Automatic Speech Recognition System for Dictating Medical Findings

Branislav Popović, *Member, IEEE*, Edvin Pakoci and Darko Pekar

***Abstract* — The paper presents an automatic speech recognition (ASR) system for dictating medical findings, developed by AlfaNum – Speech Technologies Ltd for the Pension and Disability Insurance Fund of the Republic of Serbia. The system is developed in a form of a distributed client-server architecture. The training of acoustic models is performed using a "chain" sub-sampled deep time-delay neural network (TDNN), while language models training is conducted using recurrent neural networks (RNNs), composed of "relu-renorm" layers followed by long short-term memory projection (LSTMP) components. The client application sends recorded user data to the server, where recognition of speech samples is performed in real time. The data is stored locally as well as in the central database, and can be exported in an appropriate form upon request. Recognition accuracy of 97% on a vocabulary of up to 50000 words is achieved.**

***Index Terms*—Automatic speech recognition, dictation, medical, Serbian, Latin.**

I. Introduction

AUTOMATIC speech recognition is a widely used technology for converting spoken words by users into text, i.e., for creating the transcription of the given conversation. Many human-machine interaction systems exist in a variety of different areas. ASR applications include dictation systems, voice assistant applications, smart homes, call centres, tools for aiding people with disabilities, and so on. As for the Serbian language, the state-of-the-art systems are constantly being upgraded. The recent research was mostly directed towards language modelling, because the previous systems had a lot of trouble dealing with the inflectivity of the Serbian language (i.e., having different cases, grammatical numbers or grammatical genders for words, which are all differentiated only by short word suffixes). The state-of-the-art Serbian language models involve deep recurrent neural networks that use embedding vectors as word representations and incorporate sub-word features, as well as additional lexical and morphological features for each word, on top of the usual "1-of-*N*" vector representation of words [1]-[2]. The state-of-the-art acoustic models, on the other hand, for several years now involve different variations of purely sequence-trained deep time-delay neural networks with subsampling, specifically designed to better model long temporal contexts [3]-[4]. These acoustic models also include accent-specific vowel models, Mel-frequency cepstral coefficients (MFCCs), pitch features and speaker identity vectors, or *i*-vectors [5], for the purpose of adaptation to different speakers and channels.

The need for training an ASR system for dictation or transcription in the medical field is not a new need – some solutions were suggested back at the end of the 20th century [6]. Most of the suggested uses include automatic speech recognition and a following transcriptionist review, and sometimes even a final physician review [7]. The main goal is to minimize clinically relevant errors, and provide reasonable general quality in practice [7]-[8]. Depending on the use (acoustic and linguistic complexity and variability, expected vocabulary size, etc.), different accuracy rates are reported, as well as different changes in department productivity. Sometimes there are changes for the worse too, as some doctors found that even though automatic speech recognition helps the overall department productivity (high gains), the length of time it takes to finish the dictation and produce the final report often increases [9]-[10]. Generally, physician surveys usually do find that most of them agree that the usage of ASR technology is a good idea, even though only a part of them (e.g. about half) report time savings [11].

The system presented in this paper is so far the only medical ASR system in existence for the Serbian language. It was created using the best available acoustic and language models mentioned above, as well as additional textual medical data obtained directly from the eventual user of the system, the Serbian Pension and Disability Insurance Fund. The system is implemented using a classic client-server architecture. The client application was specifically developed for this purpose from scratch, as it needed to fulfil very specific requirements for the Fund as well as to provide additional functionalities, like creating a final report in the specific legal form and exporting it to the central database.

The remainder of this paper is organized as follows. Section II discusses the system architecture, including the applied techniques and the used training databases. Section III is about the client application interface. Section IV describes the testing procedure and the accuracy of the system. Finally, section V gives a short recapitulation and conclusion.

Branislav Popović is with the Department for Power, Electronic and Telecommunication Engineering, Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia, Department for Music Production and Sound Design, Academy of Arts, Alfa BK University, Nemanjina 28, 11000 Belgrade, Serbia, and Computer Programming Agency Code85 Odžaci, Železnička 51, 25250 Odžaci, Serbia (phone: +381613207989; e-mail: bpopovic@uns.ac.rs).

