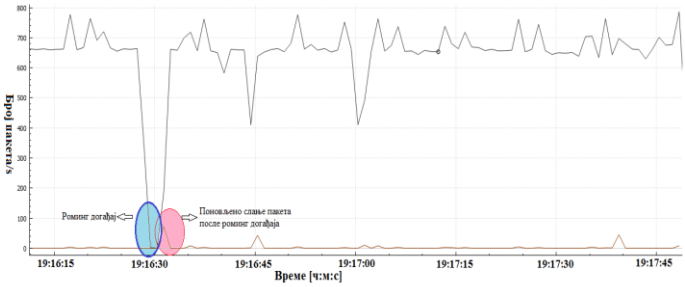
Основни задатак рачунара за надзор саобраћаја на мрежи (станица за надзор) је био да прикупи и сними целокупну размену рамова у слоју везе података између мобилне станице и AP уређаја на који је мобилна станица везана. Рачунар који је коришћен као мониторинг рачунар је био преносни рачунар марке Dell Vostro 1540 на којем је инсталиран Windows 7 Ultimate, 2,4 GHz и 4 GB RAM меморије који у себи има интегрисану мрежну картицу Broadcom 802.11n Network Adapter и Microsoft Virtual WiFi Miniport Adapter.

У свим сценаријима, рачунар за надзор је коришћен заједно са мобилном станицом која је вршила генерисање саобраћаја и снимање њеног пакетског саобраћаја према Интернету и при томе улазила у извршења роминг процеса. Овај рачунар се налазио непосредно поред мобилне станице и кретао се истом трасом и по истим условима као и мобилна станица. Апсолутно време на оба рачунара је постављено да буде идентично. На рачунару за надзор мреже је био активиран софтвер CommView for WiFi.

C. *Поступак мерења*

Поступак мерења на изабраној локацији у циљу експерименталне карактеризације роминга у IEEE 802.11n мрежи Електронског факултета у Нишу детаљно је описан у [5].

Анализом пакетског саобраћаја између мобилне станице и AP уређаја на транспортном нивоу који је снимљен приликом његовог кретања на траси одређује се зависност протока од временског положаја корисника на траси и уочава се повезаност временског тренутка роминг догађаја са временским интервалом у коме проток драстично пада. Типичан пример зависности броја пренетих пакета од временског положаја корисника на траси дат је на Сл. 4. На истој слици је браон бојом представљено поновно слање пакета (ретрансмисија) након што мобилна станица изврши процес роминга.



Сл. 4. Типичан пример зависности броја пренетих пакета и поново послатих пакета од временског положаја корисника на траси у условима интезивног саобраћаја

На Сл.4 се може уочити драстичан пад протока TCP сегмената непосредно пре него што мобилни клијент изврши роминг и нагло повећање броја пакета који су поново послати одмах након што мобилни клијент заврши роминг. Ово се објашњава тиме да је пре него што је клијент извршио роминг он дошао у просторну област (роминг област) где је сигнал AP уређаја на који је клијент везан веома слаб (ко што је већ претходно речено ниво сигнала AP уређаја пада испод -70 dBm). При тим вредностима сигнала долази до губитка рама у слоју везе података (рамови се шаљу, али се због лоше бежичне везе за њих не добијају ACK рамови потврде о њиховом успешном пријему). Пошто сваки изгубљени рам података носи у себи инкапсулиран IP пакет, а сваки пакет носи у себи TCP сегмент, тиме се проузрокују и губици у преносу TCP сегмената. Након што клијент изврши роминг и успостави стабилну конекцију са другим AP уређајем, поново се шаљу TCP сегменти који су изгубљени у преносу непосредно пре роминга.

V. Мерење особина саобраћаја током извршења роминга на траси "Траса 1 - Приземље"

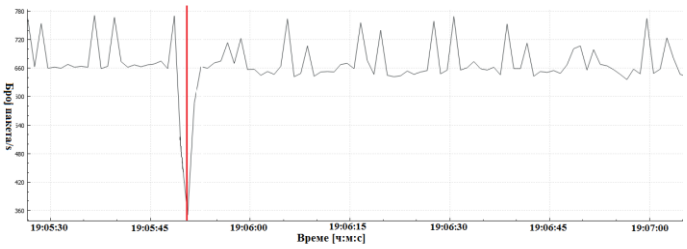
У првом мерењу особина мреже, мобилни корисник се кретао од тачке А ка тачки Б (Сл. 5). Са Сл. 6. се може уочити да је број пренетих пакета у јединици времена стандардан у току целе трасе кретања са појединим пиковима када је сам број пренетих пакета био висок.

Сл. 5. Деоница на којој је вршено мерење особина мреже на „Траси 1- Приземље“

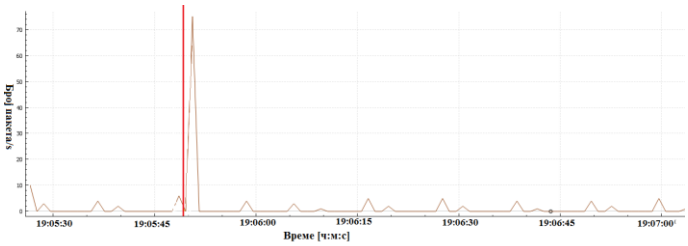
1	MNGT/PROBE REQ.	Mobilna stanica	Broadcast	19:05:50.310420
2	MNGT/PROBE REQ.	Mobilna stanica	Broadcast	19:05:50.569382
3	MNGT/PROBE REQ.	Mobilna stanica	Broadcast	19:05:50.597203
4	MNGT/PROBE REQ.	Mobilna stanica	Broadcast	19:05:50.598152
5	MNGT/PROBE RESP.	Ubiquiti:C4:51:40	Mobilna stanica	19:05:50.609363
6	MNGT/PROBE REQ.	Mobilna stanica	Broadcast	19:05:50.627328
7	MNGT/PROBE REQ.	Mobilna stanica	Broadcast	19:05:50.628274
8	MNGT/PROBE RESP.	F0:9F:C2:2A:0A:7C	Mobilna stanica	19:05:50.632701
9	MNGT/PROBE RESP.	Ubiquiti:C4:51:40	Mobilna stanica	19:05:50.643525
10	MNGT/AUTH	F0:9F:C2:2A:0A:7C	Mobilna stanica	19:05:50.717345
11	MNGT/REASS. RESP	F0:9F:C2:2A:0A:7C	Mobilna stanica	19:05:50.746168
12	MNGT/REASS. RESP	F0:9F:C2:2A:0A:7C	Mobilna stanica	19:05:50.748026
13	MNGT/ACTION	F0:9F:C2:2A:0A:7C	Mobilna stanica	19:05:50.750027
14	MNGT/ACTION	F0:9F:C2:2A:0A:7C	Mobilna stanica	19:05:50.751874
15	MNGT/ACTION	F0:9F:C2:2A:0A:7C	Mobilna stanica	19:05:50.751881

Сл. 6. Размена рамова између мобилне станице и АР уређаја који одговара роминг догађају T1_RD1, пролаз на траси AB1

На основу мерења реализованих у софтверском пакету CommView for WiFi који су приказани на Сл. 6. може се јасно видети да је у времену од 19:05:50,310420 до 19:05:50,748026 дошло до значајне деградације броја пренетих пакета. На ову деградацију броја пренетих пакета је утицало то што је мобилна станица ушла у процес извршења роминга, тј. мобилна станица је уместо сигнала са претходног АР уређаја почела да користи сигнал са другог АР уређаја (F0:9F:C2:2A:0A:7C). Тренутак у којем је дошло до преласка са једног на други АР уређај је обележен црвеном линијом, просечан број пренетих пакета у јединици времена је износио 685 пакета/s. (Сл. 7)



Сл. 7. Број пренетих TCP пакета у јединици времена приликом кретања на траси AB1, где је уочен роминг саидентификатором роминга T1_RD1 на „Траса 1- приземље“



Сл. 8. Број поново послатих пакета у јединици времена приликом проласка на траси AB1, где је уочен роминг са идентификатором роминга T1_RD1 на „Траса 1- приземље“

Током мерења карактеристика саобраћаја пажњу нам је привукао и податак који се односи на број поновно послатих рамова у току једног мерења од тачке А ка тачки Б и у супротном смеру (Сл. 8). Када се погледа број поновно послатих пакета, јасно се уочава чињеница да је овај број висок у тренуцима након што је укупан проток пакета саобраћаја јако низак. Другим речима, примећује се да у временском интервалу непосредно пре роминг догађаја укупан број пренетих пакета у јединици времена (проток пакета) непосредно нагло опао да би се одмах након роминга опоравио и вратио на просечну вредност од око 685 пакета/s. Са друге стране, одмах након роминга, број пакета који су поново послати у јединици времена нагло расте, што траје кратак временски период, а онда се поново враћа на малу вредност након роминга. Ово се објашњава губитком рамова података, а самим тим и TCP сегмената непосредно пре роминга (због ниског нивоа сигнала базног АР уређаја) и за време роминга (због прекида слања рама података у handoff интервалу).

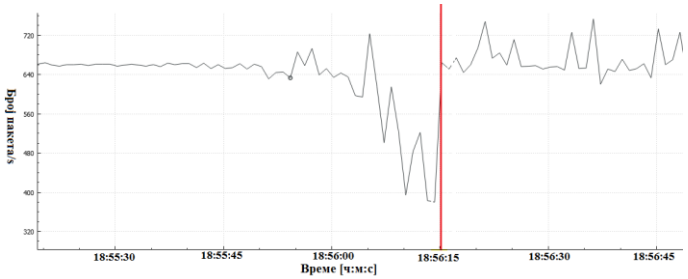
У другом мерењу особина мреже кретали смо се од тачке Б ка тачки А (Сл. 5). Са Сл. 10 се може уочити да је број пренетих пакета константан из разлога што се тачка Б налази у амфитеатру А1, па је и сам ниво сигнала са АР врло висок. Током реализације другог мерења такође смо се кретали брзином од 3 km/h.

4912	CTRL/RTS	Mobilna stanica	26:A4:3C:BB:D8:D3	18:56:07.327457
4913	CTRL/ACK	N/A	Mobilna stanica	18:56:07.570501
4914	CTRL/RTS	Mobilna stanica	26:A4:3C:BB:D8:D3	18:56:15.125848
4915	CTRL/RTS	Mobilna stanica	26:A4:3C:BB:D8:D3	18:56:15.126912
4916	MNGT/PROBE REQ.	Mobilna stanica	Broadcast	18:56:16.072677
4917	MNGT/PROBE REQ.	Mobilna stanica	Broadcast	18:56:16.075857
4918	MNGT/PROBE RESP.	F0:9F:C2:2A:0A:7C	Mobilna stanica	18:56:16.079020
4919	MNGT/PROBE RESP.	F0:9F:C2:2A:09:EA	Mobilna stanica	18:56:16.083836
4920	MNGT/PROBE REQ.	Mobilna stanica	Broadcast	18:56:16.102823
4921	MNGT/PROBE REQ.	Mobilna stanica	Broadcast	18:56:16.103785
4922	MNGT/PROBE RESP.	F0:9F:C2:2A:09:EA	Mobilna stanica	18:56:16.108121
4923	MNGT/PROBE RESP.	F0:9F:C2:2A:09:2A	Mobilna stanica	18:56:16.115138
4924	MNGT/PROBE RESP.	F0:9F:C2:2A:09:EA	Mobilna stanica	18:56:16.117321
4925	MNGT/PROBE RESP.	F0:9F:C2:2A:09:2A	Mobilna stanica	18:56:16.119518
4926	MNGT/AUTH	F0:9F:C2:2A:09:2A	Mobilna stanica	18:56:16.158217
4927	MNGT/ACTION	F0:9F:C2:2A:09:2A	Mobilna stanica	18:56:16.161467
4928	MNGT/REASS. RESP	F0:9F:C2:2A:09:2A	Mobilna stanica	18:56:16.163529
4929	CTRL/RTS	F0:9F:C2:2A:09:2A	Mobilna stanica	18:56:16.472967
4930	CTRL/RTS	F0:9F:C2:2A:09:2A	Mobilna stanica	18:56:16.478289
4931	CTRL/RTS	F0:9F:C2:2A:09:2A	Mobilna stanica	18:56:16.478826

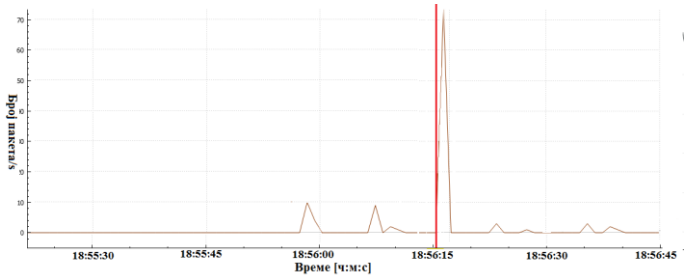
Сл. 9. Размена рамова између мобилне станице и АР уређаја који одговара роминг догађају T1_RD4, пролаз на траси BA2

Како се прилази тренутку роминга, број пренешених пакета је почиње да опада. На основу мерења реализованих у софтверском пакету CommView for WiFi који су приказани на Сл. 9 може се јасно видети да је у времену од 18:56:16,072077 до 18:56:16,163529 дошло до тога да је мобилна станица ушла у процес извршења роминга, тј. мобилна станица је уместо сигнала са АР уређаја (26:A4:3C:BB:D8:D3) почела да користи сигнал са

другог АР уређаја (F0:9F:C2:2A:09:2A). Временски тренутак у којем је дошло до преласка са једног на други АР уређај је обележен црвеном линијом, просечан број пренетих пакета у јединици времена (проток) у току мерења у условима интезивног саобраћаја је износио 615 пакета/s. (Сл. 10).



Сл. 10. Број пренетих TCP пакета у јединици времена приликом кретања на траси BA2, где је уочен роминг са идентификатором роминга T1_RD4 на „Траса 1- приземље“



Сл. 11. Број поново послатих пакета у јединици времена приликом проласка на траси BA2, где је уочен роминг са идентификатором роминга T1_RD4 на „Траса 1- приземље“

Број поново послатих пакета у току мерења од тачке Б ка тачки А је значајан у тренутку након извршења роминга (Сл. 11). У складу са пртходном дискусијом, овакав податак је очекиван због тога што у периоду непосредно пре одлуке о вршењу роминга мобилна станица због лошег сигнала почиње да губи рамове података (нема потврде за послате рамове), а самим тим и TCP сегменте. Када мобилна станица успостави сигурну комуникацију са новим АР уређајем и изврши се оговарајућа размена порука између претходног и новог АР уређаја, тада се тек може поновити слање пакета који су изгубљени.

VI. ЗАКЉУЧАК

Мерењем карактеристика пакетског саобраћаја у транспортном слоју уочен је драстичан пад протока TCP сегмената непосредно пре него што мобилна станица изврши роминг и нагло повећање броја пакета који су поново послати одмах након што мобилна станица заврши роминг процедуру. Ово се објашњава тиме да је пре него што је мобилна станица извршила роминг, мобилна станица дошла у просторну област (роминг област) где је сигнал АР уређаја на који је клијент везан веома слаб (ко што је већ претходно речено ниво сигнала АР уређаја пада испод -70 dBm). При тим вредностима сигнала долази до губитка рамова у слоју везе података. Пошто сваки изгубљени рам података носи у себи енкапсулиран IP пакет, а сваки пакет носи у себи TCP сегмент, тиме се проузрокују и губици у преносу TCP пакета. Након што мобилна станица изврши роминг и успостави стабилну конекцију са другим АР уређајем, поново се шаљу TCP пакети који су изгубљени у преносу непосредно пре роминга. Ова директна повезаност драстичног пада протока TCP пакета непосредно пре извршења роминга и ретрансмисије пакета одмах након завршетка роминга указује на то да праћење процента губитка рама података у слоју везе података може бити један од битних елемената алгоритма о доношењу одлуке о извршењу роминга. Посебан изазов у наредним истраживањима биће анализа саобраћаја у Wi-Fi мрежама са напредним управљањем роминг процесима где се разменом података и дирекном координацијом АР уређаја и клијента, АР уређаји неспоредно укључују у доношењу одлуке о оптималном тренутку извршења роминга.

LITERATURA

[1] Martinovic, Ivan & A Zdarsky, Frank & Bachorek, Adam & Schmitt, Jens. (2009). *Measurement and Analysis of Handover Latencies in IEEE 802.11 i Secured Networks*.

[2] S. Pack, J. Choi, T. Kwon, Y. Choi: *„Fast Handoff Support in IEEE 802.11 WirelessNetworks“*, IEEE communications surveys and tutorials, the first quarter, 2007.

[3] Mishra, Arunesh & Shin, Minho & Arbaugh, William. (2003). *An empirical analysis of the IEEE 802.11 MAC layer handoff process*. Computer Communication Review. 33. 10.1145/956981.956990.

[4] https://blogs.arubanetworks.com/industries/client-roaming-triggers/, приступано април 2019.

[5] Д. Лазовић, З. Станковић, Ј. Бајчетић, „Роминг у IEEE 802.11 мрежама“, ETRAN 2018

ABSTRACT

The main goal of this paper is to present the basic principles of the experimental characterization of wireless computer networks, as well as the development of methods and procedures for measuring and analyzing measured results related to the monitoring of client roaming events and their impact on packet data transmission in the IEEE 802.11 network environment. Accordingly, starting from some existing solutions for measuring the handoff interval, combining, upgrading and adapting to the real conditions in the network, the process of measuring and analyzing results related to identification, basic roaming event characterization and monitoring of its impact on the network performance was developed and applied in the real environment of the IEEE 802.11 network, which might be considered as the main contribution of this experiment.

Roaming in 802.11 networks and its experimental characterization

Danilo Lazović, Zoran Stanković, Jovan Bajčetić

Analiza uticaja arhitekture mreže na kvalitet signala u okviru LTE tehnologije

Ivana Stojanović, Mladen Koprivica, *Senior Member, IEEE*, Nenad Stojanović, Aleksandar Nešković, *Senior Member, IEEE*

***Apstrakt*—U radu je analiziran uticaj arhitekture mreže na kvalitet signala u okviru četvrte generacije javne mobilne mreže. Analiza je izvršena pomoću parametara RSRP *(Reference Signal Received Power)*, RSRQ (*Reference Signal Received Quality*) i realno ostvarivog protoka podataka. Parametri kvaliteta signala su prikupljeni merenjem pomoću TEMS *Investigation* i TEMS *Pocket* softvera. Merenja su sprovedena na Elektrotehničkom fakultetu u prizemlju zgrade Tehničkih fakulteta za scenario makro i mikro ćelije. Ustanovljeno je da se bolji kvalitet signala obezbeđuje u mikro ćelijama. Kvalitet signala je razmatran i po različitim servisima koji se obezbeđuju korisniku.**

***Ključne reči*—LTE, makro ćelija, mikro ćelija, kvalitet servisa.**

I. Uvod

Mobilne komunikacije su postale deo svakodnevnice. U poslednje dve decenije su evoluirale od skupe tehnologije koju su sebi mogli da priušte samo pojedinci, do danas kada su postale sveprisutni sistemi koje koristi većina svetske populacije. Tehnologija LTE (*Long Term Evolution*), predstavlja evoluciju postojećih sistema i nudi nekoliko prednosti za korisnike i operatere, u poređenju sa tehnologijama prethodnih generacija. Ova tehnologija dovela je do poboljšanja performansi i kapaciteta sistema, bolje iskorišćenosti radio resursa i smanjenja potrošnje energije.

Operateri planiraju radio mrežu raspoređivanjem baznih stanica. Nakon toga se najčešće vrši optimizacija mreže, za šta je neophodno izvršiti testiranje karakteristika na aktivnoj mobilnoj mreži. Iz tog razloga proizvođači telekomunikacione merne opreme imaju zadatak da nađu najjednostavniji način za merenje parametara potrebnih za analizu.

U ovom radu je izvršena eksperimentalna analiza performansi LTE tehnologije na aktivnoj mobilnoj mreži u prizemlju zgrade Elektrotehničkog fakulteta. Testirani su sledeći servisi: *web browsing* statičke stanice, *ping* veličine paketa 32 B, *video streaming*, *ping* veličine paketa 800 B, *file download* i *file upload*. Ovi servisi predstavljaju tipične servise LTE tehnologije. Takođe, posmatrani servisi su najčešće korišćeni servisi od strane studenata koji provode vreme u zgradi Elektrotehničkog fakulteta. Parametri koji su razmatrani su: *Reference Signal Received Power* (RSRP), *Reference Signal Received Quality* (RSRQ) i realno ostvariv protok prenosa podataka (*Throughput*). Prilikom merenja razmatrana su dva scenarija, kada se analizirani servis izvršavao preko makro i mikro ćelije. Makro scenario predstavlja slučaj kada mobilni terminal opslužuje makro ćelija bazne stanice, a mikro scenario kada to čini mikro ćelija bazne stanice. Izvršena je uporedna analiza rezultata dobijenih prilikom makro i mikro scenarija.

Slična istraživanja vršena su i u okviru UMTS (*Universal Mobile Telecommunication Systems*) tehnologije (treća generacija javne mobilne mreže). U [1] je pokazano da je uvođenjem mikro ćelija na odgovarajućim rastojanjima od makro ćelija, moguće poboljšanje funkcionisanja mreže. Poboljšanje mreže je prikazano kroz ostvaren protok prenosa podataka, odnos signal-šum (*Signal to Noise Ratio*, SNR) i nivo prijemnog signala kao parametara kvaliteta signala u UMTS mreži.

Ostatak rada organizovan je na sledeći način. U drugom poglavlju se govori o LTE tehnologiji, njenom nastanku i karakteristikama. Treće poglavlje sadrži opis korišćenog mernog postupka. U četvrtom poglavlju predstavljeni su rezultati prikupljenih podataka i uporedna analiza dobijenih rezultata. Konačno, u petom poglavlju dati su zaključci sa pravcima daljeg istraživanja.

II. Osnove LTE tehnologije i njenih parametara

A. Osnove LTE tehnologije

LTE je označena kao radio tehnologija četvrte generacije. Pojava LTE tehnologije je dovela do nekoliko tehnoloških rešenja koja su do tada bila nepoznata. Ciljevi prelaska na LTE mrežnu tehnologiju su: obezbeđenje većeg protoka podataka korisnicima, poboljšanje spektralne efikasnosti, realizacija znatno efikasnije paketske komutacije, poboljšanje i povećanje broja servisa i njihova implementacija, prevođenje mobilne mreže na isključivo paketsku mrežu i bolja integracija sa postojećim standardima prenosa i obrade signala [2].

Na *downlink* strani komunikacije LTE koristi OFDMA (*Orthogonal Frequency Division Multiple Access*) sistem višestrukog pristupa, što za rezultat ima robusniji sistem sa povećanim kapacitetom. Povećanje kapaciteta telekomunikacionog kanala se postiže multipleksiranjem

Ivana Stojanović – Telekom Srbija a.d., Bulevar umetnosti 16a, 11000 Beograd, Srbija i Elektrotehnički fakultet, Univerzitet u Beogradu, Bulevar kralja Aleksandra 73, 11120 Beograd, Srbija (e-mail: ivnvukanic@gmail.com).

Mladen Koprivica – Elektrotehnički fakultet, Univerzitet u Beogradu, Bulevar kralja Aleksandra 73, 11120 Beograd, Srbija (e-mail: kopra@etf.rs).

Nenad Stojanović – Vojna akademija, Univerzitet odbrane u Beogradu, Generala Pavla Jurišića Šturma 33, 11000 Beograd, Srbija (e-mail: nivzvk@hotmail.com).

Aleksandar Nešković – Elektrotehnički fakultet, Univerzitet u Beogradu, Bulevar kralja Aleksandra 73, 11120 Beograd, Srbija (e-mail: neshko@etf.rs).

korisničkih podataka malog protoka na širem kanalu, dok se robusnost postiže raspoređivanjem korisničkog saobraćaja po učestanostima kako bi se izbegla uskopojasna interferencija i feding usled višestruke propagacije. *Uplink* strana komunikacije koristi SC-FDMA (*Single Carrier-Frequency Division Multiple Access*) sistem višestrukog pristupa, čiji je najznačajniji faktor energetska efikasnost, kako bi se povećala pokrivenost i smanjili troškovi terminala i potrošnja energije. SC-FDMA ima nizak odnos vršne i srednje snage signala (*Peak to Average Power Ratio*, PAPR) što predstavlja glavni razlog korišćenja ove tehnike na *uplink* strani komunikacije [3].

Ovakav pristup kada se koriste dva različita sistema na *downlink* i *uplink* strani obezbeđuje ortogonalnost između korisnika, smanjenje interferencije i poboljšanje kapaciteta mreže [3]. Propusni opseg se može izabrati u intervalu od 1.4 MHz do 20 MHz u zavisnosti od raspoloživog spektra. Propusni opseg od 20 MHz obezbeđuje do 150 Mbps protoka na *downlink* strani komunikacije kada je u upotrebi 2x2 MIMO (*Multiple Input Multiple Output*) i 300 Mbps kada je u upotrebi 4x4 MIMO sistem. Protok na *uplink* strani može dostići i do 75 Mbps.

B. Opis korišćenih parametara LTE tehnologije

Radi procene kvaliteta performansi jednog digitalnog telekomunikacionog sistema najčešće se koristi verovatnoća bitske greške (*Bit Error Rate*, BER). U konkretnim telekomunikacionim sistemima koriste se i neki drugi načini procene kvaliteta signala kako bi se korisnicima obezbedila što bolja usluga. RSRP i RSRQ su pokazatelji kvaliteta signala koji se često koriste u optimizacji javne mobilne mreže LTE tehnologije.

RSRP predstavlja nivo snage signala na prijemu, ali ne govori o njegovom kvalitetu, već predstavlja indikator pokrivenosti ćelije [4]. Vrednost RSRP se izražava u dBm i služi kao metrika za odluku o reselekciji i *handover*–u.

RSRQ predstavlja parametar koji ukazuje na kvalitet primljenog signala [4]. Najpre se izmeri ukupni signal (i interferencija i šum) koji se prima u jednom OFDM (*Orthogonal Frequency Division Multiplex*) simbolu, čime se dobija vrednost RSSI (*Received Strength Signal Indicator*). Uporedo se meri i ukupna vrednost snage signala RSRP. Odnos ova dva parametra pomnožen sa brojem resurnih blokova daje konačnu vrednost RSRQ što se može predstaviti izrazom:

$$RSRQ = \frac{RSRP}{RSSI} N_{RB}, \qquad (1)$$

gde N_{RB} predstavlja broj resursnih blokova koji zavisi od propusnog opsega koji se koristi u LTE tehnologiji.

Ukoliko se korisnik nalazi u ruralnom području, posmatra se samo nivo RSRP zbog malog broja baznih stanica koje pokrivaju tu oblast. Međutim, ukoliko se korisnik nalazi u urbanom području sa zadovoljavajućim nivoom RSRP, ali se javlja velika interferencija zbog postojanja više baznih stanica u toj oblasti, na osnovu nivoa RSRQ se donosi zaključak kada je potrebno da mobilni terminal opslužuje druga bazna stanica (*handover*), a kada mobilni terminal može ostati na onoj koja ga trenutno opslužuje [5].

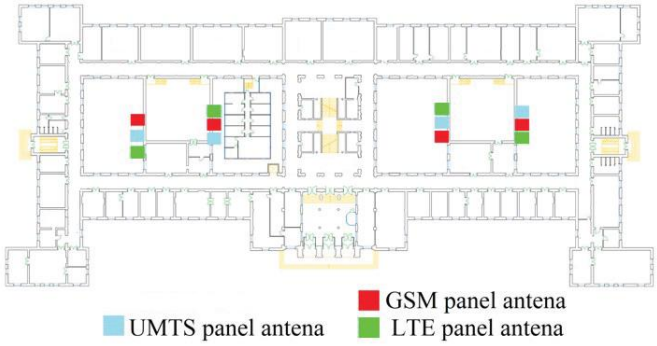
Binarni protok (*Throughput*), predstavlja brzinu prenosa bita u jedinici vremena u digitalnom telekomunikacionom sistemu. Realno ostvariv protok na liniji veze direktno utiče na performanse javne mobilne mreže, a posebno na kvalitet servisa baziranih na komutaciji paketa [6].

III. OPIS MERNOG POSTUPKA

Za potrebe merenja korišćen je TEMS *Investigation* softver i mobilni uređaj *Sony Xperia Z3 D6603* koji služi za prikupljanje podataka. TEMS *Investigation* je softverski alat za optimizaciju, verifikaciju, rešavanje problema i održavanje javne mobilne mreže. On prikuplja i obrađuje podatke, a zatim daje analizu u realnom vremenu. Tehnnologije koje podržava su: LTE (FDD (*Frequency Division Duplex*), TDD (*Time Division Duplex*)), WCDMA (*Wideband Code Division Multiple Access*) / HSPA (*High Speed Packet Access*) / HSPA+, GSM (*Global System for Mobile Communication*) / GPRS (*General Packet Radio Servis*), Wi-Fi (*Wireless-Fidelity*)…

Postupak merenja je obavljen na Elektrotehničkom fakultetu u prizemlju zgrade Tehničkih fakulteta Univerziteta u Beogradu za scenario makro i mikro ćelija. Merenja su vršena tokom radnog dana zbog najveće opterećenosti mreže u tom periodu. Za testiranje servisa LTE tehnologije korišćena je mobilna mreža Telekoma Srbije. Merenja su vršena u *indoor* okruženju, sprovedena je analiza za slučaj kada mobilni terminal opslužuje makro ćelija bazne stanice i za slučaj kada to čini mikro ćelija bazne stanice.

Tokom merenja član mernog tima se kretao prizemljem noseći sa sobom prenosivi računar i mobilni terminal na kojima su instalirani TEMS *Investigation* i TEMS *Pocket*, respektivno. Na računaru je u okviru korišćenog softvera pokretan skript koji se sastojao od niza servisa od interesa. Član mernog tima je jednom prošao kroz prizemlje i tom prilikom je celokupan skript ponovljen tri puta.



Sl. 1. Mapa prizemlja Elektrotehničkog fakulteta sa rasporedom antena baznih stanica GSM, UMTS i LTE tehnologija.

Položaj baznih stanica i izgled prizemlja u kome je vršeno merenje prikazan je na Sl. 1. Crvenom bojom su predstavljene pozicije GSM (*Global System for Mobile Communication*)

panel antena, plavom bojom UMTS (*Universal Mobile Telecommunication System*) pozicije panel antena i zelenom bojom su predstavljene pozicije LTE panel antena.

Merna procedura sastojala se u testiranju različitih servisa za scenario makro i mikro ćelije. Prikupljeni podaci su unošeni u *Matlab* i *Microsoft Office Excel* softverske alate pomoću kojih je potom vršena obrada merenih rezultata.

IV. REZULTATI

Na osnovu sprovedenih merenja grafički su prikazani dobijeni rezultati računanjem kumulativne funkcije raspodele verovatnoće (*Cumulative Distribution Function*, CDF) [7].

U Tabeli I [3] su prikazane referentne vrednosti opsega nivoa RSRP i RSRQ parametara LTE tehnologije, na osnovu kojih se utvrđuje stepen kvaliteta signala, odnosno kvalitet pružene usluge korisniku.

Na Sl. 2 predstavljen je nivo RSRP za makro i mikro scenario. Vrednosti nivoa RSRP se kreću u opsegu od -120 dBm do -95 dBm za scenario makro ćelije, tako da signal ulazi u tri opsega prikazana u Tabeli I. Oko 25% izmerenih vrednosti signala je nezadovoljavajućeg kvaliteta, 5% je signal dobrog kvaliteta, dok preostalih 70% čini signal zadovoljavajućeg kvaliteta. Vrednosti nivoa RSRP za scenario mikro ćelije se kreću u opsegu od -90 dBm do -55 dBm, što predstavlja odličan nivo signala na osnovu pokazatelja iz Tabele I. Dolazi se do zaključka da je značajno bolji nivo signala u slučaju mikro ćelije, što je i očekivano, jer se mikro bazna stanica nalazi na zidu zgrade Elektrotehničkog fakulteta, pa pruža bolju pokrivenost signalom.

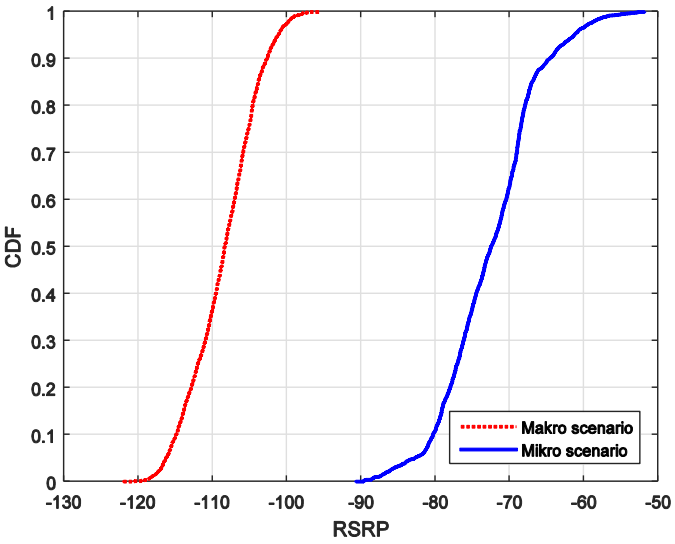
Tabela I
Vrednosti nivoa signala u odnosu na kvalitet

	Parametri	RSRP (dBm)	RSRQ (dB)
Kvalitet signala	Odličan	> -84	> -5
	Dobar	-102 do -85	-11 do -6
	Zadovoljavajući	-111 do -103	
	Nezadovoljavajući	< -112	< -12

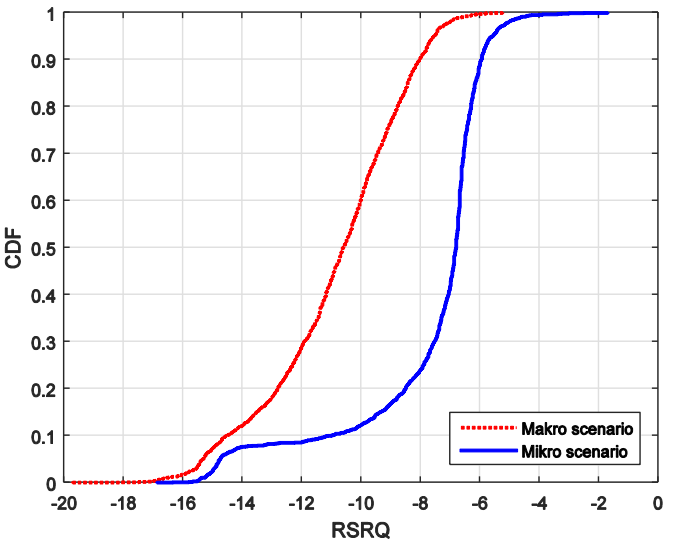
Za nivo RSRQ se može reći da je dobar ako se nalazi u opsegu od -11 dB do -6 dB. Za scenario makro ćelije nivo RSRQ se kreće u opsegu od -20 dB do -5 dB, dok se za mikro scenario nivo RSRQ kreće u opsegu od -16 dB do -2 dB, što je ilustrovano na Sl. 3. Vidi se da je nivo RSRQ bolji u mikro scenariju što je i očekivano zbog blizine mikro bazne stanice korisniku, ali i da u oba scenarija signal obuhvata sva tri nivoa kvaliteta signala predstavljena u Tabeli I.

Poređenje protoka za dva scenarija, kada se analizirano merenje izvršavalo preko makro i mikro ćelije, izvršeno je pomoću Sl. 4 i Sl. 5, gde su prikazani protoci u *downlink* i *uplink* smeru komunikacije. Na Sl. 4 predstavljene su normalizovane vrednosti protoka na *downlink* smeru komunikacije, za servis *download* u trajanju od 10 s. Vrednosti su normalizovane tako što je svaka od njih podeljena sa najvećom izmerenom vrednošću za servis *download* u trajanju od 10 s u okviru mikro ćelije. Viši protoci se ostvaruju u slučaju mikro scenarija i to više od dvostruke vrednosti brzine prenosa podataka u odnosu na makro scenario.

U slučaju brzine prenosa podataka na *uplink* smeru komunikacije za servis *upload* podataka u trajanju od 10 s takođe se viši protoci podataka dostižu u okviru mikro scenarija. Razlika u brzini prenosa podataka između makro i mikro scenarija je nešto manja nego u slučaju *downlink* smera komunikacije. I u ovom slučaju su vrednosti protoka podataka normalizovani maksimalnom vrednošću protoka kod servisa *upload* u intervalu od 10 s, na *uplink* smeru komunikacije. Navedeni protoci podataka na *upload* smeru komunikacije prikazani su na Sl. 5.



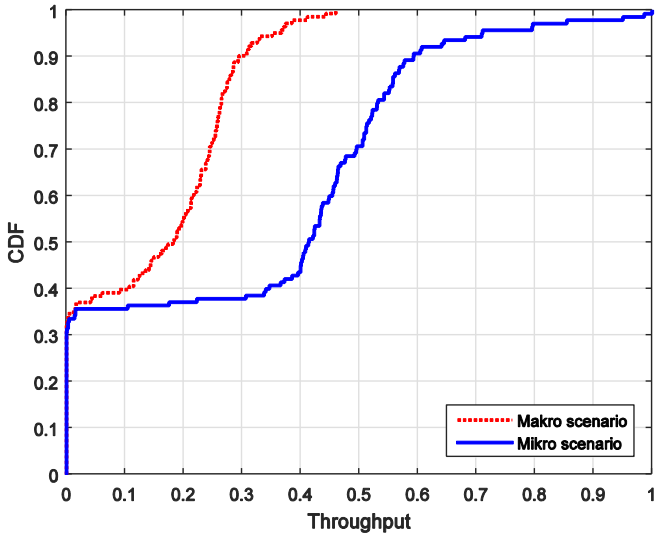
Sl. 2. Poređenje vrednosti RSRP za scenario makro i mikro ćelije.



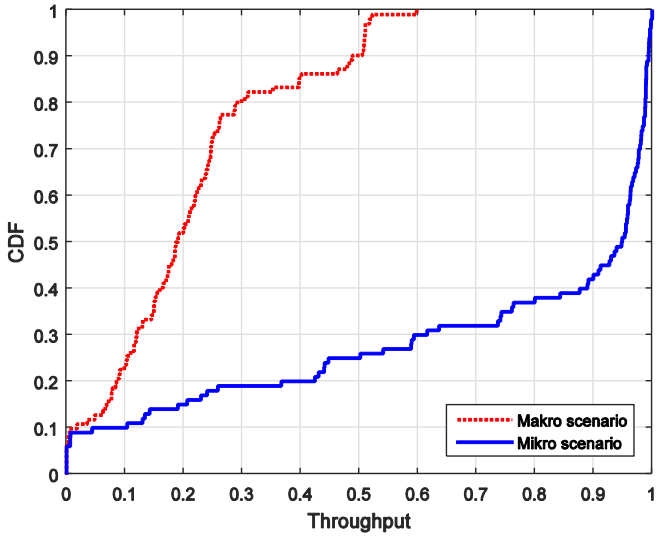
Sl. 3. Poređenje vrednosti RSRQ za scenario makro i mikro ćelije.

Na Sl. 6 i Sl. 7 prikazan je kvalitet signala u vidu RSRP i RSRQ u odnosu na najčešće servise koji se obezbeđuju

korisnicima. Prikazane vrednosti dobijene su tako što je najpre vršeno usrednjavanje nivoa signala nakon svakog od tri ponavljanja skripta tokom merenja, a zatim su usrednjene i tri novodobijene vrednosti. Za prikaz i analizu izabrani su *download* u trajanju od 10 s, *upload* u trajanju od 10 s, *download* 3 MB, *upload* 1 MB, učitavanje *web* stranice *youtube* i pregled videa sa *web* stranice *youtube*. Navedeni servisi su izabrani, jer se najčešće koriste od svih servisa koji su mereni tokom istraživanja. S obzirom na visoke protoke prenosa podataka koje nam omogućava LTE tehnologija, vreme potrebno za *download* 3 MB podataka je značajno kraće u odnosu na samo testiranje servisa *download* u trajanju od 10 s. Isto važi i za *upload* servis.



Sl. 4. Protok za servis *download* u trajanju od 10 sekundi na *downlink* smeru komunikacije za scenario makro i mikro ćelije.

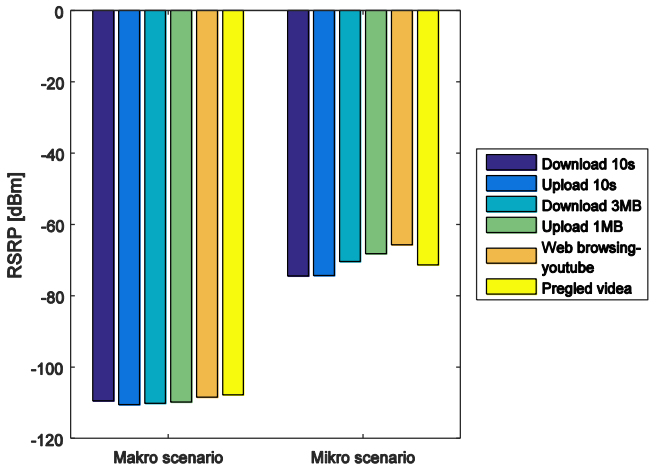


Sl. 5. Protok za servis *upload* u trajanju od 10 sekundi na *uplink* strani komunikacije za scenario makro i mikro ćelije.

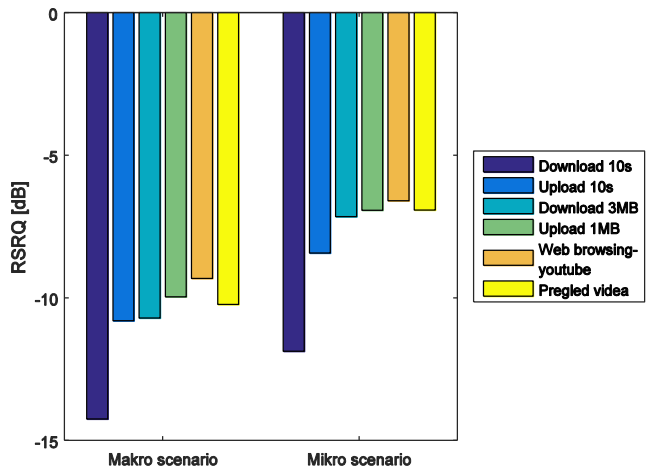
Vrednosti RSRP za makro scenario pokazuju da je za sve servise kvalitet signala na granici zadovoljavajućeg što značajno utiče na obavljanje komunikacije. Takođe, parametar RSRQ kod makro scenarija, za većinu testiranih servisa ima vrednosti oko granice između zadovoljavajućeg i nezadovoljavajućeg. Lošiji kvalitet signala javlja se prilikom *download* smera komunikacije u intervalu od 10 s, dok je prilikom učitavanja stranice *youtube* signal bio nešto boljeg kvaliteta u odnosu na signale kod ostalih servisa.

Kod mikro scenarija su dobijeni drugačiji rezultati. Posmatranjem parametra RSRP uočen je veoma visok stepen kvaliteta signala za sve testirane servise. Sve vrednosti RSRP su pokazale da se ostvaruje odličan nivo signala kada bazna stanica mikro ćelije opslužuje mobilni terminal na bliskom rastojanju. Parametar RSRQ i u ovom slučaju daje nešto drugačije rezultate. Servis *download* podataka u trajanju od 10 s i u ovom slučaju ima nezadovoljavajući nivo signala, dok se svi ostali servisi mogu svrstati u kategoriju signala sa dobrim kvalitetom kada se vrednosti dobijene merenjem uporede sa vrednostima iz Tabele I.

Sve prikazane vrednosti za parametre RSRP i RSRQ dobijene su usrednjavanjem svih izmerenih vrednosti, prilikom čega nije bilo značajnijih varijacija od srednje vrednosti merenih veličina.



Sl. 6. Vrednosti RSRP za scenario makro i mikro ćelije za različite servise.



Sl. 7. Vrednosti RSRQ za scenario makro i mikro ćelije za različite servise.

V. ZAKLJUČAK

U ovom radu je prikazan uticaj arhitekture mreže na kvalitet signala LTE sistema prenosa. Kroz izmerene vrednosti RSRP i RSRQ pokazano je da se bolji kvalitet dobija za mikro scenario. Kada se posmatra ostvareni protok podataka, pokazano je da se i prilikom *download* i *upload* komunikacije postižu viši protoci u okviru mikro scenarija, gde su postignute dvostruko veće brzine prenosa podataka nego u slučaju makro scenarija. Što se tiče posmatranih servisa najlošiji kvalitet signala se javlja kod *download* smera komunikacije u trajanju od 10 s, dok je najbolji kvalitet uočen prilikom učitavanja *web* stranice *youtube*.

U daljem radu moguće je posmatrati druge servise kao i druge lokacije na kojima bi se vršilo merenje. Takođe, merenje se može obavljati više dana kako bi se uočile neke pravilnosti u skladu sa brojem korisnika koji se nalaze na posmatranoj lokaciji. Na teritoriji Republike Srbije se očekuje povećanje saobraćaja u okviru LTE tehnologije, jer je implementacija ove tehnologije i dalje u toku, pa se očekuje da se eventualno dobiju i drugačiji rezultati. Na kraju, treba napomenuti da su u toku pripreme za uvođenje VoLTE (*Voice over LTE*) tehnologije koja će obezbediti servis govora, pa bi se i taj servis, najmanje otporan na bilo kakve smetnje u bežičnom telekomunikacionom kanalu, nekom narednom analizom trebao razmatrati.

LITERATURA

[1] D. Pouhè, E. Driton, M. Salbaum, "The use of microcells as a means of optimizing UMTS networks." *Proc. of WFMN*, pp. 127-132, 2007.

[2] B. Krenik, "4G wireless technology: When will it happen? What does it offer?", *IEEE Asian Solid-State Circuits Conference*, pp. 141-144, Fukuoka, Japan, November 3-5, 2008.

[3] Holma H, Toskala A, "*LTE for UMTS: Evolution to LTE-Advanced*", John Wiley and Sons, Nokia Simens Networks, Finland, 2011.

[4] F. Afroz, R. Subramanian, R. Heidary, K. Sandrasegaran, S. Ahmed, "SINR, RSRP, RSSI and RSRQ measurements in long term evolution networks", *International Journal of Wireless & Mobile Networks*, vol. *7*, no. *4*, pp. *113-123*, August 2015.

[5] D. Aziz, R. Sigle, "Improvement of LTE handover performance through interference coordination", In *VTC Spring 2009 – IEEE 69th vehicular technology conference*, pp. *1-5*, 2009.