Edvin Pakoci is with AlfaNum Speech Technologies, Bulevar vojvode Stepe 40, 21000 Novi Sad, Serbia (e-mail: edvin.pakoci@alfanum.co.rs).

Darko Pekar is with AlfaNum Speech Technologies, Bulevar vojvode Stepe 40, 21000 Novi Sad, Serbia (e:mail: darko.pekar@alfanum.co.rs).

II. System Architecture

The system is implemented in the form of a standard client-server architecture. Speech samples are sent to the ASR server to be processed and recognized. Voice activity detection is carried out implicitly, based on the most probable phoneme sequence for a given frame and the calculated signal energy. The recognition results are conveyed to the client in real time.

A. Server Side

The server recognizes chunks of audio samples. Audio content recognition is enabled for up to 15 users in parallel. Different grammars (i.e., language models) are provided, depending on the currently chosen textual field (domain of interaction), e.g., one of the fields allows dictation of medical findings in Latin, which is why a special grammar had to be trained for that specific purpose.

The baseline model is a "chain" sub-sampled time-delay deep neural network. The network is trained using cross-entropy training and a sequence-level objective function [3], [4], [12], while the training procedure consists of the pre-DNN and the DNN phase. For the pre-DNN phase, static features, including 14 Mel-frequency cepstral coefficients (MFCCs), energy and 3 pitch-related features – probability of voicing, log-pitch and delta-pitch, as well as their first and second order derivatives are extracted (the final feature vector is 54-dimensional). This phase consists of an initial flat-start monophone HMM-GMM training, triphone HMM-GMM training (targeting 3500 HMM states and 35000 Gaussians both for the first and second triphone pass), and speaker adaptive training (SAT, targeting the same model complexity). The final pre-DNN HMM-GMM model is used to provide input data alignments for the deep neural network (DNN) training. For this phase, 40 high-resolution MFCCs together with the 3 previously described pitch-based features, and a 100-dimensional speaker identity vector (*i*-vector) are used as features, producing a 143 dimensional feature vector.

The TDNN consists of 10 hidden layers, each of them containing 1024 neurons. The lower layers are trained using temporal context windows that include the preceding, the current and the following frame. The training of higher layers is conducted using also windows of 3 frames, but with 3-frame-long gaps between them. Acoustic models are trained using the recently expanded speech database for the Serbian language. The database consists of audio book recordings (recorded in a studio environment, spoken by professional speakers, 32 male and 64 female speakers, 168 hours of data), radio talk show recordings (179 hours of data, 21 male and 14 female speakers) as well as mobile phone recordings from interactions between humans and machines (requests, questions, and other inquiries, 61 hours of data, 169 male and 181 female speakers). Audio data is sampled at 16 kHz, 16 bits per sample, mono PCM. The number of speakers is increased artificially, using various combinations of speech speed and pitch modifications for similarly long chunks of data for each speaker which had enough data (398 and 420 distinct sub-speakers are obtained for audio books and radio shows, respectively, while the mobile phone speakers didn't need to be broken up). A version of the original database with a predetermined amount of added background noise was also created and incorporated into the acoustic model training. The noise recordings varied in type, from traffic and "cocktail party" noises, to construction noises, wind noises, etc [1], [12].

The language models are trained using previously anonymized real-life document examples from the Serbian Pension and Disability Fund and additional Serbian corpuses in the administrative, scientific, literary and journalistic functional styles [1], [12]. Recurrent neural network language models (RNNLMs) are used for this purpose. The network consists of 3 layers with Rectified Linear Unit (ReLU) activation functions, followed by a renormalization block (i.e., "relu-renorm"), each one containing 512 embedded neurons. LSTMP layers are injected between consecutive relu-renorm layers, while both recurrent and non-recurrent projection dimensions are set to 256. Max *n*-gram order is set to 4, therefore approximating lattice rescoring by merging histories in the lattice if they share the same 4-gram history, which prevents the lattice from exploding exponentially. The language model training is run for 30 epochs – 210 iterations based on the amount of input data. The best iteration is calculated based on the objective function value on the previously extracted validation dataset, which does not take a part in RNNLM training.

After successful initialization and authorization, the ASR server is ready to communicate with client applications. During its operation, the ASR server will print out various information in its console, such as recognized users' commands, speech detection times, confidence scores, etc. Alternatively, the ASR server can also be started as a service, without displaying the console. All the information can also be written to log files.

B. Client Side

The client interface contains several cards for the header and all the 7 standardized textual fields from the Pension and Disability Insurance Fund legal form:

1. Personal data
2. Significant allegations of the compliant
3. Medical history, physical, laboratory and other findings
4. Diagnosis (in Latin)
5. Assessment and opinion on disputable issues, as well as opinion on significant facts and circumstances not considered in the previous proceedings
6. Assessment and opinion on the correctness of findings, opinions and evaluation of expert authorities in the first instance proceedings
7. Explanation of the assessment and opinion on the correctness of the findings, opinions and evaluation of the expert authority in the first instance proceedings

Switching among the cards can be done via the appropriate keyboard shortcut, or by clicking on the desired card name (below the application toolbar). In addition to the voice input, all standard options for working with text are enabled, such as

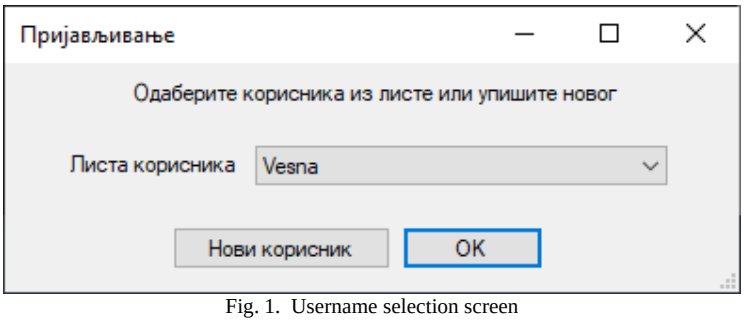

Fig. 1. Username selection screen

selecting, copying, cutting and pasting the text, changing the font size, bolding the text, as well as its conversion to the appropriate format (upper, lower, sentence or title case letters as well as letter spacing). Entries are saved automatically in the predefined folder on a local machine, as well as in the central database on a remote server. Separate subfolder is created for each of the cards, containing textual data in rich text format (RTF), together with the corresponding audio files (speech samples sent to the ASR server for recognition) and additional information about recognized words (JSON file). When saving the document, the application stores data about the medical worker (name, title, affiliation and codes) who is the current user, and creates the appropriate folder structure. This data will be linked with the username specified in the application and automatically withdrawn from the database each time a user opens the application. The interface of the client application is described in more details in the following section.

III. Application Interface

When the application is started, a username needs to be selected by choosing the appropriate name from the drop-down list presented in Fig. 1, or by entering a new username. The drop-down list is formed based on the entries previously stored in the remote database as well as the usage history of the concrete application. This name is used to identify the user communicating with the ASR server, in order to allow adaptation to the voice of a particular speaker (by linking its adaptation parameters to the selected username), therefore allowing multiple users to use the same client application with their own parameters. Speaker adaptation allows the speech recognizer to adapt the acoustic model parameters for a specific user, regardless of the gender and tone of the speaker's voice. Adaptation for any speaker should be conducted before the first recognition task. During adaptation, the user utters a predefined sequence lasting only a few seconds (a sequence of numbers in our case specifically, but other sequences would work as well). In addition to the acoustic characteristics of the voice, the signal energy level is also recorded, as well as the confidence measure for the recognized words (based on frame-level acoustic scores in the decoder). These two additional parameters are used for voice activity detection in the provided audio.

Graphical user interface of the client application is shown in Fig. 2 (personal data) and Fig. 3 (diagnosis). The recognition process on the ASR server begins by clicking on the "Start dictation" button ("Započni diktiranje" in Serbian), or by selecting the appropriate keyboard shortcut. When initiating recognition, it is important to be known which card is currently selected, so that recognition could be initiated with the appropriated grammar (i.e., language model), trained for the specific domain of interaction. If another card is clicked during recognition, the ASR server is automatically sent information to change the active language model to the one associated with the newly selected card, without having to manually stop the recognition and then restart. Recorded audio samples are sent in chunks to the ASR server, in order to enable online recognition – processing of samples begins as soon as the server accumulates enough data, and before the end of the signal, therefore allowing recognition in real time. The user ends the recognition by pressing the same button (whose text has now been changed to "End recognition" ("Kraj diktiranja")).