[6] Đ. Lukić, M. Koprivica, N. Nešković, A. Nešković, "Eksperimentalna analiza performansi 2G/3G/4G javne mobilne mreže", *24. telekomunikacioni forum Telfor 2016, Zbornik radova*, str. 238-241, Beograd, Srbija, 2016.

[7] T. Mazloum, S. Aerts, W. Joseph, J. Wiart, "RF-EMF exposure induced by mobile phones operating in LTE small cells in two different urban cities," *Annals of Telecommunications 74, 1-2*, pp. *35-42*, 2019.

ABSTRACT

In this paper the impact of the network arhitecture on signal quality in the fourth generation of the public mobile network is analyzed. The analysis was performed using RSRP (Reference Signal Received Power), RSRQ (Reference Signal Received Quality) and throughput parameters. The signal quality parameters were collected by measurement using TEMS Investigation and TEMS Pocket software. The measurements were carried out at the School of Electrical Engineering on the ground floor of the Technical Faculty building for the macro and micro cell scenario. It has been found that better signal quality is ensured in micro cells. The quality of the signal is also considered by the various services provided to the user.

Analysis of the impact of network architecture on signal quality in LTE technology

Ivana Stojanović, Mladen Koprivica, Nenad Stojanović, Aleksandar Nešković

Uporedna analiza klasa *range-free* postupaka za lokalizaciju u bežičnim senzorskim mrežama

Kristina Josifović, Marko Matić, Gorana Crnobrnja, Dragana Lemaić, Goran Marković, *Member IEEE*

***Apstrakt* — Lokalizacija u bežičnim senzorskim mrežama (WSN, *Wireless Sensor Network*) predstavlja proces određivanja prostornih koordinata senzorskih čvorova (SN, *Sensor Nodes*), kao stvarnih, relativnih ili apsolutnih, pozicija u prostoru. Potreba za lokalizacijom SN u WSN javlja se u cilju proširenja funkcionalnosti na različite aspekte otkrivanja, praćenja, predikcije ili sprečavanja pojave određenih događaja u različitim primenama WSN. Predložene su mnoge tehnike za sprovođenje postupaka za lokalizaciju i definisane mere kvaliteta lokalizacije u vidu tačnosti i preciznosti lokalizacije. U radu je predstavljena numerička analiza performansi referentnih predstavnika klasa postupaka za *range-free* lokalizaciju, koja je izvršena krišćenjem razvijenog simulacionog modela u MatLab okruženju, za potrebe određivanja greške lokalizacije posmatranog skupa postupaka za različite scenarije primene. Na osnovu dobijenih rezultata, izvedeni su osnovni zaključci o mogućnosti primene posmatranog skupa *range-free* postupaka za lokalizaciju u WSN.**

***Ključne reči* — Bežične senzorske mreže, lokalizacija čvorova, centralizovani i distribuirani postupci, *range-free* lokalizacija.**

I. Uvod

LOKALIZACIJA predstavlja specifičan proces u okviru bežičnih senzorskih mreža (WSN, *Wireless Sensor Networks*) čija je svrha određivanje pozicije (lokacije) senzorskih nodova (SN, *Sensor Nodes*) u prostoru. Poznavanje pozicije SN u prostoru neophodno je u cilju ostvarivanja osnovne senzorske funkcije SN u WSN, i to onda kada je od interesa informacija o lokaciji na kojoj je izvršeno merenje posmatranog fizičkog fenomena. Dodatno, lokacija SN je bitna i za potrebe rutiranja senzorskih podataka kroz mrežu, njihove agregacije, kao i za izvođenje ostalih mrežnih protokola. Dodatno, poznavanje lokacije SN značajno je i za potrebe obavljanja detekcije, praćenja, lokalizacije, kontrole i prevencije pojave događaja (*event*) ili objekata u određenim primenama WSN. Proces lokalizacije podrazumeva estimaciju vrednosti prostornih koordinata svakog SN u mreži, ali se kao rezultat lokalizacije može definisati i prostorna oblast, kao deo fizičkog prostora, okarakterisana skupom koordinatnih tačaka. Ukoliko SN sam sprovodi proces svoje lokalizacije, često se koristi naziv samo-lokalizacija (eng. *self-localization*). Kao osnovne mere kvaliteta sprovedenog postupka za lokalizaciju SN najčešće se posmatraju procenat lokalizovanih SN, tačnost i preciznost lokalizacije. Kvalitet lokalizacije zavisi od različitih uticaja okruženja, topologije mreže, gustine prostornog rasporeda SN, kao i od degradirajućih efekata koji utiču na propagaciju signala koje mogu uneti greške pri merenjima parametara na osnovu kojih se obavlja postupak za lokalizaciju [1, 2]. Stoga, od interesa je da se različiti postupci za lokalizaciju analiziraju pri različitim scenarijima primene i uticaja okruženja, što je i bio jedan od osnovnih motiva istraživanja čiji su rezultati prikazani u okviru ovog rada.

Do sada je predložen veliki broj postupaka putem kojih se na različite načine određuju prostorne pozicije (lokacije) SN u WSN. Proces lokalizacije SN može se izvesti putem primene centralizovanog postupka, kada se na jednom mestu, najčešće pristupnom uređaju WSN, tj. *sink*-u, prikupljaju sve potrebne informacije (parametri) od svih SN, i na osnovu njih obavlja lokalizacija svih SN u mreži. Osim toga, postupak za lokalizaciju može biti distribuiran, kada svaki SN prikuplja potrebne informacije i sam određuje svoju lokaciju, a zatim o tome obaveštava ostatak mreže. U ovom slučaju, proces za lokalizaciju se često može podeliti u dve faze: prvu, u kojoj svaki SN određuje rastojanje ili konektivnost do svojih suseda i/ili referentnih čvorova (za koje je lokacija unapred poznata sa zadatom tačnošću), i drugu, kada svaki SN koristi te informacije kao i istovetne informacije dobijene od svojih suseda kako bi izvršio estimaciju svoje lokacije.

U postupku lokalizacije SN u okviru WSN, veoma često se koristi određeni broj referetnih čvorova (čije su koordinate *a priori* poznate), a koji znatno olakšavaju i povećavaju kvalitet procesa lokalizacije. Ovi referentni čvorovi imaju iste funkcije kao i ostali SN u mreži, ali mogu imati složeniji hardver, npr. opremljeni su GPS (*Global Positioning System*) prijemnicima za potrebe pozicioniranja, ili se planski postavljaju na tačno određenim, unapred poznatim, pozicijama u prostoru. Ovakvi čvorovi nazivaju se *beacon* čvorovi (BN, *Beacon Nodes*), pri čemu je dokazano da njihovo prisustvo u WSN omogućava velika poboljšanja kvaliteta lokalizacije ostalih SN, kao i da je neretko njihovo prisustvo u WSN čak i neophodno kako bi se lokalizacija SN u mreži uspešno obavila, [1-5].

U ovom radu, najpre je data osnovna klasifikacija i pregled postupaka za lokalizaciju, sa posebnim osvrtom na veoma jednostavnu grupu postupaka, tzv. *range-free* postupaka, koja za svoj rad ne zahteva poseban hardver kao ni određivanje međusobnih rastojanja SN u mreži. Nakon toga, prikazani su

Kristina Josifović – Elektrotehnički fakultet, Univerzitet u Beogradu, Bulevar Kralja Aleksandra 73, 11020 Beograd, Srbija (e-mail: kristinajos@etf.bg.ac.rs).

Marko Matić – MsC student, Elektrotehnički fakultet, Univerzitet u Beogradu, Bulevar Kralja Aleksandra 73, 11020 Beograd, Srbija (e-mail: mm173386m@student.etf.bg.ac.rs).

Gorana Crnobrnja – Telenor d.o.o. Beograd, Omladinskih brigada 90, 11070 Novi Beograd, Srbija (e-mail: gorana.crnobrnja@telenor.rs).

Dragana Lemaić – Allied testing Serbia d.o.o. Beograd, Jurija Gagarina 26v, 11070 Novi Beograd, Srbija (e-mail: lemaicdragana@gmail.com).

Goran Marković – Elektrotehnički fakultet, Univerzitet u Beogradu, Bulevar Kralja Aleksandra 73, 11020 Beograd, Srbija (e-mail: gmarkovic@etf.bg.ac.rs).

rezultati uporedne analize tipičnih predstavnika klasa *range-free* postupaka za lokalizaciju SN za određen skup scenarija primene, uz navođenje osnovnih izvedenih zaključaka analize.

II. KLASIFIKACIJA I PREGLED POSTUPAKA ZA LOKALIZACIJU

Osnovna podela postupaka za lokalizaciju može se izvršiti na *range-based* i *range-free* postupke, [1-3]. U prvu grupu, *range-based*, spadaju oni postupci za lokalizaciju zasnovani na određivanju međusobnog rastojanja SN, u kojima se potom putem primene trilateracije, triangulacije ili multilateracije određuje lokacija za sve SN u mreži. Druga grupa, *range-free* postupaka za lokalizaciju, [3], zasniva se najčešće na primeni informacija o međusobnoj konektivnosti SN, tj. povezanosti SN u okviru mreže, a lokacija SN određuje se bez direktnog merenja rastojanja između SN ili posrednog određivanja na osnovu nivoa snage signala na prijemu (RSSI, *Received Signal Strength Idicative*). Kod *range-free* postupaka za lokalizaciju, BN imaju izuzetno bitnu ulogu, kao referentne tačke za potrebe procene pozicije SN.

Merenje međusobnog rastojanja između SN može se vršiti na osnovu više različitih parametara [1, 2]. Najjednostavniji postupak je merenje vremena propagacije signala (ToA , *Time of Arrival*), sa pretpostavkom da je rastojanje između SN proporcionalno vremenu za koje radio signal stigne od jednog SN do drugog. Sa druge strane, može se meriti razlika vremena propagacije signala između dva čvora (TDoA, *Time Difference of Arrival*), gde se posmatraju trenuci prispeća dva signala koja se prenose različitim brzinama, pri čemu se ne zahteva sinhronizacija početnog i krajnjeg čvora. Dodatno, primena RSSI zasnovanih postupaka podrazumeva merenje nivoa snage signala na predaji i prijemu kod SN, kao i poznavanje modela slabljenja radio signala usled propagacije, pri čemu se ne zahteva dodatni hardver jer SN poseduju radio prijemnike. Konačno, merenje upadnog ugla pri prijemu radio signala (AoA, *Angle of Arrival*), gde se posmatra ugao prispeća signala u SN u odnosu na referentni pravac, može koristiti kao parametar za lokalizaciju SN. Određivanjem prethodno navedenih parametara, uz poznavanje koordinata BN, u okviru *range-based* postupaka lokalizacije obavljaju se proračuni lokacije SN uz primenu tehnika triangulacije, trilateracije ili multilateracije, [2].

U okviru *range-free* postupaka lokalizacije takođe se koriste BN, a mogu se koristiti i RSSI vrednosti koje se, za razliku od *range-based* pristupa, koriste samo za indikaciju relativnog odnosa, ali ne i samu estimaciju rastojanja između čvorova. Rad *range-free* postupaka za lokalizaciju znatno se više oslanja na međusobnu komunikaciju SN i BN, i razmenu potrebnih informacija, čime SN jedni drugima pomažu u postupku lokalizacije. Važno je napomenuti da se *range-free* postupci za lokalizaciju mogu realizovati i kao centralizovani, i kao distribuirani postupci. Pri tome, kod centralizovanih postupaka podrazumeva se da u mreži postoji centralni uređaj, najčešće *sink*, kojem se šalju podaci svih SN, a koji te podatke obrađuje, i obavlja procenu lokacije za svaki SN. Prednost centralizovanog pristupa je to da *sink* dobija potpun skup informacija potrebnih za lokalizaciju, te ima znatno više podataka za potrebe donošenja preciznih odluka o lokaciji svih SN u datoj mreži. Nedostaci centralizovanog postupka lokalizacije su jako složena implementacija, visoki zahtevi u smislu procesorske snage uređaja koji obavlja lokalizaciju (usled obrade velike količine podataka i složenog postupka optimizacije), i veoma velika potrošnja energije za potrebe dostavljanja potrebnih informacija od svih SN u mreži. Sa druge strane, distribuirani postupci za lokalizaciju rade na principu samo-lokalizacije, u kome svaki SN sam određuje svoju lokaciju na osnovu razmene podataka sa susednim SN i BN. Distribuirani postupci za lokalizaciju su jednostavniji za implementaciju, za njihov rad se zahteva znatno manji nivo komunikacije u mreži, a posledično se troši i značajno manja količina energije. Ipak, ove pogodnosti se postižu po cenu slabije tačnosti i preciznosti lokalizacije, [1, 2].

U radu su, kao posebno interesantni, analizirani *range-free* postupci za lokalizaciju. Pri tome, najpre su opisani neki tipični predstavnici centralizovanih i distribuiranih postupaka za lokalizaciju, a potom su ovi postupci razmatrani u okviru procesa analize čiji su rezultati prikazani u ovom radu.

Approximate Point in Triangulation (APIT) postupak za lokalizaciju, [4, 5], je tipičan predstavnik distribuiranih postupaka, i on zahteva primenu manjeg broja BN u mreži. BN periodično šalju *beacon* signale ka susednim SN u mreži, pa su u cilju ostvarivanja većeg dometa radio-komunikacije, opremljeni predajnicima veće snage. Kako bi se obavila lokalizacija posmatranog SN, on mora biti u dometu barem tri BN koji emituju *beacon* signale ka svim susednim čvorovima u mreži. Svaki SN, ili drugi BN, prima *beacon* signale koji nose informaciju o koordinatama predajnika (BN), zapisuje tu informaciju, kao i vrednost RSSI parametra signala, i razmenjuje ih lokalnom komunikacijom sa susednim SN. Formiraju se sve moguće kombinacije trouglova, čija temena čine po tri BN, od svih dostupnih BN u mreži. Za svaki posmatrani SN, izdvajaju se trouglovi u kojima se on nalazi, tako što se upoređuju nivoi snage primljenih *beacon* signala za taj SN i za sve njegove susede koji su primili signale istog skupa BN, a težište preseka svih takvih trouglova predstavlja koordinate datog SN. Opisani postupak se obavlja posebno za svaki SN u mreži i naziva se APIT test, [4, 5].

Drugi predstavnik distribuiranih postupaka lokalizacije je *Distance Vector – Hop* (DV-Hop) postupak, [6], koji se realizuje u tri faze, i koji funkcioniše na osnovu korišćenja multilateracije. U prvoj fazi se obavlja plavljenje mreže tzv. *beacon* paketima, pri čemu se kao parametri koriste broj *hop*-ova (koraka) pri prosleđivanju paketa kroz mrežu i koordinate BN koji su izvor paketa. Brojač koraka je inicijalno postavljen na nulu i inkrementira se svaki put kad dođe do drugog čvora, pri čemu se svaki novi paket prosleđuje sa inkrementiranim brojem koraka. Druga faza postupka, podrazumeva da BN određuju prosečnu dužinu *hop*-a do svakog drugog BN proračunom srednje vrednosti rastojanja, d_i, [6],

$$d_i = \frac{\sum_{i \neq j}^{N} \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}}{\sum_{i \neq j}^{N} h_{ij}}, \quad (1)$$

gde je h_{ij} broj koraka između i-tog i j-tog BN od N BN, a (x_i, y_i) i (x_j, y_j) su koordinate i-tog i j-tog BN, respektivno. U trećoj fazi, svaki SN sam određuje svoju lokaciju korišćenjem metode najmanjih kvadrata (LS, *Least Square*). Pri tome, SN pamti minimalnu vrednost koraka do BN, tj. bira se optimalna putanja do onog BN za dati SN sa najmanjim brojem posredničkih čvorova. Procenjuje se prosečna dužina koraka kroz optimalne putanje, uz proračun korekcionog faktora koji se propagira kroz mrežu i koristi kao dodatni parametar za proračun lokacije SN, [6].

Centroid Localization Algorithm (CLA), [7], je distribuiran postupak za lokalizaciju koji se zasniva na poznavanju konektivnosti posmatranog SN i skupa dostupnih BN. Pri tome, koristi se *broadcast* princip slanja *beacon* signala od BN koji sadrže koordinate BN, a procena pozicije SN obavlja se prostim usrednjavanjem koordinata skupa BN u čijem se dometu nalazi posmatrani k-ti SN, odnosno imamo da je lokacija određena sa, [7],

$$(x_k, y_k) = \left(\frac{\sum_{j=1}^{N_k} x_j}{N_k}, \frac{\sum_{j=1}^{N_k} y_j}{N_k}\right), \quad (2)$$

gde su (x_k, y_k) koordinate posmatranog SN, a (x_j, y_j) koordinate BN, dok je N_k broj BN sa kojima je dati SN konektivan. CLA predstavlja jedan od najjednostavnijih *range-free* postupaka za lokalizaciju, pri čemu je *Weighted* CLA (W-CLA) [8, 9] verovatno najznačajnija varijanta ovog postupka. Naime, CLA postupak ne uzima u obzir uticaj udaljenosti pojedinačnih dostupnih BN od posmatranog SN, dok W-CLA postupak koristi i informacije o težinskim koeficijentima koji označavaju procenat učešća koordinata svakog od BN pri lokalizaciji SN. Kao težinski koeficijent najčešće se koriste RSSI parametri, a koji direktno zavise od rastojanja, na osnovu sledećih izraza, [8, 9],

$$w_{jk} = \left(d_{jk}\right)^{-g} = \left(P_{ref} \times 10^{(RSSI_{jk}/20)}\right)^{g}, \quad (3)$$

$$(x_k, y_k) = \left(\frac{\sum_{j=1}^{N_k} w_{jk} \times x_j}{\sum_{j=1}^{N_k} w_{jk}}, \frac{\sum_{j=1}^{N_i} w_{jk} \times y_j}{\sum_{j=1}^{N_k} w_{jk}}\right) \quad (4)$$

gde je w_{jk} težinski koeficijent k-tog SN i j-tog BN, d_{jk} je njihovo međusobno rastojanje, P_{ref} referentna snaga BN, g je usvojena vrednost propagacionog slabljenja, dok je $RSSI_{jk}$ nivo snage signala j-tog BN na prijemu u k-tom SN, pri čemu je N_k broj BN sa kojima je dati SN konektivan.

Još jedna predložena izmena CLA postupka za lokalizaciju je *Compensated* CLA (C-CLA), [9], zasnovana na poznavanju konektivnosti posmatranog SN i skupa BN u dometu, kao u osnovnom CLA postupku i W-CLA modifikaciji. Kod C-CLA postupka uvodi se određena razlika u načinu procene težinskih koeficijenata, odnosno posmatra se broj dostupnih BN za posmatrani SN kao dodatni parametar, i uvodi se pravilo da težinski koeficijenti dostupnih BN imaju veću vrednost ukoliko njih ima više, a na osnovu toga što se očekuje da je u tom slučaju rastojanje između njih i posmatranog SN manje. U okviru C-CLA postupka koristi se izraz (3) za procenu ovih koeficijenata isto kao i W-CLA, dok se izraz, [9],

$$w'_{jk} = w_{jk} \times {N_k}^{2 \times w_{jk}}, \quad (5)$$

koristi za procenu novog težinskog koeficijenta w'_{kj} na osnovu vrednosti osnovnog koeficijenta w_{jk} iz izraza (4), pri čemu je N_k ukupan broj dostupnih BN za posmatrani SN.

Mnoge implementacije WSN ne zahtevaju veliku tačnost i preciznost, već samo grubu procenu pozicije svih SN. Jedno rešenje, koje koristi centralizovani pristup za lokalizaciju je *Convex Position Estimation* (CPE) postupak, [10]. CPE postupak se zasniva na primeni *beacon* signala koje šalju BN i SN, putem kojih SN dobijaju informacije o konektivnosti svih čvorova u mreži. Svaku WSN posmatramo kao graf sa n čvorova. Prvih m čvorova su BN čije su koordinate poznate, a preostalih $n - m$ su SN koje treba lokalizovati. Pri tome, proces optimizacije koji se koristi u okviru CPE postupka zahteva određivanje takvih vrednosti skupa koordinata $(x_i, y_i), i = 1, \cdots, n$, da rastojanje čvorova koji su konektivni bude manje ili jednako maksimalnom dometu svakog SN pri komunikaicji između SN, ili dometu BN pri komunikaciji BN i SN. Svaki SN mora da popuni tabelu povezanosti sa drugim SN/BN u mreži, tj. upisuje podatke o rastojanju između njega i ostalih čvorova sa kojima komunicira. Svaki čvor zatim šalje ovakvu tabelu ka *sink*-u koji procesira prikupljene informacije na osnovu kojih donosi odluku o lokacijama SN. CPE postupak koristi *semidefinite* programiranje za određivanje optimalnog rešenja, a u ovom radu korišćen je dostupni alat mosek za optimizaciju, [11].

Recursive Position Estimation (RPE), [12, 13], predstavlja rekurzivan postupak za estimaciju lokacije SN, koji se može primeniti na lokalizaciju SN u WSN velikih dimenzija. RPE postupak se zasniva na korišćenju informacije o međusobnoj konektivnosti SN, ali se za razliku od CPE koriste samo lokalne informacije (distribuiran postupak). Prva faza RPE postupka podrazumeva određivanje skupa lokalno dostupnih BN, dok se u drugoj fazi se primenom postupka sličnim sa CPE postupkom, ali ovog puta u samom SN sprovodi samo-lokalizacija datog SN. U četvrtoj, završnoj fazi, SN koji su imali dovoljno lokalnih informacija za samo-lokalizaciju, pridružuju se skupu referentnih čvorova, pa se oblast u kojoj je moguća lokalizacija iterativno povećava. RPE postupak omogućava visoku fleksibilnost pri realizaciji lokalizacije u WSN, ali po cenu manje tačnosti pošto greška određivanja lokacije SN iz prethodnih iteracija RPE postupka, koji ulaze u skup referentnih za naredne iteracije, negativno utiče na kvalitet lokalizacije SN u narednim iteracijama, [12, 13].

III. SIMULACIONI MODEL I REZULTATI SPROVEDENE ANALIZE

Za potrebe uporedne analize prethodno navedenih tipičnih predstavnika pojedinih klasa *range-free* postupaka razvijen je poseban simulacioni model u MatLAB okruženju. Posmatrana je WSN dimenzija senzorskog polja 250 m×250 m, koju čine 5×5 kvadratne ćelije, sa $N_{BC} \in \{1,2,4\}$ pravilno raspoređenih BN i 8 SN slučajnog rasporeda po ćeliji (ukupno 200 u WSN).

Posmatrani su scenariji rada WSN u kojima je domet BN u komunikaciji sa SN bio vrednosti $R_{BC} \in \{35\,m, 50\,m, 100\,m\}$, dok je domet između SN $R_{SN} \in \{25\,m, 35\,m, 50\,m\}$. Posmatran je idealan slučaj propagacije, u kojoj su svi SN u dometu BN/SN međusobno konektivni, kao i to da su potrebni podaci o konektivnosti čvorova u mreži i vrednosti RSSI (koja se koristi kod nekih postupaka, npr. APIT, W-CLA, C-CLA) bile dostupne u svim SN, pri čemu je za W-CLA i C-CLA postupke predpostavljena vrednost parametra $g = 1.5$.

Generisan je skup od 100 nezavisnih postavki WSN za svaki scenario (*Monte-Carlo* eksperimenata), uz primenu skupa postupaka za lokalizaciju (APIT, CLA, C-CLA, W-CLA, CPE, RPE i DV-Hop) realizovanih u skladu sa opisom iz literature. Poređenjem sa tačnim vrednostima, određena je srednja vrednost i varijansa greške lokalizacije za svaki scenario i postupak, pri čemu su rezultati u pogledu srednje greške lokalizacije prikazani u **TABELA 1**, odnosno varijanse greške u **TABELA 2** (prikazana je najveća vrednost varijanse zabeležena na skupu svih eksperimenata - postavki WSN). Pri tome, u cilju preglednosti nisu dati rezultati za CLA i C-CLA postupke, kojima su dobijani nešto lošiji rezultati u odnosu na W-CLA postupak, uz vrlo slično ponašanje sa promenom parametara scenarija. Dodatno nisu dati rezultati za domet BN od 100 m za koji je za sve postupke i domete SN ostvaren lošiji kvalitet lokalizacije u odnosu na manje domete.

TABELA 1 - SREDNJA GREŠKA LOKALIZACIJE [M] U ZAVISNOSTI OD BROJA BN PO ĆELIJI (1, 2 ILI 4), DOMETA BN (35, 50) I DOMETA SN (25, 35, 50)

N_{BC} = 1	**APIT**	**W-CLA**	**CPE**	**RPE**	**DV-Hop**
R_{BC}=35,R_{SN}=25	13.7131	13.0713	12.6019	--	19.0880
R_{BC}=35,R_{SN}=35	13.8748	13.2986	14.2951	25.5867	16.2088
R_{BC}=35,R_{SN}=50	13.5601	12.9891	17.0464	27.0686	31.3221
R_{BC}=50,R_{SN}=25	42.5284	7.2990	15.5655	21.0223	18.1854
R_{BC}=50,R_{SN}=35	43.8940	7.0622	17.0698	24.2348	15.6736
R_{BC}=50,R_{SN}=50	44.7816	7.1734	20.8031	29.1971	18.5762
N_{BC} = 2	**APIT**	**W-CLA**	**CPE**	**RPE**	**DV-Hop**
R_{BC}=35,R_{SN}=25	47.2995	7.4486	9.3699	14.0129	12.4056
R_{BC}=35,R_{SN}=35	51.0214	7.5757	10.3391	16.4946	16.4121
R_{BC}=35,R_{SN}=50	49.1076	7.6309	12.4914	17.8936	35.7326
R_{BC}=50,R_{SN}=25	37.2471	5.7823	11.4317	14.6355	17.4117
R_{BC}=50,R_{SN}=35	33.0340	5.9433	13.0799	16.7549	15.1720
R_{BC}=50,R_{SN}=50	30.3560	6.0346	15.6587	20.1590	21.9678
N_{BC} = 4	**APIT**	**W-CLA**	**CPE**	**RPE**	**DV-Hop**
R_{BC}=35,R_{SN}=25	26.1304	3.4488	6.6174	8.5485	18.6379
R_{BC}=35,R_{SN}=35	25.5173	3.5488	7.2821	9.7934	20.1496
R_{BC}=35,R_{SN}=50	24.9662	3.4258	8.3603	10.3014	39.9884
R_{BC}=50,R_{SN}=25	21.7050	3.9173	9.0742	11.0431	27.4979
R_{BC}=50,R_{SN}=35	17.9319	3.9461	9.7577	11.8677	26.1945
R_{BC}=50,R_{SN}=50	17.1553	3.9343	11.3036	14.0260	24.5250

Analizom dobijenih rezultata, može se uočiti da za većinu postupaka srednja greška lokalizacije opada sa povećanjem broja BN u mreži (tj. porasta gustine prostornog rasporeda BN u mreži), osim za APIT za manje domete BN i DV-Hop za veće vrednosti dometa SN i/ili izuzetno veliki broj BN kada se smanji prosečan broj *hop*-ova između BN, odnosno kada je veći broj BN direktno konektivan. Može se zaključiti da APIT postupak lokalizacije karakterišu znatno lošije performanse u odnosu na sve ostale razmatrane postupke. Za distribuirane RPE i DV-Hop postupke, porast dometa SN kao i povećanje broja BN, generalno dovodi do opadanja vrednosti srednje greške lokalizacije (osim za suviše veliki broj BN u mreži kod DV-Hop postupka), dok povećanje dometa BN, kada ih ima manji broj u okviru WSN, pozitivno utiče na tačnost lokalizacije. Kod DV-Hop postupka se može zapaziti i to da se najbolji rezultati ostvaruju za slične iste vrednosti dometa SN i BN (tj. za iste vrednosti za manje domete BN odnosno nešto manje vrednosti dometa SN za veće domete BN). Ovo je posledica principa rada DV-Hop protokola, kod koga se procenjuje srednja dužina *hop*-a u mreži pa suviše različiti dometi SN i BN mogu usloviti porast greške procene (ne vodi se računa o tome da li se radi o vezi SN sa SN ili SN sa BN). Sa druge strane, u slučaju centralizovanog CPE postupka, kao i distribuirane postupke CLA, W-CLA i C-CLA, povećanje broja BN uvek uslovljava opadanje srednje vrednosti greške lokalizacije. Kod ovih postupaka, povećanjem dometa BN osim za izuzetno veliku gustinu prostornog rasporeda BN (tj. $N_{BC} = 4$), ostvaruju se veća tačnost i preciznost lokalizacije. Pri tome, smanjivanje dometa SN za datu grupu postupaka doprinosi povećanju tačnosti i preciznosti lokalizacije. Ovu grupu postupaka, kao što vidi u **TABELA 2** karakteriše i znatno veća preciznost lokalizacije (manja varijansa greške) u odnosu na ostale postupke na čiji rad u ovom pogledu značajno utiče raspored SN u okviru senzorskog polja date WSN, odnosno u odnosu na skup BN koji se koristi za lokalizaciju datog SN.

TABELA 2 - VARIJANSA LOKALIZACIJE [M] U ZAVISNOSTI OD BROJA BN PO ĆELIJI (1, 2 ILI 4), DOMETA BN (35, 50) I DOMETA SN (25, 35, 50)

N_{BC} = 1	**APIT**	**W-CLA**	**CPE**	**RPE**	**DV-Hop**
R_{BC}=35,R_{SN}=25	229.65	238.72	108.11	--	283.19
R_{BC}=35,R_{SN}=35	50.28	57.21	127.29	701.65	208.68
R_{BC}=35,R_{SN}=50	50.96	57.77	171.10	712.14	423.40
R_{BC}=50,R_{SN}=25	1438.07	49.27	152.34	724.44	149.69
R_{BC}=50,R_{SN}=35	1463.73	45.82	186.39	704.39	237.90
R_{BC}=50,R_{SN}=50	1475.31	47.55	256.08	747.74	225.13
N_{BC} = 2	**APIT**	**W-CLA**	**CPE**	**RPE**	**DV-Hop**
R_{BC}=35,R_{SN}=25	2074.48	26.86	62.86	644.18	63.90
R_{BC}=35,R_{SN}=35	2192.06	29.07	87.13	669.91	241.50
R_{BC}=35,R_{SN}=50	2171.73	27.76	104.46	405.66	521.03
R_{BC}=50,R_{SN}=25	1945.61	33.83	102.56	628.50	128.97
R_{BC}=50,R_{SN}=35	1754.39	36.41	130.34	702.35	110.76
R_{BC}=50,R_{SN}=50	1465.45	38.83	169.11	730.47	293.25
N_{BC} = 4	**APIT**	**W-CLA**	**CPE**	**RPE**	**DV-Hop**
R_{BC}=35,R_{SN}=25	1765.97	11.76	37.75	615.05	319.47
R_{BC}=35,R_{SN}=35	1946.87	12.89	52.90	592.68	247.63
R_{BC}=35,R_{SN}=50	1843.95	10.93	62.07	621.41	520.83
R_{BC}=50,R_{SN}=25	798.03	27.53	75.71	641.14	576.09
R_{BC}=50,R_{SN}=35	512.91	22.14	85.24	677.65	409.27
R_{BC}=50,R_{SN}=50	85.97	21.41	100.03	680.73	336.59

Sprovedena je i analiza u cilju određivanja osetljivosti datih postupaka na grešku poznavanja lokacije BN sa varijansom greške reda 0.25 m, čiji su rezultati dati u **TABELA 3**. Uočava se da *range-free* postupci imaju veoma malu osetljivost (u smislu povećanja srednje greške lokalizacije) na grešku poznavanja lokacija BN (radi preglednosti nisu dati rezultati

za 4 BN po ćeliji – dobijaju se slične vrednosti kao u **TABELA 1** za iste parametre scenarija), dok se kod CPE postupka javlja velika osetljivost kada imamo veći broj BN u mreži uz manje vrednosti dometa BN i veće vrednosti dometa SN.

TABELA 2 - SREDNJA GREŠKA LOKALIZACIJE [m] U ZAVISNOSTI OD BROJA BN PO ĆELIJI (1, 2 ILI 4), DOMETA BN (35, 50) I DOMETA SN (25, 35, 50), U SLUČAJU POSTOJANJA GREŠKE POZNAVANJA LOKACIJE BN

N_{BC} = 1	APIT	W-CLA	CPE	RPE	DV-Hop
R_{BC}=35,R_{SN}=25	13.7176	13.0754	12.6108	--	19.0666
R_{BC}=35,R_{SN}=35	13.8605	13.2827	14.2847	25.5607	16.1351
R_{BC}=35,R_{SN}=50	13.5638	12.9892	17.0649	27.0822	31.3169
R_{BC}=50,R_{SN}=25	42.5323	7.3056	15.5662	21.0293	18.1510
R_{BC}=50,R_{SN}=35	43.9364	7.0726	16.9153	24.2228	15.7471
R_{BC}=50,R_{SN}=50	44.8218	7.1797	20.8088	29.2215	18.6091
N_{BC} = 2	APIT	W-CLA	CPE	RPE	DV-Hop
R_{BC}=35,R_{SN}=25	47.3024	7.4545	9.3672	14.0137	12.3757
R_{BC}=35,R_{SN}=35	51.0108	7.5657	28.1485	16.6395	16.4017
R_{BC}=35,R_{SN}=50	49.1069	7.6509	30.0863	17.9598	35.7450
R_{BC}=50,R_{SN}=25	36.2796	5.7761	11.4525	14.6543	17.3803
R_{BC}=50,R_{SN}=35	31.7370	5.9405	13.0720	16.7427	15.2435
R_{BC}=50,R_{SN}=50	29.1862	6.0274	15.6772	20.1785	22.0195

IV. ZAKLJUČAK

Na osnovu rezultata sprovedene analize, zaključuje se da se dobar kvalitet lokalizacije SN, za određene uslove (parametre scenarija primene WSN) može ostvariti kako korišćenjem distribuiranih (npr. W-CLA), tako i centralizovanih postupaka (npr. CPE) za lokalizaciju. Pri tome, pri projektovanju WSN treba odrediti pogodan izvor u smislu dometa BN i broja BN (tj. gustine rasporeda), kao i dometa SN u skladu sa izabranim postupkom za lokalizaciju (ili obrnuto), kako bi se postigla prihvatljiva greška lokalizacije SN sa što manjim brojem BN.

Centralizovani postupci za lokalizaciju zahtevaju znatno veću potrošnju energije, pa imajući u vidu da se sličan kvalitet lokalizacije može ostvariti i distribuiranim postupcima, za WSN većih dimenzija treba razmotriti primenu distribuiranih postupaka. Naime, u WSN ovog tipa javljaju se znatno veća rastojanja između SN, kao i veća prosečna rastojanja SN do pristupnog uređaja (*sink*) u kome se prikupljaju informacije. Posledično, primenu centralizovanih postupaka zato odlikuju znatno veći zahtevi po pitanju količine bežične komunikacije i potrošnje energije neophodne za potrebe lokalizacije.

Može se očekivati da se daljom analizom može odrediti povoljna kombinacija postupaka za lokalizaciju čijom bi se sekvencijalnom primenom u više koraka, ili u sklopu nekog iterativnog kooperativnog postupka u kome SN sarađuju u cilju poboljšanja kvaliteta lokalizacije, obezbedilo optimalno ponašanje u smislu uspešnosti lokalizacije, potrošnje energije pri komunikaciji za potebe lokalizacije i optimizaciju (uz minimizaciju) broja i dometa BN koji bi se koristio u tako dobijenom postupku lokalizacije.

ZAHVALNICA

Rad je delimično finansiran od Ministarstva prosvete, nauke i tehnološkog razvoja RS, kroz projekte tehnološkog razvoja 32028 i 32037.

LITERATURA

[1] I. F. Akyildiz, and M. C. Vuran, *Wireless Sensor Networks*, NewYork, USA, John Wiley & Sons Inc., 2010.
[2] W. Dargie, and C. Poellabauer, *Fundamentals of Wireless Sensor Networks Theory and Practice*, NewYork, USA, John Wiley & Sons Ltd, 2010.
[3] T. He; C. Huang, B. M. Blum, J. A. Stanković and T. Abdelzaher, "Range-Free Localization Shemes for Large Scale Sensor Networks", Proc. MobiCom 2003, San Diego, California, USA, pp. 81-95, 14-19 September, 2003.
[4] S. M. Hosseininirad, M. Niazi, J. Pourdeiliami, S. K. Basu, and A. A. Pouyan, "On improving APIT algorithm for better localization in WSN", *Journal of AI and Data Mining*, Vol.2, No 2, pp. 97-104, 2014.
[5] W. Tie-Zhou, Z. Yi-Shi, Z. Hui-Jun, L. Biao, "Wireless Sensor Network Node Location Based on Improved APIT", *Journal of Surveying and Mapping Engineering*, Vol. 1, Issue 1, pp. 15-19, June 2013.
[6] H. Chen, K. Sezaki, P. Deng, and H. C. So, "An Improved DV-Hop Localization Algorithm for Wireless Sensor Networks," Proc. 3rd IEEE Conference on Industrial Electronics and Applications, Singapore, pp. 1557-1561, 2008.
[7] X. Shen, Z. Wang, P. Jiang, R. Lin, and Y. Sun, "*Connectivity and RSSI Based Localization Scheme for Wireless Sensor Networks*". In: Huang DS., Zhang XP., Huang GB. (eds) Advances in Intelligent Computing. ICIC 2005, Lecture Notes in Computer Science, vol. 3645, Berlin, Heidelberg, Springer, 2005.
[8] M. Arun, N. Sivasankari, P. T. Vanathi, and P. Manimegalai, "Analysis of Average Weight Based Centroid Localization Algorithm for Mobile Wireless Sensor Networks," *Advances in Wireless and Mobile Communications*, Vol. 10, No. 4, pp. 757-780, 2017.
[9] Q. Dong, and X. Xu, "A Novel Weighted Centroid Localization Algorithm Based on RSSI for an Outdoor Environment," *Journal of Communications*, Vol. 9, No. 3, pp. 279-285, 2014.
[10] L. Doherty, K. S. J. Pister, and L. E. Ghaoui, "Convex Position Estimation in Wireless Sensor Networks," Proc. 20th IEEE INFOCOM 2001, Vol. 3, Anchorage, USA, pp. 1655-1663, April, 2001.
[11] https://www.mosek.com/, poslednji pristup 12.04.2019. godine.
[12] J. Albowicz, A. Chen, and L. Zhang, "Recursive Position Estimation in Sensor Networks," Proc. 9th International Conference on Network Protocols, Riverside, CA, USA, pp. 35-41, November 2001
[13] P. Kristalina, W. Hendrantoro, and G. Hendrantoro, "Improve the Robustness of Range-Free Localization Methods on Wireless Sensor Networks using Recursive Position Estimation Algorithm", *Journal of ICT Research and Applications*, Vol. 5, No. 3, pp. 203-222, 2011.

ABSTRACT

Localization in Wireless Sensor Networks (WSN) is the process of determining the spatial coordinates of Sensor Nodes (SN), in order to determine their actual position, relative or absolute. The requisite for SN localization arises to extend the WSN functionality to the aspects of detection, tracking, prediction and prevention of the occurrence of certain events in various WSN applications. Many localization techniques have been proposed and the measures of quality of localization in the form of accuracy and precision have been defined. In this paper, we present the numerical analysis of referent *range-free* localization procedures performance, executed by using simulation model developed in the MatLAB program package, which we used to calculate the localization error in different application scenarios for here considered localization procedures. Based on the obtained results, we drawn the conclusions on the application of these localization procedures in the WSN.

Comparative analysis of several classes of range-free localization algorithms in wireless sensor network

Kristina Josifović, Gorana Crnobrnja, Marko Matić, Dragana Lemaić and Goran Marković

Unapređenje postupaka za lokalizaciju u WSN sa kombinovanjem DV-Hop i Centroid rešenja

Gorana Crnobrnja, Kristina Josifović i Goran Marković, *Member, IEEE*

***Apstrakt* — Lokalizacija, kao postupak određivanja pozicije senzorskih čvorova u okviru bežičnih senzorskih mreža (WSN, *Wireless Sensor Network*), karakteriše se tačnošću i preciznošću estimacije prostornih koordinata (lokacija) na kojima se čvorovi nalaze. Predložen je veliki broj rešenja za potrebe lokalizacije u WSN, na osnovu različitih pristupa, u cilju poboljšanja kvaliteta lokalizacije. Jedno od predloženih rešenja je *Distance Vector – Hop* (DV-Hop) postupak lokalizacije. U ovom radu posmatran je originalan DV-Hop postupak, kao i modifikacije ovog postupka zasnovane na kombinovanju DV-Hop postupka i jednostavnog *Centroid* postupka. Predložena je dodatna modifikacija postupka ovog tipa u cilju daljeg unapređenja performansi ove klase postupaka za lokalizaciju, odnosno izvršena je uporedna analiza performansi DV-Hop zasnovanih postupaka, *Centroid* postupka, i skupa modifikovanih postupaka zasnovanih na kombinovanju ova dva osnovna postupka za lokalizaciju. Za potrebe analize performansi postupaka lokalizacije za različite scenarije primene u WSN samostalno je razvijen specifičan simulacioni model, čijom primenom su dobijeni rezultati na osnovu kojih je izvršena analiza mogućnosti primene skupa postupaka namenjenih za lokalizaciju u okviru WSN u slučaju mreža većih dimenzija.**

***Ključne reči* — Bežične senzorske mreže, postupci za lokalizaciju, DV-Hop lokalizacija, *Centroid* postupak.**

I. Uvod

BRZI razvoj bežičnih komunikacionih tehnologija, kao i razvoj tehnologija za izradu senzora doveli su do pojave bežičnih senzorskih mreža (WSN, *Wireless Sensor Networks*). U okviru ovih mreža koriste se uređaji malih dimenzija i ograničenih mogućnosti (ograničeno napajanje električnom energijom, mala procesorska snaga i raspoloživa memorija,..), senzorski nodovi (SN, *Sensor Nodes*), koji omogućavaju prikupljanje informacija o okruženju (primenom senzora) i njihovo dostavljanje (primenom bežične komunikacije) kroz mrežu. WSN imaju brojne i značajne primene u životnom i radnom okruženju, industriji, ekologiji i mnogim drugim sferama, [1, 2]. Pri tome, ove mreže predstavljaju pogodno rešenje za primene u kojima postoji potreba za prikupljanjem podataka iz fizičkog okruženja za potrebe kontrole, nadzora i upravljanja sredinom. Za praktičnu primenu WSN često je neophodno da se, uz prikupljene senzorske podatke, da i tačna lokacija na kojima je obavljeno merenje, ali zbog zahteva za malom cenom, potrošnjom energije i veličinom, SN najčešće nisu opremljeni GPS (*Global Positioning System*) modulima za potrebe određivanja pozicije u prostoru (lokacije SN).

Pod lokalizacijom SN podrazumeva se proces određivanja njihove nepoznate pozicije u prostoru (lokacije). Za potrebe lokalizacije SN u WSN do sada je predložena primena velikog broja različitih postupaka posebno razvijenih za ove potrebe, [1, 2]. Za potrebe lokalizacije u WSN, ali i drugim bežičnim mrežama, često se koriste referentni (*beacon*) čvorovi (BN, *Beacon Node*) za koje je poznata tačna lokacija, a koji kroz komunikaciju sa ostalim SN omogućavaju lokalizaciju ovih SN sa većom tačnošću i preciznošću. Pri tome, tačna lokacija referentnih čvorova određuje se primenom ugrađenih GPS modula ili postavljanjem na unapred definisane pozicije.

Postupci za lokalizaciju SN u WSN mogu se klasifikovati na više načina. Jedan način klasifikacije je na distribuirane i centralizovane postupke. Distribuirani postupci su oni kod kojih se procena lokacije odvija zasebno u svakom SN mreže, tj. svaki SN određuje svoju lokaciju na osnovu podataka koje dobija od susenih čvorova ili referentnihh čvorova. Nasuprot tome, kod centralizovanih postupaka definiše se glavni čvor, najčešće *sink,* koji sakuplja sve informacije na nivou mreže i obavlja procenu lokacija svih SN. Generalno, centralizovani postupci omogućavaju bolje rezultate u pogledu tačnosti i preciznosti lokalizacije na račun povećane potrošnje energije i složenosti postupka, [1]. Distribuirani postupci predstavljaju znatno jednostavnija rešenja sa stanovišta implementacije, a koje odlikuju znatno manji zahtevi u smislu razmene podataka (komunikacije) i potrošnje energije, ali po cenu nešto lošijih performansi lokalizacije, [1, 2]. Sa druge strane, klasifikacija postupaka za lokalizaciju može se izvršiti na *range-based* i *range-free* postupke za lokalizaciju. U *range-based* postupke ubrajamo one u kojima se primenom dodatnih hardverskih komponenti i/ili posebnih postupaka merenja, pod određenim pretpostavkama, određuje međusobno rastojanje SN, a nakon toga, primenom trilateracije ili multilateracije, i pozicija SN. Postupci za lokalizaciju u kojima se ne koriste dodatne hardverske komponente, niti se obavlja merenje međusobnih rastojanja SN za potrebe lokalizacije nazivamo *range-free* postupcima. Kod *range-based* postupaka za lokalizaciju, [1], procena rastojanja između SN i BN, odnosno međusobnog rastojanja SN obavlja se na osnovu merenja vremena prispeća signala (TOA, *Time of Arrival*), razlike vremena prispeća signala između SN (DTOA, *Difference Time of Arrival*), merenja ugla nailaska signala (AOA, *Angle of Arrival*) ili merenja snage signala na prijemu (RSSI, *Received Signal Strength Indicator*). U slučaju *range-free* postupaka, [1, 2] ne

Gorana Crnobrnja – Telenor d.o.o. Beograd, Omladinskih brigada 90, 11070 Novi Beograd, Srbija, (e-mail: gorana.crnobrnja@telenor.rs).

Kristina Josifović – Elektrotehnički fakultet, Univerzitet u Beogradu, Bulevar Kralja Aleksandra 73, 11020 Beograd, Srbija (e-mail: kristinajos@etf.bg.ac.rs).