In addition to the final recognition result, the ASR server also allows continuous recognition – partial recognition results are provided during processing, i.e., upon voice activity detection (VAD), although these results can be modified by the decoder as new samples arrive. The final result for the previous VAD segment is determined after each long enough pause in speech (about one second or more). Both alphabets (Cyrillic and Latin) are supported, both during dictation and typing.

The client application supports a wide range of punctuation marks that are automatically converted from words during result printing, e.g., period, comma, colon, semicolon, question mark, exclamation mark, hyphen or dash, open and closed parenthesis, quotation marks, slashes, etc. – provided that the "*period*" word (i.e., "*tačka*" in Serbian) is converted only if recognized at the end of the speech segment. Automatic conversion of recognized digits, base and ordinal numbers, as well as dates is also supported. For example, the sequence "*sedmi oktobar hiljadu devetsto osamdeset prve*" (Eng. "*the seventh of October nineteen eighty-one*") will be converted into "*7.10.1981.*".

Numbers containing a decimal point can also be dictated. Furthermore, ordinal numbers (from "*first*" to "*hundredth*") followed by a slash are converted into a Roman numeral – this is used particularly in sequences such as "*po Glavi drugoj kroz B*" (Eng. "*according to Chapter second slash B*"), which will be converted into "*po Glavi II/B*" (Eng. "*according to Chapter II/B*").