Goran Marković – Elektrotehnički fakultet, Univerzitet u Beogradu, Bulevar Kralja Aleksandra 73, 11020 Beograd, Srbija (e-mail: gmarkovic@etf.bg.ac.rs).

koriste se navedeni parametri, već se kao osnovna informacija koriste podaci o konektivnosti SN sa drugim čvorovima mreže, pri čemu se svi podaci upisuju u tabelu rutiranja u svakom SN. Ovakvi postupci ne zahtevaju primenu dodatnog hardvera, što ih čini pogodnijim za implementaciju u WSN. Iako ova grupa postupaka za lokalizaciju ne obezbeđuje podjednako dobar kvalitet lokalizacije SN kao *range-based* postupci, oni zbog jednostavnosti i cene implementacije ipak predstavljaju dobar izbor u slučaju mreža u kojima se zahteva samo gruba procena pozicije SN. U najpoznatije *range-free* postupke mogu se svrstati sledeći postupci za lokalizaciju: *Centroid Localization Algorithm* (CLA), *Convex Position Estimation* (CPE), *Approximate Point in Triangulation* (APIT)*, Distance Vector – Hop* (DV-Hop), [1, 3-11].

Centroid Localization Algorithm (CLA), [3], predstavlja jedan od najjednostavnijih *range-free* postupaka. U CLA se koriste podaci o konektivnosti posmatranog SN i dostupnih BN. Prostorne koordinate SN određuju se kao centar oblasti formirane od BN, u čijem dometu se nalazi posmatrani SN, i to prostim usrednjavanjem prostornih koordinata ovih BN. Osim osnovnog CLA postupka predloženi su i unapređeni postupci zasnovani na CLA. *Weighted* CLA (WCLA) pripada ovom skupu unapređenih algoritama, ali kod njega se, za razliku od osnovnog CLA postupka, uvode dodatni težinski koeficijenti za određivanje procenta učešća dostupnih BN u proceni lokacije SN u zavisnosti od rastojanja BN i posmatranog SN, [12, 13]. Ipak, u WCLA postupku se obavlja merenje rastojanja BN i SN, odnosno procena rastojanja na osnovu vrednosti RSSI i predpostavljenog modela propagacije radio signala (tzv. *path-loss* model), te on spada u *range-based* postupke lokalizacije.

Osnovni DV-Hop (DVH) postupak, [6, 7], zasniva se na pristupu distribuirane lokalizacije SN. Osim osnovnog DVH postupka, predložene su različite modifikacije bazirane na ovom osnovnom DV-Hop postupku, a koje odlikuju neke prednosti u pogledu performansi lokalizacije, [7-9]. Osim toga, predložena je i posebna grupa postupaka za lokalizaciju koja je bazirana na kombinaciji jednostavnog CLA postupka i DVH postupka. Motiv za razvoj ove grupe postupaka je da se kombinovanjem prevaziđu nedostaci i istaknu prednosti DVH i CLA postupaka. Naime, CLA postupak ne uzima u obzir uticaj udaljenosti između BN i SN, odnosno činjenicu da što je BN bliži SN, to njegov uticaj na procenu lokacije SN treba da bude relativno veći, odnosno da takve BN treba smatrati prikladnijim za estimaciju lokalizacije datog SN, kao što je to slučaj u pomenutom WCLA postupku. Osim toga, za uspešnu primenu CLA postupka, kao i za prethodno navedene *range-based* postupke zasnovane na CLA postupku (npr. WCLA), [12, 13], neophodna je relativno velika gustina prostornog rasporeda BN kao i/ili njihov veliki domet, kako bi za svaki SN u mreži postojao dovoljno veliki skup BN u čijem se dometu dati SN nalazi, kako bi lokalizacija mogla da se izvrši. U slučaju WSN velikih dimenzija sa malom gustinom BN u senzorskom polju WSN, odnosno kada je domet BN relativno mali u odnosu na dimenzije mreže (što je slučaj posmatran u ovom radu) CLA i na njemu zasnovane postuke odlikuju relativno loše performanse lokalizacije pri čemu se javlja veliki procenat SN koje nije moguće lokalizovati.

Kako bi se pravazišao problem nedovolje konektivnosti SN sa dostupnim skupom BN, odnosno da bi se omogućilo da se pri primeni osnovnog principa lokalizacije na bazi CLA postupka u obzir uračuna uticaj rastojanja između BN i SN, predloženo je kombinovanje DVH i CLA postupaka. Tipični postupci ovog tipa su WCA (*Weighted Centroid Algorithm*) postupak, [10, 11], i *Improved* WCA (IWCA) postupak, [11]. U ovim modifikovanim postupcima procena rastojanja između BN i SN određuje se korišćenjem mehanizma DVH postupka, i proširuje mogući skup BN koji se mogu koristiti za potrebe lokalizacije na osnovu konektivnosti posmatranog SN kome se određuje lokacija i BN ostvarene *multi-hop* komunikacijom preko drugih SN u mreži.

Analizom prethodno navedenih WCA i IWCA postupaka za lokalizaciju uočeno je da oni ne obavljaju na najbolji mogući način procenu rastojanja SN i BN primenom DVH mehanizma, i stoga ne postižu očekivana poboljšanja kvaliteta lokalizacije. Daljom modifikacijom IWCA postupka, koja je predložena u ovom radu kao modifikovani IWCA (MIWCA) postupak, ostvarene su nešto bolje performanse lokalizacije SN u WSN većih dimenzija u scenarijima primene u kojima ne postoji dovoljna konektivnost SN sa BN neophodna za primenu CLA postupka. Komparativna analiza performansi svih ovde razmatranih *range-free* postupaka za lokalizaciju obavljena je korišćenjem numeričke analize izvedene putem specijalno razvijenog simulacionog modela WSN u okruženju programskog paketa MatLab, koja je izvršena za različite scenarije primene u pogledu broja BN (tj. gustine rasporeda BN u senzorskom polju), dometa BN i dometa SN. Rezultati sprovedene analize pokazali su da za analizirani slučaj niske konektivnosti postoje određene prednosti ovde predloženog MIWCA postupka u odnosu na druge posmatrane postupke.

U drugom poglavlju rada, dat je sažeti opis osnovnog (originalnog) DVH postupka i IDVH postupka za lokalizaciju, dok su u trećem opisani WCA, IWCA i MIWCA postupci na bazi kombinovanja CLA i DVH pristupa za potrebe *range-free* lokalizacije. Opis simulacionog modela, razvijenog i korišćenog za analizu performansi razmatranih postupaka za lokalizaciju za različite scenarije primene u WSN i osnovni rezultati analize, prikazani su u četvrtom poglavlju, dok su u poslednjem, petom poglavlju data zaključna razmatranja.

II. OSNOVNI I UNAPREĐENI DV-HOP POSTUPAK

A. Osnovni DV-Hop postupak

Osnovni DV-Hop postupak za lokalizaciju sadrži tri faze rada kroz koje vrši estimaciju lokacija svih SN u mreži. U prvoj fazi svaki BN obavlja plavljenje (*flooding*) mreže *beacon* paketima koji sadrže podatak o poznatim i tačnim lokacijama tog BN. Svaki čvor u mreži, BN ili SN, proverava koji od primljenih paketa od istog BN sadrži najmanji broj *hop*-ova, tj. koraka, od datog BN i taj paket zadržava, dok ostale odbacuje. Čvor inkrementira broj *hop*-ova iz izabranog paketa za 1 i prosleđuje paket dalje kroz mrežu. Cilj ove faze je da svaki čvor mreže formira tabelu rutiranja sa najmanjim brojem koraka (*hop*-ova) ka svim ostalim čvorovima mreže.

Sledeća, druga, faza služi za proračun prosečne dužine *hop*-a između BN u mreži, kao aritmetičke sredine svih *hop*-ova zabeleženih u tabeli rutiranja koji čine najkraću putanju između svaka dva BN. Svi BN šalju podatak o prosečnoj dužini *hop*-a ka svim SN u mreži, dok SN čuva i prosleđuje podatak o paketu koji je prvi primio, što osigurava da je sačuvao informaciju onog BN koji mu je najbliži. Nakon toga, svaki, k-ti SN određuje rastojanje od svih BN u mreži, d_{ik}, na osnovu relacije (2), koristeći kao parametre prosečnu dužinu *hop*-a za najbliži BN (koju je sačuvao), Hop_size_i, i broj hopova do ostalih BN u mreži, [6, 7],

$$Hop\_size_i = \frac{\sum_{j \neq i} \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}}{\sum_{j \neq i} h_{ji}}, \quad (1)$$

$$d_{ik} = Hop\_size_i \times h_{ik}, \quad (2)$$

gde je h_{ik} broj hop-*ova* (koraka) između i-tog BN i k-tog SN, a (x_i, y_i) i (x_j, y_j) su koordinate i-tog i j-tog BN, respektivno.

U trećoj fazi koristi se *Least Square* (LS) metod za potrebe određivanja lokacije, koju svaki SN sprovodi za sebe, a na osnovu izabranog skupa BN, i procene rastojanja do tih BN.

Iako poseduje određene kvalitete, DV-Hop algoritam ima i bitne nedostatke, a koji se ogledaju u tome da stvarni razmak između čvorova mreže bitno odstupa od procenjenih vrednosti u izrazima (1) i (2). Naime, procene rastojanja SN i BN iz izraza (2), i međusobnih rastojanja BN na osnovu izraza (1), predstavljaju samo grube procene, odnosno greška u proceni rastojanja nastaje usled toga što se koristi srednja dužina *hop*-samo za najbliži BN datom SN, dok je realna dužina svakog *hop*-a u mreži različita i zavisi od lokacija BN i SN.

Kako bi se smanjio uticaj navedene greške, razvijeni su modifikovani DV-Hop postupci. Jedan takav postupak opisan je u nastavku ovog poglavlja, dok su oni koji predstavljaju kombinaciju DVH i CLA postupaka opisani u sledećem poglavlju, u kome je definisan i MIWCA postupak.

B. Improved DV-Hop (IDVH) postupak

Improved DV-Hop (IDVH) postupak, [8], razvijen je u cilju poboljšanja tačnosti određivanja lokacije u odnosu na osnovni DVH postupak uz minimalne izmene koje su jednostavne za implementaciju. U odnosu na osnovni DVH postupak, izmene se kod IDVH postupka uvode u drugoj i trećoj fazi. U drugoj fazi, SN, ne određuje samo srednju dužinu *hop*-a do najbližeg BN, već usrednjava srednje dužine *hop*-a za sve BN, odnosno određuje se prosečna dužina *hop*-a $Hop\_size_{average}$, na nivou cele WSN, kao, [8],

$$Hop\_size_{average} = \frac{\sum_{i=1}^{N} Hop\_size_i}{N}, \quad (3)$$

gde N predstavlja ukupan broj BN, a Hop_size_i prosečnu dužinu *hop*-a za svaki BN iz izraza (1). Dodatno, u izrazu (2) za proračun rastojanja do svih BN u mreži koristi se samo prosečna dužina *hop*-a dobijena korišćenjem izraza (3). Treća faza se razlikuje po tome što se ne koristi LS metod, već se za lokalizaciju primenjuje 2-D *Hyperbolic* algoritam, [9].

III. MODIFIKOVANI POSTUPCI ZASNOVANI NA KOMBINACIJI DVH I CLA POSTUPAKA I PREDLOG NOVOG MIWCA POSTUPKA

A. WCA i IWCA postupci bazirani na DV-Hop postupku

WCA postupak uveden je u cilju prevazilaženja nedostatka CLA postupka vezanog za to da se pri estimaciji lokacije SN koriste lokacije svih BN sa kojima je konektivan bez obzira na rastojanje BN i datog SN. U slučaju WCA postupka uvode se težinski koeficijenati, koji uzimaju u obzir udaljenost BN od SN. Odnosno, procena lokacije k-tog SN u slučaju WCA postupka data je na osnovu izraza, [10],

$$x_k = \frac{\sum_{i=1}^{N_k} w_{ik} \times x_i}{\sum_{i=1}^{N_k} w_{ik}}, \quad y_k = \frac{\sum_{i=1}^{N_k} w_{ik} \times y_i}{\sum_{i=1}^{N_k} w_{ik}}, \quad (4)$$

gde su (x_i, y_i) koordinate i-tog BN, w_{ik} težinski koeficijenti, dok je N_k je ukupan broj BN sa kojima je dati k-ti SN konektivan. Postupak se može primeniti kada u WSN postoji manji broj BN, jednostavan je za implementaciju i može se primeniti za svaki SN u mreži. Pri tome, težinski koeficijenti w_{ik} se određuju na osnovu primene prve faze DVH postupka, tj. nakon što svaki k-ti SN prikupi podatak o minimalnom broju *hop*-ova h_{ik} do i-tog BN, na osnovu izraza, [10, 11],

$$w_{ik} = \frac{1}{h_{ik}}, \quad (5)$$

Na osnovu opisa WCA postupka, očigledno je da što je veći broj *hop*-ova, to je BN udaljeniji od SN i tada je njegov relativan uticaj na procenu lokacije SN manji.

U slučaju kada je broj *hop*-ova između SN i BN u proseku znatno veći, a stvarno rastojanje između njih manje, dolazi do povećanja greške određivanja lokacije u slučaju primene WCA, [11]. Ovo se događa u slučaju kada u mreži postoji veći broj SN (tj. kada se poveća gustina prostornog rasporeda SN), ili kada se smanji broj BN u odnosu na broj SN u mreži. Stoga je predložen IWCA postupak koji treba da u navedenim uslovima da bolje rezultate lokalizacije u odnosu na WCA postupak. U slučaju IWCA postupka, u trenutku kada SN dobije podatak o minimalnom broju *hop*-ova do svakog BN, obavlja se sortiranje u rastućem poretku i bira se nekoliko BN sa najmanjim brojem koraka. Prosečna dužina *hop*-ova se računa na osnovu tih nekoliko izabranih BN, kao, [11],

$$Hop\_size_{k,IWCA} = \frac{\sum_{j=1}^{N_k} Hop\_size_j}{N_k}, \quad (6)$$

gde je N_k broj prethodno izabranih BN, a Hop_size_j je prosečna dužina *hop*-a za j-ti BN iz izraza (1). Težinski koeficijenti za k-ti SN i i-ti BN definisani su izrazom, [11],

$$w_{ik,IWCA} = \left(\frac{1}{h_{ik}}\right)^{1/n_k} \quad (7)$$

pri čemu je $n_k = Hop\_size_{k,IWCA}/r$, a r je domet SN. Sama estimacija prostornih koordinata SN obavlja se korišćenjem izraza (4) samo uz primenu $w_{ik,IWCA}$ umesto w_{ik}.

B. Predlog dodatne modifikacije IWCA postupka (MIWCA)

Ovde je predložena dodatna modifikacija IWCA postupka, označena kao MIWCA postupak, za potrebe lokalizacije SN u WSN velikih dimenzija. Na osnovu rezultata sprovedene numeričke analize za prethodno opisane DVH, IDVH, CLA, WCA i IWCA postupke, a koji su dati u narednom poglavlju, uočeno je da srednja greška lokalizacije pri primeni težinskih koeficijenata definisanih u WCA i IWCA postupcima ne opada na očekivani način u odnosu na CLA i DVH postupke. Dodatno, zapaženo je da jednostavan CLA postupak ostvaruje znatno bolje performanse u odnosu na DVH i IDVH postupke, u slučaju velike gustine i većeg dometa BN u mreži, dok se za mreže sa malim brojem (gustinom prostornog rasporeda) BN najbolji rezultati postižu primenom IDVH postupka, pri čemu CLA postupak za malu gustinu prostornog rasporeda i manje domete BN ne omogućava klasifikaciju značajnog procenta SN u mreži (SN nisu konektivni sa dovoljnim brojem BN). Pri tome, na osnovu detaljnije analize ponašanja CLA, DVH i IWCA postupaka uvedene su određene izmene.

U MIWCA postupku se na kraju prve faze DVH postupka, nakon što SN dobije informaciju o minimalnom broju *hop*-ova do svakog BN, obavlja sortiranje po rastućem broju *hop*-ova, i za dalji postupak lokalizacije odabira skup BN koj sadrži N_k članova. Pri tome, definisan je maksimalan broj BN u skupu, N_{max}, uz maksimalni broj od 3 *hop*-a do datog SN. Dodatno je postavljen uslov da se skup BN sa 2 ili 3 *hop*-a koristi samo ukoliko je njihov broj veći od broja BN na rastojanju od 1 *hop*-a (skup BN koji se koristi u okviru CLA postupka) pošto se zapaža da je ovaj uslov pretežno ispunjen samo ako je SN udaljen od ivice senzorskog polja (za SN bliske obodu senzorskog polja se pri korišćenju većeg broja udaljenih BN javlja porast greške lokalizacije – estimirana lokacija se pomera ka centru senzorskog polja). Nakon toga, umesto da se usrednjeva prosečne dužine hopova za N_k odabranih BN, izraz (6) iz IWCA postupka, usrednjava se broj *hop*-ova između N_i odabranih BN i posmatranog SN. U tom slučaju, težinski koeficijenti za k-ti SN i i-ti BN određuju kao,

$$n_{k,MIWCA} = Hop_{k,average} = \frac{\sum_{j=1}^{N_k} h_{kj}}{N_k}, \quad (8)$$

$$w_{ik,MIWCA} = \left(\frac{1}{h_{ik}}\right)^{1/n_{k,IMWCA}} \quad (9)$$

Nakon toga, proces estimacije lokacija SN sprovodi se na isti način kao i kod IWCA postupka.

IV. SIMULACIONI MODEL I REZULTATI NUMERIČKE ANALIZE

Za potrebe analize performansi navedenih postupaka za lokalizaciju razvijen je poseban simulacioni model u MatLab okruženju. Posmatrana je WSN dimenzija senzorskog polja 250 m×250 m, koju čine 5×5 kvadratne ćelije, sa $N_{BC} = 1, 2$ ili 4 pravilno raspoređenih BN i 8 SN slučajnog rasporeda po ćeliji (ukupno 200 u okviru WSN). Posmatrani su scenariji rada WSN u kojima je domet BN u komunikaciji sa SN bio vrednosti $R_{BC} \in \{35\text{ m}, 50\text{ m}, 100\text{ m}\}$, odnosno skup dometa komunikacije između SN $R_{SN} \in \{25\text{ m}, 35\text{ m}, 50\text{ m}\}$. Posmatran je model sa idealnom propagacijom, u kome su svi BN/SN na rastojanju manjem od dometa bežične komunikacije (dometa BN ili SN) konektivni, odnosno smatrano je da su svi potrebni podaci o konektivnosti čvorova dostupni svim SN.

Generisan je skup od 200 međusobno nezavisnih postavki WSN (*Monte-Carlo* eksperimenata) za svaki od scenarija, uz primenu skupa algoritama za lokalizaciju (DVH, IDVH, CLA, WCA, IWCA i MIWCA) realizovanih u potpunosti u skladu sa opisom u literaturi, odnosno opisom MIWCA datim u poglavlju III.B. Poređenjem estimiranih lokacija SN sa tačnim lokacijama, za sve analizirane postupke za lokalizaciju određena je srednja vrednost i varijansa greške lokalizacije za svaki scenario, pri čemu su rezultati u pogledu srednje greške lokalizacije prikazani u **TABELA 1**. U cilju preglednosti nisu dati rezultati za IWCA postupak, kojim su za sve scenarije dobijani nešto lošiji rezultati u odnosu na WCA postupak, uz slično ponašanje sa promenom parametara scenarija. Dodatno nisu dati rezultati za domet BN od 100 m za koji je za sve postupke za lokalizaciju i domete SN ostvaren lošiji kvalitet lokalizacije u odnosu na manje vrednosti dometa.

TABELA 1 - SREDNJA GREŠKA LOKALIZACIJE [m] U ZAVISNOSTI OD BROJA BN PO ĆELIJI (1, 2 ILI 4), DOMETA BN (35, 50) I DOMETA SN (25, 35, 50)

N_{BC} = 1	DVH	IDVH	CLA	WCA	MIWCA
R_{BC}35,R_{SN}25	19.0880	14.2422	-	58.3605	11.8822
R_{BC}35,R_{SN}35	16.2088	12.4639	13.8748	61.5776	12.3632
R_{BC}35,R_{SN}50	31.3221	33.0579	13.5601	67.5890	13.3322
R_{BC}50,R_{SN}25	18.1854	14.2433	13.4178	61.8151	13.5041
R_{BC}50,R_{SN}35	15.6736	11.6049	13.3510	61.7988	13.2818
N_{BC} = 2	**DVH**	**IDVH**	**CLA**	**WCA**	**MIWCA**
R_{BC}35,R_{SN}25	12.4056	14.1987	10.0848	59.3484	9.2323
R_{BC}35,R_{SN}35	16.4121	18.2241	10.1249	61.9049	8.9665
R_{BC}35,R_{SN}50	35.7326	37.4694	10.1550	67.9338	11.5163
R_{BC}50,R_{SN}25	17.4117	18.4477	10.1087	63.5789	10.3011
R_{BC}50,R_{SN}35	15.1720	16.2655	10.4007	63.7606	11.3761
N_{BC} = 4	**DVH**	**IDVH**	**CLA**	**WCA**	**MIWCA**
R_{BC}35,R_{SN}25	18.6379	19.7999	5.8622	61.3969	6.1920
R_{BC}35,R_{SN}35	20.1496	23.1325	5.9546	61.9694	7.1469
R_{BC}35,R_{SN}50	39.9884	40.7533	5.9111	68.1051	9.2172
R_{BC}50,R_{SN}25	27.4979	31.5759	8.5091	64.9559	8.5383
R_{BC}50,R_{SN}35	26.1945	29.8578	8.5418	64.9027	9.8999

Na osnovu dobijenih rezultata utvrđeno je da kod DVH, IDVH, IWCA postupaka vrednost srednje greške lokalizacije povećava ako se povećava broj BN u mreži, a što je posledica smanjenja prosečnog broja *hop*-ova između BN, odnosno između SN i skupa bliskih BN, što izaziva povećanje greške estimacije srednje dužine *hop*-a. Nasuprot tome, za CLA i MIWCA postupke srednja greška lokalizacije se smanjuje pri povećanju broja BN u WSN, zato što se javlja veći broj BN na malom rastojanju do svakog SN. Osim toga, u slučaju CLA postupka porast vrednosti dometa BN za isti broj BN u WSN omogućava poboljšanje kvaliteta lokalizacije samo kada

imamo manji broj BN u mreži (usled povećanja broja BN koji su konektivni sa SN), dok za veći broj BN (gustinu BN u mreži) srednja greška lokalizacije počinje da se povećava sa porastom dometa BN. U slučaju MIWCA postupka povećanje dometa BN i SN uglavnom izaziva pogoršanje uspešnosti lokalizacije (osim za slučaj srednje gustine i malog dometa BN - $N_{BC} = 2$ i $R_{BC} = 35m$ u **TABELA 1** kada se najbolji rezultati ostvaruju za srednji domet SN). Kada domet BN postane suviše veliki (100 m) broj *hop*-ova između BN drastično opada ili su oni direktno konektivni, dok broj BN koji je direktno konektivan sa posmatranim SN drastično raste, pa se ostvaruju izuzetno loši rezultati lokalizacije za sve razmatrane postupke. Analizaran je i uticaj dometa SN. Pokazuje se da povećanje dometa SN nema značajan uticaj na grešku lokalizacije kod CLA posupka, što je posledica toga što se u okviru ovog postupka ne ostvaruje komunikacija između SN, dok se kod DVH i IDVH postupaka uočava da su srednje vrednosti dometa (35 m) pogodne za mali broj BN u mreži ($N_{BC} = 1$) ili veće domete BN (npr. 50 m), dok se za manje vrednosti dometa BN (25 m) bolji rezultati ostvaruju kada imamo veći broj BN u mreži. Kao što je već navedeno, kod MIWCA postupka, povećanje dometa SN generalno izaziva pad kvaliteta lokalizacije, zbog pojave većih rastojanja do skupa BN koji se koriste za lokalizaciju datog SN.

Treba naglasiti da se u slučaju CLA postupka, ako se posmatra samo srednja greška lokalizacije ostvaruju veoma dobri rezultati, osim u slučaju kada je gustina rasporeda BN u mreži mala ($N_{BC} = 1$), a kada se za manje domete BN ($R_{BC} = 35m$) javlja slučaj da određen procenat SN ne može da bude lokalizovan (u **TABELA 1** je data srednja greška samo za one SN koji su uspešno lokalizovani). Dodatno, zapaža se da se za SN bliske središtu senzorskog polja ostvaruje mnogo bolji kvalitet lokalizacije, dok se za SN bliske ivicama senzorskog polja greška lokalizacije drastično povećava. Slično ponašanje se može zapaziti i u slučaju MIWCA postupka.

Procenjena vrednost srednje varijanse greške lokalizacije za posmatrani skup postupaka i parametre simulacionog modela prikazan je u **TABELA 2.** Uočava se da CLA i MIWCA postupci u pogledu uticaja parametara scenarija imaju slično ponašanje u pogledu varijanse greške, a koje je usklađeno sa onim koje se javlja za srednju grešku lokalizacije. Ostali posmatrani postupci imaju značajno veću (za neke scenarije i drastično veću) varijansu greške lokalizacije u odnosu na onu zabeleženu kod CLA i MIWCA postupaka.

Dodatno, sprovedena je i analiza performansi u pogledu osetljivosti na grešku poznavanja lokacije BN sa definisanom varijansom greške reda 0.25 m, čiji su rezultati dati u **TABELA 3**. Poređenjem rezultata u **TABELA 1** i **TABELA 3** za iste parametre scenarija (broj BN, domet BN i domet SN), ukazuje na to da posmatrani *range-free* algoritmi, imaju veoma malu osetljivost (u smislu povećanja greške) na grešku lokacija BN.

Ukupno gledano, za scenarije primene sa malim i srednjim brojem BN (mala i srednja gustina prostornog rasporeda BN), a pogotovu u slučaju manjih vrednosti dometa BN, MIWCA postupak za lokalizaciju predstavlja najbolje rešenje od ovde posmatranih postupaka, pri čemu se za veće domete BN ostvaruju slične ili malo lošije karakteristike u poređenju sa CLA postupkom. Sa druge strane, za veliku gustinu BN u mreži, CLA postupak predstavlja najbolje rešenje, sa nešto lošijim performansama u slučaju većih vrednosti dometa BN kada postaje poredIv po kvalitetu sa MIWCA postupkom, odnosno obezbeđuje najbolji kvalitet lokalizacije SN u odnosu na ostale ovde posmatrane *range-free* postupke.

TABELA 2 - VARIJANSA GREŠKE LOKALIZACIJE [m^2] U ZAVISNOSTI OD BROJA BN PO ĆELIJI (1, 2 ILI 4), DOMETA BN (35, 50) I DOMETA SN (25, 35, 50)

N_{BC} = 1	DVH	IDVH	CLA	WCA	MIWCA
$R_{BC}35,R_{SN}25$	153.81	84.10	-	657.82	53.48
$R_{BC}35,R_{SN}35$	90.25	49.99	43.91	705.10	44.53
$R_{BC}35,R_{SN}50$	270.18	202.58	44.68	788.58	50.04
$R_{BC}50,R_{SN}25$	119.33	64.98	40.69	750.69	43.56
$R_{BC}50,R_{SN}35$	90.86	44.58	40.01	734.69	55.06
N_{BC} = 2	**DVH**	**IDVH**	**CLA**	**WCA**	**MIWCA**
$R_{BC}35,R_{SN}25$	51.71	72.32	28.82	680.55	31.63
$R_{BC}35,R_{SN}35$	108.87	145.05	29.75	701.31	35.41
$R_{BC}35,R_{SN}50$	355.76	445.65	29.05	813.89	54.61
$R_{BC}50,R_{SN}25$	101.14	147.67	38.59	744.43	41.99
$R_{BC}50,R_{SN}35$	79.28	103.42	38.88	770.38	45.51
N_{BC} = 4	**DVH**	**IDVH**	**CLA**	**WCA**	**MIWCA**
$R_{BC}35,R_{SN}25$	163.86	195.13	18.97	704.12	22.66
$R_{BC}35,R_{SN}35$	177.51	204.81	19.64	739.29	34.26
$R_{BC}35,R_{SN}50$	424.12	1680.3	20.42	832.75	52.99
$R_{BC}50,R_{SN}25$	318.53	174.14	45.04	753.56	46.26
$R_{BC}50,R_{SN}35$	304.85	222.56	44.84	762.06	63.94

TABELA 3 - SREDNJA GREŠKA LOKALIZACIJE [m] U ZAVISNOSTI OD BROJA BN PO ĆELIJI (1, 2 ILI 4), DOMETA BN (35, 50) I DOMETA SN (25, 35, 50), U SLUČAJU POSTOJANJA GREŠKE POZNAVANJA LOKACIJE BN

N_{BC} = 1	DVH	IDVH	CLA	WCA	MIWCA
$R_{BC}35,R_{SN}25$	19.0666	14.2335	-	58.3642	12.0627
$R_{BC}35,R_{SN}35$	16.1351	12.4345	13.8605	61.5708	12.3515
$R_{BC}35,R_{SN}50$	31.3169	33.0854	13.5638	67.5971	13.3420
$R_{BC}50,R_{SN}25$	18.1510	14.2367	13.4186	61.8090	13.5045
$R_{BC}50,R_{SN}35$	15.7471	11.6154	13.3619	61.8012	13.2881
N_{BC} = 2	**DVH**	**IDVH**	**CLA**	**WCA**	**MIWCA**
$R_{BC}35,R_{SN}25$	12.3757	14.2157	10.0930	59.3394	9.2446
$R_{BC}35,R_{SN}35$	16.4017	18.0945	10.1409	61.8967	8.9847
$R_{BC}35,R_{SN}50$	35.7450	37.4738	10.1566	67.9406	11.5280
$R_{BC}50,R_{SN}25$	17.3803	18.4719	10.1278	63.5921	10.3216
$R_{BC}50,R_{SN}35$	15.2435	16.2664	10.3958	63.7618	11.3714
N_{BC} = 4	**DVH**	**IDVH**	**CLA**	**WCA**	**MIWCA**
$R_{BC}35,R_{SN}25$	18.5963	19.8319	5.8582	61.4001	6.1874
$R_{BC}35,R_{SN}35$	20.0501	23.1473	5.9681	61.9742	7.1612
$R_{BC}35,R_{SN}50$	40.0328	40.9195	5.9287	68.1075	9.2344
$R_{BC}50,R_{SN}25$	27.5092	31.6167	8.5125	64.9596	8.5414
$R_{BC}50,R_{SN}35$	26.2151	29.9389	8.5415	64.8982	9.8950

V. ZAKLJUČAK

Na osnovu rezultata izvršene numeričke analize zaključuje se da nije moguće dati istovetan opšti zaključak o uticaju parametara scenarija primene za sve analizirane postupke za lokalizaciju, prvenstveno zato što određene grupe postupaka rade na različitim principima. Iako je očekivano (na osnovu literature) da ranije predloženi modifikovani postupci (IDVH,

WCA, IWCA) ostvaruju bolje rezultate od originalnog DVH postupka, tokom same numeričke analize pokazalo se da za određene scenarije primene (prevashodno u funkciji gustine prostornog rasporeda BN i dometa BN) postoje značajna odstupanja. Delimično i usled specifičnog pravilnog rasporeda BN u modelu mreže korišćenog za potrebe analize, predloženi MIWCA postupak ostvaruje bolje rezultate u pogledu srednje greške lokalizacije u odnosu na sve posmatrane postupke, osim u odnosu na CLA postupak. Pri tome, MIWCA postupak ima nešto bolje performanse za mreže sa manjim brojem i dometom BN, dok se u slučaju velike gustine BN u mreži najbolje performanse ostvaruju primenom CLA postupka. Kod oba ova postupka za lokalizaciju javlja se znatno manja greška lokalizacije za SN bliže centru a znatno veća za SN bliže ivicama senzorskog polja. MIWCA/CLA postupci za lokalizaciju za posmatrane scenarije primene poseduju znatno bolje performanse u odnosu na WCA/IWCA postupke, i vidno nadmašuju DVH i IDVH postupke.

Treba imati u vidu, da je CLA postupak najjednostavniji od ovde razmatranih postupaka, kao i to da ne zahteva dodatnu komunikaciju između SN, a samim tim ni potrošnju energije za potrebe razmene podataka koji se koriste za lokalizaciju. U tom smislu, on predstavlja izuzetno pogodno rešenje za WSN u kojima čvorovi imaju ograničenu raspoloživu energiju, ali se pri tome mora voditi računa da je njegova uspešna primena moguća samo ukoliko postoji veći broj gusto raspoređenih BN sa relativno velikim dometom. Svi navedeni postupci spadaju u klasu *range-free* postupaka i stoga za svoj rad ne zahtevaju dodatni hardver, što je takođe pogodno za primenu u WSN.

Pokazano je da je moguće izvršiti dodatne modifikacije ranije predloženih rešenja na bazi kombinovanja jednostavnog CLA postupka i originalnog DVH postupka, kojim se može generisati rešenje za potrebe lokalizacije pogodno za scenarije primene u kojima nije moguće na zadovoljavajući način primeniti originalne postupke. Ipak, treba naglasiti da je ovde prikazano rešenje (MIWCA postupak), iako omogućava uspešnu lokalizaciju u određenim uslovima primene u WSN, sa nešto boljim kvalitetom lokalizacije od drugih razmatranih protokola, prvenstveno razvijeno kako bi se ispitalo da li je moguće uspešno kombinovati pristupe na bazi CLA i DVH postupaka tako da se dobije rešenje pogodno za primenu u različitim uslovima rada. Konačno, može se zaključiti da je nophodno dalje istraživanje u cilju razvoja jednostavnog rešenja za lokalizaciju u WSN, kojim bi se između ostalog rešio problem lokalizacije SN bliskih ivicama senzorskog polja a koji postoji kod svih postupaka zasnovanih na primeni CLA i DVH koncepata lokalizacije.

ZAHVALNICA

Rad je delimično finansiran od Ministarstva prosvete, nauke i tehnološkog razvoja RS, kroz projekte tehnološkog razvoja 32028 i 32037.

LITERATURA

[1] I. F. Akyildiz, and M. C. Vuran, *Wireless Sensor Networks*, NewYork, USA, John Wiley & Sons Inc., 2010.

[2] T. He, C. Huang, B. M. Blum, J. A. Stanković and T. Abdelzaher, "Range-Free Localization Shemes for Large Scale Sensor Networks", Proc. MobiCom 2003, San Diego, California, USA, pp. 81-95, 14-19 September, 2003.

[3] X. Shen, Z. Wang, P. Jiang, R. Lin, and Y. Sun, "*Connectivity and RSSI Based Localization Scheme for Wireless Sensor Networks*". In: Huang DS., Zhang XP., Huang GB. (eds) Advances in Intelligent Computing. ICIC 2005, Lecture Notes in Computer Science, vol. 3645, Berlin, Heidelberg, Springer, 2005.

[4] L. Doherty, K. S. J. Pister, and L. E. Ghaoui, "Convex Position Estimation in Wireless Sensor Networks," Proc. 20th IEEE INFOCOM 2001, Vol. 3, Anchorage, USA, pp. 1655-1663, April, 2001.

[5] S. M. Hosseininirad, M. Niazi, J. Pourdeiliami, S. K. Basu, and A. A. Pouyan, "On improving APIT algorithm for better localization in WSN", *Journal of AI and Data Mining*, Vol.2, No 2, pp. 97-104, 2014.

[6] F. Wang, C. Wang, Z. Z. Wang, and X-Y. Zhang, "A Hybrid Algorithm of GA + Simplex Method in the WSN Localization," *International Journal of Distributed Sensor Networks*, Vol. 2015, Article ID 731894, pp. 1-9, July 2015, https://doi.org/10.1155/2015/731894.

[7] X. Du, *DV-Hop Localization Algorithms in Wireless Sensor Network*, Master of Engineering Thesis, Dalian University of Technology, 2012

[8] H. Chen, K. Sezaki, P. Deng, and H. C. So, "An Improved DV-Hop Localization Algorithm for Wireless Sensor Networks," Proc. *3rd IEEE Conference on Industrial Electronics and Applications,* Singapore, pp. 1557-1561, 2008.

[9] X. Chen, B. Zhang, "Improved DV-Hop Node Localization Algorithm in Wireless Sensor Networks," *International Journal of Distributed Sensor Networks*, Vol. 2012, Articele ID 213980, pp. 1-7, July 2012, https://doi.org/10.1155/2012/213980

[10] B. Zhang, M. Ji, and L. Shan, "A Weighted Centroid Localization Algorithm Based on DV-Hop for Wireless Sensor Network," Proc. 8th International Conference on Wireless Communications, Networking and Mobile Computing, Shanghai, China, pp. 1-5, 2012.

[11] A. Kaur, P. Kumar, and G. P. Gupta, "A weighted centroid localization algorithm for randomly deployed wireless sensor networks," *Elsevier Journal of King Saud University - Computer and Information Sciences*, Vol. 31, Issue 1, pp. 82-91, January 2019.

[12] M. Arun, N. Sivasankari, P. T. Vanathi, and P. Manimegalai, "Analysis of Average Weight Based Centroid Localization Algorithm for Mobile Wireless Sensor Networks," *Advances in Wireless and Mobile Communications*, Vol. 10, No. 4, pp. 757-780, 2017.

[13] Q. Dong, and X. Xu, "A Novel Weighted Centroid Localization Algorithm Based on RSSI for an Outdoor Environment," *Journal of Communications*, Vol. 9, No. 3, pp. 279-285, 2014.

ABSTRACT

Localization, as a process that determine the positions of sensor nodes (SN) within wireless sensor networks (WSN), is characterized by the SN coordinates (locations) estimation accuracy and precision. Various solutions for SN localization have been proposed, with the aim of improving quality measures in different ways. One of them is Distance Vector-Hop (DV-Hop) localization procedure. In this paper we give an overview of the original DV-Hop procedure, some modifications based on combining original DV-Hop procedure and a simple Centroid localization. We also propose a modified solution that should introduce further performance improvement. The detailed performance analysis for the observed localization procedures was performed. A specific simulation model is developed, in order to perform the performance analysis for the different WSN application scenarios. Based on the numerical analysis results the applicability of here considered *range-free* localization procedures in large-scale wireless sensor networks was examined.

The Improvement of Localization Procedures for WSN based on Combining of DV-Hop and Centroid Algorithms

Gorana Crnobrnja, Kristina Josifović and Goran Marković

Srednja verovatnoća greške po bitu pri prenosu modulisanog signala u FSO sistemu

Jelena Todorović, Branimir Jakšić, Petar Spalević, Mile Petrović i Ana Tošković

***Apstrakt*—U radu je izračunata i analizirana srednja verovatnoća greške po bitu (Average Bit Error Rate) signala u Free Space Optical sistemu modulisanim sa Differential Phase Shift Keying i Binary Phase Shift Keying šemom. Srednja verovatnoća greške po bitu je određena za slučaj atmosferskog kanala modelovanim sa Gamma-Gamma raspodelom u funkciji od intenziteta signala. Rezultati su osim u analitičkoj formi, predstavljeni i grafički, u zavisnosti od odnosa srednje optičke snage na prijemu i varijanse kanalnog šuma, a za različite dužine Free Space Optical linka i jačine atmosferske turbulencije. Analiziran je kvalitet signala na osnovu srednje verovatnoće greške po bitu za slabu, umerenu i jaku atmosfersku turbulenciju.**

***Ključne reči*— ABER - Average Bit Error Rate; BPSK - Binary Phase Shift Keying; DPSK - Differential Phase Shift Keying; FSO - Free space optical; Gamma-Gamma.**

I. UVOD

Neprekidan razvoj različitih servisa bežičnih telekomunikacionih sistema, dovodi do potrebe za proučavanjem i unapređenjem njihovih performansi. Osnovni zahtevi koji su prisutni u procesu unapređenja performansi bežičnih telekomunikacionih sistema su: obezbeđivanje velikih brzina prenosa, veliki kapacitet kanala i što veći domet veze sa što manjom verovatnoćom greške.

Free Space Optical (FSO) je komunikaciona tehnologija koja omogućuje bežični full-duplex gigabitni prenos podataka u okruženju [1,2]. FSO sistemi omogućuju prenos signala sa protokom od nekoliko Gb/s, dok mikrotalasne veze omogućuju protok od nekoliko Mb/s. FSO koristi širok frekvencijski spektar, otporan je na elektromagnetne smetnje i interferenciju sa susednim kanalima (zbog dobrog definisanog uskog snopa signala bez slabljenja snage), a pruža i visok stepen bezbednosti [1,3]. FSO linkovi su kratki i kreću se do 2.5 km. Na kvalitet prenosa signala u FSO sistemima najveći uticaj imaju atmosferske prilike. Sneg i kiša znatno imaju manji uticaj na kvaltet prenosa u odnosu na atmosferske turbulencije i maglu. Atmosferski efekti na FSO sistem mogu uzrokovati pojave kao što su širenje snopa, fluktuacije intenziteta i faze signala, scintilacija [4].

Postoji više modela za opisivanje prostiranja signala u FSO sistemima sa atmosferskom turbulencijom, među kojima se izdvajaju: Gamma-Gamma, Lognormal, K-raspodela, Rician, Rayleigh i Nakagami-m koje se modeluju u funkciji od intenziteta signala ili odnosa signal-šum (SNR - Signal Noise Ratio) [5]. Model Gamma-Gamma raspodele je predložen kao pogodan matematički model za slabu i jaku turbulenciju i pruža dobre rezultate u poređenju sa eksperimentalnim [6].

Za opisivanje kvaliteta prenosa signala u FSO sistemima koriste se mnoge performanse među kojima se izdvajaju odnos signal šum (SNR), verovatnoća otkaza (OP - Outage Probability), kapacitet kanala (CC - Channel Capacity) i srednja verovatnoća greške po bitu (ABER - Average Bit Error Rate). Da bi se odredio ABER potrebno je poznavati funkciju gustine verovatnoće (PDF - Probability Density Function). PDF primljenog signala je po prirodi nestacionarna i zavisi od parametara atmosferske turbulencije, tako da ABER pruža dobru sliku o kvalitetu prenosa signala u FSO sistemima.

Za prenos signala u FSO sistemima se koristi više modulacionih formata, među kojima su najpopularniji On-Off keying (OOK), Binary Phase Shift Keying (BPSK), Differential Phase Shift Keying (DPSK) i Frequency Shift Keying (FSK) [7]. OOK format je relativno jednostavan, ali ne daje superiorne performanse. BPSK format zahteva složenu implementaciju u demodulatoru, ali daje veoma dobre performanse. DPSK format zahteva manje složeniju implementaciju od BPSK, performanse su lošije od BPSK ali znatno bolje od OOK [8]. U literaturi [3, 5-9] su predstavljene mnogobrojne statističke karakteristike signala opisanih različitim modelima raspodele i različitim modulacionim šemama u FSO sistemima. U radu [7] izračunat je ABER signala modelovanog Gamma-Gamma raspodelom u funkciji od SNR-a za više modulacionih formata.

U ovom radu određeni su analitički izrazi za ABER za Gamma-Gamma model kanala u funkciji od intenziteta signala, a pri prenosu u FSO sistemima korišćenjem DPSK i BPSK modulacione šeme. U drugom poglavlju dat je model sistema koji se razmatra i polazni izrazi za računanje ABER-a. U trećem poglavlju su predstavljeni izračunati izrazi za ABER u zatvorenom obliku za oba modela modulacionih šema. U četvrtom poglavlju dati su i komentarisani numerički rezultati dobijeni za ABER za signale modulisanim DPSK i BPSK šemama, a za različite stepene atmosferske turbulencije i dužine FSO linka. Peto poglavlje je Zaključak.

Jelena Todorović – Fakultet tehničkih nauka, Univerzitet u Prištini, Knjaza Miloša 7, 38220 Kosovska Mitrovica, Srbija (e-mail: jelena.todorovic@pr.ac.rs).

Branimir Jakšić – Fakultet tehničkih nauka, Univerzitet u Prištini, Knjaza Miloša 7, 38220 Kosovska Mitrovica, Srbija (e-mail: branimir.jaksic@pr.ac.rs).

Petar Spalević – Fakultet tehničkih nauka, Univerzitet u Prištini, Knjaza Miloša 7, 38220 Kosovska Mitrovica, Srbija (e-mail: petar.spalevic@pr.ac.rs).

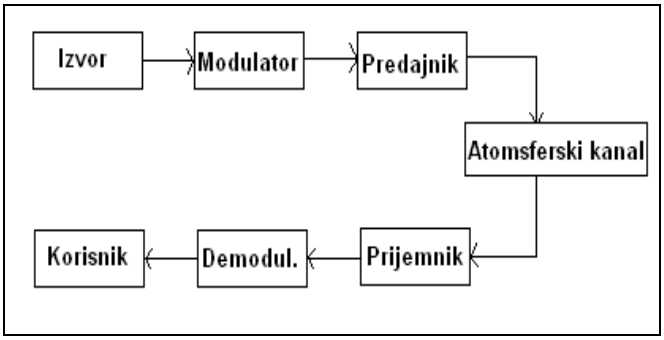
Mile Petrović – Fakultet tehničkih nauka, Univerzitet u Prištini, Knjaza Miloša 7, 38220 Kosovska Mitrovica, Srbija (e-mail: mile.petrovic@pr.ac.rs).

Ana Tošković – Visoka tehnička škola strukovnih studija, Nušićeva 6, 38227 Zvečan, Srbija (e-mail: a.toskovic@vts-zvecan.edu.rs).

II. MODEL SISTEMA

Razmatran je tipičan FSO sistem koji se sastoji od modulatora, predajnika, atmosferskog kanala, prijemnika i demodulatora, kao što je prikazano na Sl. 1.

Signal se iz izvora dovodi na modulator u kome se primenjuje jedan obilk digitalne modulacione šeme. U ovom slučaju su razmatrane dve modulacione šeme: DPSK i BPSK. Predajnik šalje modulisani signal koji se prenosi kroz atmosferski kanal modelovan sa Gamma-Gamma raspodelom. Za oba slučaja modulacije na prijemnoj strani sistema određuje se srednja verovatnoća greške po bitu – ABER.



Sl. 1. Model FSO sistema.

Gustina verovatnoće - PDF za Gamma-Gamma model u funkciji od intenziteta je data izrazom [9]:

$$f(I)=\frac{2(\alpha\beta)^{\frac{\alpha+\beta}{2}}}{\Gamma(\alpha)\Gamma(\beta)}I^{\frac{\alpha+\beta}{2}-1}K_{\alpha-\beta}\left(2\sqrt{\alpha\beta I}\right), \quad (1)$$

gde $\Gamma(\cdot)$ predstavlja Gamma funkciju [10, Eq. 8.310.1], a $K\upsilon(\cdot)$ je modifikovana Beselova funkcija druge vrste υ-tog reda [10, Eq. 8.407]. Parametri α i β predstavljaju efektivan broj vrtloga velikih i malih razmera, respektivno. To su parametri atmosferske turbulencije koji se za propagaciju ravanskih talasa i "zero inner scale" mogu izraziti kao [9]:

$$\alpha=\left[e^{\frac{0.49\sigma_R^2}{\left(1+1.11\sigma_R^{12/5}\right)^{7/6}}}-1\right]^{-1}, \quad \beta=\left[e^{\frac{0.51\sigma_R^2}{\left(1+0.69\sigma_R^{12/5}\right)^{5/6}}}-1\right]^{-1} \quad (2)$$

gde σ_R^2 predstavlja Rojtovu varijansu koja se koristi za određivanje intenziteta optičkog signala usled atmosferske turbulencije, a definiše se kao:

$$\sigma_R^2=1.23C_n^2k^{7/6}L^{11/6}. \quad (3)$$

Parametar C_n^2 označava indeks refrakcije koji se koristi kao mera za jačinu turbulencije. Za horizontalne putanje propagacije parametar C_n^2 se smatra konstantnim sa srednjim vrednostima od 10^{-17} $m^{-2/3}$ do 10^{-13} $m^{-2/3}$ za kanale od slabe do jake turbulencije, respektivno. Parametar k je talasni broj, koji se definiše kao $k = 2\pi/\lambda$ sa talasnom dužinom λ, dok je L rastojanje između predajnika i prijemnika, odnosno dužina propagacije optičkog signala.

Srednja verovatnoća greške po bitu u zavisnosti od fluktuacije intenziteta optičkog signala kada se prenos u FSO sistemu vrši preko DPSK može se izraziti kao [11]:

$$P_e=\frac{1}{2}\int_0^\infty e^{-\frac{P_T}{\sigma_N}}f(I)dI, \quad (4)$$

gde P_T predstavlja srednju optičku snagu na prijemu, dok je σ_N varijansa kanalnog šuma. Funkcija $f(I)$ predstavlja gustinu verovatnoće (PDF) primljenog signala određene raspodele koja se koristi za opisivanje modela kanala. U ovom slučaju je u pitanju Gamma-Gamma raspodela predstavljena sa (1).

Srednja verovatnoća greške po bitu u zavisnosti od fluktuacije intenziteta optičkog signala kada se prenos u FSO sistemu vrši preko BPSK može se izraziti kao [11]:

$$P_e=\frac{1}{2}\int_0^\infty \operatorname{erfc}\left(\frac{P_T}{\sigma_N}I\right)f(I)dI, \quad (5)$$

gde je $\operatorname{erfc}(\cdot)$ komplementarna funkcija greške [10, Eq. 8.250.4], a $f(I)$ predstavlja funkciju gustine verovatnoće Gamma-Gamma raspodele.

III. ANALITIČKI REZULTATI

Zamenom PDF za Gamma-Gamma model (1) u (4) dobija se:

$$P_e=\frac{(\alpha\beta)^{\frac{\alpha+\beta}{2}}}{\Gamma(\alpha)\Gamma(\beta)}\int_0^\infty I^{\frac{\alpha+\beta}{2}-1}e^{-\frac{P_T}{\sigma_N}I}K_{\alpha-\beta}\left(2\sqrt{\alpha\beta I}\right)dI. \quad (6)$$

Kako bi se izračunao izraz u zatvorenom obliku za ABER, modifikovana Beselova funkcija druge vrste $K\upsilon(\cdot)$ se predstavlja preko Meijer G funkcije na sledeći način [10, Eq. 9.34.3]:

$$K_\nu(x)=\frac{1}{2}G_{0,2}^{2,0}\left[\frac{x^2}{4}\middle|\begin{matrix}-\\(\nu-2),\ -(\nu-2)\end{matrix}\right]. \quad (7)$$

Takođe, koristi se i relacija za transformaciju eksponencijalne funkcije u Meijer G funkciju [12, Eq. 8.4.3]:

$$e^{-x}=G_{0,1}^{1,0}\left[x\middle|\begin{matrix}-\\0\end{matrix}\right]. \quad (8)$$

Primenom (7) i (8), izraz za izračunavanje ABER-a pri DPSK (6) se svodi na:

$$P_e = \frac{(\alpha\beta)^{\frac{\alpha+\beta}{2}}}{\Gamma(\alpha)\Gamma(\beta)} \int_0^\infty I^{\frac{\alpha+\beta}{2}-1} \times G_{0,1}^{1,0}\left[\frac{P_T}{\sigma_N} I \middle| \begin{matrix} - \\ 0 \end{matrix} \right] \times \times \frac{1}{2} G_{0,2}^{2,0}\left[\alpha\beta I \middle| \begin{matrix} - \\ \frac{\alpha-\beta}{2}, -\frac{\alpha-\beta}{2} \end{matrix} \right] dI \quad . \tag{9}$$

Primenom [13, Eq. 07.34.21.0011.01] u (9) dobija se rešenje u zatvorenom obliku za ABER za slučaj prenosa korišćenjem DPSK modulacione šeme:

$$P_e = \frac{(\alpha\beta)^{\frac{\alpha+\beta}{2}}}{2\Gamma(\alpha)\Gamma(\beta)} \left(\frac{P_T}{\sigma_N} \right)^{-\frac{\alpha+\beta}{2}} \times \times G_{1,2}^{2,1}\left[\frac{\alpha\beta}{\frac{P_T}{\sigma_N}} \middle| \begin{matrix} 1-\frac{\alpha+\beta}{2} \\ \frac{\alpha-\beta}{2}, -\frac{\alpha-\beta}{2} \end{matrix} \right] \quad . \tag{10}$$

Zamenom PDF za Gamma-Gamma model (1) u (5) dobija se:

$$P_e = \frac{(\alpha\beta)^{\frac{\alpha+\beta}{2}}}{\Gamma(\alpha)\Gamma(\beta)} \int_0^\infty I^{\frac{\alpha+\beta}{2}-1} \operatorname{erfc}\left(\frac{P_T}{\sigma_N} I \right) K_{\alpha-\beta}\left(2\sqrt{\alpha\beta I} \right) dI \, . \tag{11}$$

Kako bi se izračunao izraz u zatvorenom obliku za ABER, komplementarna funkcija greške erfc(·) se predstavlja preko Meijer G funkcije [14, Eq. 06.27.26.0003.01]:

$$\operatorname{erfc}(x) = 1 - \frac{x}{\sqrt{\pi}} G_{1,2}^{1,1}\left[x^2 \middle| \begin{matrix} 1/2 \\ 0, \ -1/2 \end{matrix} \right] \, . \tag{12}$$

Primenom (7) i (12), izraz za izračunavanje ABER-a pri BPSK (11) se svodi na:

$$P_e = \frac{(\alpha\beta)^{\frac{\alpha+\beta}{2}}}{2\Gamma(\alpha)\Gamma(\beta)} \int_0^\infty I^{\frac{\alpha+\beta}{2}-1} \times \times \left[1 - \frac{P_T I}{\sigma_N \sqrt{\pi}} \times G_{1,2}^{1,1}\left[\left(\frac{P_T}{\sigma_N} I \right)^2 \middle| \begin{matrix} 1/2 \\ 0, \ -1/2 \end{matrix} \right] \right] \times \times G_{0,2}^{2,0}\left[\alpha\beta I \middle| \begin{matrix} - \\ \frac{\alpha-\beta}{2}, -\frac{\alpha-\beta}{2} \end{matrix} \right] dI \quad . \tag{13}$$

Primenom [13, Eq. 07.34.21.0012.01] dobija se rešenje u zatvorenom obliku za ABER u slučaju BPSK modulacione šeme:

$$P_e = \frac{1}{2} - \frac{1}{2\sqrt{\pi}\alpha\beta\Gamma(\alpha)\Gamma(\beta)} \frac{P_T}{\sigma_N} \times \times H_{3,2}^{1,3}\left[\left(\frac{P_T}{\alpha\beta\sigma_N} \right)^2 \middle| \begin{matrix} \left(\frac{1}{2},1\right), \ (-\alpha,2), \ (-\beta,2) \\ (0,1), \ \left(-\frac{1}{2},1\right) \end{matrix} \right] \tag{14}$$

gde $H_{p,q}^{m,n}(\cdot)$ predstavlja Fox H funkciju, koja je, ustvari, generalizacija Maijer G funkcije. Fox H funkcija se u specijalnom slučaju svodi na Maijer G funkciju [13, Eq. 07.34.26.0008.01].

IV. NUMERIČKI REZULTATI

Za potrebe numeričkog proračuna posmatran je FSO sistem na talasnoj dužini λ = 875 nm. Razmatrana su tri tipa atmosferske turbulencije: slaba, umerena i jaka, sa indeksima refrakcije $C_n^2 = 6 \cdot 10^{-15}$ m$^{-2/3}$, $C_n^2 = 2 \cdot 10^{-14}$ m$^{-2/3}$ i $C_n^2 = 1.2 \cdot 10^{-13}$ m$^{-2/3}$, respektivno. Posmatrana su dva slučaja za rastojanja prijemnika od predajnika L=1 km i L=2 km.

U Tabeli 1 date su vrednosti parametara α i β dobijene za slučajeve koji se razmatraju.

TABELA I
VREDNOSTI PARAMETARA ATMOSFERSKE TURBULENCIJE.

C_n^2		$6 \cdot 10^{-15}$	$2 \cdot 10^{-14}$	$1.2 \cdot 10^{-13}$
L=1 km	σ_R^2	0.23	0.77	4.65
	α	10.28	4.80	4.49
	β	8.77	3.08	1.25
L=1.5 km	σ_R^2	0.48	1.63	9.79
	α	6.05	4.02	5.64
	β	4.47	1.89	1.10
L=2 km	σ_R^2	0.82	2.76	16.58
	α	4.68	4.07	6.85
	β	2.93	1.47	1.05

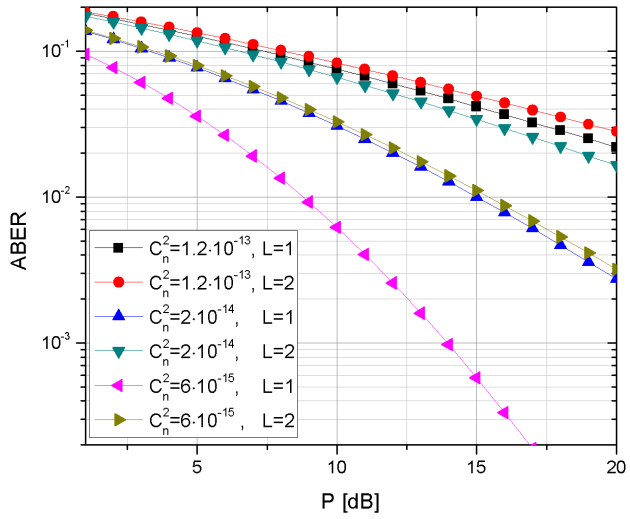
Na osnovu dobijenih analitičkih izraza u zatvorenom obliku za ABER pri DPSK i BPSK modulaciji, (10) i (14), respektivno, predstavljeni su grafici promene ABER u zavisnosti od odnosa $P = P_T/\sigma_N$. Grafici za ABER Gamma-Gamma modela kanala za DPSK i BPSK modulaciju su dati na Sl. 2 i Sl. 3, respektivno.

Sa datih slika se može videti da sa porastom odnosa P dolazi do smanjenja srednje verovatnoće greške po bitu. ABER brže opada za niže stepene atmosferske turbulencije u odnosu na jaku turbulenciju. Sa povećanjem turbulencije dolazi do povećenja ABER-a. Takođe, sa Sl. 2 i Sl. 3 se može videti da veće vrednosti dužine FSO linka dovode do povećanja ABER-a.

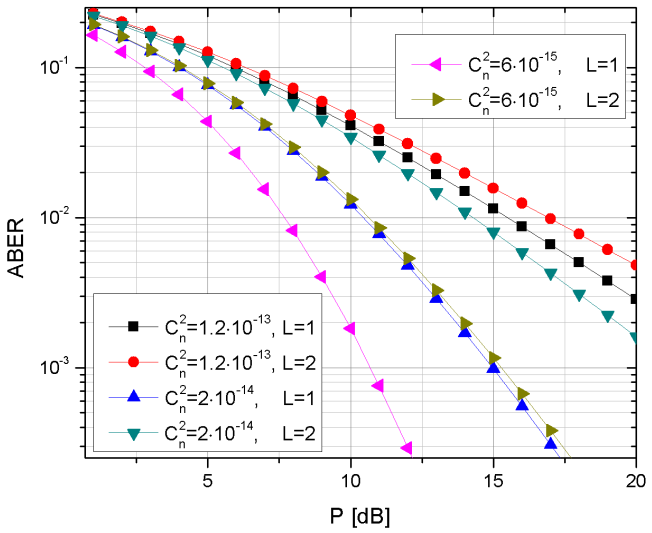
Kod DPSK modulacione šeme vrednosti ABER-a su reda 10^{-1} i kreću se sve do P > 20 dB pri jakoj turbulenciji, pri

umerenoj turbulenciji istog reda su do 15 dB, a pri slaboj turbulenciji do 9 dB. Da bi se ABER smanjio na red 10^{-3} potrebna je znatno viša snaga, za jaku i umerenu turbulenciju znatno više od 20 dB. Za slabu turbulenciju je potrebna snaga od 14 dB kako bi se ABER spustio na red 10^{-3}.

Takođe, može se videti da je srednja verovatnoća greške po bitu kod BPSK modulacije reda 10^{-1} do snage od 15 dB za jaku atmosfersku turbulenciju, a za umerenu turbulencije je do 10 dB. Kod slabe turbulencije ABER je reda 10^{-1} do snage od 7 dB. ABER prelazi iz reda 10^{-2} u red 10^{-3} kod jake turbulencije tek za $P > 20$ dB, kod umerene turbulencije za 15 dB, a kod slabe turbulencije za 12 dB.



Sl. 2. ABER za Gamma-Gamma model kanala pri DPSK modulaciji.

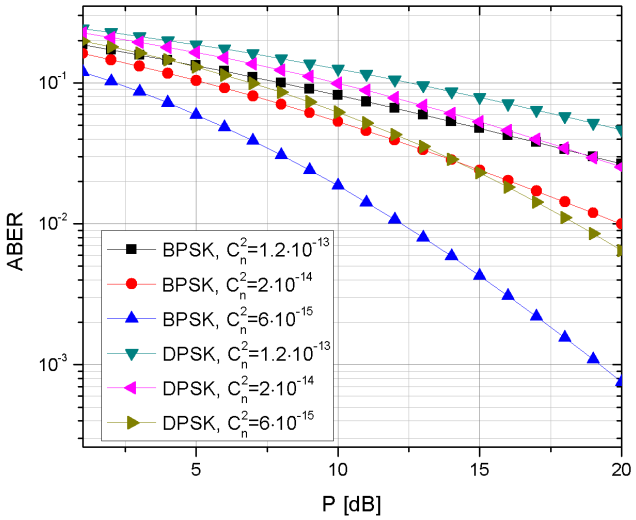


Sl. 3. ABER za Gamma-Gamma model kanala pri BPSK modulaciji.

Na osnovu dobijenih rezultata za srednju verovatnoću greške po bitu može se zaključiti da se bolje karakteristike prenosa dobijaju za BPSK modulacionu šemu. Kod BPSK modulacije je potrebno mnogo manje snage (tj. odnosa srednje optičke snage na prijemu i varijanse kanalnog šuma) nego kod DPSK modulacije da bi se dobio isti red ABER-a.

Na Sl. 4 je prikazana promena ABER-a u funkciji od $P = P_T/\sigma_N$ usled slabe, umerene i jake turbulencije za DPSK i BPSK modulacionu šemu. Rastojanje od predajnika do prijemnika iznosi L=1.5 km.

Sa Sl. 4 se može videti da se veće vrednosti ABER-a dobijaju korišćenjem DPSK nego BPSK modulacione šeme. Odnosno, BPSK daje bolje performanse FSO sistema. Takođe, može se videti da kod BPSK, ABER znatno brže opada sa porastom snage nego kod DPSK modulacione šeme. Pri višim odnosima $P = P_T/\sigma_N$ postoje dominantnije razlike ABER-a za DPSK i BPSK modulacionu šemu. Ta razlika je veća za niže stepene atmosferske turbulencije, dok je za jake atmosferske turbulencije razlika približno konstantna duž celog opsega odnosa $P = P_T/\sigma_N$.



Sl. 4. Poređenje ABER-a za DPSK i BPSK modulaciju.

V. ZAKLJUČAK

U radu su predstavljene i analizirane dobijene vrednosti ABER-a za signale modulisane DPSK i BPSK modulacionom šemom koji se prostiru kroz FSO kanal modelovan sa Gamma-Gamma raspodelom u funkciji od intenziteta. Dobijeni rezultati su grafički predstavljeni kako bi se video uticaj slabe, umerene i jake atmosferske turbulencije i dužina FSO linka na kvalitet prenosa signala.

Koristeći predstavljene rezultate može se predvideti ponašanje realizacija FSO sistema za različite modele modulacionih formata i u različitim propagacionim okruženjima, što omogućava projektantima sistema mobilnih prenosa da za željene performanse sistema naprave racionalna sistematska rešenja.

ZAHVALNICA

Ovaj rad je rađen u okviru istraživačkih projekta Ministarstva nauke i tehnološkog razvoja Republike Srbije: TR32023 i TR35026.

LITERATURA

[1] V. Stamatios, "Next Generation Intelligent Optical Networks - From Access to Backbone," USA: Springer, 2008

[2] I. I. Kim, B. McArthur, and E. Korevaar, "Comparison of laser beam propagation at 785 nm and 1550 nm in fog and haze for optical wireless communications," Proc. SPIE Opt. Wireless Communications, vol. 4214, pp. 26-37, 2001.

[3] K. Wakafuji and T. Ohtsuki, "Performance analysis of atmospheric optical subcarrier-multiplexing systems and atmospheric optical subcarrier-Modulated code-division multiplexing systems," Journal of Lightwave Technology, vol. 23, no. 4, pp. 1676-1682, Apr. 2005.

[4] L. C. Andrews, and R. L. Phillips, "Laser beam propagation through random media," 2nd ed. Bellingham, Wash.: SPIE Press, 2005.

[5] B. Barua, T. A. Haque, and Md. R. Islam, Error Probability Analysis of Free-Space Optical Links with Different Channel Model under Turbulent Condition," International Journal of Computer Science & Information Technology (IJCSIT), vol 4, no 1, pp. 245-258, Feb 2012. DOI: 10.5121/ijcsit.2012.4119

[6] H. E. Nistazakisa, V. D. Assimakopoulos, and G. S. Tombras. "Performance estimation of free space optical links over negative exponential atmospheric turbulence channels," Optik, vol. 122, no. 24, pp. 2191-2194, 2011.

[7] M. I. Petkovic, N. M. Zdravkovic, C. M. Stefanovic, G. T. Djordjevic, "Performance analysis of SIM-FSO system over Gamma-Gamma atmospheric channel," Proceedings of International Scientific Conference on Information, Communication and Energy Systems and Technologies ICEST 2014, pp. 19-22, 2014

[8] H. Zhang, H. Li, C. Hao, "Performance Analysis for BPSK, DPSK and OOK-Based FSO System in Atmospheric Turbulence Conditions," International Journal of Simulation - Systems, Science & Technology, vol. 17, Iss. 36, pp. 371-376, 2016. DOI 10.5013/ IJSSST.a.17.36.37

[9] H. G. Sandalidis, T. A. Tsiftsis, and G. K. Karagiannidis, "Optical Wireless Communications With Heterodyne Detection Over Turbulence Channels With Pointing Errors," Journal of Lightwave Technology, vol. 27, Iss. 20, pp. 4440-4445, 200. DOI 10.1109/JLT.2009.2024169

[10] I. S. Gradshteyn and I. M. Ryzhik, "Table of Integrals, Series, and Products," 7th Ed., USA: Elsevier Academic Press, 2007.

[11] Dj. Bandjur, B. Jaksic, S. Panic, M. Bandjur, A. Matovic, and E. Mekic, "Transmission Over Kappa-Mu Fading Channels with Gamma Distributed Random Line-Of-Sight Components," Rev. Roum. Sci. Techn.– Électrotechn. et Énerg., vol. 62, no. 2, pp. 179–184, 2017

[12] A. P. Prudnikov, Y. A. Brychkov, and O. I. Marichev, "Integral and Series," 2nd Ed., Moskva: Fizmatlit, 2003.

[13] The Wolfarm Functions Site: MeijerG functions. [Online] Available: http://functions.wolfram.com/PDF/MeijerG.pdf

[14] The Wolfarm Functions Site: Erfc functions. [Online] Available: http://functions.wolfram.com/PDF/Erfc.pdf

ABSTRACT

In this paper, the average bit-error rate of the signal in the free-space optical system modulated with differential phase shift keying and binary phase shift keying scheme is calculated and analyzed. The average bit-error rate is determined in the case of an atmospheric channel modeled with a Gamma-Gamma distribution in function of signal intensity. The results are presented both analytically and graphically, depending on the rate of the average received optical power and the channel noise variance for different lengths of the free-space optical link and the strength of the atmospheric turbulence. The quality of the signal was analyzed based on the average bit-error rate for weak, moderate and strong atmospheric turbulence.

Average Bit Error Rate at Modulated Signal Transmission in FSO System

Jelena Todorović, Branimir Jakšić, Petar Spalević, Mile Petrović and Ana Tošković

Achilles - MARS: Modular Chess System

Vladan Vučković

***Abstract*— This paper has intention to describe the original construction and infrastructure of the modular chess system Achilles (version 2019) - MARS consisting and connecting several complex subsystems. The central part of this system is high-performance 30 million per position distributed chess engine Achilles 2019. The system MARS is an old system renewed in 2014 and now connected to Achilles as the interactive module and option to play against professional chess players.**

***Index Terms*—Artificial Intelligence, Computer Chess, Electronics.**

I. Introduction

During May, 1997 the field of research in artificial intelligence happened to be an important event. In a direct duel IBM chess supercomputer Deep Blue beat world champion grandmaster Garry Kasparov. It was the first event of its kind, the first sign of possible machine intelligence supremacy over the human intelligence in certain specific areas of human activity. Considering the doubts about the result, many researchers in this field started to scrutinize this very interesting field. After the match, IBM showed no more interest in computer chess, so there were no more ways to test this chess supercomputer and its chess power.

IBM Deep Blue vs. Kasparov rematch has never happened again. After 1997, some researchers started to construct machines with intention to repeat the success of IBM Deep Blue. A joint feature of many of these projects is the orientation of standard PC technology and the X86 machine instruction set as the basis of the chess computer. With intensive development of multithread and multicore technologies which enable the use of internal parallelism in the CPU has become possible to construct programs with grandmaster chess rating on standard PC machines. On the other hand, local area network PC clusters opened up new possibilities of distributed processing data and thereby enhanced power compared to the software on a single processor. At any rate, the simultaneous development of technology, accelerating CPU speed and architecture in the sense of parallelization has brought huge benefits to computer chess.

Development of the FPCGA logic circuits opened the possibility to implement some critical parts of chess algorithms in hardware, which is largely utilized in the system Hydra. Hydra was a leading chess system in the certain period of development, and its successes were crowned with a convincing victory over the grandmaster Michael Adams, with score of 5.5 - 0.5 (2005). The direction to which lead the combination of PC architecture and FPCGA hardware circuits with this system became very popular. However, similar to the IBM Deep Blue, the Hydra proved to be a complex and expensive system for development and maintenance. PC technology which primarily have processing power only concentrated in CPUs, has considerably more intense development comparing to FPCGA, so it was proved in a short time that software modifications are significantly faster than hardware. After a while, mainstream direction for the optimal chess system was entirely based on multiprocessor and network PC technology. Today's computer chess technology development especially in the ICGA computer chess championships shows the absolute domination of this direction. Also, the modern PC chess engines have reached the grandmaster playing strength.

However, 1997 chess match was undoubtedly the initial for the author, and gave the directions for further research and the start of working on creating its own system of chess grandmaster strength. To start the project of such great scale with no initial material resources and organized institutional support, or opportunities for fast commercialization, for an individual seemed like an insurmountable task. However, in the laboratory environment, with PII 400 Mhz, somewhere around 2000 year, author had begun this extensive work. In subsequent years the author has developed all of the components for the complex cluster chess system, the entire structure, and application and communication software. Since 2001, when the program *Geniss Axon* debuted on single processor, the system evolved in parallel multiprocessor program *Achilles* and in January 2007 definitely overcome the chess grandmaster strength and scored its first victory against a grandmaster in the official tournament. If we look at the ELO rating directly, Axon-Achilles progresses with 1400 ELO points in its 15 years of development.

In the meantime, the program has played a lot of official matches and tournaments, with rating FIDE chess players, connected to the world's largest chess server *Chessbase* and through it have played 3 *FreeStyle* tournaments. At last tournament, Achilles haven`t lose a match (with one win and all others matches draws). It is also important to mention the collaboration with German company *Arena* following the authors PhD theses, which was defended on 24.10.2006, in the field of artificial intelligence and computer chess – the first of that kind in author's country.
The next big event was a match against grandmaster Igor Miladinovic in July 2007, which ended with the 3.5-0.5 victory of Achilles system. The author continued developing his cluster chess system to the newest version which defatted GM Miroslav Miljkovic in two games TV match (1.5-0.5).

This paper has intention to describe the original construction of modular cluster chess system Achilles (version 2019) consisting and connecting several complex subsystems, including MARS. The system core is high-

Vladan Vučković is with the Faculty of Electronic Engineering, University of Niš, 14 Aleksandra Medvedeva, 18000 Niš, Serbia (e-mail: vladanvuckovic24@gmail.com; vladan.vuckovic@elfak.ni.ac.rs).

performance 30 million per position distributed chess engine and details of its construction is presented in Section 2. The second module is presented in Section 3 - named MARS. It is an old system renewed in 2014 and now connected to *Achilles* as the module and option to play against professional chess players. This paper will be concluded with final remarks.

II. ACHILLES 2019 SYSTEM

Nowadays, there are a set of standard algorithms including MiniMax, Alfa-Beta, Negamax, PVS, Null Move for constructing powerful chess engines [1]. Also, there are other supporting and background algorithms responsible for maximizing the engine play strength. Common feature of these algorithms is that all of them have MiniMax procedure as basic code. Also, there are some pioneering efforts to mimic the grandmaster way of thinking in chess [2].

By definition the MiniMax function is recursive search-tree algorithm supporting the artificial decision-maker in logic games generally. Using algorithms to search that content implementing MiniMax procedure, the value of the terminal nodes is prolonged in reverse direction - through the tree to the root node. Even in the earliest stages of computer chess research a significant negative effect that may occur in the final stage of evaluation nodes was observed. This effect is called *horizon effect* because of fatal errors that could be generated trying to evaluate terminal dynamic positions [3].

To minimize time needed for calculation of the complex tree search in parallel computer environment, **we introduce the method of the candidate moves analogue to the human thinking process in chess.** Simplified, we reduce the set of playable moves at the root node only to few moves using the complex heuristic. Then the parallel machines process only the sub-trees generated from these moves.

The search trees are almost completely independent, in order that is possible to implement the parallel search for each candidate move using the authors' Achilles cluster chess engine as the basic infrastructure [4]. This implementation accelerates the complete process and increases the depth of the searching at the same time.

A. The Parallelization Method Using Original Methods Of The Candidates Move

Transition on working in multiprocessor environment led to the need of solving the sequence of problems on conception, implementation and communicational levels [5],[6]. The use of the parallel system, which consists of gathering of standard single processor PC machines connected by LAN network, implements new solutions which are not covered enough in the literature.

Exploring this area, the author discovered the new theoretical conception which was the base for making a completely new parallelization algorithm on distributive machines using vector parallelization methods. If we observe the basic chess engine plan [5], which is covered on multiple levels in this paper, we can register clear separation of 3 phases: the main, ALFA_BETA/PVS/null-move algorithm [7], relies on quiescence tracer terminal knots, which invokes evaluator in its every knot.

(basic tree >> quiescence searcher >> evaluator)

In fundamental plan the linear conception of the chess engine is clearly visible and that conception is actually based on the reduction principle – the basic position is reduced through the main tree to the sequence of the terminal nodes, which are further reduced with quiescence procedure which then declines to the sequence of static nodes which are evaluated.

But, the human chess master has different approach:

(basic tree >> candidate move selection >> calculation of limited selected lines)

Our new conception, that follows last direction, can be efficient only if new processors are used in different lines - therefore in multiprocessor environment. The main problem here lies in the fact that communication problems approximately grows with the square of number of processors and that is not favorably. On the other hand, according to the suggested new conception operation of all processors is undependable and asynchronous in a way that from a hardware perspective it is only necessary to establish a classic distributive network between the processors, without the need for multiprocessor motherboards.

B. Achilles – parallel chess system structure

Achilles chess system represents the parallel system of the new generation which is based on distributed parallelization with preselecting of candidate moves [6]. The first version, developed in 2005, consisted of 8-processor system, based on AMD X64 machines, operated on 1,8 Ghz and on Axon II program in separate knots, generates over 95% of cases, one external vector level, which means intensification the basic machines play level in practical game for over 200 ELO points. The parallel machine plays on the level which can achieve ratings of over 2800 ALO points in matches with shorter time control (up to 25 minutes per player). From the theoretical aspect, if the network with 82=64 processors would be at disposal, a machine with 2 external levels could be implemented by the same method of parallelization, so that increase of the power in relation with the rating of the Axon II program would be approximately 400 ELO points; this means that a parallel machine of such type could play with strength of over 3000 ELO points. The visual design of the Achilles system which performs distributed parallelization is shown in Fig. 1:

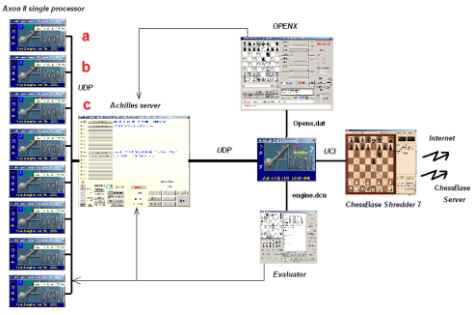


Fig. 1. Achilles engine diagram.

The connection with separate Axon II machines which operate on single processor systems is performed via LAN. This form of communication and synchronization [8] allows

that machines do not have to be on the network with the same IP address, which also allows practical possibility for development of parallel systems with denary processors (Fig.2). Solving communicational problems for Achilles program system, the author conceived a universal system for coded chess transfer using standard UDP (*Universal Datagram Protocol*) protocols.

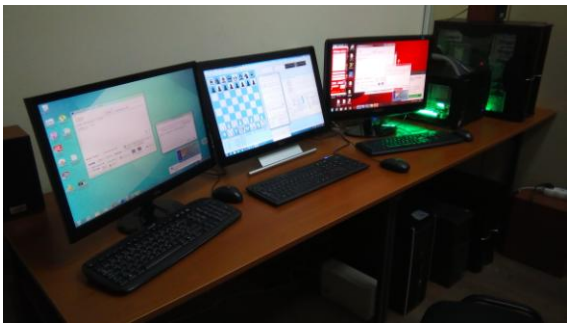


Fig. 2. Internal LAN connected Achilles cluster core.

C. Candidate moves distribution

In Achilles.2019 parallel infrastructure, candidate move principle implementation is rather simple. For example, if we use the game tree with three candidate moves (a, b, c) from the MiniMax tree root are simply redistributed to the first 3 nodes in Achilles system (Fig.1.).

In previous section we analyzed the game tree, so all three branches following the candidate moves could be calculated independently and concurrently. The acceleration is almost linear, depending on position. The Achilles could run concurrently many of candidate moves depending on the numbers of physical PC nodes in system [9]. The system GUI presented in Fig.3 is able to run up to 5 candidate moves in parallel.

Fig. 3. Achilles GUI with cluster control panels.

III. CHESS DEMONSTRATION SYSTEM MARS

Electronic devices for displaying scores in many sport branches, including athletic, swimming; winter and other sports which are in connection with surveying first appeared in the last quarter of the 20th century, based on technology to develop to updated digital electronics nowadays.

Approximately at the same time, the idea about automatic translation and demonstration in real time chess game, from the stage of important matches was born. On the first few there was an electronic chess semaphore supplied by sensor table. It was used at the world championship match Spassky–Petrossian in Moscow, 1969, and it was designed by Hungarian engineer Barfai. At the same time this match was demonstrated in Central Chess Club on one other board, designed by prof. Gleb Hlebutin from The Perm University. Displayed chess pieces between two boards were very different, much better than on the Russian display, but Hungarian system was supplied by the quickest sensor table.

After the match SSSR–The Rest of the World, organized in Belgrade 1970, an original idea to design chess demonstration board by the same principle as composition of the digits on the 7-segments deflector was born. We have found acceptable solution for chess esthetics and technology reproductive in one 10-segments deflector rosette. See down shown figures; each segment is supplied by two colored bulbs to represent white and black piece symbols (Fig.4).

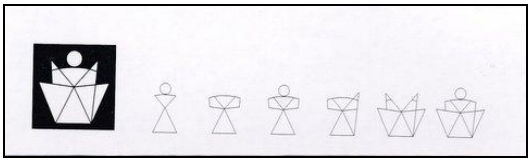


Fig. 4. Original solution for MARS demo-board pieces.

The first experimental used during the match Korchoy–Spassky was very successful, but for my assistants and me here at Electronic factory in Nish, Serbia, a problem to find satisfactory solution for sensor chess table appeared. The experiment have lead from different essays – resistance squares, read-relay, Hall sensors and inductive coupled oscillators circuits in the end. One of the main conditions was not to attain any difference between ordinary wooden chess pieces and at the same time provide invisible built-in sensors. The same rule was applied in designing of the table, respecting FIDE rules of proportion pieces diameter on square base – approximately 2 : 3. At the same time it was presented by newspaper and announced in one sensor TV chess demonstration system in Central Italian Chess Club in Milan, designed by Dottore De Bellis. Subsystems of the MARS (*Multifunctional Automatic Running Chess System*) are:

1) *Display*, implemented in classical track bulb technology, size 2 by 2 meters on pedestal with black and white collapsed times and move counters, and in-built power sources for energetic and electronic circuits.

2) *Tournament chess wooden table* with sensors manufactured by first class wood and well contrasted veneer in accordance with FIDE rules, supplied with hidden sensor system for piece recognition, control card and linked to control unit; and PC-486 as the Control Unit, with additional A/D convertor, as well as conventional SW for testing, initialization system and the main function – transmission and displaying tournament chess game in real time and authentic tournament conditions. There are possibilities for immediately replaying and printing games. MARS system has been well known for forty years now and it had been used on many chess events, including two matches at The Word Championship and two Olympics, too.

A. Conception of the digital sensor

The recognition of twelve (6 white + 6 black) different chess pieces is based on twelve resonant frequencies arranged by oscillator circuits built in each chess piece. The same kind of pices is trimmed at the same frequency. The

experiment defined range for the given problem conditions is from 490 to 610 kHz, by increment of 10 kHz. This is the most linear part of the frequency U(f) characteristics for all twelve piece oscillators and sensors matrix 8 x 8 squares positioned under chess table at distance of 7±0.3 mm. The best electromagnetic energy interchange between passive oscillators and sensors happens when the both spools are of the similar spool inductivity (in our case 1 mH approximately). Oscillators in the pieces are designed by high frequency tin phone wires on self keeping spool with one core in the middle and soft ferrite core for resonant frequency trimming. Sampling rate is 2000 per sec or 3 moves per sec. During experiments with different models of spool a few problems appeared. The first one was the impossibility of manufacturing circuits with similar varying characteristics. This problem was solved by careful sensor selection and finding the best position on table matrix to avoid any interferences. The second problem was to solve recognitions gaps between neighboring squares. The temperature stabilization of each passive oscillator in the chess pieces was of the greatest importance, because resonant frequencies depend a lot on temperature stabilization. The proper solution was found by ceramic heat stabilized capacitor, all heat isolated by wax. The last, but not the simplest problem appeared in reproduction of several hundreds of chess pieces in low series laboratory manufactured.

One of the best SW improvements was designed with one initial system utility, which assured 12 x 64 referent levels for all different pieces and chess table square, saved and simply modified by system on reset procedure. This utility is provide by real-time control of this referent level matrix to avoid unregistered unstable signals and blinking of the chess symbols on PC screen and simultaneous on large display, too. The inductive coupled sensor schema, showed bellow (Fig.5):

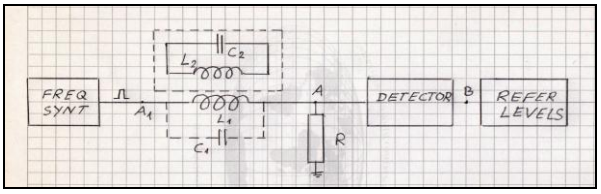


Fig. 5. Original sensor solution for MARS chess board.

IV. Conclusion

This paper presented some of original improvements embedded into the basic MiniMax procedure with direct implementation in author's Achilles computer chess program. Parallel version of the MiniMax procedure, based on candidate move principle, improves basic function and increases the total playing strength of the chess engine. There are some similar theoretical approaches, but according to the author's knowledge, there are no published data for the implementation of the results.

This idea exploits the independences in game tree calculating, as shown in the second Section, to accelerate the game tree calculation in distributed environment. To simplify, after determination of the candidate moves, all of them are distributed to different nodes in Achilles cluster.

After that, the game tree for each candidate move is calculated concurrently. This method is similar to human grandmaster way of thinking and makes decisions during the game of chess. The solution is proved in direct implementation in *Achilles* tournament practice (Fig.5). The future prospect will be to increase the amount of embedded chess knowledge in the phase of candidate move generating.

According to some tests, parallel system *Achilles.2019* has ELO rating above 3100. According to modern tendencies in computer chess, there is plenty of room for further improvement of the system by adding more processors up to maximum capacity, increasing the speed of operation of basic modules using faster multiprocessors, complete switching to 1Gbit network, and improving the level of the *Axon.2019* software, especially its evaluation function.

Appendix

Fig. 5. Complete Achilles-Mars connection and system demonstration.

Acknowledgement

This paper is supported by the interdisciplinary project III44006 of the Ministry of Science and Technology of the Republic of Serbia. Also, the author thanks Eng. Andro Mošić, the constructor of the MARS system, for the presenting some original details about his system.

References

[1] Slate D. J. and Atkin. L. R. Chess 4.5 – "The Northwestern University Chess Program", *Chess Skill in Man and Machine* (ed. P. W. Frey), pp. 82-118. Springer-Verlag, New York, N. Y. 2nd ed. 1983. ISBN 0-387-90815-3., 1977.

[2] Barry Smyth, Padraig Cunningham "Advances in Case-Based Reasoning: 4th European Workshop" EWCBR'98, Springer, Dublin, Ireland, September 1998. ISBN 3-540-64990-5.

[3] Vučković V. "The Theoretical and Practical Application of the Advanced Chess Algorithms", *PhD Theses*, The Faculty of Electronic Engineering, The University of Nis, Serbia, 2006.

[4] Vučković V. *Axon/Achilles* experimental chess engines information could be find at: http://axon. elfak. ni. ac. rs , http://chess. elfak. ni. ac. rs, 2007.

[5] Vučković V. "The Compact Chessboard Representation", ICGA Journal,Volume 31, Number 3, Tilburg, The Netherlands, ISSN 1389-6911. pp. 157- 164., 2008.

[6] Vučković V. "The Method of the Chess Search Algorithms Parallelization using Two-Processor Distributed System", The Scientific Journal Facta Universitatis, Series Mathematics and Informatics, Volume 22, Number 2, Niš, ISSN 0352-9665. pp. 175-188., 2007.

[7] Althöfer, I. An Incremental Negamax Algorithm. *Advances in Computer Chess 5* (ed D.F. Beal), pp. 31-41. Elsevier, Amsterdam. ISBN 0-444-87159-4, 1989.

[8] Šolak R. and Vučković V. ,"Time Management During a Chess Game", *ICGA Journal*, Volume 32, Number 4, Tilburg, The Netherlands, ISSN 1389-6911. pp. 206- 220., 2010.

[9] Vučković V. "Advanced Chess Algorithms and Systems", Monography, Zadužbina Andrejević, Biblioteka Dissertatio, ISSN 0354-7671, Belgrade, 2011.

The Potential of Using EEG Data in Evaluation of Visual Short-Term Memory Test Results

Milos Antonijevic, Miodrag Zivkovic, Sladjana Arsic, Aleksandar Jevremovic

***Abstract*— In this research we tried to determine the possible correlation between participant' emotional state while doing the Visual short-term working memory test in regard to the test results he achieved. We have analyzed data gathered from thirteen subjects doing the test with two images. First results show that, by analyzing EEG data, we can achieve accuracy of 61.66% in classifying instances in regard to Correctness of the answer class. Also, by comparing overall changes in participant's emotional state, we were able to conclude that there is a significant decrease in stress, engagement, relaxation and focus when participant switched from viewing the image to answering the questions.**

***Index Terms*—visual short-term memory, human-computer interaction, classification, EEG.**

I. Introduction

The main goal of this research is to determine the possible correlation between participant' emotional state while doing the Visual short-term working memory test in regard to the test results he achieved.

For the purpose of this paper, we have gathered data of human-computer interaction by using two sensors: EEG device and computer mouse. For synchronized sensor data processing we used developed HCI-MAP (Human-computer Interaction Monitoring and Analytics Platform) architecture. In this paper, we described our first results of the quantitative and qualitative analyses based on EEG data.

Visual short-term memory (VSTM) is defined as the ability to remember a small amount of visual information, such as colours, shapes and similar, during a short period of time [1]. The multicomponent model of working memory was introduced by Baddeley and Hitch [2,3]. The latest version of the model [4] consists of three systems, which include components for keeping and processing information.

Neuropsychological assessment enables the testing and assessment of VSTM. The most frequently used practices include classic direct and indirect numbers test from Wechsler's scale, NEPSY test by Korkman, Kirk and Kemp (from 1998), continuous performance test (CPT), memory malingering test (TOMM), visual organization test (VOT), test of variables of attention (TOVA) and Tower of London test. These tests measure not only visual short-term memory, but also short-term memory, reaction speed, working memory, visual scanning, the perception of the environment, remembering of the context, naming, distinguishing and speed of data processing [5].

VISMEM is very important and frequently used test and it is created using the classic TOMM test [6]. The test consists of showing an image to the subject for the limited period of time. During this time, the subject has a task to look at the image and memorize the context and remember as many details from the image as possible. After the given time expires, the image is removed and the subject gives answers to different questions about the image [7].

Electroencephalography (EEG) is a non-invasive method of tracking changes in electrical voltage of brain neurons during a defined time interval. EMOTIV EPOC+ device [8] was used for measuring the variability of emotional characteristics of the subject, depending on the changes in the surrounding environment. The manufacturer has developed the algorithms for extracting values of six emotional states (Interest, Engagement, Excitement, Stress, Relaxation, and Focus) from raw EEG data that we used in this research [9].

II. Related Work

In the study in [10], the authors investigated the emotional states of students, represented in the form of frustration and excitement, that occurred as a result of feedback information gathered from intelligent tutoring systems (ITS). By analysing the obtained data, the authors developed a system that enabled student emotions to be anticipated and the feedback information to be modified accordingly. Similarly, emotional reactions to different visual stimuli were examined in two independent works [11,12].

The goal of the experiment described in [13] was the detection of the level of pleasure. The authors also tested a method for correcting the robot's behaviour in order to increase the pleasure level. Although the experiment was carried out on a small number of participants (four males), a correct classification was achieved in an average of 79.2% of cases.

In one of our papers, we described some steps towards applying artificial intelligence and EEG signals for the improvement of electronic assessments [14]. The first analyses point to the possibility of using certain question types in electronic tests in order to influence the psychological state of students during assessments. For example, by inserting "funny" questions with one obvious correct answer,

M. Antonijevic is PhD Candidate with the Informatics and Computing Department, University Singidunum, Danijelova 32, 11000 Belgrade, Serbia (e-mail: mantonijevic@singidunum.ac.rs).

M. Zivkovic is with the Informatics and Computing Department, University Singidunum, Danijelova 32, 11000 Belgrade, Serbia (e-mail: mzivkovic@singidunum.ac.rs).

S. Arsic is with Academy of Vocational Studies, 35230 Cuprija, Serbia (e-mail: sarsic101@gmail.com).

A. Jevremovic is with the Informatics and Computing Department, University Singidunum, Danijelova 32, 11000 Belgrade, Serbia (e-mail: ajevremovic@singidunum.ac.rs).

this system can decrease the students' stress. Furthermore, the most interesting questions are the easy ones, while the focus of the student can be increased by using "impossible" questions with no correct answers.

III. Applications

For collecting and synchronizing data from different sensor devices and client applications, in this research we used HCI-MAP (Human-Computer Interaction and Monitoring Platform) platform [15]. We have developed a separate application for each sensor with the possibility to send data to the platform over the HTTP(S) protocol and the HCI-MAP API. Each application is implemented as a web application which uses the same interface for sending data (Fig. 1).

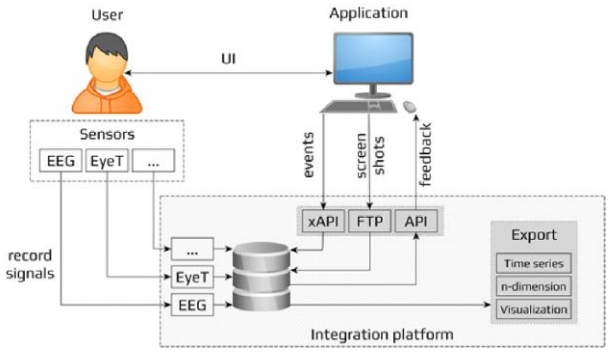


Fig. 1. The HCI-MAP architecture.

HCI-MAP platform is used for synchronization of gathered data from client applications and various sensors; data aggregation and processing in real time; and returning of obtained results in suitable formats for further analysis by computer or interpretation by humans. Besides the sensors, using the same interface, a platform can receive information from user applications.

VSTM application has been implemented as a modern interactive Web application. Used technologies include HTML5, CSS and JavaScript. The application contains three main sections: initialization screen, a screen with the image to be memorized and questions. User has a limited time to remember as many details from the image as possible. After the available time expires, the image will be removed and questions will be shown (Fig. 2).

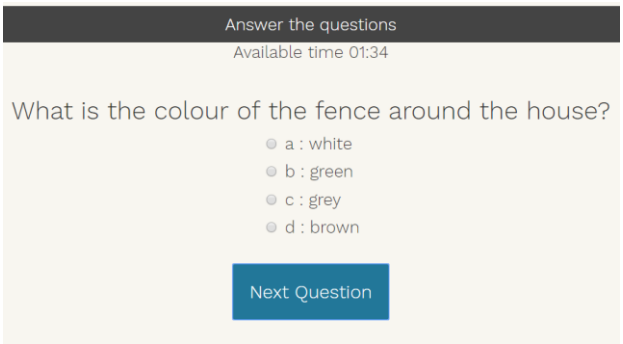


Fig. 2. Screen with a question in VSTM application.

IV. Experiment

Thirteen subjects, divided into two groups, took part in the experiment. Each subject was presented with two images during two recording sessions, without taking a break. The first image is shown in an isolated environment, and the second image is shown in front of the audience. The first group of subjects was presented with image A as the first image, and image B as the second image. The second group of subjects was presented with reversed order of images (image B first, image A second).

Fig. 3. Participants wearing EEG equipment during the experiment.

At the beginning of the testing, the subject is presented with an image for one minute. After that, the subject has two minutes to answer 10 questions about image details.

Due to a communication problems, we have 23 valid data sets for analyses (eleven full sets and one containing partial data from participant 12 when viewing image A second in a row).

V. Results

Gathered data was organized by participant number, question number, and correctness of the answer.

In this paper, we tried to classify data using Correctness of the answer class. In this case, an instance can belong to the one of three classes - viewing the IMAGE, correct answer (TRUE) or wrong answer (FALSE). We did not use *time duration* feature in any of classification algorithms because the time interval for viewing the image (which is 60 seconds) is greatly bigger from the intervals of answering the questions. It should also be noted that we are dealing with imbalanced dataset (number of instances belonging to different classes is significantly different).

Having that in mind, it can be said that there are not any significant classification results achieved using *Correctness of the answer* attribute as class. The best result is achieved by using **K-nearest neighbors classifier**, **with k = 3**, where we have more correct than incorrect classifications in two classes - IMAGE and TRUE (Table 1), while with all other used classification algorithms there are more correct than incorrect classifications in only one class (TRUE). K-nearest neighbors

classifier gives 61.66% of correctly classified instances using cross-validation and 61.63% of correctly classified instances with 66% training set.

TABLE I
CORRECTNESS OF THE ANSWER, K-NEAREST NEIGHBORS CLASSIFIER (K = 3) RESULTS

A. Cross-validation

IMAGE	TRUE	FALSE	<-- classified as
16	5	2	IMAGE
5	119	32	TRUE
3	50	21	FALSE

B. Training set with 66% of data

IMAGE	TRUE	FALSE	<-- classified as
4	1	1	IMAGE
3	43	12	TRUE
2	14	6	FALSE

After performing the classification attempts, we have tried to analyze the changes in emotional state features by doing qualitative data analysis. When analyzing EEG data, we have tried to determine the relation between individual feature and the class by comparing its average values for each of three possible classes (IMAGE, TRUE, FALSE). After that, we have selected the features with the biggest difference in average values and tried to group them based on belonging to the same emotional state. We made a comparison between the class *Viewing the image*, on one side, and *Wrong* and *Correct answer* classes on the other (Table 2).

TABLE II
COMPARISON OF FEATURE' AVERAGE VALUES FOR CLASS VIEWING THE IMAGE AND CLASSES WRONG AND CORRECT ANSWER

Average Difference in %	Feature	Emotional state
-32.28	Stress Max	STRESS
-25.29	Eng Max	ENGAGEMENT
-15.85	Rel Max	RELAXATION
-13.58	Foc Max	FOCUS

The results from the Table 2. show that there is a significant decrease in **stress**, **engagement**, **relaxation** and **focus** when the participant switched from viewing the image to answering the questions. For example, maximal stress value was, in average, 32.28% higher while user was viewing the picture.

VI. CONCLUSION

For this paper, we gathered data from thirteen subjects and two used sensors for HCI data collection - EEG device and computer mouse. Raw EEG data was then transformed to six emotional states (Interest, Engagement, Excitement, Stress, Relaxation, and Focus) that were used in different classification attempts in regard to the *Correctness of the answer* class. The best result is achieved by using K-nearest neighbors classifier with the accuracy of 61.66%.

In our qualitative analysis, we were able to conclude that there is a significant decrease in (average) stress, engagement, relaxation and focus when participant switched from viewing the image to answering the questions.

Our future work will include both quantitative and qualitative analysis of possible correlation between EEG data and the order of the image that was presented to the user, i.e. if the user was in the presence of the audience during the testing or not. Also, we will try to determine if there is correlation between participant's emotional state and the type of the image he was viewing during the test (image A or image B).

REFERENCES

[1] A. D. Baddeley, "The episodic buffer: a new component of working memory", Trends in Cognitive Science, vol. 4, no. 11, pp. 417-423, 2000.

[2] A.D Baddeley, G.J. Hitch, "Working memory", G.H. Bower (ed.): The psychology of learning and motivation, New York: Academic Press., pp. 47-90, 1974.

[3] A.D. Baddeley, "Working memory", New York, Oxford University Press, 1986.

[4] A. D. Baddeley, "The episodic buffer: A new component of working memory?", Trends in Cognitive Sciences, vol. 4, pp. 417-423, 2000.

[5] T.P. Alloway, G.A. Alloway, "Investigating the predictive roles of working memory and IQ in academic attainment", Journal of Experimental Child Psychology, vol. 106, no. 1, pp. 20-29, 2009.

[6] T. N. Tombaugh, "Test of memory malingering: TOMM", North Tonawanda, NY: Multi-Health Systems, 1996.

[7] M. R.Asato, J. A. Sweeney, B. Luna, "Cognitive processes in the development of TOL performance", Neuropsychologia, vol. 44, no. 12, pp. 2259-2269, 2006.

[8] "Emotiv." [Online]. Available: https://www.emotiv.com.

[9] "Advanced EEG Technology - Backed by Science." [Online]. Available: https://www.emotiv.com/the-science/.

[10] P. S. Inventado, R. Legaspi, M. Suarez, and M. Numao, "Predicting Student Emotions Resulting From Appraisal of Its Feedback," Res. Pract. Technol. Enhanc. Learn., vol. 6, no. 2, pp. 107–133, 2011.

[11] P. Bobrov, A. Frolov, C. Cantor, I. Fedulova, M. Bakhnyan, and A. Zhavoronkov, "Brain-computer interface based on generation of visual images," PLoS One, vol. 6, no. 6, 2011.

[12] M. K. Petersen, C. Stahlhut, A. Stopczynski, J. E. Larsen, and L. K. Hansen, "Smartphones get emotional: Mind reading images and reconstructing the neural sources," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 6975 LNCS, no. PART 2, pp. 578–587, 2011.

[13] E. T. Esfahani and V. Sundararajan, "Using Brain–Computer Interfaces To Detect Human Satisfaction in Human–Robot Interaction," Int. J. Humanoid Robot., vol. 08, no. 01, pp. 87-101, 2011.

[14] M. Antonijevic, G. Shimic, A. Jevremovic, M. Veinovic, S. Arsic, "The Potential for the Use of EEG Data in Electronic Assessments", Serbian Journal of Electrical Engineering, vol. 15, no. 3, pp. 339-351, 2018.

[15] A. Jevremovic, S. Arsic, M. Antonijevic, A. Ioannou, N. Garcia, "Human-Computer Interaction Monitoring and Analytics Platform - Wisconsin Card Sorting Test Application", HealthyIoT 2018 - 5th EAI International Conference on IoT Technologies for HealthCare, Guimaraes, Portugal, November 21- 23, 2018.

Deep Learning in Development of Model-Dependent Diagnostic: Recognition of Detector Characteristics in Measured Responses

Miroslava Jordović Pavlović, Dragan Markushev, Slobodanka Galović, Marica Popović

***Abstract*— Deep learning has successfully been implemented in various domains, including photoacoustics. The collection and creation of massive datasets creates new possibilities. Deep learning methods, when applied on massive datasets, are able to extract very useful patterns. This can lead to solutions to many problems. In this paper we discuss and develop deep learning application for the recognition of a detector influence pattern on recorded responses of a measurement chain in model-dependent experimental measurements. This enables the fast calibration of the method, which is necessary for its further application in the characterization or scanning of the examined objects with satisfactory accuracy. Frequency gas-microphone photoacoustic measurements were taken as the case study. The paper presents three models for the solution of instrument influence on true signals in photoacoustic experiments. We analyze the influence of neural network depth and the number of outputs on the prediction accuracy, and then we discuss the choice of the optimal solution.**

***Index Terms*—Deep learning; regression; massive dataset; photoacoustics; model-dependent diagnostic; microphone.**

I. INTRODUCTION

Deep learning is the area of machine learning which has seen the most intensive growth in the past few years. By bringing new techniques, algorithms and implementations, deep learning has produced impressive results. These methods have dramatically improved the state-of-the-art in speech recognition, visual object recognition, object detection, and many other domains [1]. Generally, deep learning is applicable in many fields of science, medicine, business, and in other real-world problems. Deep learning algorithms can potentially be used in every field of medicine, from drug discovery to clinical decisions. In those applications, like many others, deep learning is far ahead of other machine learning algorithms. Deep convolutional networks have been proven as a good solution for medical image classification, localization, detection, segmentation, and registration [2].

Deep learning in bioinformatics has many applications: sequence analysis, biomolecular property prediction, biomedical image processing and diagnosis [3].

Another reason for the present and future successes of deep learning is that it requires very little engineering by hand, so it can easily take advantage of increases in the amount of available computation and data. Deep-learning methods are representation-learning methods with multiple levels of representation, obtained by composing simple but non-linear modules, each of which transform a representation at one level (starting with the raw input) into a representation at a higher, slightly more abstract level. With the composition of a significant number of such transformations, very complex functions can be learned. Methods may very well discover interesting structures in large datasets [1][4][5]. Those are the reasons why they are very suitable for application in many domains.

This paper will show that deep learning is applicable in model-dependent diagnostic techniques with no calibration method, which enables the exclusion of measurement chain influence and in particular the influence of the microphone characteristics. This influence is not at all simple and eludes the usual kinds of differential calibration and standardization on the referential sample. Having said that, characterization done by those methods cannot give the exact properties which could satisfy fundamental scientific research. The case study was carried out for gas-microphone frequency photoacoustic technique. Justification of deep neural network application in photoacoustics relative to shallow neural networks is presented.

II. DEEP LEARNING IN PHOTOACOUSTICS

Physical parameters that configure in the physical model of the photoacoustic response are mostly nonlinearly dependent, very often with unknown and unavailable characteristics of the transformation process of the examined physical quantity in the electric signal. It is expected that the development and application of neural networks in photoacoustics and all similar model-dependent methods which use detectors of a common purpose is a good decision because deep learning is able to

Miroslava Jordović Pavlović is affiliated with the College of Applied Sciences Užice, Trg svetog Save 34, 31000 Uzice, Serbia (e-mail: miroslava.jordovic-pavlovic@ vpts.edu.rs).

Marica Popović is affiliated with the Vinča Institute of Nuclear Sciences, P.O. Box 522, 11001 Belgrade, Serbia (e-mail: maricap@vin.bg.ac.rs).

Dragan Markushev is affiliated with the Institute of Physics, University of Belgrade, Pregrevica 118, 11080 Pregrevica, Serbia (e-mail: dragan.markushev@ipb.ac.rs).

Slobodanka Galović is affiliated with the Vinča Institute of Nuclear Sciences, P.O. Box 522, 11001 Belgrade, Serbia (e-mail: bobagal@vin.bg.ac.rs)

approximate any nonlinear mapping with very high accuracy and reliability, and in that manner recognize and extract pattern of influence of individual parameters on true signal enabling calibration of the method. Also, classification models of different nonlinear mappings could be designed with very high accuracy using deep neural networks, which are applicable in the selection of different source influences on the true signal.

Neural networks are present in photoacoustics and other model-dependent measurement techniques which have been developed for scientific research and biomedical diagnostics for a very long time.

In [6], a shallow neural network is used for the reconstruction of the optical profile of optically gradient materials based on the frequency, magnitude and phase of the measured PT (photothermal) response.

In paper [7] shallow neural network with forward signal propagation was designed and used to simultaneously determine main physical parameters, such as: thermal diffusivity, thermal expansion coefficient and thickness, from transmission, frequency modulated photoacoustic response of the sample.

Examples of deep learning application in photoacoustics for the past few years are numerus.

Paper [8] presents deep convolutional network application for noise removal in photoacoustic recognition of images. Photoacoustic imaging is a method for the visualization of point-like targets. Using this method, detection of anatomical features or metal implants in the human body is possible, which can further be used in cancer detection, monitoring blood vessel flow, detecting and guiding surgeries, etc. Laser beam transmission in the presence of highly echogenic structures has consequences for the creation of a reflection artifact that may appear as a true signal. Deep convolutional networks turn out to be a good solution for the classification of a true signal from other artifacts with high accuracy and reliability.

A deep learning framework for image reconstruction in photoacoustic tomography (PAT) is presented in [9]. A sparse data problem is discussed. A direct and highly efficient reconstruction algorithm based on a deep convolutional neural network was developed. Neural network weights are adjusted prior to the actual image reconstruction based on a set of training data. The proposed reconstruction approach can be interpreted as a network that uses the PAT filtered backprojection algorithm for the first layer, followed by the U-net architecture for the remaining layers. Numerical results demonstrate that the proposed deep learning approach reconstructs images with a quality comparable to the state-of-the-art iterative approaches for PAT.

In [10] the authors used an MLP (Multy Layer Perceptron) for the simultaneous determination of the laser beam spatial profile and relaxation time of the polyatomic molecules in gases in real time within trace atmosphere gas monitoring. The spatial profile of the laser beam is variable, so its simultaneous determination contributes to the precision of the photoacoustic experiment, because it will correct the resulting variations. The same authors go a step forward in [11], so a feedforward MLP recognizes both the spatial profile of the laser beam and the values of the laser fluence, which contribute to additional precision in the measurement of different pollutant concentration in a wide range of values in a urban and rural environment.

However, as far as we know, neural networks have not yet been applied to the recognition of the influence of processes that are happening inside the detector. Data used for characterization are dependent on those processes. This influence cannot be understood as noise, but as a systematic influence which depends on the detector, and two completely identical detectors do not exist in practice. Accordingly, detector recognition is a kind of measurement set calibration, particularly in situations of detector changes when higher gain or a different measurement range is needed for different materials and structures or because of the failure of the existing detector in the serial measurements of the same sample. Such a calibration is a necessary step for a further inverse problem solution, apropos the determination of the examined sample characteristics with an accuracy required for fundamental research, which is significantly higher than the accuracy required for the application of some materials and structures or for biomedical diagnostics. In this paper the methods are based on deep neural networks which are able to effectively and very quickly recognize detector influence so that the calibration of the used experimental set can be done as suggested.

III. MASSIVE DATASET REGRESSION MODELS FOR DETECTOR PARAMETER PREDICTION

Our aim is to incorporate computational intelligence, especially deep learning, in the so-called "intelligent measurement system", which will be able to perform complex commands. We expect that such a system will be able to learn and to adapt to specific problems and to maintain high accuracy, reliability, and measurement rate. In the beginning, our intelligent measurement system will have the possibility of signal autocorrection relative to instrument influence. Although the case study was done on gas-microphone photoacoustic frequency measurements, the method application can easily be extended to a great number of model-dependent measurement systems with variable detectors.

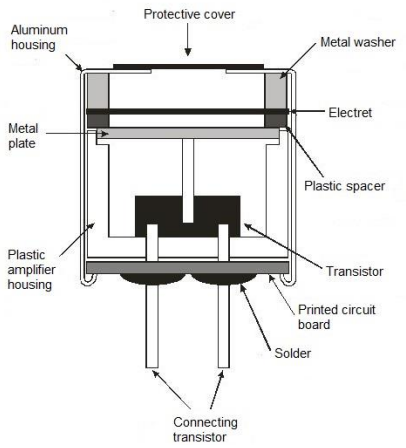


Fig.1. Schematic diagram of a cell of minimal volume.

We previously created a simple and cheap photoacoustic

measurement device [12], Fig.1. The common characteristic of of experimental signals due to the electronic or acoustic properties of the used instruments in the frequency domain [13], Fig.2. Based on the analysis of a great number of executed measurements on different materials and a comparison with the theoretical predictions which assume the detector to be ideal, it can be concluded that the microphone as the basic part of the detector measurement system brings most of the disturbances into the experiment.

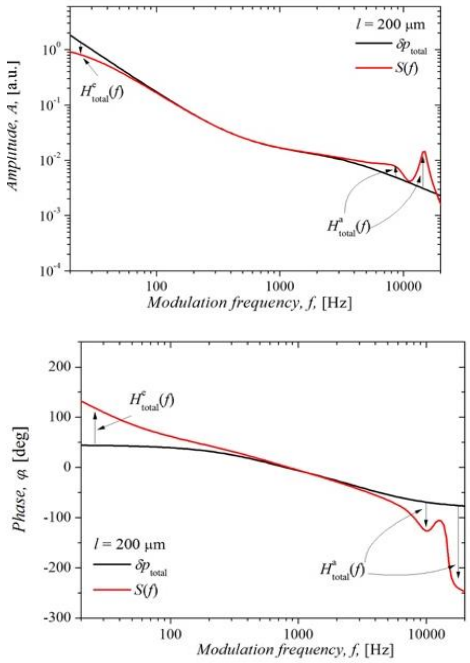


Fig. 2. Simulated a) amplitude and b) phase of the total photoacoustic signal (black line) and distorted experimental signal (red line)

A database of 67500 records was obtained from a well-known theoretical model. We obtained a massive dataset, and it is in precisely such datasets that deep learning recognizes interesting and useful structures, as well as patterns of non-linear dependence. The dataset was structured and labeled. The theoretical data corresponded to the commercial microphone ECM30B. Based on the statistical analysis of the collection of experimental measurements, it was concluded that frequency f_2 is the most stable one compared to the observed parameters, and three values were taken for network training: the central value 25 Hz and two values which are ± 5 % of the central value (23,75 Hz and 26,25 Hz). Also based on the statistical analysis, it was concluded that 10 values should be taken for each of the frequencies f_3 and f_4, distributed at equal distances in the range 8930-9866 Hz and 13965-15432 Hz, respectively. The least stable parameters are the damping factors of the second order low-pass filter, and they were presented with 15 values which were irregularly distributed from 0.015 to 0.99. Some of the curves from this dataset are presented in Fig. 3 (2250 lines). Every curve is presented with 200 points, and every point is presented with two characteristics, an amplitude and phase. In this way, one record in the database is presented with 400 all photoacoustic measurement systems is the high distortion features, 200 amplitudes and 200 phases. The dataset was first shuffled and then divided into a training set of a total of 57500 records or 82.6% of the total number of recordings, a validation set and a test set both of 5000 records or 8,7% of the total number of recordings. In this way, the training, validation and test sets were obtained randomly.

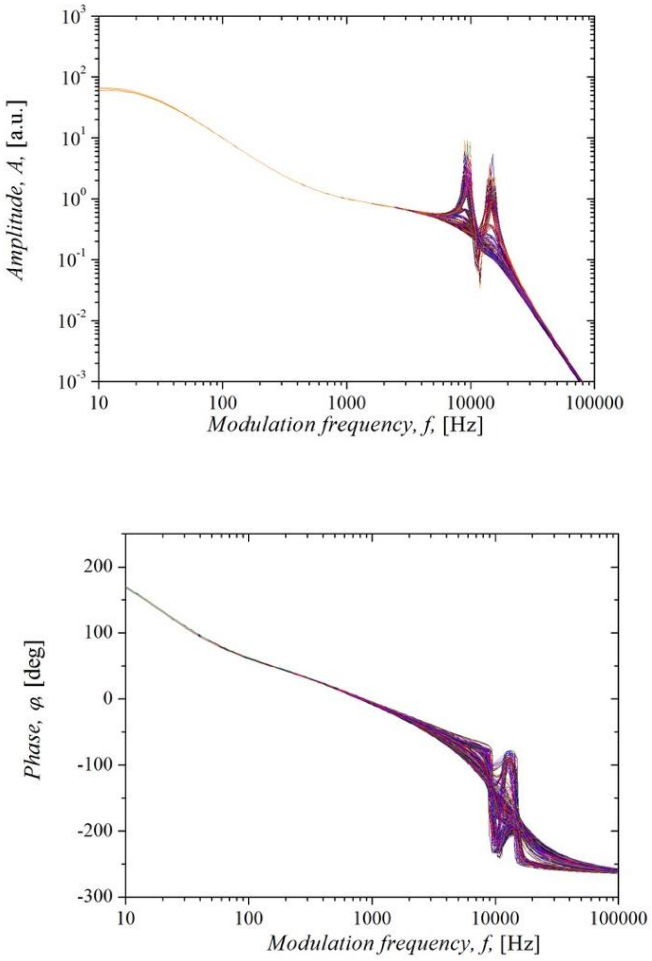


Fig.3. Curves: a) amplitude and b) phase of distorted photoacoustic signals with different microphone characteristics from the dataset used for network training [14].

Our aim is the development of a regression model for the prediction of five specific microphone parameters connected to its electronic and geometric features, which are not determined by the producers and could not be found in the specifications for the particular microphone. Based on our analysis of the theoretical models of the microphone as a sensor and a converter of pressure changes into an electrical signal, as well as those carried out on electrical measurements, it was shown that a five-parameter description of the detector influence on every detected signal is enough. In that way, microphone influence on the experimental signal can be determined. Now it is possible to correct the experimental signal in order to reach a “pure” signal, generated only from the excited sample. An MLP was our choice because of higher accuracy.

In this paper we present three regression models. The first model has an MLP with three layers, two of which are hidden, the first one with 30 neurons, and the other with 17 neurons, and one output layer with 5 neurons. The second model has 5 MLPs with three layers, two hidden, the first one with 30 neurons and the second with 17 neurons, and one output layer with 1 neuron. The third model has an ANN (Artificial Neural Network) with one hidden layer with 47 neurons, and is a shallow neural network. The network outputs represent the targeted microphone parameters: f_2, the characteristic microphone frequency connected to its RC characteristics, f_3, f_4 characteristic acoustic resonances of the microphone, and ξ_3 and ξ_4 reciprocal quality factors. The characteristics of a lock-in amplifier, whose role is played by the sound card described with parameter f_1, is considered known (f_1=15Hz). The input vector has 400 features, Fig. 4

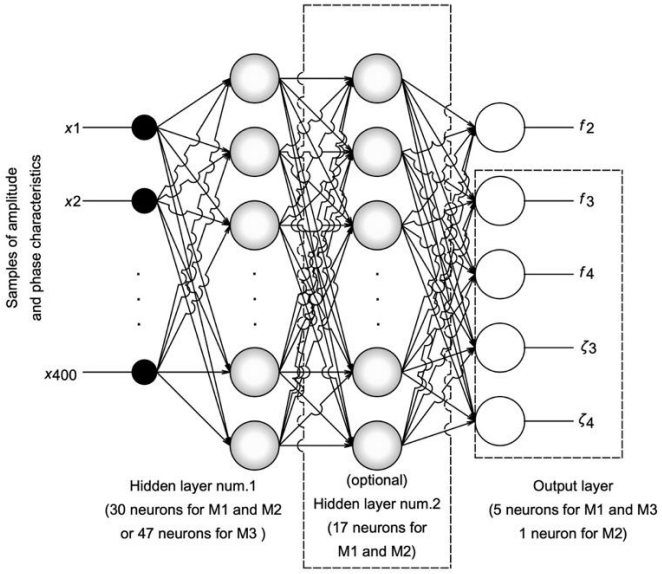


Fig. 4. Structure of Model1, Model2 and Model3

The normalization of the input vector was achieved by dividing each element x_i of the input vector by the maximum absolute value, determined over all the examples at the "i" frequency. This normalization type proved itself as the best solution for our model, then some others. With the application of this normalization type and without a value change in the other parameters, an acceptable value for the accuracy of the model was obtained in the iterative process of model parameter selection. We tested two more types of normalization, normalization obtained by subtracting the mean value of all examples at the "i" frequency from each element x_i and by dividing it by a standard deviation, as well as N2 or the Frobenius norm, but the results were not acceptable. The output vector y is normalized in a similar manner.

We chose a tanh activation function and the Xavier algorithm for weight initialization. [15]

We applied supervised learning for the model training. The Adam algorithm for error function evaluation was applied in order to achieve optimal weight values in the backprop [16]. The optimization is intensified by the Mini-batch technique, which is applied when the dataset is big enough (as it is in this case). This technique provides visible results of parameter optimization even in the first epoch, thanks to the division of the given dataset into smaller ones, which are treated as a whole, and applies error function evaluation on these smaller datasets. The learning rate for all the models has the same value of 10^{-4}.

The open source platform for machine learning, Tensorflow, was used for the realization of the models. Tensorflow is very popular for the realization of deep neural networks. It is based on a data flow graph. The graph nodes represent mathematical operations, while the graph edges represent the multidimensional data arrays (tensors) that flow between them.

Metrics for the models were defined, the same for all three models: deviation from the accurate value is less or equal to 5%, and that corresponds to the photoacoustic experiment. Model accuracy was analyzed relative to the set metric. The error function value on the training set and the validation set are similar, so we can conclude that model generalizes well and is not overfitted.

In accordance with Table1, it was concluded that the best results were achieved with Model 2, and the worst with Model 3. By dividing the same number of neurons into two layers, we got an accuracy approximately 2% higher in Model 1 than in Model 3. This difference will be bigger if the network is trained on experimental values, because theoretical models present idealizations, i.e. an approximation of real conditions. Is 2% small enough not to make a difference between the models? It depends on the application, for the photoacoustic experiment for industrial application it is significantly high. The importance of neural network depth for learning was proven by the increase of the accuracy of the model with more layers, under the same conditions (learning rate, number of epochs). Model 2 proves the fact that the deeper the neural network is, the better the recognition of behavior patterns in the data. Reducing the deviation from the accurate values in Model 2 in relation to Model 1 shows that the multilayer neural network can approximate the nonlinear output quite well. In Model 2 the neural network concentrates all its power on one output and achieves a very high accuracy for 3 of the 5 microphone parameters, as much as 99.99%, while training lasts for a far smaller number of epochs.

TABLE I:
A COMPARATIVE ANALYSIS OF THREE REGRESSION MODELS FOR THE PREDICTION OF THE MICROPHONE PARAMETER

	Accuracy		Cost		Numbers of epochs
Model 1	98.59%		0.000001		5000
Model 2	99.99%,99.99%,99.99%,99.61%, 99.596%		<0.000001,<0.000001,<0.000001,0.000001, 0.000001		1500,1500,1500,3000,3000
Model 3	0.969833		0.000003		5000
Average deviation from the accurate value expressed in the percentage of the accurate value on the training set					
Parameter	f_2	f_3	f_4	ξ_3	ξ_4
Model 1	0.02025029	0.08571574	0.03485037	1.0117933	0.59135133
Model 2	0.00367374	0.04975093	0.02321718	0.44225055	0.28317332
Model 3	0.03180477	0.1299281	0.06766562	1.9676312	1.1875261
Average deviation from the accurate value expressed in the percentage of the accurate value on the validation set					
Parameter	f_2	f_3	f_4	ξ_3	ξ_4
Model 1	0.02028082	0.08540299	0.03530468	0.998583	0.60733956
Model 2	0.00359443	0.05007159	0.02289878	0.4521597	0.31682032
Model 3	0.03162626	0.12913962	0.06732392	1.9448547	1.1696345
Average deviation from the accurate value expressed in the percentage of the accurate value on the test set					
Parameter	f_2	f_3	f_4	ξ_3	ξ_4
Model 1	0.02026834	0.0861348	0.0351213	0.9855594	0.5777709
Model 2	0.0037558	0.04898341	0.02315352	0.45252872	0.2967481
Model 3	0.03270168	0.13082047	0.06930758	1.9330658	1.2145972
Prediction time					
	CPU time		Computation_time(CPU +load time)		
Model 1	14 ms		31 ms		
Model 2	14 ms		5x30 ms		
Model 3	12ms		29ms		

In Model 1 the neural network splits its power to the five outputs, so the accuracy of this model is smaller. The same MLPs in Model 2 for different microphone parameters achieve different accuracies. This difference is just more proof that the network approximates the real situation very well. Parameters $\xi_3\, i\, \xi_4$ are very unstable photoacoustic quantities which depend on many other parameters. The theoretical model is not able to approximate the parameters very well. The neural network discovered this instability in the data.

IV. Conclusion

In this paper we discussed deep learning application in the calibration of model-dependent measurement techniques with nonlinear detector influence on the measured signal. Few examples of successful application were presented. The analysis of three regression models for microphone parameter predictions in photoacoustic experiments was presented. It was shown that higher accuracy was achieved by models with two hidden layer neural networks compared to a model with one hidden layer neural network, for the same total number of neurons. Based on this, it can be concluded that it is the depth, not the size of the neural network that matters. In the development of the regression model for the purpose of correcting the measuring chain influence in photoacoustic experiments, we selected a two hidden layer neural network structure, not one with more hidden layers, because the achieved accuracy was satisfactorily high. We accomplished the set metric. However, we intuitively know that models with a higher depth will be our actual research direction for some of future applications in model-dependent measurements, especially for the case of complex nonlinear dependences of input and output quantities.

For the purpose of a "smart" measurement system development, we chose Model 1 as the most practical solution for our needs. The application of this regression model for calibration of experimental set could be generalized on similar problems in other measurement or transmission problems. We consider Model 1 and Model 2 as the real choices relative to the given requirements.

ACKNOWLEDGEMENTS

We would like to thank to Ministry of Education, Science, and Technological Development of the Republic of Serbia for the financial support for this research through Project Nos. III45005. Special thanks also to college Drasko Furundjic, Institute Mihailo Pupin, for practical suggestions during the research.

REFERENCES

[1] Y. LeCun, Y. Bengio, G. Hinton, "Deep learning," *Nature*, vol. 13, no. 1, p. 35, 2015.

[2] J. Ker, L. Wang, J. Rao, and T. Lim, "Deep Learning Applications in Medical Image Analysis," *IEEE Access*, vol. 6, pp. 9375–9379, 2017.

[3] Y. Li, C. Huang, L. Ding, Z. Li, Y. Pan, and X. Gao, "Deep learning in bioinformatics: introduction, application, and perspective in big data era," *Methods*, In Press, Available online 22 April,2019.

[4] G. E. Hinton, S. Osindero, and Y. W. Teh, "A Fast Learning Algorithm for Deep Belief Nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–54, 2006.

[5] J. Schmidhuber, "DEEP Learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[6] M. N. Popovic, D. Furundzic, and S. P. Galovic, "Photothermal Depth Profiling Of Optical Gradient Materials By Neural Network," *Publ. Astron. Obs. Belgrade*, vol. 89, no. May 2015, 2010.

[7] K. Djordjevic, D. Markushev, Z. Cojbasic, S. Galovic "Photoacoustic measurements of thermal and elastic properties of n-type silicon using neural networks," *Silicon J.*, unpuslished

[8] D. Allman, A. Reiter, and M. A. L. Bell, "Photoacoustic Source Detection and Reflection Artifact Removal Enabled by Deep Learning," *IEEE Trans. Med. Imaging*, vol. 37, no. 6, pp. 1464–1477, 2018.

[9] S. Antholzer, M. Haltmeier, and J. Schwab, "Deep learning for photoacoustic tomography from sparse data," *Inverse Probl. Sci. Eng.*, pp. 1–22, 2018.

[10] M. Lukić, Ž. Ćojbašić, M. D. Rabasović, and D. D. Markushev, "Computationally intelligent pulsed photoacoustics," *Meas. Sci. Technol.*, vol. 25, no. 12, p. 125203, 2014.

[11] M. Lukić, Ćojbašić, M. D. Rabasović, D. D. Markushev, and D. M. Todorović, "Laser Fluence Recognition Using Computationally Intelligent Pulsed Photoacoustics Within the Trace Gases Analysis," *Int. J. Thermophys.*, vol. 38, no. 11, 2017.

[12] M. D. Rabasovic, M. G. Nikolic, M. D. Dramicanin, M. Franko, and D. D. Markushev, "Low-cost, portable photoacoustic setup for solid samples," *Meas. Sci. Technol.*, vol. 20, no. 9, p. 95902, 2009.

[13] S. Aleksic, D. Markushev, D. Pantic, M. Rabasovic, D. Markushev, and D. Todorovic, "Electro-acustic influence of the measuring system on the photoacoustic signal amplitude and phase in frequency domain," *Facta Univ. - Ser. Physics, Chem. Technol.*, vol. 14, no. 1, pp. 9–20, 2016.

[14] M. Jordovic-Pavlovic, M. Stankovic, M. Popovic, Z. Cojbasic, S. Galovic, D. Markushev, "Artificial Neural Networks Application In Solid State Photoacoustics Based on Microphone Response Recognition in the Frequency Domain,"*Journal of Computational Electronics*, unpublished .

[15] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010, vol. 9, pp. 249–256.

[16] D. P. Kingma and J. Ba, "Adam: {A} Method for Stochastic Optimization," *CoRR*, vol. abs/1412.6, 2014.

Player Skill Modeling and Feature Selection for a Video Game

Zoran Ž. Ćirović and Nataša A. Ćirović

***Abstract*—E-sports has increased significance, with growing number of professional players and monetization potential, thus the player skill modeling in strategy games has long been a subject of interest for researchers. In this paper we use publicly available video game telemetry data from StarCraft 2 to explore the player skill modeling and development of expertise. The analysis includes selection of statistical features for original dataset and their application to a classifier. The statistical features are obtained from the original data, applying statistical functions: mean, standard deviation, minimum, maximum, inter-quartile range, skewness, kurtosis. On such obtained set of statistical features, we applied feature ranking using t-test hypothesis, then feature subset selection is done using suboptimal searching techniques. We tested the obtained results on standard *k*-NN classifier.**

***Index Terms*—Video game skill learning; Feature selection; Feature ranking; *k*-NN classifier**

I. Introduction

HAVING in mind the enormous popularity of e-sports today, and that these skills must be acquired by practice, the significance of prediction of parameters that can establish the level of acquired skills is clearly of great importance for the gain in interaction with players. Additionally, e-sports is already generating significant revenues and the potential of further monetization in the next few years can reach up to $3 billion, [1].

Thus, the significance of player modeling, expression of human characteristics and prediction that can establish the level of acquired skills is clearly of great importance for the gain in interaction, [1], [2].

In real-time strategy (RTS) games the players do not play in turns, thus motor skills with a keyboard and mouse are an integral component of the game. One such game is StarCraft 2, that supports semi-professional and professional players, that often have a decade or more of RTS experience, [3].

Furthermore, it is reasonable to expect that the development of skill learning expertise would resemble the development of expertise in less related domains, such as education (e.g. VR surgery skills training, [4]) or health (e.g. personalizing games for rehabilitation of patients [2]).

Great popularity, as well as profitability, of gaming yields the increased research of the player modeling, [2], [5]. Some of the research directions belong to the cognitive science [4], [6]. Others are in the domain of machine learning, for e.g. Ravari et al. compare the relative importance of match and player skill and style features for the purpose of winner prediction [7], and Avontuur et al. deal with determining the player's league as early in the game as possible, [8].

Chen et al. [9] researched classification applying various clustering techniques and dimension reduction, over the dataset in [4].

With the increased amount of information available today, the problem of handling big datasets appears. Reducing the number of dimensions in an adequate way has become extremely important. This is done by using feature selection techniques, [10], [11]. In order to choose a good feature set, we require a way of measuring the ability of a feature set to discriminate accurately between two or more classes.

In our research we use the video game telemetry data from StarCraft 2 from the publicly available dataset collected by Thompson et al. [4].

The goal of our research is to determine the features that best describe the skill levels, for modeling player of StarCraft 2, using the dataset from [4]. Having in mind previous research, we applied a new process of feature selection by combining ranking and searching algorithms. Such hybrid technique of feature selection can include many statistical functions for feature extraction, considers the nature of the features and it can be implemented to similar applications, e.g. skill acquirement. We tested it with the standard *k*-NN model and gained significant accuracy for relatively small number of features.

II. Data Processing

Data processing block diagram is shown in Figure 1. Input vectors are first collected, then in the preprocessing part, transformed into feature vectors that are more suitable for further processing.

Next, feature selection is performed. The last part of processing is the classifier, where the selected features are used in two separate phases: (i) in the training phase – for modeling activities, and (ii) in the testing phase. If the dataset is not big enough, then the procedure of cross validation is applied.

Zoran Ž. Ćirović is with the School of Electrical and Computer Engineering of Applied Studies, 283 Vojvode Stepe, 11000 Belgrade, Serbia (e-mail: zoran.cirovic@viser .edu.rs).

Nataša A. Ćirović is with the School of Electrical Engineering, University of Belgrade, 73 Bulevar Kralja Aleksandra, 11000 Belgrade, Serbia (e-mail: natasa@etf.bg.ac.rs).

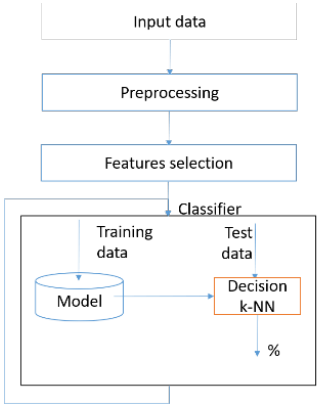


Fig. 1. Data processing block diagram.

A. *Preprocessing*

Data preprocessing block diagram is shown in Figure 2. In order to achieve equal contribution of all features in classification, normalization of input vectors should be implemented [13]. If the expected feature distribution is standard Gaussian, and the variance is not too small, standardization should be implemented instead [14].

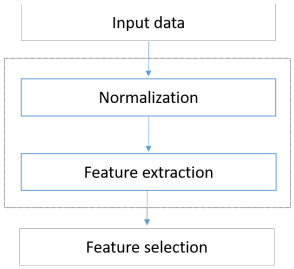


Fig. 2. Block diagram of preprocessing.

If statistical data is formed in the preprocessing phase, then framing of input vectors is performed. Framing is grouping of vectors in order to obtain statistical feature vectors. In framing overlapping is mandatory, in order to include the sudden changes in data.

B. *Statistical features*

A collection of statistical functions used for creation of statistical features is shown in Table I [15], [16]. A more precise definition of some of the features is given by (1) to (3).

TABLE I
SET OF ANALYZED FEATURES

	Description
Avg	Mean value
Std	Standard deviation
max	Maximum value
min	Minimum value
iqr	Interquartile Range
SK	Skewness
K	Kurtosis

$$iqr = Q_3 - Q_1, \tag{1}$$

$$SK = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - Avg}{Std} \right)^3, \tag{2}$$

$$K = \frac{n(n+1)\sum (x_i - Avg)^4 - 3\left(\sum (x_i - Avg)^2\right)^2 (n-1)}{(n-1)(n-2)(n-3)\,Std^4}, \tag{3}$$

where Q_1 and Q_3 are mean values of the first and last quartile, respectfully.

III. FEATURE SELECTION

Feature selection reduces the dimensionality of feature space, removes redundant, irrelevant or noisy data. Some of the benefits for the applications are: (i) model simplification; (ii) more efficient use of the model, both in long-term training and in testing or exploitation phase; (iii) generalization of features is emphasized, increasing robustness and reducing the influence of excessive overfitting, [10]-[12]

The straightforward algorithm is to test all possible subsets of features thus determining the one that gives the best results for the classifier. This is the process of complete search and it is not feasible in the process of calculation and it is can be specific to the used classifier. The alternative is to use a combination of some feature selection techniques, i.e. determining a subset of the complete set of features by applying a metric that defines the quality of the selected subset.

There are several classifications of feature selection methods. Based on the number of variables considered:

- Univariate methods, variable ranking - consider the input variables one by one;
- Multivariate methods, variable subset selection - consider entire groups of variables jointly (e.g. greedy search algorithms).

Based on the usage of the classification model in the feature selection process:

- Filter: evaluate quality of selected features, independent from the classification algorithm that will use them;
- Wrapper: require application of a classifier (which should be trained on a given feature subset) to evaluate this quality;
- Embedded methods: the feature selection method is built in the model (or rather its training algorithm) itself (e.g. decision trees).

A. *Feature selection based on ranking methods*

Feature selection based on ranking methods can be used as a method for individual independent features selection. Thus, certain criterion is used. By ranking we can determine the significance of individual features, but not the significance of a subset of features. Some of the criterion that can be used for ranking are:

- T-test – Absolute value two-sample t-test with pooled

variance estimate;
- Entropy – Relative entropy, also known as Kullback-Leibler distance or divergence;
- Bhattacharyya – Minimum attainable classification error or Chernoff bound.

B. *Sequential feature selection search*

Sequential feature selection search algorithms are a family of greedy search algorithms, which are of multivariate type.

Some of the methods are:
- Sequential Forward Selection;
- Sequential Backward Selection;
- Sequential Floating Forward Selection;
- Sequential Floating Backward Selection.

The methods are iterative, and in each step, the best feature is selected, based on the applied criterion. The process is repeated as many times as the number of features we want to select or get appropriate accuracy.

IV. Dataset

For obtaining the experimental results the dataset was collected by Thompson et al. [4], from 3,340 players of RTS video game - StarCraft 2, across 7 distinct levels of skill called leagues (Bronze, Silver, Gold, Platinum, Diamond, Masters, Professional). Every record contains 18 different features. The data is gathered by remote measurement without affecting the players in any way. Besides the first two inputs in the record that are used for identification (GameID and LegueIndex), the features are [4]:

Age: Age of each player (integer);

2. HoursPerWeek: Reported hours spent playing per week (integer);

3. TotalHours: Reported total hours spent playing (integer);

4. APM: Action per minute (continuous);

5. SelectByHotkeys: Number of unit or building selections made using hotkeys per timestamp (continuous);

6. AssignToHotkeys: Number of units or buildings assigned to hotkeys per timestamp (continuous);

7. UniqueHotkeys: Number of unique hotkeys used per timestamp (continuous);

8. MinimapAttacks: Number of attack actions on minimap per timestamp (continuous);

9. MinimapRightClicks: number of right-clicks on minimap per timestamp (continuous);

10. NumberOfPACs: Number of PACs per timestamp (continuous);

11. GapBetweenPACs: Mean duration in milliseconds between PACs (continuous);

12. ActionLatency: Mean latency from the onset of a PACs to their first action in milliseconds (continuous);

13. ActionsInPAC: Mean number of actions within each PAC (continuous);

14. TotalMapExplored: The number of 24x24 game coordinate grids viewed by the player per timestamp (continuous);

15. WorkersMade: Number of SCVs, drones, and probes trained per timestamp (continuous)

16. UniqueUnitsMade: Unique unites made per timestamp (continuous);

17. ComplexUnitsMade: Number of ghosts, infestors, and high templars trained per timestamp (continuous);

18. ComplexAbilitiesUsed: Abilities requiring specific targeting instructions used per timestamp (continuous).

Here PAC stands for Perception Action Cycle. Think of a PAC as when a player moves the camera and takes at least one action. Within a PAC there is a calculated number of actions, latency time between those actions, and gap of time between the next PAC.

V. Experimental Results

For examining if the selected subset of features is suitable, we use the standard *k*-NN classifier. For normalization of the input data we use minmax scaling function.

A. *Experiment 1*

Firstly, we created a *k*-NN classifier, using the 18 original input features, after normalization, from the dataset. The achieved error was 33.96%. The used classifier is defined for $k = 3$, where k is the number of nearest neighbors. For other values of k the results did not vary significantly (around 1%). The model is tested on different samples then the once used for training, using cross validation with 5 folds. Even though the number of features is small, we tried to throw out less significant features and got a gain of only 3.33%, i.e. the error is 30.63%.

B. *Experiment 2*

Statistical data processing is realized by grouping normalized input data in frames of length 30, with overlapping between frames. We used statistical functions shown in Table I for creation of statistical features. Thus, we got total of $18 \cdot 7 = 126$ statistical features. Next, we investigated 7 statistical groups separately, and all together, using the same classifier. The error rates are shown in Fig. 3.

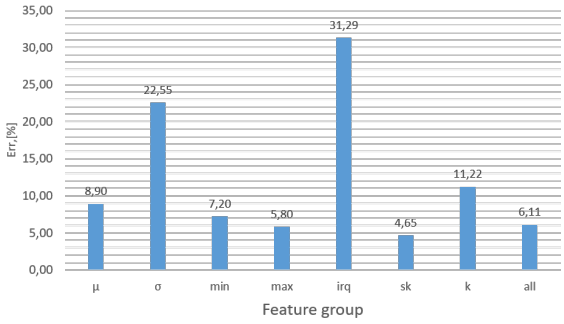


Fig. 3. Error rates for statistical features groups.

C. *Experiment 3*

Next, the analysis of feature significance for each skill level was done with T-test as the ranking criterion. Fig. 4 and fig. 5 show the 40 best ranked features for input variables, for skill levels 2 and 6, respectfully.

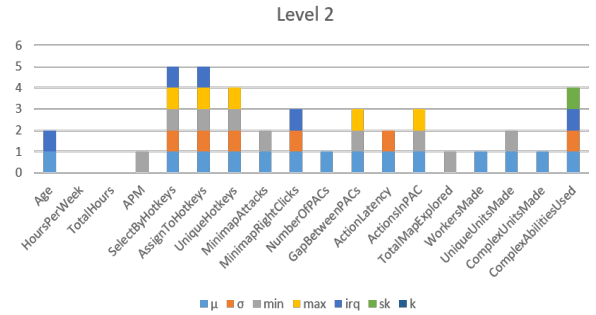


Fig. 4. Best 40 features for Level 2, for input variables.

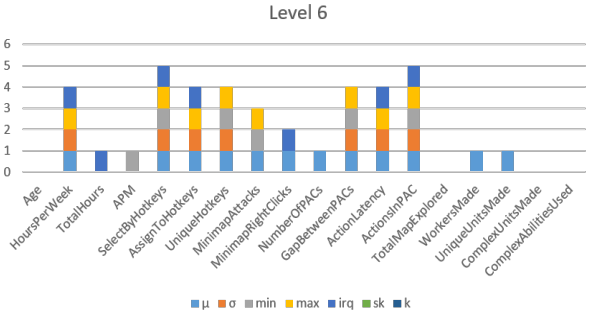


Fig. 5. Best 40 features for Level 6, for input variables.

The obtained results provide an individual ranking for each feature. For example, the Age feature is important for the starting level, but it is totally insignificant for higher levels. On the contrary, the feature SelectByHotkeys is highly ranked for all levels.

Fig. 6 and fig. 7 show the 40 best ranked features for statistical functions, for skill levels 2 and 6, respectfully.

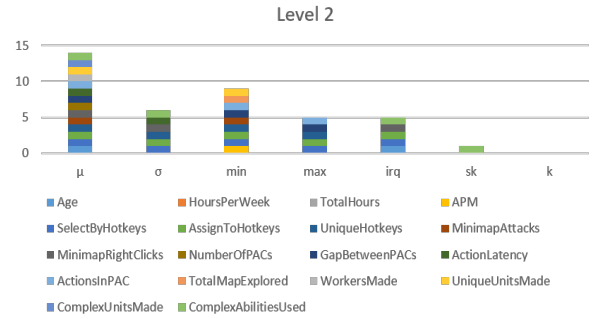


Fig. 6. Best 40 features for Level 2, for statistical functions.

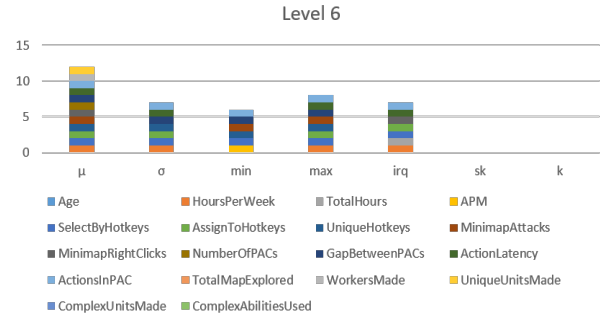


Fig. 7. Best 40 features for Level 6, for statistical functions.

We can observe that the group of features based on irq, skewness and kurtosis show different significance, depending on other features involved in the entire feature vector and skill levels.

Obviously, in this way it is not possible to define a unique group of features for all skill levels. For this reason, we investigated joint set of selected features for all skill levels in the next experiment.

D. *Experiment 4*

For every skill level we chose smaller number of best ranked features, *M*. Final feature set is a union of selected features for each level, *N*. Results for different *M* are shown in Table II. By implementing feature selection for all skill levels, we gain higher accuracy.

TABLE II
ERROR OF CLASSIFICATION FOR BEST *N* FEATURES

M	3	5	7	10	15
N	17	19	23	36	47
Err, [%]	2.58	2.42	4.03	4.03	3.89

E. *Experiment 5*

Last experiment is refining the selection of features obtained in the previous experiment. Using Sequential Forward Selection SFS for $M = 15$, ($N = 47$), we choose the best subset of 17 features. The error rate in this case is 2,03%.

The selected features are:

- Mean: 1) APM: Action per minute; 2) SelectByHotkeys; 3) AssignToHotkeys; 4) UniqueHotkeys; 5) GapBetweenPACs; 6) ActionLatency;
- Std. dev.: 7) Age 8) SelectByHotkeys; 9) GapBetweenPACs;
- Min.: 10) HoursPerWeek 11) SelectByHotkeys; 12) AssignToHotkeys; 13) UniqueHotkeys; 14) MinimapAttacks;
- Max.: 15) MinimapAttacks; 16) NumberOfPACs; 17) ActionLatency.

VI. CONCLUSION

We investigated input variables gathered from players of one RTS game, with the aim to determine the characteristics of players that can best describe their skill level. We compared the experimental results to the initial experiment implemented using the standard *k*-NN classifier for $k = 3$, on the original dataset. The achieved error in this case was 33.96%.

We implemented the statistical processing of the input variables with 7 statistical functions, thus we got statistical features. Based on this, we considered classification on 7 groups of statistical features, and all together. The statistical extraction provided gain in accuracy, with the error rate going from 4.65% to 31.29% for individual groups, and 6.11% for all features together. The obtained results show the significance of the statistical feature extraction in this case.

Using the feature selection ranking method, we selected the

best features for each level. The obtained results provide an individual ranking for each feature. Results show different significance for different groups of features, Fig. 4-7. In this way it is not possible to define a unique group of features for all skill levels.

In order to determine the significant features for all skill levels, we created a union of previously selected 15 best features of each level, obtaining the set of 47 different features. On this set we implemented the greedy search algorithms to obtain the subset of 17 best features. On this subset the *k*-NN classifier provides the error rate of 2.03%. We can observe that among the best 17 features there are no features from all statistical groups.

By combining selection methods of ranking and greedy search we can get significant improvements in classification of skill levels.

With the obtained feature subset, we can determine the skill level of the player with accuracy of approx. 98%.

The procedure of feature selection is efficient, considers the nature of the features and it can be implemented to similar applications, e.g. skill acquirement.

In future work we plan to verify our results with other classifiers. Also, we plan to do a comparative analysis using other feature selection techniques.

REFERENCES

[1] C. D. Merwin, M. Sugiyama, P. Mubayi, T. Hari, H. P. Terry, A. Duval, "The World of Games eSports From Wild West to Mainstream," The Goldman Sachs Groups, Inc., New York, NY, USA, Rep. *Oct.*, 2018.

[2] S. C. J. Bakkes, P. H. M. Spronck. G van Lankveld, "Player behavioral modelling for video games", Entertainment Comp., vol. 3, pp. 71-79, 2012.

[3] K. Adil, F. Jiang, S. Liu, W. Jifara, Z. Tian, Y. Fu "State-of-the-art and open challenges in RTS game-AI and Starcraft", Int. J. Adv. Comp. Sci. Appl., vol. 8, no. 12, pp.16-24, 2017.

[4] J. J. Thompson, M.R. Blair, L. Chen, A.J. Henrey "Video Game Telemetry as a Critical Tool in the Study of Complex Skill Learning", *PLOS ONE*, vol. 8, no. 9, *e75129*, Sep, 2013.

[5] G. N. Yannakakis, J. Togelius, *Artifcial Intelligence and Games,* Berlin, Germany: Springer, 2018.

[6] J. J. Thompson, C. M. McColeman, E. R. Stepanova, M. R. Blair, "Using Video Game Telemetry Data to Research Motor Chunking, Action Latencies, and Complex Cognitive-Motor Skill Learning", Top. Cogn. Sci., vol. 9, pp. 467-484, 2017.

[7] Y. N. Ravari, S. Bakkes, P. Spronck, "StarCraft Winner Prediction", Proc. 12th AAAI Conf. Artificial Intelligence and Interactive Digital Entertainment, AIIDE 2016, Burlingame, California, USA, pp. 2-8, Oct. 8-12, 2016.

[8] T. Avontuur, P. Spronck, M. Van Zaanen, "Player skill modeling in starcraft II," Proc. 9th AAAI Conf. Artificial Intelligence and Interactive Digital Entertainment, AIIDE 2013, Boston, USA, pp. 2-8, Oct. 14-18, 2013.

[9] P. Chen, Z. Qi, Y. Pan, S. Cheng, "Multivariate and Categorical Analysis of Gaming Statistics," Proc. *18th Int. Conf. Network-Based Infor. Syst.* Taipei, Taiwan, pp. 286-293, Sep. 2-4, 2015.

[10] I. Guyon, A. Elisseeff, "An Introduction to Variable and Feature Selection", J. Mach. Learn. Res., vol. 3, pp 1157-1182, 2003.

[11] G. Chandrashekar, F. Sahin, "A survey on feature selection methods", Comp. El. Eng, vol. 40, no. 1, pp. 16-28, 2014.

[12] J. Miao, L. Niu, "A Survey on Feature Selection", Procedia Comp. Sci., vol. 91, pp. 919-926, 2016.

[13] G. Nadarajoo, N. F. Aziz, N. A. Rahmat, Z. M. Yasin, N. A. Wahab, N. A. Salim, "Impact of Data Transformation and Preprocessing in Supervised Learning Algorithm", J. Fund. Appl. Sci., vol. 10, no. 5S, pp. 551-561, 2018.

[14] P. Trebuňa, J. Halčinová, M. Fil'o, J. Markovič, "The importance of normalization and standardization in the process of clustering," IEEE 12th Int. Symp. Appl. Mach Intellig. Inform. (SAMI), Herl'any, Slovakia, pp. 381-385, 2014.

[15] B. Esmael, A. Arnaout, R. K. Fruhwirth, G. Thonhauser, "A Statistical Feature-based Approach for Operations Recognition in Drilling Time", Series. Int. J. Comput. Inf. Syst. Ind. Manag. Appl., vol. 5, pp. 454-461, 2015.

[16] J. P. Verma, A-S. G. Abdel-Salam, *Testing Statistical Assumptions in Research*, Hoboken, New Jersey, USA: John Wiley & Sons, Inc., 2019.

Semantic Technology-Based Platform for Automated Assessment of Information Systems Course Projects

Nenad Petrovic, Milorad Tosic and Valentina Nejkovic

***Abstract*— When it comes to computer science education, it has been recognized that practical programming assignments (projects and exercises) are essential element. However, in recent years, the number of students enrolled to bachelor degree computer science and information technology courses has increased dramatically, bringing new challenges for universities and educational institutions. Many problems in courses with huge number of participants are related to the quality of student assignment/exercise evaluation, as it is time-consuming and error-prone process. Therefore, the automation of student assignment evaluation is seen as a solution. While there are many existing tools and platforms that tackle this problem, most of them are limited only to exercises for introductory courses. In this paper, we focus on aspects of automated assessment of student projects based on Java and MySQL, developed in context of the third year bachelor degree Information Systems course at Faculty of Electronic Engineering, University of Nis in Sebia. As an outcome of this research, we propose the platform architecture based on semantic technology and present a tool for automated project assessment based on static code analysis aiming the usage in blended learning environments.**

***Index Terms*—education; semantic technology; static code analysis**

I. Introduction

Assessment is of utmost importance for educational activities, as it provides feedback whether the goals of education are achieved or not to both the students and teachers.

Continuous assessment during a course can be highly beneficial, as it can directly enhance the learning process by providing early feedback on the quality of students' exercise solutions [1, 2, 3]. However, providing timely feedback on student programming exercises and helping them think about the mistakes they made is time-consuming and requires huge effort, especially when it comes to classes with large number of students, as it is the case nowadays at universities where computer science and information technology courses are taught. In recent years, we have witnessed a huge growth of number of computer science and information technology university students worldwide. Exactly the same trend is notable locally in Serbia as well, reflecting the global trends [4]. As the class size grows, the amount of work related to assessment is often limited in some way, which may affect the evaluation quality. Therefore, the automation of assessment process is seen as a possible solution [3]. Automatic assessment tools not only speed-up the evaluation process, but could also improve the quality and fairness of the assessment.

In past, most platforms for automatic programming exercise assessment were based on the externally observable behavior of the code, treating it as a black box. While this approach might be acceptable for exercises in introductory-level courses where the correct output for particular test inputs is known in advance, it is not suitable for more complex student projects where there might be countless possibilities of the outputs which still might be considered valid and positively evaluated. This is the case in courses where the assignments have more open formulation that leaves a certain degree of freedom and creativity to students, such as Software Engineering, Information Systems and Web Programming. In assignments for these courses, it is often required that students implement some business logic aspects that could require in-depth semantic analysis of the code meaning if assessed automatically, rather than simple comparison to pre-defined correct outputs for given inputs. Moreover, most of the existing automatic assessment platforms are often limited to one programming language, while the projects in previously mentioned courses might involve usage of many tools, programming languages and technologies.

In this paper, we focus on automated assessment of Information Systems course students' projects on third year of bachelor degree (Faculty of Electronic Engineering, University of Nis, Serbia) that are based on Java and MySQL. As an outcome, we propose a semantic framework, platform architecture and present a case study where our tool prototype for static semantic code analysis (process of examining source code without the execution) is used to confirm the effectiveness of the proposed approach in blended learning environment.

II. Related work

There are many existing solutions for automatic assessment of student's Java program code. Some of them rely only on the results obtained during the execution of pre-defined tests, the others only take into account the static code analysis, while there are also solutions that involve both the dynamic

Nenad Petrovic, Milorad Tosic and Valentina Nejkovic are with the Faculty of Electronic Engineering, University of Niš, Aleksandra Medvedeva 14, 18000 Niš, Serbia (e-mail: nenad.petrovic@elfak.ni.ac.rs, milorad.tosic@elfak.ni.ac.rs, valentina.nejkovic@elfak.ni.ac.rs).

and static aspects, which is the case in more recent works. However, it is identified that the most of these solutions are targeting introductory computer science courses (Algorithms and Programming, Data Structures, Programming Languages, Object-oriented Programming) where student project structure is quite simple. Moreover, most of these are mainly designed to be integrated with MOOCs, where the project assessment is solely based on the automatic assessment, while we aim to provide a solution that will support the educational process in blended learning environment where the final assessment would still involve teacher at certain points. Despite the fact that MOOC platforms have achieved great success around the world, especially for low-cost corporate training and certification, when it comes to the teaching practice in universities, this model does not necessarily have good results [5]. Blended learning and flipped classroom learning techniques that combine online resources and technology in various ways with personal engagement between faculty and students have shown much better outcome [4]. In what follows, some of the existing solutions for automated students' programming exercise assessment will be mentioned.

In [6] and [7], grading systems for Java introductory programming courses was presented that grades submissions both dynamically (based on JUnit framework) and statically (based on graph representation of the program and quality measured by software metrics). These solutions performed well, but they are limited to basic procedural programming exercises. In [3], a similar approach was used, providing more sophisticated feedback. The solution for static analysis of students' Java programs presented in [8] is based on software engineering metrics and structural similarity. However, that framework does not implement semantic analysis and is still limited to introductory exercises. Moreover, the framework presented in [9] offers libraries for both the static and dynamic Java program assessment and it copes well with object-oriented programming exercises, but still does not cover the assessment of the assignments that have more complex structure and involve the usage of many different tools and technologies. On the other side, the approach shown in [2] is based model-based black-box testing to assess both the functional correctness and algorithmic complexity of students' exercises in the introductory-level programming courses.

The main advantage of our solution, compared to previously mentioned work is the ability to perform in-depth semantic analysis of students' program code based on semantic code annotations in order to detect program properties and support for courses with projects that have more complex structure (not limited only to introductory-level courses) that might involve the usage of many different programming languages, tools and technologies.

The idea of the adopted approach originates from [10, 11], where parsing and semantic code annotation were used for design-time analysis of robotic experiments written in a domain-specific language. For parser development, Xtext[1] framework was used. It enables the generation of corresponding parser using a simple, but powerful grammar definition language.

[1] https://www.eclipse.org/Xtext/

III. Semantic Framework

Semantic analysis of digital artifacts is a process of understanding their meaning based on context. In cases when applied to program code, it is required both to parse it and store the knowledge extracted from the code in suitable way, so the reasoning mechanisms can be applied to infer knowledge from the existing facts, while traditional databases just enumerate all available facts.

For this purpose, we use ontologies. An ontology is an explicit and formal description of knowledge about a domain of interest, the core of which is a machine-processable specification with a formally defined meaning [12]. Resource Description Framework (RDF) is a formal language for specification of structured information that is often used to formally describe ontologies [12, 13]. It is a data model that provides a way to express simple statements about resources, using named properties and values [12, 13]. In context of RDF, classes are used to define types of things and categories. In addition to defining the classes of things within the ontology, it is also possible to define specific properties that characterize those classes of things. Relations and facts about the classes, their properties and relations are stored as triplets within the RDF triple store for both the ontology definition and its instances. SPARQL is a semantic query language that provides possibility to retrieve and manipulate the data stored within the RDF triple store [12, 13].

Within our framework, the ontologies are used to represent the knowledge about various aspects about both the static and dynamic properties of student projects. Static properties of the project refer to the presence of the necessary project artifacts and their meaning in project context. On the other side, the dynamic properties are related to the behavior of the application during the execution. In Fig. 1, an overview of the ontologies behind our framework is given.

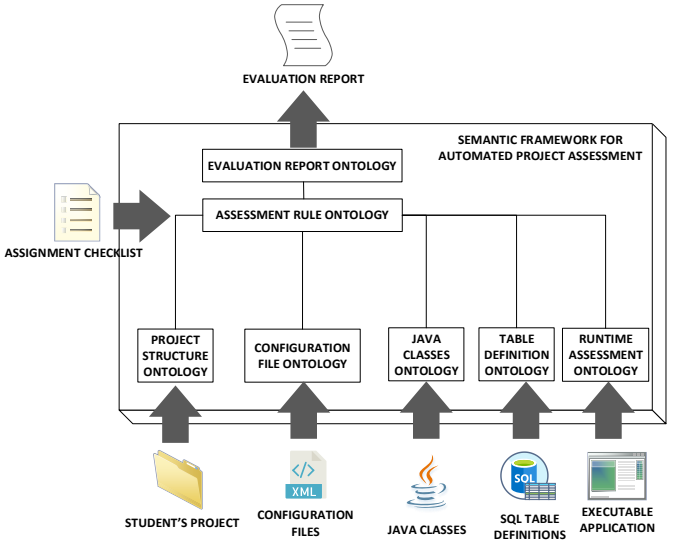


Fig. 1. Ontologies behind the semantic framework for automated assessment of student projects

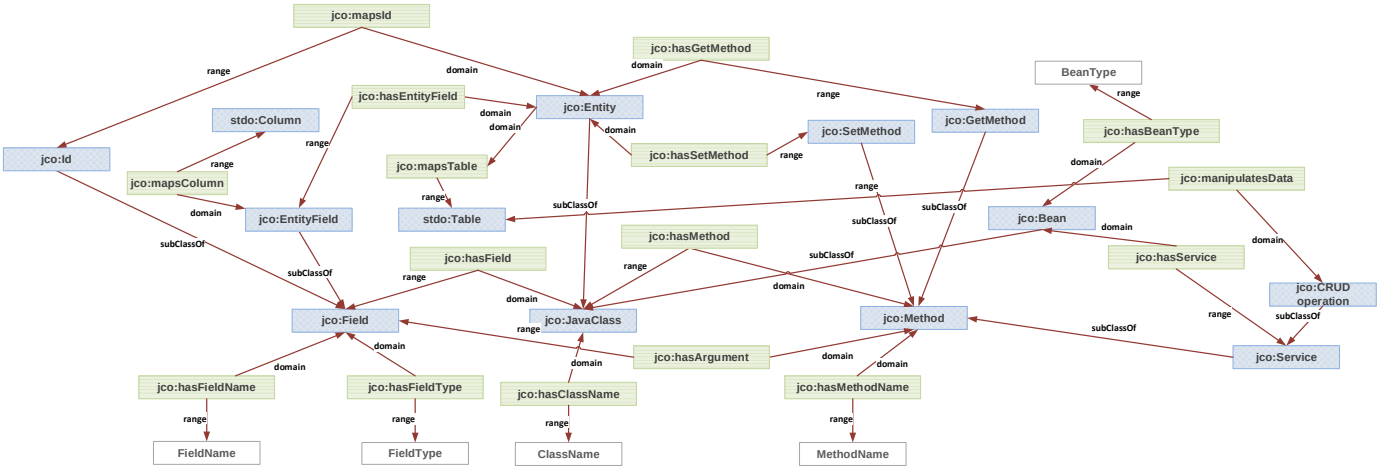


Fig. 2. An excerpt from the Java Classes Ontology that is used for static code analysis

Project Structure Ontology represents the knowledge about the structure of student's project and which are the assets that need to be contained within the project. In scope of this framework, we focus on Java EE Maven projects that are using Hibernate for object persistence and Enterprise Java Beans for business logic. Configuration File Ontology is used to store the knowledge about the structure of specific configuration files, such as pom.xml for Maven projects and persistence.xml. Java Classes Ontology describes the structure of various types of Java classes, such as Entity classes used for JPA/Hibernate ORM mapping and EJB for business logic implementation. For example, Entity classes are used for object persistence and must map to a valid database table and corresponding columns, identifier and contain getter/setter methods. This way, we provide the means for static, semantical Java code analysis based on pre-defined queries in order to detect whether certain conditions are satisfied within the student's code. In Fig. 2, an excerpt from the Java Classes Ontology is given. Table Definition Ontology holds the knowledge about the representation of SQL data definition statements that are used for creation of the database that is a part of the project. While all the previously mentioned ontologies refer to the static aspects of project assessment, Runtime Assessment Ontology should contain the knowledge about the application behavior obtained during the execution, based on pre-defined tests. Moreover, the Assessment Rule Ontology (depicted in Fig. 3) represents the course instructor-defined evaluation criteria for both the static and dynamic project aspects used for assessment.

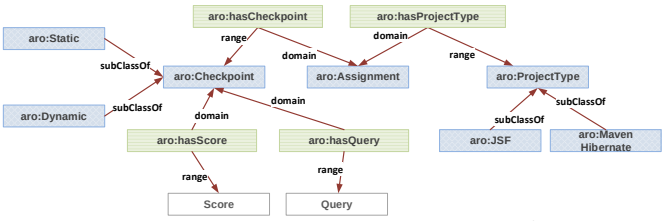


Fig. 3. Assessment Rule Ontology holding the teacher-defined evaluation criteria

These evaluation rules are defined as a list of checkpoints that project needs to satisfy. Each checkpoint has its score value and the corresponding assert written as a SPARQL query that needs to be executed against the RDF triple store to determine whether the considered student project satisfies the given criterion either static or dynamic. Additionally, the information about the project type can be used in order to check if a particular project complies with the required structure requirements and contains all the necessary elements for the specific project type (Java Web Application, Maven Hibernate etc.). An example SPARQL query for a checkpoint that determines whether the Java entity class correctly maps the table name and identifier to SQL table, for an assignment that involves object-relational mapping, is given in Listing 1.

```
PREFIX stdo: <http://www.tasorone.com/tsc/resources/stdo/>
PREFIX jco: <http://www.tasorone.com/tsc/resources/jco/>
SELECT ?entity ?table
WHERE {
  GRAPH <http://www.example.com/project1> {
   ?entity jco:mapsId ?id.?table stdo:hasIdentifier ?id.
    ?entity jco:mapsTable ?table_name.?table stdo:hasName
?table_name.
  }
}
```

Listing 1. An example SPARQL query for a checkpoint in assignment that involves object-relational mapping

Finally, the Evaluation Report Ontology stores the knowledge necessary to generate the assessment feedback, such as the list of checkpoints that are satisfied/not satisfied within the student's project and achieved scores for each of them.

IV. Assessment Platform Overview

In this section, the complete overview of the platform for automated assessment of student projects (illustrated in Fig. 4) is given, its main components are described and underlying mechanisms explained.

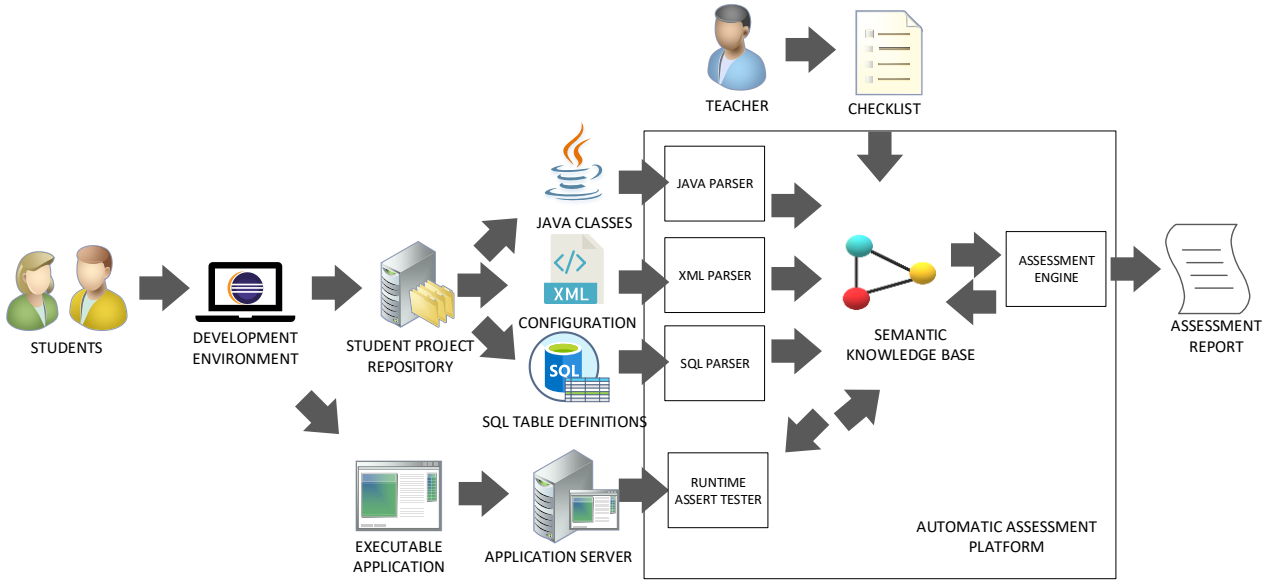


Fig. 4. Overview of the platform for automated student project assessment

The students develop the project artifacts using an integrated development environment (such as Eclipse) and, optionally, some additional tools (DBMS, Web/database server) depending on the concrete course requirements. The

projects with all the related artifacts are stored in remote student project repository (SVN, for example) that is synchronized with the student's development environment. Once the student decides to submit the assignment, the automated assessment process starts.

First, all the project artifacts that represent a source code (Java classes, XML configuration files, SQL table definitions) are recognized within the project directory and parsed using the corresponding parser. For each of them, during the parse tree traversal, the information of interest is extracted and stored within the semantic knowledge base, so it can be later used by the reasoning mechanisms for project assessment. This way, the static assessment of student project source code is performed. On the other side, the executable artifacts of the project are transferred to the application server that is able to run them. The execution is performed according to the pre-defined tests and led by the runtime assert tester component that inserts the triplets about the runtime application behavior to the knowledge base.

The project assessment (both static and dynamic) is performed according to the checklist defined by the course teacher. It consists of checkpoints that represent the conditions that both the source code and executable project artifacts need to satisfy, while each of them affects the final project score at certain percentage.

The final project score is calculated by the assessment engine. The assessment engine executes the queries corresponding to the assessment checkpoints and sums up the score obtained for each of them. Moreover, the corresponding triplet about the checkpoint assessment is inserted in triple store. The score calculation algorithm is given as simplified Java-alike code in Listing 2. Once the final score assessment is done, the results are showed as an assessment report which contains the scores for each of the checkpoints, so the students can be aware of their mistakes and improve their work for next submission or further assignments.

```
int assessment(String student_id,String assignment_name){
Checkpoint[] checkpoints=getAllCheckpoints(assignment_name);
int total_score=0;
foreach(Checkpoint checkpoint in checkpoints)
{
  int current_cpoint_score=0;
  boolean condition_satisfied=executeQuery(checkpoint.query);
  if(condition_satisfied)
  {
     total_score=total_score+checkpoint.score;
     current_checkpoint_score=checkpoint.score;
  }
insertTriplet(checkpoint,"isAssessed",current_cpoint_score);
}
insertTriplet(student_id,"hasAchieved",total_score);
return total_score;
}
```

Listing 2. Automatic project assessment algorithm

V. Case Study

As a case study, we consider automated assessment of student assignments developed as an outcome of previous cycle (spring 2018) of Information Systems Course that we teach on third year of bachelor degree. Each student receives a problem from random domain as an assignment, such as e-commerce, ticket booking, information system for school/university/library, mobile operator software platforms and many others. The task is to design and develop an information system according to the given requirements, using Java EE and MySQL technology. When it comes to object-relational mapping, JPA and Hibernate are used. However, the mapping issues are not the main focus of this course, so the students were not obliged to do complex relationship mappings. It was acceptable to implement just two tables, all the CRUD operations implemented and some of them used within the business logic functions. Despite the fact that the solutions' requirements and complexity may vary from case to

case, the project structure requirements common for all the assignments are:

- The project is based on Java EE technology[2]; it should be managed and built using Maven[3]
- MySQL[4] database server is used for data storage
- Hibernate/JPA should be used for object persistence using annotated entity classes
- Wildfly[5] is used as an application server
- Enterprise Java Beans (EJB)[6] is used for implementation of business logic
- A console client application that implements a usage scenario covering all the functional aspects of the application

For each of the previously given checkpoints, SPARQL queries were defined by the project evaluator to determine whether the conditions at the level of provided source code assets are satisfied. The architecture of the information system that has to be developed by students in our course is illustrated in Fig. 5.

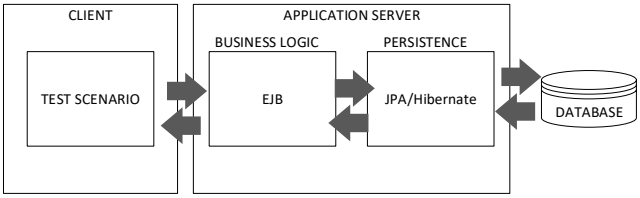


Fig. 5. Overview of the assignment for Information Systems course

VI. Evaluation and Results

In academic year 2017/18, we had 226 enrolled students. However, not all of them were able to complete the project, while some of them were disqualified due to plagiarism. Therefore, 200 students submitted the complete project, while two students among them had different assignments related to research work and were not relevant for this case study. For evaluation purposes, we have selected 50 student projects randomly from this set (out of 198) and imported into the automatic assessment platform.

In Table I., the evaluation results are given. The first column presents the name of each checkpoint that was scored within the project. The second column provides information about the maximum score that can be achieved for each of the checkpoints. The third column shows the average score for each of the checkpoints obtained during manual project evaluation by course instructor. On the other side, the fourth column displays the average results of automatic assessment platform for each of the checkpoints. Finally, the last column shows the absolute difference of the grades obtained manually and using automated assessment.

According to the results presented in Table I, it can be immediately noticed that the overall average score is lower in case of automatic assessment. The same stands for each checkpoint separately, except the last one, where the score obtained using our platform is higher than the case of manual evaluation. The automatic assessment platform was generally stricter than the evaluator due to fact there were some student files which were not parsed properly, due to wrong character formatting and syntax errors, especially when it comes to XML configuration files (the first two checkpoints), while these errors were still acceptable for human evaluator. On the other side, for the test scenario, the evaluation platform was less strict, due to fact that this part dramatically varies from case to case, making very difficult to automatically detect whether the student has implemented the most appropriate scenario for the considered information system and is graded more subjectively than other parts, when it comes to manual evaluation. Furthermore, in case of manual evaluation, the instructor has also considered the runtime aspects of the applications in the test scenario (running properly or no), while our platform performed solely static code analysis. Entity classes and ORM mapping checks perform quite well, with slightly lower grade than in case of manual assessment, also due to fact that some files were not parsed properly due to wrong text formatting. When it comes to EJB and business logic checkpoint, the highest grading difference has been noticed, as in case of human-based assessment also the complexity of the project was also considered, so the students have achieved some extra points despite the small errors in code.

The fact that the average grade is approximately 12,3% lower in case of automated assessment is acceptable, especially if we take into account that manual assessment by human evaluator lasted (in most cases) from 6 to 10 minutes for a single project, while it is around 90 seconds for the automatic assessment. Therefore, the assessment speed-up is at least 4 times. This is what makes the platform suitable for continuous assessment, as course instructors are not available 24 hours for tutoring of the students' projects, while the automatic assessment platform would be always available online.

TABLE I
Comparison of Average Grade Obtained Manually and Using Automatic Assessment Tool

Checkpoint	Max. points	Manual eval.	Auto-eval.	Grade diff.
Maven config	**3**	2.81	2.16	0.65
Deploy. config	**3**	2.32	1.75	0.57
Entity classes	**5**	4.12	3.78	0.34
EJB/business logic	**5**	4.41	2.92	1.49
Test scenario	**4**	3.08	3.67	0.69
Overall score	**20**	16.74	14.28	2.46

VII. Conclusion and Future Work

In this paper, a semantic approach to automatic assessment of students' programming assignments was presented. According to the evaluation results, it was shown that this approach was effective when it comes to assessment of student projects developed in context of Information Systems

[2]https://www.oracle.com/technetwork/java/javaee/overview/index.html
[3] https://maven.apache.org/
[4] https://www.mysql.com/
[5] http://wildfly.org/
[6] https://www.oracle.com/technetwork/java/javaee/ejb/index.html

course (third year of bachelor degree) at Faculty of Electronic Engineering, University of Nis. Despite the fact that the solution seems quite promising, there is still space for improvements in future.

First, the platform still lacks the implementation of the module that is responsible for runtime program evaluation. It is planned in near future to implement and integrate this module at it will be most probably based on JUnit framework. Despite that our platform performs quite well even relying only on static code analysis, our expectation is that the assessment accuracy will be increased, as we will include also the evaluation of behavioral aspects of the program.

Secondly, it is planned to adopt data mining techniques within the learning environment (as described in [14]) in order to enable adaptability and provide more sophisticated feedback to students based on their behavior patterns, combined with approach similar to [3].

Moreover, we would like to provide support for more programming languages and technologies, especially when it comes to Web programming. The efforts needed for adding the support for new technologies would involve parser construction and corresponding language-specific ontologies that would enable the semantic annotation of program code. It is also required to consider integration with plagiarism detection mechanisms at least within the scope of student project repository (including previous course cycles) that would ensure that students' solutions are original.

Finally, it is planned to use the extended version of the platform during the next cycle of Information Systems Course we teach (spring 2020), so the students will be able to have continuous feedback during the project development before the final evaluation done by the teachers. After that, we will compare the final course outcome with previous cycles where the automatic assessment tool was not used.

ACKNOWLEDGMENT

This work has received funding from the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Grant Agreement No 645220, project RAWFIE (Road-, Air- and Water- based Future Internet Experimentation) and the Serbian Ministry of Education, Science and Technological Development (project III47003).

REFERENCES

[1] D. Insa and J. Silva, "Semi-Automatic Assessment of Unrestrained Java Code: A Library, a DSL, and a Workbench to Assess Exams and Exercises", *Proceedings of 2015 ACM Conference on Innovation and Technology in Computer Science Education (ITiCSE '15)*, Vilnius, Lithuania, pp. 39-44, 2015.

[2] C. Earle, L. Fredlund and J. Hughes, "Automatic Grading of Programming Exercises using Property-Based Testing", *Proceedings of 2016 ACM Conference on Innovation and Technology in Computer Science Education (ITiCSE '16)*. Arequipa, Peru, pp. 47-52, 2016.

[3] G. Shimic and A. Jevremovic, "Problem-based learning in formal and informal learning environments", Interactive Learning Environments, 20(4), pp. 351–367, 2012. doi:10.1080/10494820.2010.486685

[4] N. Petrovic, V. Nejkovic and M. Tosic, "Dealing with scalability of laboratory sessions in computer science university courses", *Proceedings of 26th Telecommunication Forum (TELFOR 2018)*. Belgrade, Serbia, pp. 1-4, 2018. https://doi.org/10.1109/telfor.2018.8612090

[5] P. Guo, "MOOC and SPOC, which one is better?" Eurasia Journal of Mathematics, Science and Technology Education, vol. 13, pp. 5961-5967, 2017.

[6] F. Al Shamsi and A. Elnagar, "An Intelligent Assessment Tool for Students' Java Submissions in Introductory Programming Courses", *Journal of Intelligent Learning Systems and Applications*, 2012, vol. 4, pp. 59-69, 2012.

[7] M. A. Pinto and A. L. Lopes, "ACode: Web-based System for Automatic Evaluation of Java Code", pp. 1-4, 2012. Retrieved from: http://inforum.org.pt/INForum2012/docs/20120084.pdf/at_download/file

[8] N. Troung, P. Roe, P. Bancroft, "Static Analysis of Students' Java Programs", *Proceedings of the Sixth Australasian Conference on Computing Education – Volume 30.* Dunedin, New Zealand, pp. 317-325, 2004.

[9] D. Insa and J. Silva, "Automatic assessment of Java code", *Computer Languages, Systems & Structures 53*, pp. 59-72, 2018.

[10] V. Nejkovic, N. Petrovic, N. Milosevic, M. Tosic, "The SCOR Ontologies Framework for Robotics Testbed", 2018 26th Telecommunication Forum (TELFOR), Belgrade, pp. 1-4, 2018. https://doi.org/10.1109/telfor.2018.8611841

[11] N. Petrovic, V. Nejkovic, N. Milosevic, M. Tosic, "A Semantic Framework for Design-Time RIoT Device Mission Coordination", 2018 26th Telecommunication Forum (TELFOR), Belgrade, pp. 1-4, 2018. https://doi.org/10.1109/telfor.2018.8611845

[12] P. Hitzler, M. Krotzschm and S. Rudolph, "*Foundations of Semantic Web Technologies*", Chapman & Hall/CRC, USA, 2009.

[13] J. Davies, R. Studer and P. Warren, "Semantic Web Technologies: Trends and Research in Ontology-based Systems", John Wiley & Sons, Ltd, 2006.

[14] N. Petrović, "Primena data mininga u platformi za laboratorijske vežbe iz oblasti informatike i računarstva", YuInfo 2019, pp. 1-6, 2019.

Secret keys generation from mouse and eye tracking signals

Milan Milosaljević, Saša Adamović, Aleksandar Jevremović

***Abstract* — This paper presents a new approach to generating cryptographic keys, based on local mutual randomness of mouse and eye move tracker sensor signals. The cross-correlation analysis of the longitudinal and transverse components of these sensor signals confirms a sufficient level of mutual randomness, which is originally the starting point for the protocol for generating secret cryptographic keys with the help of a public discussion. Experiments show that in this way you can generate random sequences of good cryptographic properties, with speed up to 20 bits/minute for the typical interaction with web pages.**

***Index Terms*— cryptographic keys, common randomness, eye tracker signals, mouse signals.**

I. Introduction

Generating random numbers has a long history in several different disciplines: from computer simulations of probabilistic phenomena [1], cryptography [2], to artificial intelligence [3]. In the latter case, the ability to generate random sequences is related to the free will of an artificial entity, and therefore its pretension to be fully human. In [3] generating random sequences is suggested as one possible test of artificial intelligence, similar to the Turing test [4]. Generating random strings for the needs of different cryptographical services is the central and most sensitive link in the cybersecurity chain [2]. Therefore, it is not surprising that the most relevant results from this domain appear in the field of cryptography. In this context, the seminal works of Ahlswede and Csiszár [5], Maurer [6] and Csiszár and Narayan [7] draw particular attention.

The basic idea of this approach is to extract signals of sufficiently large entropies on the physical layer of today's communication systems. There are two basic directions in this field, [6]:

(i) First, extracts these signals from sources independent of communication channels (source model),

(ii) Second, extracts these signals from the noise of the present communication channels (channel model).

As illustrated in Figure 1, a source model for secret-key agreement represents a situation in which three parties, Alice, Bob, and Eve, observe the realizations of a DMS - Discrete Memoryless Source (XYZ, P_{XYZ}) with three components. The DMS is assumed to be outside the control of all parties, but its statistics are known. By convention, component X is observed by Alice, component Y by Bob, and component Z by Eve. Alice and Bob's objective is to process their observations and agree on a key K about which Eve should have no information.

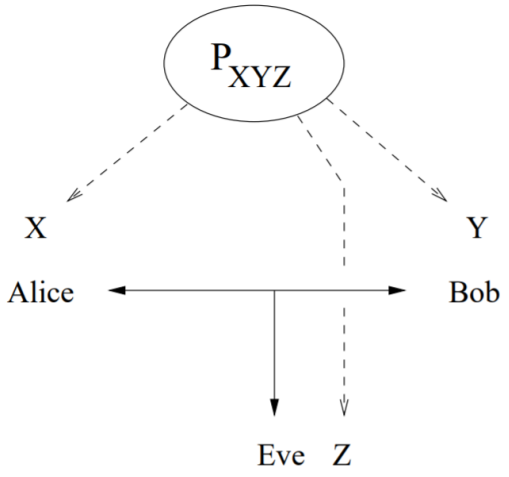


Fig.1. Secret-key agreement by Public Discussion from Common Information [7], [8].

Alice and Bob can exchange messages over a noiseless, two-way, public and authenticated channel. That is, all messages are overheard by Eve and the existence of the public channel does not provide Alice and Bob with an explicit advantage over Eve. The rules by which Alice and Bob compute the messages they exchange over the public channel and agree on a key define a four stage key distillation strategy, [9]:

1. *Randomness sharing*. Alice, Bob, and Eve observe n realizations of a DMS (XYZ, P_{XYZ}).

2. *Advantage distillation*. If needed, Alice and Bob exchange messages over the public channel to process their observations and to "distill" observations for which they have an advantage over Eve.

M. Milosavljevic is with the Technical Faculty, Singidunum University, Belgrade, Serbia E-mail: mmilosavljevic@singidunum.ac.rs

S.Adamović is with the Informatics and Computing Department, University Singidunum, Belgrade, Serbia E-mail: sadamovic@singidunum.ac.rs

A. Jevremović is with the Informatics and Computing Department, University Singidunum, Belgrade, Serbia E-mail: ajevremovic@singidunum.ac.rs

3. *Information reconciliation*. Alice and Bob exchange messages over the public channel to process their observations and agree on a common bit sequence.

4. *Privacy amplification*. Alice and Bob publicly agree on a deterministic function they apply to their common sequence to generate a secret key.

The largest achievable key rate is defined as key capacity and is given by

$$C_K=\max\{I(X;Y), I(X:Y|Z)\}, \quad (1)$$

where I(X;Y) denotes mutual information between X and Y, while I(X:Y|Z) denotes the same quantity conditioned by Z. In the special case, when Eva is totally independent from Alice and Bob, or equivalently, when Z is independent from X and Y, maximal key capacity is equal to

$$C_{k\,max} = I(X;Y). \quad (2)$$

Based on this facts, we propose application of key distillation protocol for generation of random sequences, based on DMS (XY, P_{XY}) with two components X and Y. Our primary goal is proof of proposed scheme when X and Y represents signals obtained from mouse cursor points and gaze tracker, measured during web browsing, see Fig.2.

II. Source of Randomness

To collect the eye tracking data, we used an inexpensive eye tracking device, made by a Danish startup company - The Eye Tri. The eye tracker was connected to a PC via USB3.0 interface. The technical specification of the eye tracker is the following:

- Sampling rate: 30 Hz
- Accuracy: 0.5° (average)
- Spatial resolution: 0.1° (RMS)
- Latency: < 20 ms at 60 Hz
- Calibration: 5, 9, 12 points
- Operating range: 45 cm – 75 cm
- Tracking area: 40 cm × 30 cm at 65 cm distance
- Screen sizes: Up to 24 inches
- API/SDK: C++, C# and Java included
- Data output: Binocular gaze data

The biggest advantage of the device used for our tests is its price - it costs around USD100. However, the low sampling rate makes this device unusable for analyzing the saccades, or for some other professional purpose (though, it is very effective for monitoring human-computer interaction). Some professional devices of this type provide a sampling rate of 1KHz (or even higher), so our assumption is that by using those devices the performances (in terms of time required to collect enough eye tracking data) of our solution would go up by more than sixteen times. However, having in mind that mouse movements are another limiting factor of the performance, this has to be verified experimentally.

The software used to collect the data from the eye tracking device was a combination of a custom Java-based desktop application (which was in charge of collecting gaze and mouse cursor points locally) and the HCI-MAP cloud-based platform [9]. The main purpose of using the HCI-MAP platform is to perform a time synchronization between the different sensors (eye trackers, EEG devices, applications, etc.), thus to collect the data from multiple sources in a time synchronized manner, and to use cloud computing resources to further process the data. However, for the real-life application of the approach proposed in this paper, all of the processing would be done locally, since there is no need for cloud-based processing, and the data should not leave a local computer because of the security reasons.

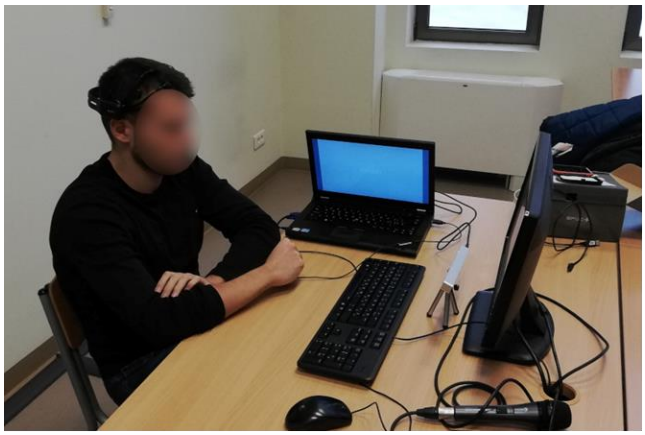


Fig. 2. Experimental settings during recording of measurement signals.

Our DMS is actually 4 dimensional (x_mouse,y_mouse,x_eye,y_eye), where (x_mouse,y_mouse) denotes coordinates of mouse pointer, while (x_eye,y_eye) measure coordinates of gaze.

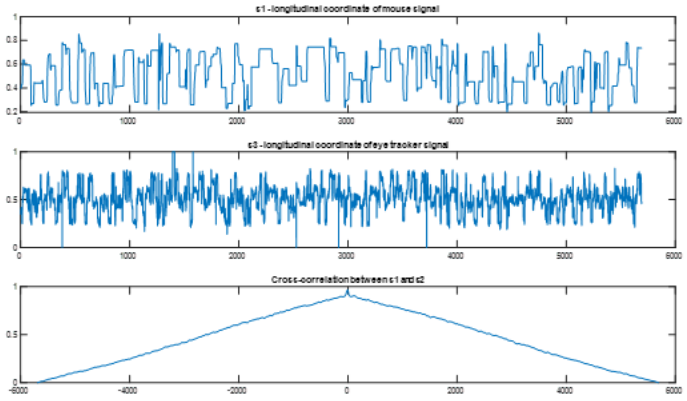


Fig. 3. Time domain behavior of tracking signals (longitudinal components) along with cross-correlation function. Time shift which maximizes cross correlation is found to be 5.

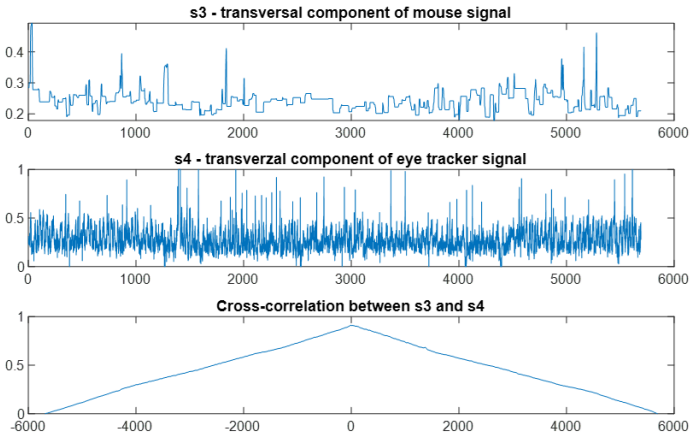


Fig. 4. Time domain behavior of tracking signals (transversal components) along with cross-correlation function. Time shift which maximizes cross correlation is found to be 4.

We show obtained signals in Fig.3 – Fig.6.

III. PROTOCOL

After measuring the tracker signals, its quantization and coding, four binary sequences are obtained at the Alice and Bob side. Then follow the standard steps of the source model secret-key agreement procedure, described at Introduction of this paper. As a measure of randomness we use Mean Shannon Block Entropy (MSBH), where average is taken over block size from 1 to 7. At the end of each round we apply a universal hash function [10], actually multiply given binary sequence with sparse random (0,1) matrix. Column-wise sparsity is 5%, while all algebraic operation are seen as operation over Galois Field GF(2).

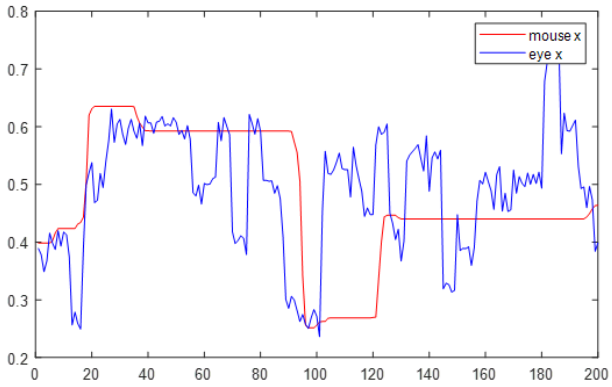


Fig. 5. Illustration of longitudinal components of tracking signals

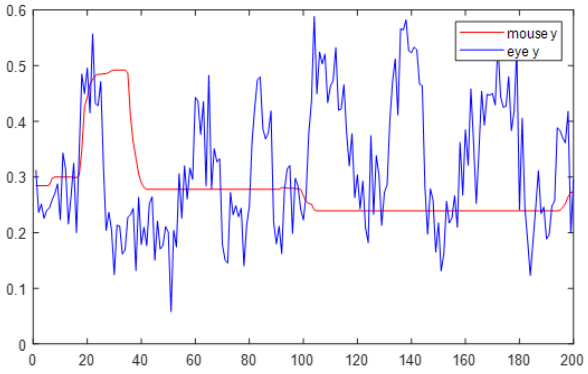


Fig. 6. Illustration of transversal components of tracking signals

IV. EXPERIMENTAL EVALUATION

Table I gives a summary of the basic results obtained for sensory signals (x_mouse, x_eye). Row named „Original“ denotes MSBH of rounds output, row named „Hashed“ stands for MSBH of universally hashed original sequences, while „Random“ denotes MSBH of random sequences generated by built in pseudorandom generator of MATLAB package of the same length.

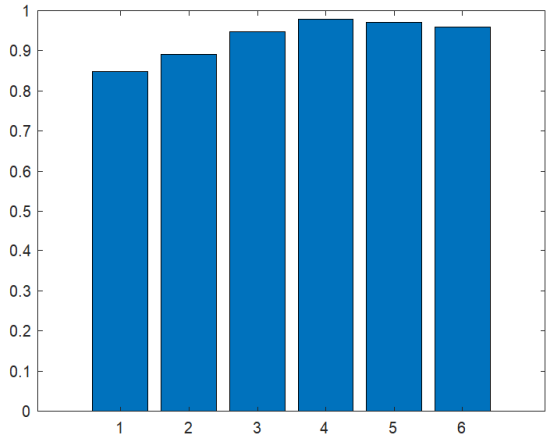


Fig. 7. Average Shannon block entropy evolution across 6 rounds of key distillation from x_mouse and x_eye signals.

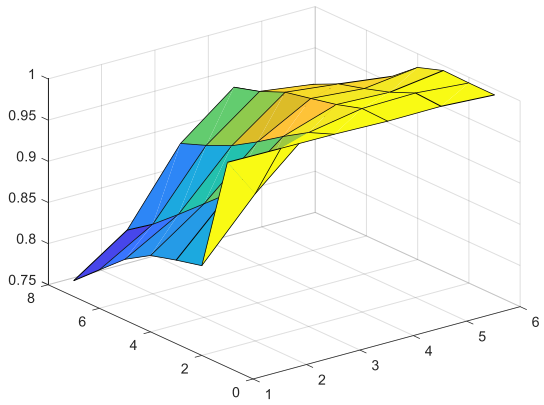


Fig. 8. Shannon block entropy evolution across 6 rounds and 1 to 7 block sizes of key distillation from x_mouse and x_eye signals.

Table I Summary of results for signals (x_mouse,x_eye)

	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
Original	0.8010	0.8140	0.8580	0.9430	0.9790	0.9740
Hashed	0.9946	0.9948	0.9943	0.9938	0.9889	0.9754
Random	0.9946	0.9947	0.9942	0.9932	0.9892	0.9723

Experimental results show that universal hashing purifies randomness. We introduce measure of gain in randomness obtained by universal hashing:

HashGain=sum((HM-HMHash)/sum(HMHash))*100 [%]

where HM means MSBE of original while HMHash denotes MSBE of hashed original.

The HahGain for pairs (x_mouse, x_eye) is - 4.9852 %. Negative sign means that hashing causes raising of randomness and consequently raising of MSBH.

From all this facts we can conclude, that proposed source of randomnes and distillation algorithm give sequences with very good random properties, even better than standard pseudorandom generators built in many famous software products, like MATLAB.

V. Conclusion

This paper presents a new protocol for generating secret cryptographic keys based on common randomness with an unlimited public discussion over an authenticated channel. The source of common randomness lies in the signals of mouse and eye tracker, which is not so rare in today everyday use. Experimental results prove feasibility of proposed approach. Obtained cryptographic keys are better at standard randomness tests compered to standard pseudo random generator approach.

Acknowledgment

This work has been supported by the Serbian Ministry of Education and Science (projects III44006, ON174008, III 47018 and TR32054). We would also like to thank prof. Panayiotis Zaphiris and prof. Andri Ioannou for providing us the eye tracker equipment and valuable advices for this research.

References

[1] Donald E. Knuth, “The Art of Computer Programming, Volume 2: Seminumerical Algorithms” - chapitre 3 : Random Numbers, Addison-Wesley, Boston, 1998.

[2] Milan M.Milosavljević, Saša Adamović, Cryptology 2, Singidunum University, 2018.

[3] Milan M.Milosavljević, „Phylosophical foundation of artificial intelligence“, In Proceedings of LII conference ETRAN-2008, Palić, jun 2008, VI.2.1-1-4

[4] Alain M. Turing, “Computing machinery and intelligence”, Computers & Thought, 11-35

[5] R. Ahlswede and I. Csiszar, “Common randomness in information theory and cryptography, Part I: Secret sharing,” IEEE Trans. Inform. Theory, vol. 39, pp. 1121–1132,

[6] U. Maurer, “Secret key agreement by public discussion from common information” IEEE Trans. Inform. Theory, vol. 39, pp. 733–742, May 1993.

[7] I. Csiszar and P. Narayan, “Secrecy capacities for multiple terminals,” IEEE Trans. Inform. Theory, vol. 50, pp. 3047–3061, Dec. 2004.

[8] Daniel Jost, Ueli Maurer, João L. Ribeiro, “Information-Theoretic Secret-Key Agreement: The Asymptotically Tight Relation Between the Secret-Key Rate and the Channel Quality Ratio”, Theory of Cryptography Conference TCC 2018: Theory of Cryptography pp 345-369, 2018.

[9] Aleksandar Jevremovic, Sladjana Arsic, Milos Antonijevic, Andri Ioannou, Nuno Garcia, "Human-Computer Interaction Monitoring and Analytics Platform – Wisconsin Card Sorting Test Application", HealthyIoT 2018 - 5th EAI International Conference on IoT Technologies for HealthCare, Guimarães, Portugal, 11/2018

[10] Carter, Larry; Wegman, Mark N., "Universal Classes of Hash Functions". Journal of Computer and System Sciences. 18 (2): 143–154, 1979.

Klasifikacija akvatičnih larvi insekata korišćenjem duboke konvolucione neuronske mreže i prenesenog učenja

Aleksandar Milosavljević, Đurađ Milošević i Bratislav Predić

***Apstrakt*—U radu je opisan metod za klasifikaciju tri vrste larvi hironomida (*Chironomidae*: *Diptera*, *Insecta*) na osnovu slika dobijenih pomoću mikroskopa i binokularne lupe. Kao klasifikator je iskorišćena duboka konvoluciona neuronska mreža ResNet-50 arhitekture koja je obučavana na 80% skupa slika, dok je preostalih 20% korišćeno za validaciju. S obzirom na relativno mali broj trening uzoraka, primenjena je tehnika prenesenog učenja (eng. *transfer learning*), tako da se krenulo od mreže koja je prethodno obučena na ImageNet trening skupu uz promenu vršnog klasifikatora mreže. Obučavanje je vršeno u dve faze. U prvoj fazi je obučavan samo vršni klasifikator na osnovu karakteristika ekstrahovanih propuštanjem slika kroz prethodno istrenirani konvolucioni deo mreže. Nakon toga je vršeno „fino podešavanje" obučavanjem celokupne mreže. Da bi se dodatno anulirao problem male količine trening podataka, primenjene su tehnike proširivanja podataka (eng. *data augmentation*) i odbacivanja (eng. *dropout*). Primenom pobrojanih tehnika ostvarena je idealna klasifikacija kako trening tako i validacionog skupa. Odgovarajući rezultati su prezentovani u radu.**

***Ključne reči*—Duboko učenje; konvolucione neuronske mreže; klasifikacija slika; preneseno učenje; pojačavanje podataka; akvatični insekti; hironomida.**

I. Uvod

Klasifikacija slika predstavlja problem gde se na osnovu skupa slika za koje je poznata odgovarajuća kategorija izgrađuje model koji je u stanju da za neke nove slike predvidi kategoriju sa određenom tačnošću. Sam zadatak nije jednostavan zbog različitih varijacija koje su prisutne na slikama. Varijacije mogu biti u položaju kamere u odnosu na objekat, veličini objekta na slici, same varijacije u izgledu jedinki neke klase, deformacije slike, zaklanjanje od strane drugih objekata, razlike u osvetljenju, kao i sadržaj pozadine.

Za rešavanje ovog problema tipično se koristi pristup zasnovan na podacima. Umesto da pokušavamo da opišemo svaku od klasa koja treba da se identifikuje, koristimo veliki broj primera za svaku od klasa kako bi obučili određeni model (klasifikator) da bude u stanju da ih identifikuje. Da bi mogli da ocenimo kvalitet obučenog modela pri klasifikaciji novih uzoraka, od polaznog skupa se uzima određeni deo koji se koristi u svrhu validacije.

U našem slučaju, problem je identifikacija vrsta akvatičnih insekata koje su važne za procenu kvaliteta površinskih voda. Jedna od najznačajnijih grupa u biomonitoringu voda su larve iz familije *Chironomidae* (*Diptera*: *Insecta*). Velika raznolikost vrsta i značajna uloga u ekosistemu čini ovu grupu ključnom u procesu bioprocene. Međutim, komplikovana identifikacija vrsta u okviru hironomida koja je vremenski zahtevna, uz obavezno angažovanje eksperata, značajno poskupljuje njihovu implementaciju u rutinske monitoring programe, propisane pravnom regulativom svake države [1].

Tradicionalni pristup problemu klasifikacije slika se zasnivao na „ručnom" projektovanju različitih ekstraktora karakteristika slike (eng. *features*) na osnovu koji bi se obučavao klasifikator [2, 3, 4, 5]. Iako su se veštačke neuronske mreže i ranije koristile [6], veliki napredak u ovoj oblasti nastupa uvođenjem konvolucionih neuronskih mreža (KNM) [7]. KNM kombinuju tri arhitekturne ideje pomoću kojih se ostvaruje određeni nivo invarijantnosti na translaciju i distorziju: lokalna receptivna polja, deljene težina i prostorno poduzorkovanje (eng. *subsampling*) [8]. KNM dolaze u žižu interesovana 2012. godine nakon pobede mreže pod nazivom AlexNet [9] na takmičenju *ImageNet Large Scale Visual Recognition Challenge* [10] (u daljem tekstu skraćeno ImageNet). AlexNet je ostvarila top-5 grešku od 15,3% što je bilo za 10,8% bolje od drugo plasiranog rešenja. Ovakav rezultat je bio moguć zahvaljujući obučavanju korišćenjem grafičkih procesora (GPU) što se smatra prekretnicom za razvoj dubokog učenja (eng. *deep learining*). U nekoliko narednih godina zabeležen je drastičan napredak dubokih KNM i poboljšanje rezultata na ImageNet takmičenju. Pobednik 2013. je mreža ZFNet [11] sa top-5 greškom 14,8%. Značaj ovog rešenja je prevashodno u tehnici za vizuelizaciju mapiranjem naučenih filtara u slike. Naredna godina donosi veliki napredak i dva značajna rešenja. VGGNet [12] postiže top-5 grešku od 7,3% uz promovisanje jednostavnosti i dubine. Pobednik 2014. godine sa top-5 greškom od 6,7% je GoogLeNet [13] koji pored značajno boljeg rezultata beleži i smanjenje parametara za 12 puta u odnosu na AlexNet. Naredne godine ResNet [14] beleži top-5 grešku od 3,6% što je gotovo duplo bolji rezultat u odnosu na prethodnu godinu. ResNet arhitektra je inspirisana filozofijom VGGNet-a uz uvođenje rezidualnog pristupa u učenju kako bi se olakšalo

Aleksandar Milosavljević – Elektronski fakultet, Univerzitet u Nišu, Aleksandra Medvedeva 14, 18000 Niš, Srbija (e-mail: aleksandar.milosavljevic@elfak.ni.ac.rs).

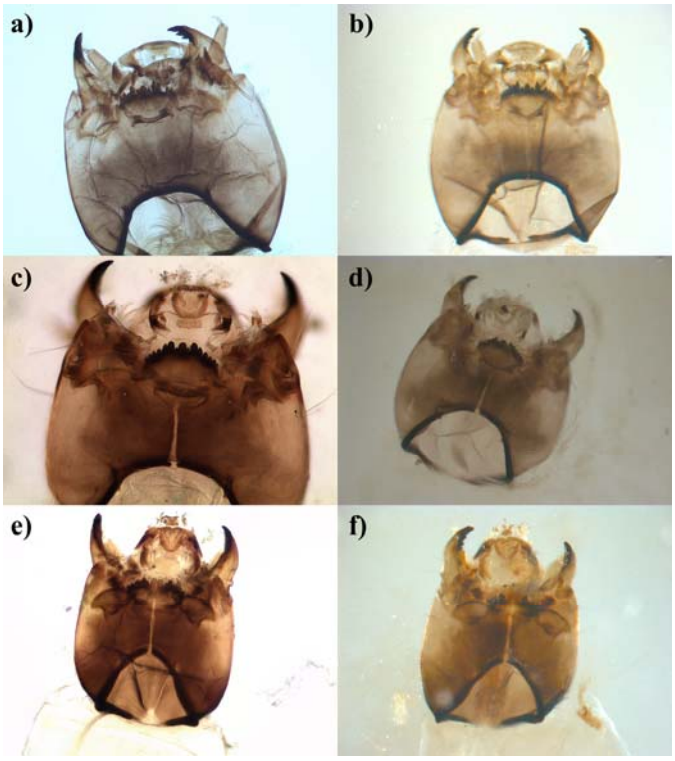
Đurađ Milošević – Prirodno-matematički fakultet, Univerzitet u Nišu, Višegradska 33, 18000 Niš, Srbija (e-mail: djuradj@pmf.ni.ac.rs).

Bratislav Predić – Elektronski fakultet, Univerzitet u Nišu, Aleksandra Medvedeva 14, 18000 Niš, Srbija (e-mail: bratislav.predic@elfak.ni.ac.rs).

obučavanje dubokih mreža. Mreža koja je pobedila 2015. je varijanta ResNet-a sa 152 sloja. Za potrebe klasifikacije vrsta larvi hironomida mi smo koristili ResNet-50 varijantu koja ima 50 slojeva.

Značaj ImageNet takmičenja nije samo u originalnom zadatku koji podrazumeva klasifikaciju slika u 1000 kategorija. S obzirom da retko koja realna primena može da po broju uzoraka parira ImageNet skupu (oko 1.200.000 trening uzoraka), ideja je da se mreže utrenirane na ImageNet skupu koriste kao polazna osnova za dalje obučavanje. Odgovarajući pristup se naziva preneseno učenje (eng. *transfer learning*) [15, 16] i predstavlja osnovu predloženog rešenja.

Rad je organizovan na sledeći način: poglavlje 2 sadrži opis skupa podataka koji je korišćen i načina na koji je formiran. U poglavlju 3 dat je opis predloženog metoda uključujući i detalje konkretne implementacije. Ostvareni rezultati i odgovarajuća diskusija dati su u poglavlju 4. Konačno, u zaključku (poglavlje 5) je dat rezime i naznačeni su pravci daljeg istraživanja.



Sl. 1. Primeri slika glavenih čaura larvi hironomida (ventralni aspekt). Slike (a), (c) i (e) su načinjene pomoću mikroskopa, dok su slike (b), (d) i (f) dobijene korišćenjem lupe. Na slikama (a) i (b) su predstavnici vrste *Chironomus riparius*, na slikama (c) i (d) vrste *Chironomus nudiventris*, dok su na slikama (e) i (f) pripadnici vrste *Microtendipes pendellus*.

II. SKUP PODATAKA

Za analizu podataka korišćene su tri različite vrste larvi hironomida, i to dve vrste istog roda: *Chironomus riparius* (Meigen, 1804) i *Chironomus nudiventris* (Ryser, Scholl & Wuelker, 1983), dok je treća vrsta *Microtendipes pedellus* (De Geer, 1776), koja pripada istoj podfamiliji (*Chironominae*) kao prethodno navedene vrste. Pre fotografisanja, larve su kuvane u kalijum hidroksidu (KOH) kako bi se povećala elastičnost i sprečilo pucanje glavenih čaura prilikom montiranja mikroskopskih preparata. Korišćeno je 200 jedinki po taksonu i sve su pozicionirane tako da se obezbedi ventralni aspekt glavene čaure koji je najinformativniji u pogledu *species*-specifičnih morfoloških karakteristika. Svi identifikacioni ključevi, koji se baziraju na morfološkim karakteristikama koriste uglavnom strukture usnog aparata (*mentum*, *mandibule* i *labrum*) koje se nalaze na ventralnom delu glave hironomida. Nakon što su mikroskopski preparati montirani, usledilo je njihovo slikanje korišćenjem dva različita optička instrumenta sa kamerom: mikroskop i binokularna lupa (DM2500 Leica System microscope; Leica DFC490 digital camera, Wetzlar, Germany). U zavisnosti od veličine larve, svaka jedinka je zumirana tako da ceo objekat (glavena čaura) stane u vidno polje optičkog instrumenta. Na mikroskopu i lupi smo prilikom slikanja koristili uveličanje u opsegu od 60 do 100 puta. S obzirom da je prilikom fotografisanja korišćen 3D objekat (glavena čaura), upotrebili smo metodu „z-stack“ kako bi slika bila oštra u svim delovima vidnog polja. Tačnije nakon formiranja 6 slika istog objekta sa različitim uveličanjem, usledilo je spajanje slojeva po *z* osi.

Broj dobijenih slika po vrstama u zavisnosti od načina akvizicije prikazan je u Tabeli I. Podela skupa podataka na trening i validacioni skup je izvršena po principu 20% podataka se odvaja za validaciju. Da bi validacioni skup po strukturi oslikavao celokupan skup, za validaciju je izdvajan svaki peti uzorak za svaku od klasa i za svaki od metoda akvizicije (lupa i mikroskop). Pošto su isti uzorci slikani mikroskopom i lupom, za validaciju su birani uzorci čiji broj pri deljenju sa 5 nema ostatak. Odgovarajući broj ukupnih trening i validacionih uzoraka je takođe dat u Tabeli I.

TABELA I
STRUKTURA SKUPA PODATAKA

Vrsta	Br. slika mikroskop	Br. slika lupa	Br. slika za obučavanje	Br. slika za validaciju
Chironomus riparius	172	196	295	73
Chironomus nudiventris	190	189	304	75
Microtendipes pendellus	179	181	289	71

III. OPIS METODA

Ostvarivanje robusnosti u režimu ograničenog broja trening uzoraka nameće korišćenje sledećih tehnika pri projektovanju i obučavanju klasifikatora:

1. Preneseno učenje
2. Odbacivanje
3. Proširivanje podataka

Preneseno učenje predstavlja osnovnu tehniku koja omogućava korišćenje dubokih KNM za rešavanje problema sa relativno malim skupom podataka. Zasniva se na korišćenju prethodno utrenirane mreže na nekom velikom skupu podataka, kao što je npr. ImageNet skup. Treniranje na ImageNet skupu obezbeđuje da mreža izgradi hijerarhiju različitih karakteristika koje se mogu naći na fotografijama generalno i koje je pogodno iskoristiti za klasifikaciju novih fotografija. Da bi takav transfer bio moguć, potrebno je zameniti vršni deo mreže zadužen za klasifikaciju i utrenirati ga koristeći niz karakteristika koje identifikuje duboka KNM. Uobičajeno je da se deo KNM ispred klasifikatora naziva enkoder. U zavisnosti od prirode novog skupa ovakav pristup može da bude i sasvim dovoljan. Međutim, u našem slučaju ulazni podaci ne predstavljaju nešto što se tipično nalazi na fotografijama, zbog toga je bilo neophodno izvršiti dvofazno obučavanje. Nakon obučavanja klasifikatora u prvoj fazi, u drugoj fazi je vršeno fino podešavanje cele mreže. Fino podešavanje nije ništo drugu do obučavanja celokupne mreže, kako klasifikatora tako i enkodera. Termin fino se koristi da naznači korišćenje vrlo malih koeficijenata učenja kako bi se što manje narušile polazne vrednosti parametara.

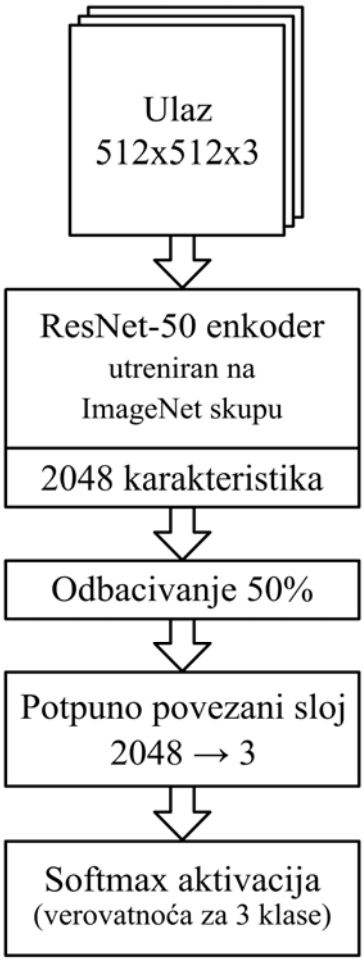
Druga tehnika koja je iskorišćena kako bi se povećala robusnost klasifikatora je odbacivanje [17]. Odgovarajući sloj je dodat nakon enkodera, tako da se u svakom koraku obučavanja odbacuje određen procenat (u našem slučaju 50%) karakteristika na osnovu kojih se vrši klasifikacija uzoraka. U fazi testiranja i eksploatacije mreže se uzimaju u obzir svi izlazi, ali se vrši skaliranje vrednosti za odgovarajući procenat odbacivanja. Na ovaj način se postiže u proseku isti nivo izlaza koji smo imali kod treniranja. Efekat koji se postiže primenom odbacivanja je da klasifikator mora da se oslanja na više različitih karakteristika pri određivanju kategorije. Na ovaj način se izbegava preterano prilagođavanje modela (eng. *overfitting*), a samim tim i bolji rezultati na validacionom skupu.

Obučavanje neuronske mreže da klasifikuje slike zahteva nalaženje karakteristika koje određuju odgovarajuće klase. Taj proces zahteva veliku količinu trening uzoraka kako bi se izolovale ključne karakteristike klasa i dobio klasifikator otporan na različite varijacije koje na slikama mogu da se jave. Kada imamo manju količinu trening podataka, a koristimo model velikog kapaciteta, dešava se model vrlo brzo „zapamti“ sve trening uzorke, ali zato daje loše rezultate na validacionom skupu. Tipično korišćena tehnika koja služi da se ovo izbegne je proširivanje podataka. Proširivanje podataka podrazumeva primenu nasumičnih transformacija nad ulaznim slikama tako da se u svakom trening ciklusu mreži predoči nešto što ranije nije „videla“. U zavisnosti od prirode skupa podataka, tipične transformacije uključuju okretanje, rotaciju, translaciju, skaliranje, zakošavanje, promenu osvetljaja i kontrasta, itd.

A. Arhitektura mreže

Vodeći se prethodno obrazloženim principima, za klasifikaciju larvi hironomida usvojena je arhitektura zasnovana na ResNet-50 [13] enkoderu prikazana na slici 2.

Na izbor ResNet-50 mreže utrenirane na ImageNet skupu kao enkodera je uticalo nekoliko faktora: dobri rezultati na ImageNet skupu, veličina mreže u pogledu broja parametara, memorijsko zauzeće u toku treniranja, brzina obučavanja, kao i dostupnost modela u korišćenom programskom okruženju. ResNet arhitektura generalno vrlo često predstavlja dobar inicijalni izbor zbog dobrog odnosa preciznosti i brzine obučavanja.



Sl. 2. Šematski prikaz korišćene arhitekture zasnovane na ResNet-50 mreži. Izlaz predstavlja raspodelu verovatnoća da je tekući uzorak primerak vrsta *Chironomus riparius*, *Chironomus nudiventris* i *Microtendipes pendellus*.

ResNet-50 enkoder poseduje 23.587.712 parametara, dok na izlazu daje 2048 karakteristika. Izlazi se dobijaju korišćenjem globalnog usrednjavanja (eng. *global average pooling*) po izlaznoj mapi karakteristika koja za korišćeni ulaz dimenzija 512x512 piksela iznosi 16x16x2048.

Na dobijenih 2048 izlaza iz enkodera se primenjuje odbacivanje sa faktorom 50%. Ovo u praksi znači da se u fazi obučavanja polovina, tj. 1024 nasumično izabranih izlaza postavi na nulu. Tako modifikovani izlazi enkodera se dovode na potpuno povezani sloj sa 3 neurona gde svaki izlaz odgovara jednoj klasi. Sloj poseduje 6.147 parametara koji se obučavaju u prvoj fazi. Na izlaze se primenjuje *softmax* aktivaciona funkcija (1) čime ovaj sloj dobija ulogu

klasifikatora karakteristika koje daje ResNet-50 enkoder. Izlazi *softmax* funkcije predstavljaju verovatnoće da je tekući uzorak pripadnik neke od klasa. Ovo praktično znači da pojedinačni izlazi imaju vrednosti iz intervala [0, 1] i da je zbir svih izlaza jednak 1.

$$S(y_i) = \frac{e^{y_i}}{\sum_j e^{y_i}}. \quad (1)$$

B. Implementacija i obučavanje mreže

Za implementaciju predložene arhitekture iskorišćen je programski jezik Python i biblioteka Keras [18]. Keras je biblioteka visokog nivoa koja definiše pojednostavljeni interfejs za implementaciju dubokih neuronskih mreža i u našem slučaju se oslanja na TensorFlow [19] biblioteku za realizaciju svih funkcionalnosti.

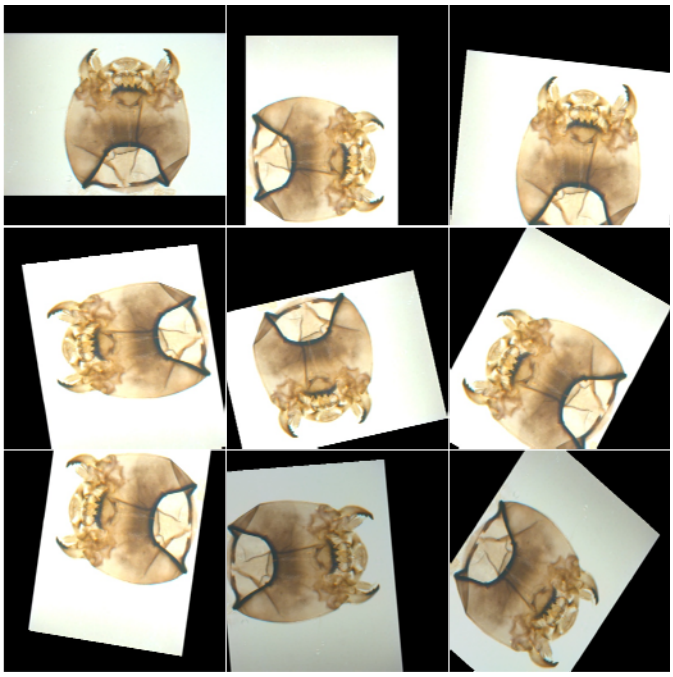
U okviru *applications* modula Keras poseduje nekoliko dubokih KNM arhitektura istreniranih na ImageNet skupu. Instanciranjem klase *ResNet50*, uz odgovarajuće parametre, dobijamo enkoder za naš model. Na izlaz enkodera se nadovezuje *Dropout* i *Dense* sloj sa *softmax* aktivacijom čime se dobija kompletan model. Kako se u prvoj fazi obučavanja težine ResNet-50 enkodera ne menjaju, potrebno je za sve konvolucione slojeve u enkoderu postaviti atribut *trainable* na *False*.

Kreirani model je kompajliran tako da koristi *Adam* [20] algoritam za optimizaciju (*optimizers* modul), *sparse_categorical_crossentropy* tip greške (*losses* modul), dok se kao metrika tačnosti koristi *sparse_categorical_accuracy* (*metrics* modul). Algoritam optimizacije je izabran zbog brze konvergencije, dok su greška za obučavanje i odgovarajuća metrika tačnosti standardni izbor za problem klasifikacije. Varijante ovih funkcija sa prefiksom *sparse_* se koriste kada se klase koje predstavljaju očekivani izlaz zadaju kao celi broj (0, 1 ili 2 u našem slučaju).

Za potrebe proširivanja podataka iskorišćena je Keras ugrađena klasa *ImageDataGenerator* koja se nalazi u *preprocessing* modulu, podmodul *image*. Pri konstruiranju odgovarajućeg generatora slika definišu se opsezi za različite transformacije koje će nasumično primenjivati. U našem slučaju korišćeno je horizontalno i vertikalno obrtanje slike, rotacija do ±90°, translacija do ±15% po oba pravca, promena osvetljaja do ±20%, zakošavanje i skaliranje do ±10%. Ilustracija je data na slici 3. Proširivanje podataka se vrši samo za trening skup, do za potrebe validacije koriste neizmenjene slike.

S obzirom da je korišćen generator slika, za obučavanje mreže koristi se metod *fit_generator*. Dodatna kontrola procesa obučavanja je u Kerasu moguća prosleđivanjem liste *callback* objekata. Odgovarajuće klase se nalaze u modulu *callbacks* i u našem slučaju iskorišćene su: *LearningRateScheduler*, *EarlyStopping*, *ModelCheckpoint* i *CSVLogger*.

LearningRateScheduler obezbeđuje definisanje proizvoljne funkcije za izmenu koeficijenta brzine obučavanja u zavisnosti od tekuće epohe obučavanja. U našem slučaju, ovaj *callback* je iskorišćen za implementaciju tzv. kosinusnog kaljenja (eng. *cosine annealing*) [21]. Kod kosinusnog kaljenja koeficijent obučavanja se smanjuje po kosinusnoj funkciji od neke inicijalne do neke minimalne vrednosti u toku određenog broja epoha koje su definisane periodom ponavljanja. U prvoj fazi obučavanja je korišćena perioda ponavljanja od 10 epoha, sa inicijalnim koeficijentom obučavanja u rasponu od 10^{-3} do 10^{-5}, uz smanjivanje 0,7 puta pri svakoj novoj periodi.



Sl. 3. Ilustracija proširivanja podataka. Prva slika (gore-levo) predstavlja ulaz, dok su ostale slike nastale primenom nasumičnih transformacija.

Kako ime sugeriše, *EarlyStopping callback* se koristi za ranije zaustavljanje procesa obučavanja ukoliko u određenom broju epoha nema napretka po određenom parametru. U našem slučaju je korišćeno 30 epoha i praćena je tačnost na validacionom skupu.

CSVLogger callback se koristi za snimanje greški i tačnosti nad trening i validacionim skupom u toku procesa obučavanja, a u cilju kasnije vizuelizacija ovog procesa.

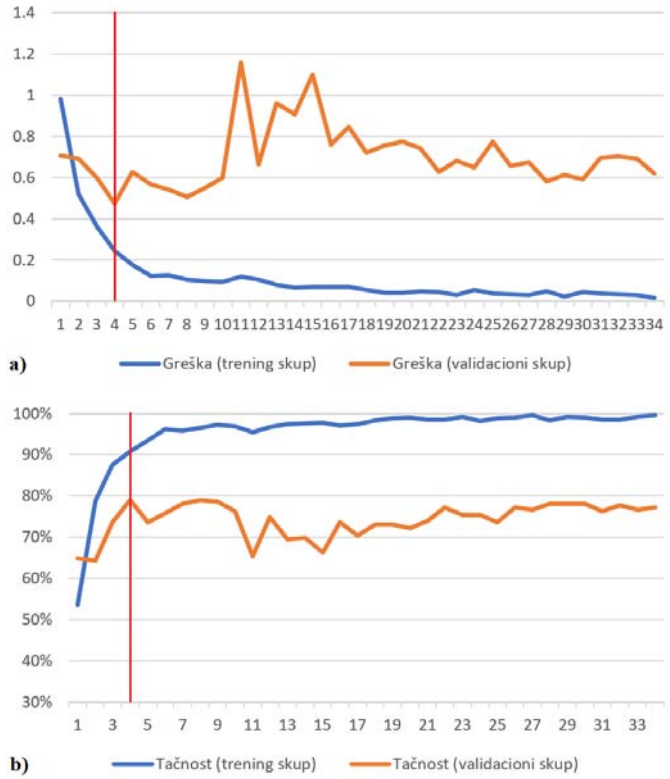
Konačno, *ModelCheckpoint* je iskorišćen za snimanje najboljeg rezultata u pogledu tačnosti postignute nad validacionim skupom. Ovako zabeležen model je iskorišćen za narednu fazu obučavanja, tj. fazu finog podešavanja.

Fino podešavanje je vršeno na gotovo identičan način, uz par sitnijih izmena. Nakon učitavanja modela dobijenog iz prve faze obučavanja, izvršeno je aktiviranje obučavanja za sve slojeve iz ResNet-50 enkodera (*trainable* atribut postavljen na *True*). Da bi se izbegla drastična izmena težina u enkoderu, koeficijent obučavanja se kretao u rasponu od 10^{-5} do 10^{-7}. Konačno, da bi se dobio model koji daje najbolje rezultate na celokupnom skupu podataka, umesto standardnog

validacionog, iskorišćen je kompletan neizmenjen skup podataka. Obučavanje je i dalje rađeno sa proširivanjem podataka trening skupa. Odgovarajući rezultati su prezentovani u narednom poglavlju.

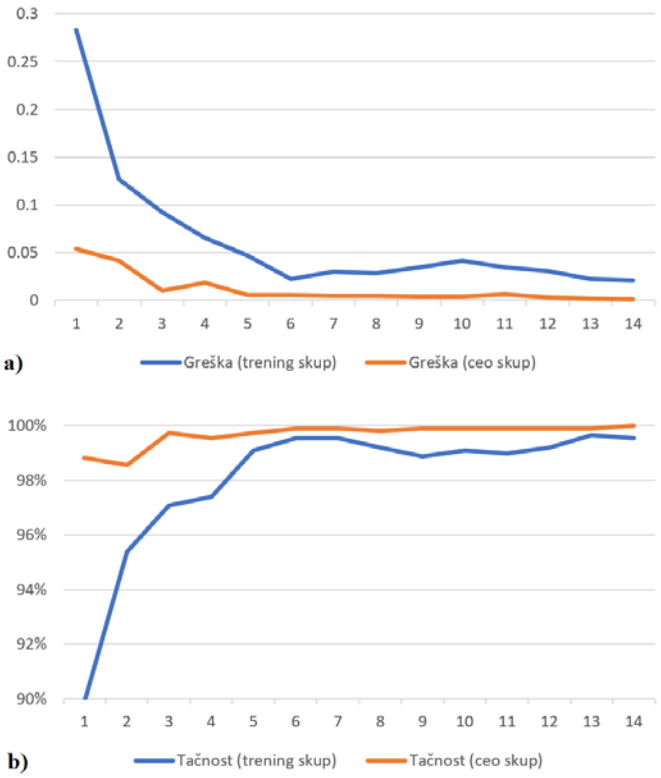
IV. Rezultati i diskusija

Tok obučavanja tokom prve faze prikazan je na slici 4. Prikazano je ukupno četiri grafikona: greška čija je minimizacija vršena i tačnost klasifikacije na trening i validacionom skupu. Obučavanje u prvoj fazi je vršeno u toku 34 epoha. Razlog za taj broj epoha je činjenica da je najbolja tačnost na validacionom skupu ostvarena već u 4 epohi, tako da je obučavanje prekinuto 30 epoha kasnije zbog *EarlyStopping*-a. Odgovarajući model je ostvario 78,99% na validacionom i 90,76% na trening skupu. Treba napomenuti da se zbog proširivanja podataka ne radi o originalnom trening skupu već nasumičnoj transformaciji istog u tekućoj epohi.



Sl. 4. Grafikon greške (a) i tačnosti klasifikacije (b) na nivou trening i validacionog skupa u toku prve faze obučavanja (obučavanje vršnog klasifikatora).

Druga faza obučavanja, fino podešavanje, kao polazni uzima prethodno dobijeni model. Inicijalno je za validaciju korišćen isti skup kao za prvu fazu, međutim pokazalo se da je mreža u stanju da ostvari idealnu klasifikaciju svih uzoraka iz ovog skupa. Ovakva situacija sprečava snimanje novih modela iako je bilo unapređenja na trening skupu. Da bi se izborili sa ovom situacijom, umesto validacije na prethodno definisanom validacionom skupu, vršena je validacija na celokupnom skupu podataka (unija validacionog i trening skupa). Za obučavanje je i dalje korišćen samo trening skup i vršeno proširivanje podataka, tako kod grafikona koji pokazuju tok obučavanja za drugu fazu (slika 5) imamo bolje rezultate za „validacioni“ skup od trening skupa. Razlog za ovu inverziju je i činjenica da se greška i tačnost na trening skupu dobija na osnovu podataka u toku obučavanja, dok se validacija vrši nakog završne epohe, pa je samim tim i mreža u datom trenutku utreniranija.



Sl. 5. Grafikoni greške (a) i tačnosti klasifikacije (b) na nivou trening i celokupnog skupa u toku druge faze obučavanja (fino podešavanje).

Gledajući grafikone za drugu fazu, moguće je uočiti da već nakon prve epohe dolazi do značajnog popravljanja tačnosti na celokupnom skupu (98,82%), dok nakon 14. epohe imamo idealnu klasifikaciju svih uzoraka (100%). Obučavanje je ručno zaustavljeno nakon toga.

V. Zaključak

U radu je opisan eksperiment čiji je cilj bio klasifikacija tri vrste larvi hironomida na osnovu slika dobijenih korišćenjem lupe i mikroskopa. Kao klasifikator iskorišćena je duboka KNM ResNet-50 arhitekture prethodno obučena na ImageNet skupu. Obučavanje je vršeno u dve faze. Prvo je obučavan samo vršni klasifikator, pa je nakon toga vršeno fino podešavanje celokupne mreže. Nakon druge faze su ostvareni idealni rezultati, tj. dobijen je model koji je u stanju da bez greške klasifikuje sve uzorke iz polaznog skupa podataka.

Sprovedeni eksperiment predstavlja samo polaznu studiju, sprovedenu na suženom skupu podataka koji sadrži uzorke

samo tri vrste (*Chironomus riparius*, *Chironomus nudiventris* i *Microtendipes pendellus*). Treba napomenuti da su prve dve vrste jako slične, te da je taj izbor namerno napravljen kako bi se ispitala mogućnost klasifikacije bliskih vrsta. Cilj eksperimenta je bio da pokaže da li primena KNM i dubokog učenja može da da dobre rezultate u detekciji larvi hironomida i kako različiti metodi akvizicije uzoraka utiču na rezultat. Ostvareni rezultati to i potvrđuju.

Dalja istraživanja će se fokusirati na metode za vizuelizaciju karakteristika na osnovu koji se KNM opredeljuje za neku klasu. Takođe, u planu je i proširivanje skupa podataka u pogledu broja klasa, ali i načina pripreme uzoraka.

ZAHVALNICA

Prikazani rezultati dobijeni su u okviru istraživanja na projektima III-43007 i III-47003 koje finansira Ministarstvo prosvete, nauke i tehnološkog razvoja Republike Srbije.

LITERATURA

[1] Đ. Milošević, V. Simić, M. Stojković, D. Čerba, D. Mančev, A. Petrović, M. Paunović, "Spatio-temporal pattern of the Chironomidae community: toward the use of non-biting midges in bioassessment programs", *Aquatic Ecology*, vol. 47, no. 1, pp. 37-55, 2013.

[2] R. M. Haralick, K. Shanmugam, I. Dinstein, "Textural Features for Image Classification", *IEEE Transactions on systems, man, and cybernetics*, vol. 6, pp. 610-621, 1973.

[3] G. Csurka, C. Dance, L. Fan, J. Willamowski, C. Bray, "Visual categorization with bags of keypoints", ECCV Workshop on Statistical Learning in Computer Vision, Prague, Czech Republic, 2004.

[4] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories", 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, vol. 2, pp: 524-531, 2005.

[5] M. L. Antonie, O. R. Zaiane, A. Coman. "Application of data mining techniques for medical image classification", Proceedings of the Second International Conference on Multimedia Data Mining, pp: 94-101, 2001.

[6] I. Kanellopoulos, G. G. Wilkinson, "Strategies and best practice for neural network image classification", *International Journal of Remote Sensing*, vol. 18, no. 4, pp. 711-725, 1997.

[7] Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, L. D. Jackel, "Handwritten digit recognition with a back-propagation network", *Advances in Neural Information Processing Systems*, pp. 396-404, 1990.

[8] Y. LeCun, Y. Bengio, "Convolutional networks for images, speech, and time series", *The handbook of Brain Theory and Neural Networks*, vol. 3361, no. 10, 1995.

[9] A. Krizhevsky, I. Sutskever, G. E. Hinton, "Imagenet classification with deep convolutional neural networks", *Advances in neural information processing systems*, pp. 1097-1105, 2012.

[10] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge", *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211-252, 2015.

[11] M. D. Zeiler, R. Fergus, "Visualizing and understanding convolutional networks", European Conference on Computer Vision, pp. 818-833, Cham, 2014.

[12] K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition", arXiv preprint arXiv:1409.1556, 2014.

[13] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, "Going deeper with convolutions", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-9, 2015.

[14] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778, 2016.

[15] L. Shao, F. Zhu, X. Li, "Transfer learning for visual categorization: A survey", *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 5, pp. 1019-1034, 2015.

[16] J. Yosinski, J. Clune, Y. Bengio, H. Lipson, "How transferable are features in deep neural networks?", *Advances in Neural Information Processing Systems*, pp. 3320-3328, 2014.

[17] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors", arXiv preprint arXiv:1207.0580, 2012.

[18] *Keras: The Python Deep Learning library*, https://keras.io, pristupano 30.3.2019.

[19] *TensorFlow: An end-to-end open source machine learning platform,* https://www.tensorflow.org/, pristupano 30.3.2019.

[20] P. D. Kingma, J. Ba, "Adam: A method for stochastic optimization", arXiv preprint arXiv:1412.6980, 2014.

[21] J. Jordan, "Setting the learning rate of your neural network", 1 March 2018, https://www.jeremyjordan.me/nn-learning-rate/, pristupano 30.3.2019.

ABSTRACT

In this paper we proposed a method for classification of three species of aquatic insect larvae based on images acquired using microscope and binocular magnifier. As a classifier we used deep convolutional neural network with ResNet-50 architecture. The network was trained using 80% of the sample images while the rest 20% were used for validation. Considering relatively small number of training samples we applied a technique called transfer learning, i.e. we started with a network pretrained on ImageNet set while only changing its top classifier. Training process went in two phases. In the first phase we only trained top classifier based on the features extracted processing images with pretrained convolutional part of the network. After that, we "fine-tuned" the network by training it as a whole. To further tackle the problem of relatively small number of training samples we applied data augmentation and dropout techniques. Using proposed method, we achieved ideal classification for both training and validation sets. The results are reported in the paper.

Classification of aquatic insect larvae using deep convolutional neural network and transfer learning

Aleksandar Milosavljević, Đurađ Milošević and Bratislav Predić

Identifikacija naslaga soli na seizmičkim snimcima korišćenjem metoda dubokog učenja za semantičku segmentaciju

Aleksandar Milosavljević

***Apstrakt*—Nekoliko oblasti u svetu koje su bogate naftom i zemnim gasom takođe poseduju velike naslage soli ispod površine. Zbog ove veze, otkrivanje preciznih lokacija naslaga soli je izuzetno značajno za kompanije koje se bave istraživanjem nalazišta ovih energenata. Lociranje naslaga soli se vrši na osnovu profesionalnih seizmičkih snimaka koje kasnije analiziraju ljudski eksperti. Rezultati ovakve analize često variraju i podložni su subjektivnosti eksperta koji analizu sprovodi. U cilju automatizacije ovog procesa i postizanja bolje preciznosti, kompanija TGS je sponzorisala takmičenje na Kaggle platformi održano u drugoj polovini 2018. godine [1]. Takmičenje je okupilo 3234 pojedinaca i timova, a u radu su prezentovani rezultati i iskustva autorovog učešća (446 pozicija). Metod predložen u radu se zasniva na obučavanju duboke konvolucione neuronske mreže za semantičku segmentaciju. Arhitektura korišćene mreže je inspirisana U-Net modelom u kombinaciji sa ResNet i DenseNet arhitekturama.**

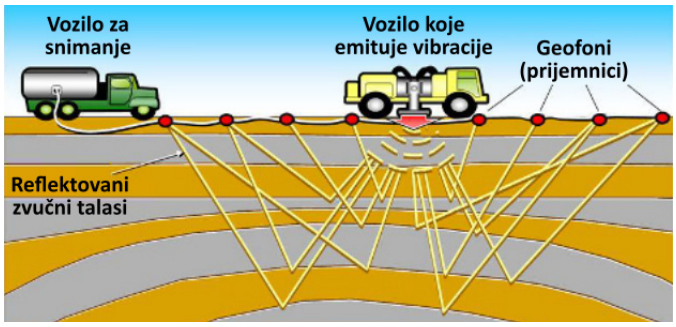
***Ključne reči*—Duboko učenje; konvolucione neuronske mreže; semantička segmentacija; seizmički snimci; naslage soli.**

I. Uvod

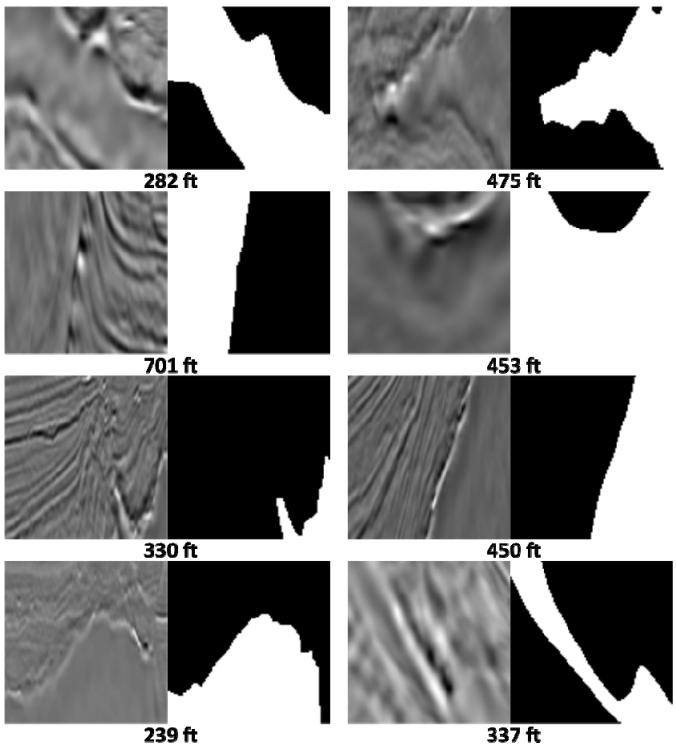
Seizmičko snimanje obezbeđuje vizuelizaciju podzemnih struktura i osnovna namena mu je otkrivanje potencijalnih nalazišta nafte i zemnog gasa. Zasniva se na emitovanju zvučnih talasa koji se odbijaju o podzemne strukture i bivaju detektovani na površini korišćenjem prijemnih uređaja koji se nazivaju geofoni (vidi sliku 1) [2]. Reflektovani zvučni signali se snimaju i kasnije obrađuju kako bi se dobili seizmički snimci koji predstavljaju 3D reprezentacije podzemnih struktura [3]. Seizmički snimci prikazuju granice između različitih tipova stena. Teorijski gledano, jačina odbijenog signala je direktno proporcionalna razlici fizičkih svojstava stena na mestu dodira. Ovo praktično znači da seizmički snimci sadrže informacije o granicama stenovitih naslaga, dok o samim stenama govore vrlo malo [1].

Otkrivanje rezervi nafte i zemnog gasa na osnovu seizmičkih snimaka se zasniva na otkrivanju rezervoarskih stena i u tom pogledu naslage soli igraju značajnu ulogu. Sedimenti soli poseduju karakteristike koje ih čine i jednostavnim i složenim za identifikaciju. Gustina soli je obično oko 2,14 g/cm^3 što je manje od većine obližnjih stena. Seizmička brzina soli je 4,5 km/s, što je obično brže od okolnih stena. Ova razlika uslovljava oštru refleksiju na granici sedimenta soli. Sediment soli je obično amorfnog oblika bez neke specijalne interne strukture, što znači da tipično nema mnogo refleksije unutar samog sedimenta sem ako ne potiče od drugih stena koje su tu zarobljene [1].

Aleksandar Milosavljević – Elektronski fakultet, Univerzitet u Nišu, Aleksandra Medvedeva 14, 18000 Niš, Srbija (e-mail: aleksandar.milosavljevic@elfak.ni.ac.rs).



Sl. 1. Ilustracija principa dobijanja seizmičkih snimaka.



Sl. 2. Primer nekoliko isečaka seizmičke slike, odgovarajućih maski koje definišu sedimente soli (bela boja) i dubina.

Zadatak Kaggle-ovog takmičenja koje je sponzorisala kompanija TGS je bio izrada algoritma koji je u stanju da na osnovu zadatih isečaka seizmičke slike dimenzija 101x101 svaki od piksela klasifikuje kao so ili ne-so. U cilju obučavanja modela obezbeđeno je 4000 isečaka seizmičkih slika i odgovarajuće informacije o tome gde se na slici nalazi so. Testiranje i ocena modela je vršena na 180000 isečaka. Pored samih slika i maski, za svaki od isečaka je obezbeđena informacija o dubini (u fitima) lokacije uzorka. Na slici 2 prikazano je nekoliko parova slika i maski.

Način na koji je postavljen, svrstava problem takmičenja u semantičku segmentaciju. Semantička segmentacija je jedan od ključnih problema računarskog vida i predstavlja naredni korak nakon klasifikacije slike i lokalizacije objekata na slici. Sva tri pobrojana problema se danas veoma uspešno rešavaju korišćenjem dubokih konvolucionih neuronskih mreža (KNM), te je i rešenje predloženo u radu zasnovano na toj tehnologiji.

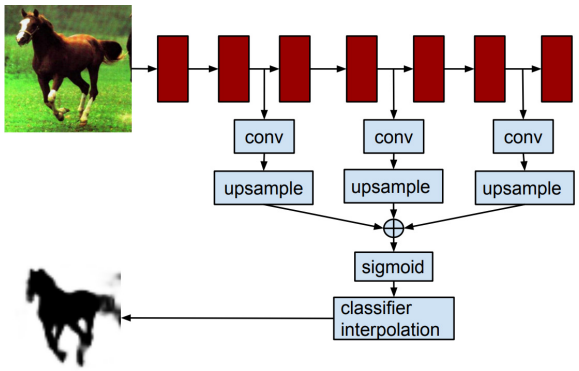
Rad je organizovan na sledeći način: poglavlje 2 sadrži opis postojećih rešenja iz oblasti semantičke segmentaciju i identifikacije naslaga soli. U poglavlju 3 dat je opis predloženog metoda u pogledu arhitekture KNM i detalja konkretne implementacije. Ostvareni rezultati i odgovarajuća diskusija dati su u poglavlju 4, nakon čega sledi zaključak u poglavlju 5.

II. Pregled postojećih rešenja

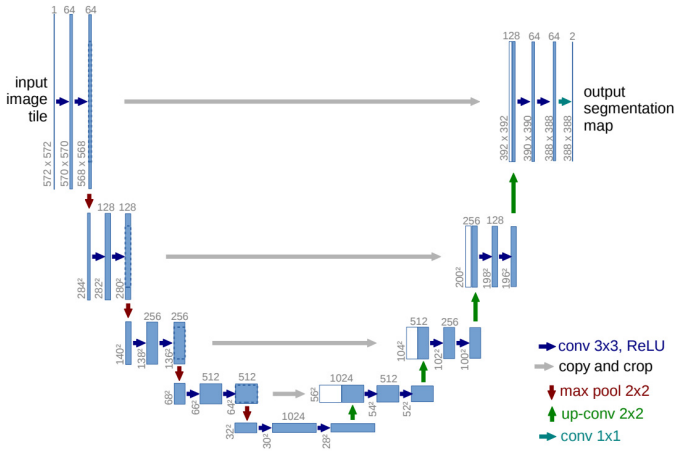
Problem analize seizmičkih slika i identifikacije soli zaokuplja pažnju velikog broja istraživača. Tipično za sve probleme računarskog vida, tradicionalni pristup analizi seizmičkih slika se zasnivao na „ručnom“ projektovanju različitih ekstraktora karakteristika (eng. *features*) i njihovoj kasnijoj analizi. U jednom od prvih radova iz ove oblasti, Pitas i Kotropoulos [4] su predložili metod zasnovan na analizi teksture u cilju segmentacije seizmičkih slika. Metodi zasnovani na korišćenju teksturnih atributa su aktuelni i u novije vreme [5, 6]. Harper i Clapp [7] su predložili proračun i korišćenje različitih seizmičkih atributa za identifikaciju naslaga soli. Izvođenje novih atributa seizmičke slike je pristup korišćen i u radu [8]. Analiza 3D seizmičkih slika je predmet istraživanja datih u [9, 10, 11]. Amin i Deriche [9] predlažu metod zasnovan na korišćenju 3D multidirekcionog detektora ivice. Wu [10] se oslanja na proračun verovatnoće da je nešto so u cilju identifikacije granice naslaga. Di *et al.* [11] predlažu višeatributnu klasterizaciju upotrebom k-means algoritma u cilju identifikacije granice sedimenta soli. Upotreba mašinskog učenja za identifikaciju četiri karakterističnih seizmičkih struktura na bazi različitih seizmičkih atributa ekstraktovanih iz slike je predložena od strane Wrona *et al.* [12].

Primena KNM [13] i dubokog učenja je dovelo do velikih pomaka u poslednjih nekoliko godina u nekoliko oblasti računarskog vida. KNM kombinuju tri arhitekturne ideje pomoću kojih se ostvaruje određeni nivo invarijantnosti na translaciju i distorziju: lokalna receptivna polja, deljene težina i prostorno poduzorkovanje (eng. *subsampling*) [14]. Duboke KNM su sposobne da izgrađuju hijerarhiju karakteristika koja ih čini pogodnim kako za klasifikaciju, tako i za lokalizaciju i semantičku segmentaciju objekata na slikama. Duboke KNM dolaze u žižu interesovana 2012. godine nakon pobede mreže pod nazivom AlexNet [15] na takmičenju *ImageNet Large Scale Visual Recognition Challenge* [16] (u daljem tekstu skraćeno ImageNet). AlexNet je ostvarila top-5 grešku od 15,3% što je bilo za 10,8% bolje od drugoplasiranog rešenja. Ovakav rezultat je bio moguć zahvaljujući obučavanju korišćenjem grafičkih procesora (GPU) što se smatra prekretnicom za razvoj dubokog učenja (eng. *deep learining*). U nekoliko narednih godina zabeležen je drastičan napredak dubokih KNM i poboljšanje rezultata na ImageNet takmičenju. Pobednik 2013. je mreža ZFNet [17] sa top-5 greškom 14,8%. Značaj ovog rešenja je prevashodno u tehnici za vizuelizaciju mapiranjem naučenih filtara u slike. Naredna godina donosi veliki napredak i dva značajna rešenja. VGGNet [18] postiže top-5 grešku od 7,3% uz promovisanje jednostavnosti i dubine. Pobednik 2014. godine sa top-5 greškom od 6,7% je GoogLeNet [19] koji pored značajno boljeg rezultata beleži i smanjenje parametara za 12 puta u odnosu na AlexNet. Naredne godine ResNet [20] beleži top-5 grešku od 3,6% što je gotovo duplo bolji rezultat u odnosu na prethodnu godinu. ResNet arhitektra je inspirisana filozofijom VGGNet-a uz uvođenje rezidualnog pristupa u učenju kako bi se olakšalo obučavanje dubokih mreža. Mreža koja je pobedila 2015. je varijanta ResNet-a sa 152 sloja.

Značaj KNM za klasifikaciju je u tome što se sa relativno malim izmenama mogu primeniti na problem semantičke segmentacije. Potpuno konvoluciona mreža [21] predstavlja upravo takvo rešenje. Zasniva se povećavanju prostorne dimenzije (eng. *upsampling*) mape karakteristika poslednjeg sloja i direktno preslikavanje u piksele mape segmenata. S obzirom da je prostorna komponenta poslednjeg konvolucionog sloja dosta gruba, u cilju dobijanja preciznijih rezultata moguće je formirati hiperkolonu [22] korišćenjem uvećanih konvolucionih mapa različitih dimenzija i obučavati izlaz na osnovu nje (vidi sliku 3). Dodatni konvolucioni sloj ispred sloja za uvećanje služi da izjednači broj kanala kako bi odgovarajuće mape mogle da se saberu i klasifikuju u na nivou svakog piksela.



Sl. 3. Semantička segmentacija korišćenjem hiperkolone (preuzeto iz [22]). Slojevi originalne KNM za klasifikaciju su prikazani crvenom bojom.



Sl. 4. U-Net arhitektura za semantičku segmentaciju (preuzeto iz [24]).

Nešto složenije arhitekture KNM za segmentaciju su predložene u [23] pod nazivom DeconvNet i u [24] pod nazivom U-Net. Na slici 4 prikazana je arhitektura U-Net mreže. Mreža poseduje enkoderski (leva strana) i dekoderski (desna strana) deo. Svaki dekoderski blok započinje konkatenacijom mape dobijene iz odgovarajućeg enkoderskog bloka i uvećane mape iz prethodnog dekoderskog bloka. Sa određenim izmenama, ovakva arhitektura je korišćena i u radu [25].

Zbog dobrih rezultata u drugim oblastima, primena dubokih KNM postaje podrazumevani izbor istraživača u oblasti segmentacije seizmičkih slika i identifikacije naslaga soli. Veliki broj radova [26, 27, 28, 29, 30, 31, 32, 33] u prošloj i tekućoj godini svedoči u korist te tvrdnje. Posebno bih istakli rad autora Babakhin *et al.* [33] koji opisuje prvoplasirano rešenje na TGS-ovom takmičenju [1].

III. Opis metoda

Pre nego se detaljnije pozabavimo opisom konkretnog rešenja, trebalo bi reći par reči o samom takmičenju. Sva Kaggle takmičenja imaju jasno definisano trajanje, pravila, podatke i metriku po kojoj se vrši evaluacija predloženih rešenja. U slučaju TGS-ovog takmičenja skup podataka za obučavanje se sastojao od 4000 slika dimenzija 101x101 i odgovarajućih segmentacionih maski. Test skup čini 180000 slika na osnovu kojih je potrebno predložiti maske čiji kvalitet se evaluira. Podaci o dubini u fitima su dostupni kako za trening tako i za test slike. Rezultati se prosleđuju u vidu CSV fajla gde se za svaku sliku iz test skupa maska kodira po *run-length encoding*-u.

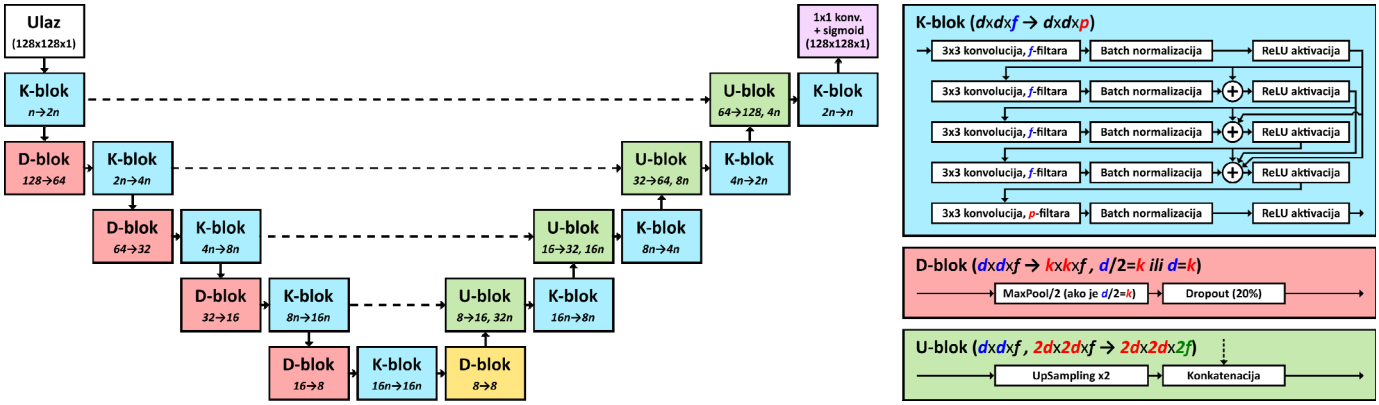
Takmičenje je trajalo od 19. jula do 20. oktobra 2018. godine. Svakog dana je moguće poslati maksimalno 5 rešenja za koje se dobija ocena po metrici koja je definisana za takmičenje. Ocena se vrši na osnovu 34% test podataka i najbolji rezultat određuje poziciju takmičara na javnoj rang listi. Konačni rezultati se objavljuju nakon završetka takmičenja na bazi preostalih 66% test podataka. Ocenjuju se samo dva predložena rešenja koje takmičar izabere. Bolja ocena određuju plasman na konačnoj (privatnoj) rang listi. Napomenimo da je do nedelju dana pre kraja takmičenja moguće vršiti spajanje učesnika u timove čime se njihovi pojedinačni rezultati objedinjuju.

Glavna prednost ovakve organizacije takmičenja je mogućnost da se za kratko vreme isproba i dobije neposredna ocena za veliki broj ideja, kao i da se vidi kvalitet sopstvenog rešenja u odnosu na ostale učesnike uz visok nivo objektivizma.

A. Arhitektura mreže

Na slici 5 je prikazana arhitektura mreže koja predstavlja konačno, tj. najbolje plasirano rešenje autorovog učešća na TGS-ovom takmičenju. Rešenje je zasnovano na U-Net [24] arhitekturi, zapravo je i nastalo modifikacijama U-Net mreže u želji da se popravi rezultat.

Kao što se sa slike 5 vidi, mreža se sastoji od tri tipa blokova (K, D i U). K-blok je najzastupljeniji i ujedno najsloženiji. U osnovi se sastoji se iz 5 konvoluciona sloja veličine kernela 3x3. Generiše se na osnovu dva parametra: ulazni broj filtara (*f*) i izlazni broj filtara (*p*). Prva četiri sloja imaju isti broj filtara (*f*) i „čvrsto su spregnuti“ operacijama sabiranja pre primene ReLU aktivacije, dok poslednji, peti, sloj ima *p* filtara i praktično služi da prilagodi broj filtara željenom izlazu. Specifična sprega slojeva sabiranjem izlaznih mapa je inspirisana ResNet [20] i DenseNet [34] arhitekturama, mada način na koji je izvedena predstavlja



Sl. 5. Prikaz arhitekture predložene KNM za segmentaciju naslaga soli. Mreža se sastoji iz tri tipa blokova (K, D i U) koji su detaljno prikazani sa desne strane. Konačni izlaz mreže se dobija 1x1 konvolucijom, sa jednim filtrom i sigmoidalnom aktivacijom na izlazu. Inicijalni broj filtara je označen sa *n*.

originalni doprinos. Pored konvolucionih slojeva, K-blok uključuje i *batch* normalizaciju i ReLU aktivaciju. Konkretan broj filtara u konvolucionim slojevima K-blokova je određen parametrom *n* koji predstavlja broj filtara u prvom konvolucionom sloju. Vrednosti sa kojima se eksperimentisalo su 16, 24 i 32.

D-blok predstavlja drugi korišćeni tip, a odgovarajući blokovi se nalaze u enkoderskog delu mreže nakon K-blokova. Namena D-bloka je duplo smanjivanje prostorne dimenzije korišćenjem *MaxPool* operacije (izuzetak je poslednji D-blok) i odbacivanje (eng. *dropout*) od 20% u cilju povećanja robusnosti mreže. Odbacivanje se primenjuje samo u fazi obučavanja mreže i predstavlja anuliranje određenog procenta elemenata mape karakteristika koja se dobija nakon operacije smanjivanja.

U-blok je pandan D-bloku u dekoderskom delu mreže. Namena mu je duplo uvećanje prostorne dimenzije mape karakteristika (*UpSampling* operacija) prostim ponavljanjem svake kolone i reda. Drugi zadatak ovog bloka je konkatenacija ovako uvećane mape se izlaznom mapom odgovarajućeg enkoderskog K-bloka (prikazano isprekidanim linijama), a u duhu U-net arhitekture.

Konačni izlaz mreže se formira na osnovu izlaza poslednjeg K-bloka primenom 1x1 konvolucije sa jednim filtrom i sigmoidalne aktivacione funkcije koja obezbeđuje izlaz u opsegu (0, 1) za svaki piksel. Pri formiranju prediktovane maske se odgovarajuće vrednosti zaokružuju na 0 ili 1.

B. Implementacija i obučavanje mreže

Za implementaciju predložene arhitekture iskorišćen je programski jezik Python i biblioteka Keras [35]. Keras je biblioteka visokog nivoa koja definiše pojednostavljeni interfejs za implementaciju dubokih neuronskih mreža i u našem slučaju se oslanja na TensorFlow [36] biblioteku za realizaciju svih funkcionalnosti. Konkretna implementacija koristi Kerasov funkcionalni API, a svaki blok je implementiran u posebnoj Python funkciji.

Od polaznog skupa trening podataka 80% slika je korišćeno za obučavanje, dok je preostalih 20% korišćeno za validaciju. Konkretno, podaci su deljeni u 5 podskupova, pa je obučavano 5 mreža, tako što je u svakoj od njih drugi podskup biran kao validacioni. Konačna predikcija je dobijena nalaženjem srednje vrednosti izlaza svih 5 mreža nakon čega je vršeno zaokruživanje. Da bi trening i validacioni skupovi sadržali reprezentativne podatke polaznog skupa, podeli na 5 podskupova je prethodilo sortiranje uzoraka po broju piksela soli na maski. Nakon toga je raspodela vršena dodelom svakog petog uzorka odgovarajućem podskupu.

Proširivanje podataka (eng. *data augmentation*) je vršeno i u fazi treniranja i u fazi pripreme rezultata. U fazi testiranja, proširivanje je vršeno na dva nivoa. Prvi nivo podrazumeva dupliranje uzoraka iz trening i validacionog skupa tako što se svaka slika i odgovarajuća maska preslikaju po vertikalnoj osi. Drugi nivo podrazumeva nasumične transformacije prostorne translacije, te skaliranje i pomeranje intenziteta. Prostorna translacija koristi osobinu da su slike veličine 101x101, dok je ulaz mreže 128x128. U 25% slučajeva slika se centrira po sredini (translacija od 13 piksela po ove ose). U ostalih 75% slučajeva se slika translira za proizvoljnu vrednost iz opsega 0–27 po jednoj i drugoj osi. Odgovarajuća translacija se vrši i sa odgovarajućom maskom. Popuna ostatka mape se vrši preslikavanjem u ogledalu vrednosti iz slike, odnosno maske. Transformacija intenziteta „skaliranje“ podrazumeva množenje intenziteta slike nasumičnom vrednošću iz opsega 0,8–1,2, dok transformacija „pomeranje“ podrazumeva dodavanje nasumične vrednosti iz opsega ±0,2.

Proširivanje podataka u fazi pripreme rezultata funkcioniše na sledeći način. Za svaku sliku iz test skupa nalazi se odziv mreže za originalnu i preslikanu sliku po vertikalnoj osi. Drugi odziv se preslikava, sabira sa originalnim i odgovarajući rezultat se deli sa 2. Ukoliko se vrši procena maski korišćenjem ansambla mreža, rezultati za svaku od mreža se snimaju (bez zaokruživanja), pa se naknadno usrednjavaju i zaokružuju kako bi se dobile konačne maske.

Obučavanje mreže je vršeno korišćenjem *Adam* [37] algoritma za optimizaciju (modul *optimizers*) i *binary_crossentropy* tipa greške (*losses* modul). Kao metrika tačnosti za izbor najboljeg rešenja korišćena je *binary_accuracy* funkcija iz *metrics* modula. Ova metrika u osnovi pokazuje koliko je piksela procentualno tačno klasifikovano. U informativne svrhe, praćena je i metrika zasnovana na odnosu preseka i unije pozitivnih piksela (eng. *Intersection over Union – IoU*).

Kontrola procesa obučavanja vršena je korišćenjem *callback* objekata. Odgovarajuće klase se nalaze u modulu *callbacks* i u našem slučaju iskorišćene su: *ReduceLROnPlateau*, *EarlyStopping*, *ModelCheckpoint* i *CSVLogger. ReduceLROnPlateau* obezbeđuje smanjivanje koeficijenta brzine obučavanja za zadati faktor ukoliko u određenom broju epoha nema napretka po određenom parametru. U našem slučaju praćena je tačnost na validacionom skupu, a koeficijent obučavanja je množen sa 0,1 ukoliko nema progresa nakon 10 epoha. Inicijalna vrednost koeficijenta je bila 10^{-3}. Za ranije zaustavljanje procesa obučavanja korišćen je *EarlyStopping callback* koji koristi sličan mehanizam. U našem slučaju je čekalo se 30 epoha ukoliko nema unapređenja tačnosti na validacionom skupu da bi se proces obučavanja zaustavio. *CSVLogger callback* se koristi za snimanje greški i tačnosti nad trening i validacionim skupom u toku procesa obučavanja, a u cilju kasnije vizuelizacija ovog procesa. Konačno, *ModelCheckpoint* je iskorišćen za snimanje najboljeg modela u pogledu tačnosti postignute nad validacionim skupom.

IV. REZULTATI I DISKUSIJA

Pre diskusije ostvarenih rezultata, potrebno je da se upoznamo sa metrikom po kojoj su vrednovani rezultati. Na takmičenju je korišćena metrika koja se zasniva na usrednjavanju IoU vrednosti po 10 pragova od 0,5 do 0,95 sa korakom 0,05. Tako, na primer, na pragu 0,5 prediktovani objekat se računa kao pogodak ukoliko je IoU veća od 0,5.

Ukoliko sa *T* označimo masku koja treba da se dobije, a sa *Y* prediktovanu masku. Na svakom pragu *t* preciznost *P(t)* se

računa korišćenjem sledećeg pravila:

$$P(t)=\begin{cases}0, & |T|=0 \wedge |Y|>0\\ 0, & |T|>0 \wedge |Y|=0\\ 1, & |T|=0 \wedge |Y|=0\\ IoU(T,Y)>t, & |T|>0 \wedge |Y|>0\end{cases} \quad (1)$$

Konačno, ocena kvaliteta prediktovane maske M se dobija usrednjavanjem rezultata po zadatih 10 pragova:

$$M=\frac{1}{10}\sum_{k=0}^{9}P(0{,}5+0{,}05k). \quad (2)$$

Pregled rezultata za različite konfiguracije u pogledu broja mreža koje čine ansambl i broja filtara u prvom sloju (parametar mreže n) je dat u Tabeli I. Pored privatnih rezultata koji su osnova za konačno rangiranje, prikazani su i javni rezultati koji su bili vidljivi i u toku trajanja takmičenja. Poređenja radi, u tabeli je prikazan i rezultat pobedničkog rešenja koje je opisano u [33], kao i autorovog prvog prijavljenog rešenja. Inače ukupan broj poslatih rešenja u toku takmičenja za pobednički tim iznosi 316, dok je u autorovom slučaju reč o 42 rešenja.

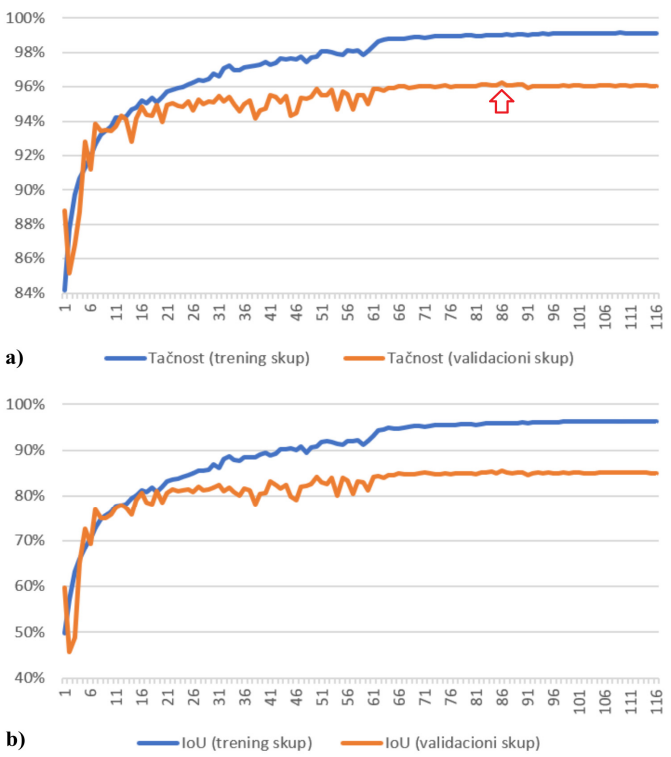
TABELA I
PREGLED REZULTATA ZA RAZLIČITE KONFIGURACIJE

Veličina ansambla	Broj filtara u prvom sloju / broj mreža	Javni rezultat	Privatni rezultat
5	16 / 5	0,82519	0,84525
5	24 / 5	0,82859	0,84771
5	32 / 5	0,83181	0,85087
10	24 / 5, 32 / 5	0,83314	0,85202
25	16 / 15, 24 / 5, 32 / 5	**0,83328**	**0,85241**
Rezultat pobedničkog rešenja		*0,88832*	*0,89646*
Rezultat prvog prijavljenog rešenja		*0,74828*	*0,76996*

U prikazanim rezultatima se mogu uočiti dva trenda. Jedan je da povećanje broja konvolucionih filtara pozitivno utiče na rezultat, a drugi je da se povećanje broja mreža u ansamblu takođe isplati. Inače, pošto korišćenje ansambla od 25 mreža nije baš uobičajena stvar, treba reći da je reč o poslednjem pokušaju da se popravi rezultat objedinjavanjem svega što je prethodno obučeno.

Ono što se iz rezultata ne vidi su svi drugi pokušaji koji nisu urodili plodom. Reč je o pokušajima sa drugim arhitekturama, korišćenjem drugačijih funkcija greške i drugačijih metrika za izbor najboljeg modela prilikom obučavanja, primena postprocesiranja korišćenjem morfoloških operacija, primena drugačijih načina za proširivanje podataka, itd.

Ilustracije radi, na slici 6 dat je prikaz toka obučavanja jedne od mreža u pogledu tačnost i IoU metrike.



Sl. 6. Grafikoni tačnosti (a) i IoU metrike (b) na nivou trening i validacionog skupa pri obučavanju pete mreže sa 32 inicijalna filtra.

V. ZAKLJUČAK

U radu je opisan metod za identifikaciju naslaga soli na seizmičkim snimcima korišćenjem duboke KNM za segmentaciju. Seizmičko snimanje i identifikacija naslaga soli je značajno zbog otkrivanja nalazišta nafte i zemnog gasa, a automatizacija ovog procesa je izuzetnog značaja za kompanije koje se bave istraživanjem nalazišta ovih energenata.

Rad je nastao učešćem autora na Kaggle takmičenju koje je sponzorisala kompanija TGS [1]. U radu je prikazano rešenje koje u konačnom plasmanu zauzelo 446 mesto od 3234 učesnika. U vreme kada je ovaj rad pisan, pojavio se i rad sa opisom prvoplasiranog rešenja [33]. Mnogo značajnije od konkretnog rešenja i plasmana je iskustvo stečeno učešćem u jednom ovakvom takmičenju.

Korišćenje takmičenja kao platforme za podsticaj i ubrzavanje naučnog i stručnog napretka već postaje oprobano rešenje. Setimo se samo do kakvog je napretka u rešavanju problema klasifikacije slika ImageNet takmičenje dovelo u svega nekoliko godina i kakav su uticaj rešenja koja su stasala na njemu imala na širu oblast. Kao učesnik TGS-ovog takmičenja, autor je bio svedok konstantnog pomeranja granica, kao opštih tako i ličnih. Čestitke pobednicima!

ZAHVALNICA

Prikazani rezultati dobijeni su u okviru istraživanja na

projektima III-43007 i III-47003 koje finansira Ministarstvo prosvete, nauke i tehnološkog razvoja Republike Srbije.

LITERATURA

[1] Kaggle.com, *TGS Salt Identification Challenge, Segment salt deposites beneath the Earth's surface*, July 2018, https://www.kaggle.com/c/tgs-salt-identification-challenge, pristupano 10.04.2019.

[2] J. F. Claerbout, *Imaging the earth's interior*, Oxford: Blackwell Scientific Publications, 1985.

[3] Chevron.com, *Seismic Imagining*, May 2015, https://www.chevron.com/stories/seismic-imaging, pristupano 13.04.2019.

[4] I. Pitas, C. Kotropoulos, "A texture-based approach to the segmentation of seismic images", *Pattern Recognition*, vol. 25, no. 9, pp. 929-945, 1992.

[5] T. Hegazy, G. AlRegib, "Texture attributes for detecting salt bodies in seismic data", SEG Technical Program Expanded Abstracts, pp. 1455-1459, 2014.

[6] M. A. Shafiq, Z. Wang, A. Amin, T. Hegazy, M. Deriche, G. AlRegib, "Detection of salt-dome boundary surfaces in migrated seismic volumes using gradient of textures", SEG Technical Program Expanded Abstracts, pp. 1811-1815, 2015.

[7] A. Halpert, R. G. Clapp, "Salt body segmentation with dip and frequency attributes", *Stanford Exploration Project,* Report SEP136, April 14, pp. 113-123, 2009.

[8] M. Shafiq, T. Alshawi, Z. Long, G. AlRegib, "SalSi: A new seismic attribute for salt dome detection", Proceedings of IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 1876-1880, Shanghai, China, March 2016.

[9] A. Amin, M. Deriche, "A new approach for salt dome detection using 3D multidirectional edge detector", *Applied Geophysics*, vol. 12, no. 3, pp. 334-342, 2015.

[10] X. Wu, "Methods to compute salt likelihoods and extract salt boundaries from 3D seismic images", *Geophysics*, vol. 81, no. 6, pp. IM119-IM126, 2016.

[11] H. Di, M. Shafiq, G. AlRegib, "Multi-attribute *k*-means clustering for salt-boundary delineation from three-dimensional seismic data", *Geophysical Journal International*, vol. 215, no. 3, pp. 1999-2007, 2018.

[12] T. Wrona, I. Pan, R. L. Gawthorpe, H. Fossen, "Seismic facies analysis using machine learning", *Geophysics*, vol. 83, no. 5, pp. O83-O95, 2018.

[13] Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, L. D. Jackel, "Handwritten digit recognition with a back-propagation network", *Advances in Neural Information Processing Systems*, pp. 396-404, 1990.

[14] Y. LeCun, Y. Bengio, "Convolutional networks for images, speech, and time series", *The handbook of Brain Theory and Neural Networks*, vol. 3361, no. 10, 1995.

[15] A. Krizhevsky, I. Sutskever, G. E. Hinton, "Imagenet classification with deep convolutional neural networks", *Advances in neural information processing systems*, pp. 1097-1105, 2012.

[16] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge", *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211-252, 2015.

[17] M. D. Zeiler, R. Fergus, "Visualizing and understanding convolutional networks", European Conference on Computer Vision, pp. 818-833, Cham, 2014.

[18] K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition", arXiv preprint arXiv:1409.1556, 2014.

[19] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, "Going deeper with convolutions", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-9, 2015.

[20] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778, 2016.

[21] J. Long, E. Shelhamer, T. Darrell, "Fully Convolutional Networks for Semantic Segmentation", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431-3440, 2015.

[22] B. Hariharan, P. Arbeláez, R. Girshick, J. Malik, "Hypercolumns for Object Segmentation and Fine-grained Localization", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 447-456, 2015.

[23] H. Noh, S. Hong, B. Han, "Learning Deconvolution Network for Semantic Segmentation", Proceedings of the IEEE International Conference on Computer Vision, pp. 1520-1528, 2015.

[24] O. Ronneberger, P. Fischer, T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation", International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 234-241, Springer, Cham, 2015.

[25] V. Badrinarayanan, A. Kendall, R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder architecture for Image Segmentation", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 12, pp. 2481-2495, 2017.

[26] J. S. Dramsch, M. Lüthje, "Deep-learning seismic facies on state-of-the-art CNN architectures", SEG Technical Program Expanded Abstracts, pp. 2036-2040. 2018.

[27] H. Di, Z. Wang, G. AlRegib, "Real-time seismic-image interpretation via deconvolutional neural network", SEG Technical Program Expanded Abstracts, pp. 2051-2055, 2018.

[28] H. Di, Z. Wang, G. AlRegib, "Deep convolutional neural networks for seismic salt-body delineation", AAPG Annual Convention and Exhibition, 2018.

[29] A. U. Waldeland, A. C. Jensen, L. J. Gelius, A. H. Schistad Solberg, "Convolutional neural networks for automated seismic interpretation", *The Leading Edge*, vol. 37, no. 7, pp. 529-537, 2018.

[30] Y. Zeng, K. Jiang, J. Chen, "Automatic Seismic Salt Interpretation with Deep Convolutional Neural Networks", arXiv preprint arXiv:1812.01101, 2018.

[31] Y. Shi, X. Wu, S. Fomel, "SaltSeg: Automatic 3D salt segmentation using a deep convolutional neural network", *Interpretation*, vol. 7, no. 3 pp. 1-36, 2019.

[32] X. Wu, L. Liang, Y. Shi, S. Fomel, "FaultSeg3D: using synthetic datasets to train an end-to-end convolutional neural network for 3D seismic fault segmentation", *Geophysics*, vol. 84, no. 3, pp. 1-36, 2019.

[33] Y. Babakhin, A. Sanakoyeu, H. Kitamura, "Semi-Supervised Segmentation of Salt Bodies in Seismic Images using an Ensemble of Convolutional Neural Networks", arXiv preprint, arXiv:1904.04445, 2019.

[34] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, "Densely Connected Convolutional Networks", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700-4708, 2017.

[35] *Keras: The Python Deep Learning library*, https://keras.io, pristupano 20.4.2019.

[36] *TensorFlow: An end-to-end open source machine learning platform,* https://www.tensorflow.org/, pristupano 20.4.2019.

[37] P. D. Kingma, J. Ba, "Adam: A method for stochastic optimization", arXiv preprint arXiv:1412.6980, 2014.

ABSTRACT

Several areas of Earth that are rich in oil and natural gas also have huge deposits of salt below the surface. Because of this connection, knowing precise locations of large salt deposits is extremely important to companies involved in oil and gas exploration. To locate salt bodies, professional seismic imaging is needed. These images are analyzed by human experts which leads to very subjective and highly variable renderings. To motivate automation and increase the accuracy of this process, a company TSG has sponsored Kaggle competition that was held in the second half of 2018 [1]. The competition was very popular gathering 3234 individuals and teams. In this paper, we present the author's contribution (446th place). The method presented in the paper relies on training a deep convolutional neural network for semantic segmentation. The architecture of the network is inspired by U-Net model in combination with ResNet and DenseNet architectures.

Identification of salt deposits on seismic images using deep learning method for semantic segmentation

Aleksandar Milosavljević

The Implementation of Peak Windowing Technique

Borisav Jovanović, Srđan Milenković

***Abstract*—. The implementation of Peak Windowing method for Peak to Average Power Reduction (PAPR) is presented in the paper. The architecture is based on a FIR filter. The results of PAPR measurement are presented for Long-Term Evolution (LTE) and Wideband Code Division Multiple Access (WCDMA) waveforms.**

***Index Terms*—Peak to Average Power Ratio; Peak Windowing.**

Borisav Jovanović and Srđan Milenković are with the Faculty of Electronic Engineering Niš, University of Niš, Aleksandra Medvedeva Street 14, Niš, Serbia (e-mail: borisav.jovanovic@elfak.ni.ac.rs. srdjan@milenkovici.net)

Improved adhesion of hybrid acrylate films by nanocrystalline polyhedral oligo silsesquioxanes (POSS)

Nataša Z. Tomić, *ICTMF*, Mustafa Kalifa, *TMF*, Marija M. Vuksanović, *VINČA Institute,* Vesna Radojević, *TMF, Belgrade, Serbia*, Radmila M. Jančić Heinemann, *TMF*, Aleksandar D. Marinković, *TMF*.

***Abstract*— The objective of this study is to investigate the influence of the polyhedral oligo silsesquioxanes (POSS) structure on the adhesion behavior of composite films onto a metallic surface. The composite films consist of UV cured Bisphenol A glycidylmethacrylate/triethylene glycol dimethacrylate (Bis-GMA/TEGDMA) as matrix and reactive POSS structures as adhesion enhancers. Composites are made with 1, 3 and 5 wt. % of POSS particles. Adhesion is evaluated using the micro Vickers hardness testing method. The contact angle of hybrid films to the brass substrate is measured and compared to the adhesion parameter from micro hardness measurements. The shape and size of the indent are analyzed and correlated to the adhesion quality. Methods used in this paper for estimation of adhesion strength and quality clearly indicate that the best adhesion enhancer of Bis-GMA/TEGDMA matrix is POSS reagent containing both hydroxyl and allyl functional group.**

***Index Terms*— polyhedral oligo silsesquioxanes, hybrid materials, adhesion, acrylate.**

Nataša Z. Tomić is with the Innovation center of Faculty of Technology and Metallurgy, Karnegijeva 4, 11070 Belgrade, Serbia (e-mail: ntomic@tmf.bg.ac.rs).

Mustafa Kalifa is with the Faculty of Technology and Metallurgy, University of Belgrade, Karnegijeva 4, 11070 Belgrade, Serbia (e-mail: mustafakalifa4@gmail.com).

Marija M. Vuksanović is with the Vinča institute of nuclear sciences, Mike Petrovića Alasa 12-14, 11351 Vinča, Belgrade, Serbia (e-mail: mdimitrijevic@tmf.bg.ac.rs)

Vesna Radojević is with the Faculty of Technology and Metallurgy, University of Belgrade, Karnegijeva 4, 11070 Belgrade, Serbia (e-mail: vesnar@tmf.bg.ac.rs).

Radmila M. Jančić Heinemann is with the Faculty of Technology and Metallurgy, University of Belgrade, Karnegijeva 4, 11070 Belgrade, Serbia (e-mail: radica@tmf.bg.ac.rs).

Aleksandar D. Marinković is with the Faculty of Technology and Metallurgy, University of Belgrade, Karnegijeva 4, 11070 Belgrade, Serbia (e-mail: marinko@tmf.bg.ac.rs).

Radon exhalation from fly-ash geopolymer mortar

Luka Rubinjoni, Igor Čeliković, Gordana Tanasijević, Miroslav Komljenović, Boris Lončar

Abstract—Geopolymers are a type of alkali activated binders, inorganic aluminosilicate polymers with amorphous cross-linked structure. Fly-ash is produced in abundance during coal firing, and poses an environmental and health risk in untreated powder form. Fly-ash geopolymer presents a sustainable alternative to Portland cement, due to lower net greenhouse gas emissions. Presence of naturally occurring radioactive elements in fly-ash is one of the factors taken into account when estimating the safety of fly-ash based building materials. Radon, a radioactive noble gas originating from the decay of radium, can leave the material and contribute to internal dose in closed spaces, so radon exhalation is of special interest. Radon exhalation for a standard sample of fly-ash geopolymer mortar was measured.

***Index Terms*— geopolymer; fly-ash; radon exhalation**

Luka Rubinjoni is with the Faculty of Technology and Metallurgy, University of Belgrade, Karnegijeva 4, 11000 Belgrade, Serbia (e-mail: rubinjoni@tmf.bg.ac.rs).

Igor Čeliković is with the Vinča Institute of Nuclear Sciences, University of Belgrade, Mike Petrovića Alasa 12-14, 11001 Belgrade, Serbia (e-mail: icelikovic@vin.bg.ac.rs).

Gordana Tanasijević is with the Institute for Multidisciplinary Research, University of Belgrade, Volgina 15, 11000 Belgrade (e-mail: gordana@imsi.bg.ac.rs).

Miroslav Komljenović is with the Institute for Multidisciplinary Research, University of Belgrade, Volgina 15, 11000 Belgrade (e-mail: miroslav@imsi.bg.ac.rs).

Boris Lončar is with the Faculty of Technology and Metallurgy, University of Belgrade, Karnegijeva 4, 11000 Belgrade, Serbia (e-mail: bloncar@tmf.bg.ac.rs).

The Strategy of Building and Using Simplified Robotic Models in Engineering Projects

Zorica Dodevska, Vladimir Kvrgic, Marko Mihic

***Abstract*—Building robotic models for the purposes of achieving goals in engineering projects, and its usage as a strategy advantage of such projects, are the main topics of this paper. Also, implementation of innovative technologies, like augmented reality (AR) and internet of things (IoT), on simplified robotic models (SRM) are at the focus in the paper. The authors find that the strategy of building and using SRM concept in complex engineering projects is fully justified, particularly before construction of the real objects that can be very expensive and time consuming. The fact that developed solutions (for example, AR apps) can be transferred to the final objects with necessary testing and modification strongly supports the authors' standing in this paper. Uniqueness of this work is reflected in the use of LEGO elements for building physical model of a centrifuge for pilot training, according to the project requirements, and developing an AR/IoT mobile app that could be applied to the real centrifuge. In addition to the strategy based on SRM, the authors stress the importance of such AR development strategy.**

***Index Terms*—Simplified robotic models; strategy; engineering projects; augmented reality; LEGO; centrifuge for pilot training.**

.

Zorica Dodevska is with the Research and Development Institute Lola Ltd, 70a Kneza Višeslava, 11030 Belgrade, Serbia (e-mail: zorica.dodevska@li.rs).

Vladimir Kvrgic is with the Institute Mihajlo Pupin, University of Belgrade, 15 Volgina, 11060 Belgrade, Serbia (e-mail: vladimir.kvrgic@pupin.rs).

Marko Mihic is with the Faculty of Organizational Sciences, University of Belgrade, 154 Jove Ilića, 11000 Belgrade, Serbia (e-mail: mihicm@fon.bg.ac.rs).

Author index

Author index

CIP - Каталогизација у публикацији
Народна библиотека Србије, Београд

621.3(082)(0.034.2), 534(082)(0.034.2), 004(082)(0.034.2), 681.5(082)(0.034.2), 621.039(082)(0.034.2), 66.017(082)(0.034.2), 57+61(048)(0.034.2), 006.91(082)(0.034.2)

INTERNATIONAL Conference on Electrical, Electronic and Computing Engineering (6 ; 2019 ; Silver Lake)
Proceedings of Papers [Elektronski izvor] = Zbornik radova / (Ic)ETRAN 2019, 6th International Conference on Electrical, Electronic and Computing Engineering in conjunction with ETRAN, 63rd National Conference on Electrical, Electronic and Computing Engineering, Silver Lake, Serbia, June 03 - 06, 2019 ; [glavni urednik Dejan Popović ; urednici, editors Slobodan Vukosavić, Boris Lončar]. - Beograd : Društvo za ETRAN : Akademska misao = Belgrade : ETRAN Society : Academic Mind, 2019 (Beograd : Akademska misao). - 1 elektronski optički disk (CD-ROM) ; 12 cm.
Sistemski zahtevi: Nisu navedeni. - Nasl. sa naslovne strane dokumenta. - Radovi na srp. i engl. jeziku. - Tekst ćir. i lat. - Tiraž 200. - Bibliografija uz svaki rad. - Abstracts.

ISBN 978-86-7466-785-9 (AM)

1. Друштво за електронику, телекомуникације, рачунарство, аутоматику и нуклеарну технику. Конференција (63 ; 2019 ; Сребрно језеро)

a) Електротехника - Зборници b) Акустика - Зборници c) Рачунарска технологија - Зборници d) Системи аутоматског управљања - Зборници e) Нуклеарна техника - Зборници f) Технички материјали - Зборници g) Биомедицина - Зборници h) Метрологија - Зборници

COBISS.SR-ID 280126476