Appropriate keywords can be used in combination with

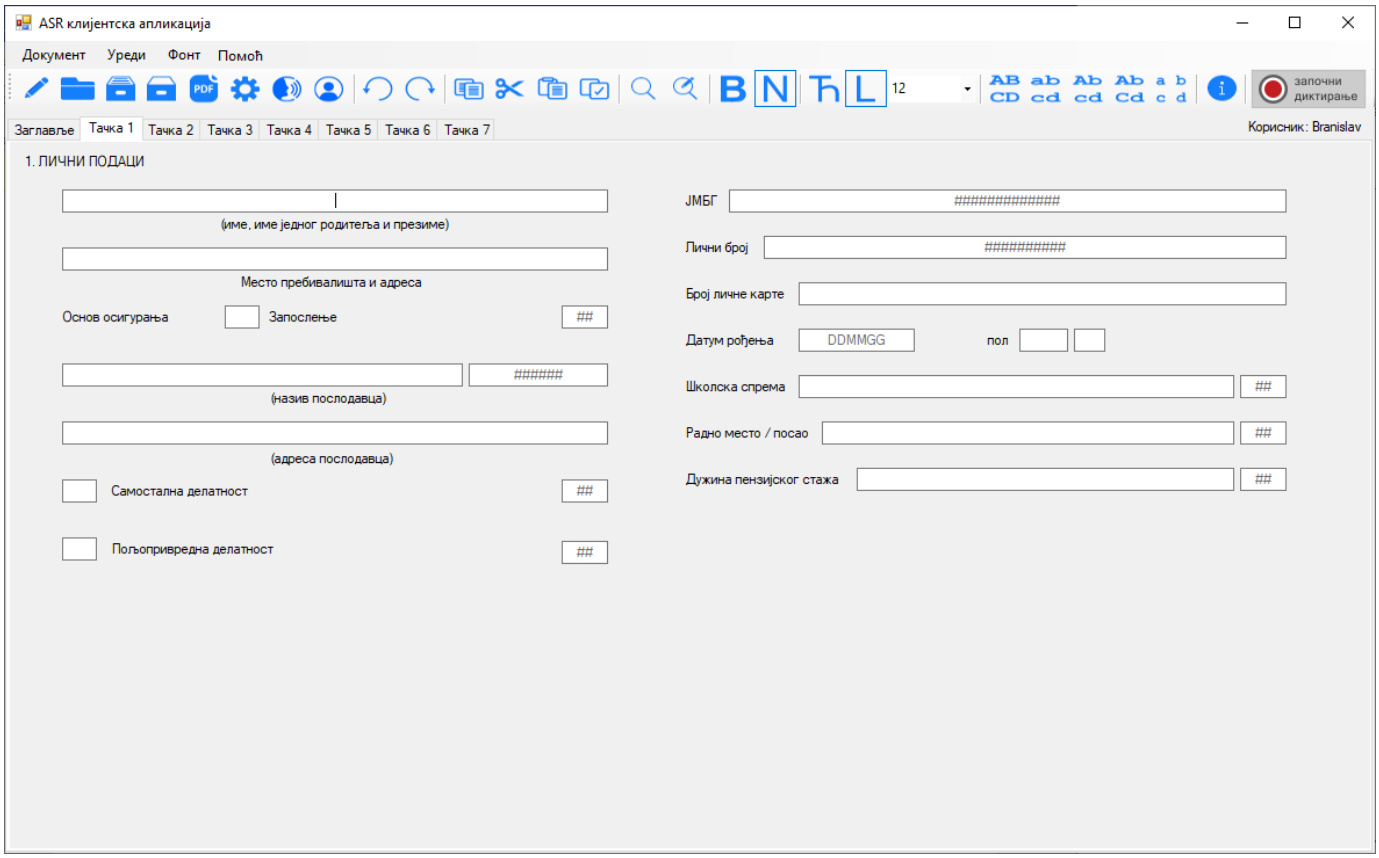

Fig. 2. Graphical user interface (personal data)

numbers. The word "*rimski*" (Eng. "*Roman*") in front of a number between 1 and 100 (either cardinal or ordinal) will make that number be written as a Roman numeral, and the command "*slovima*" (Eng. "*in text*") means that the number will be written in textual form, inside parenthesis (this needs to be done next to percentages in the Fund documents). It should also be noted that the word "*procenat*" or "*procenata*" (Eng. "*percent*") is always converted into the "%" sign, and the words "*plus*", "*minus*" and "*jednako*" (Eng. "*equals*") to the appropriate symbols "+", "-" and "=", respectively, if found next to a number. Several measurement units (meters, centimetres, millimetres, kilograms, grams, milligrams, millilitres, millimetres of mercury, per minute, per second, per litre, per square meter), are also automatically converted when recognized. Special keywords are defined for the dictation of secondary textual fields on cards 4 and 7 – "*šifra dijagnoze*" (Eng. "diagnosis code"), "*invalidnost*" (Eng. "*disability*"), "*telesno oštećenje*" (Eng. "*physical impairment*"), "*potreba za pomoći i negom*" (Eng. "*need for help and care*"), "*nesposobnost*" (Eng. "*incapacity*") and "*kontrolni pregled*" (Eng. "*control examination*"), after which one or two digits should be pronounced, or a two-digit number. A smaller set of commands such as "*obriši reč/rečenicu/paragraf*" (Eng. "*delete word/sentence/paragraph*") for correction purposes (these have to be said as a separate speech segment) and "*novi red*" (Eng. "*new line*", if said at the end of a speech segment) are also supported. There is also the "*kraj diktiranja*" (Eng. "*end dictation*") command, equivalent to clicking on the "End dictation" button – the current recording ends and all recognition results are returned to the client application.

All acronyms can be pronounced letter by letter ("*a b c d*" – Serbian Cyrillic pronunciation), or in the "singing" style, i.e. "*a be ce de*" (Serbian Latin pronunciation), and if they contain a vowel, they can be pronounced like a regular word (for example, the acronym "VOD" can be pronounced as "*v o d*", "*ve o de*", or simply "*vod*"). For some predefined acronyms and abbreviations, it is possible to pronounce whole words and still have the result written as an abbreviation, e.g. "*fundus oculi sinistri*" will be converted to "*FOS*", while "*Klinički centar*" (Eng. "*Clinical centre"*) will be converted into "*KC*".

IV. System Accuracy

The testing of the application was conducted in a relatively controlled environment (low overall and background noise), with high-quality microphones and previously prepared texts, at a time when the testers were already familiar with how to use the application. During testing, words were spoken at a normal rate, well-articulated and not overly stressed (i.e., neutral speech, no emotions in the voice). Speech speed was between 12 and 15 characters per second – neither too fast, nor too slow. Depending on the currently selected card, recognition is possible in Serbian or Latin, in real time. The recognition accuracy of about 97% on a vocabulary of about 50000 words is achieved in all domains of interaction. The accuracy was calculated in the usual way – by subtracting the word error rate (WER) from 100%, where WER is the sum of the number of word substitutions, deletions and insertions (in

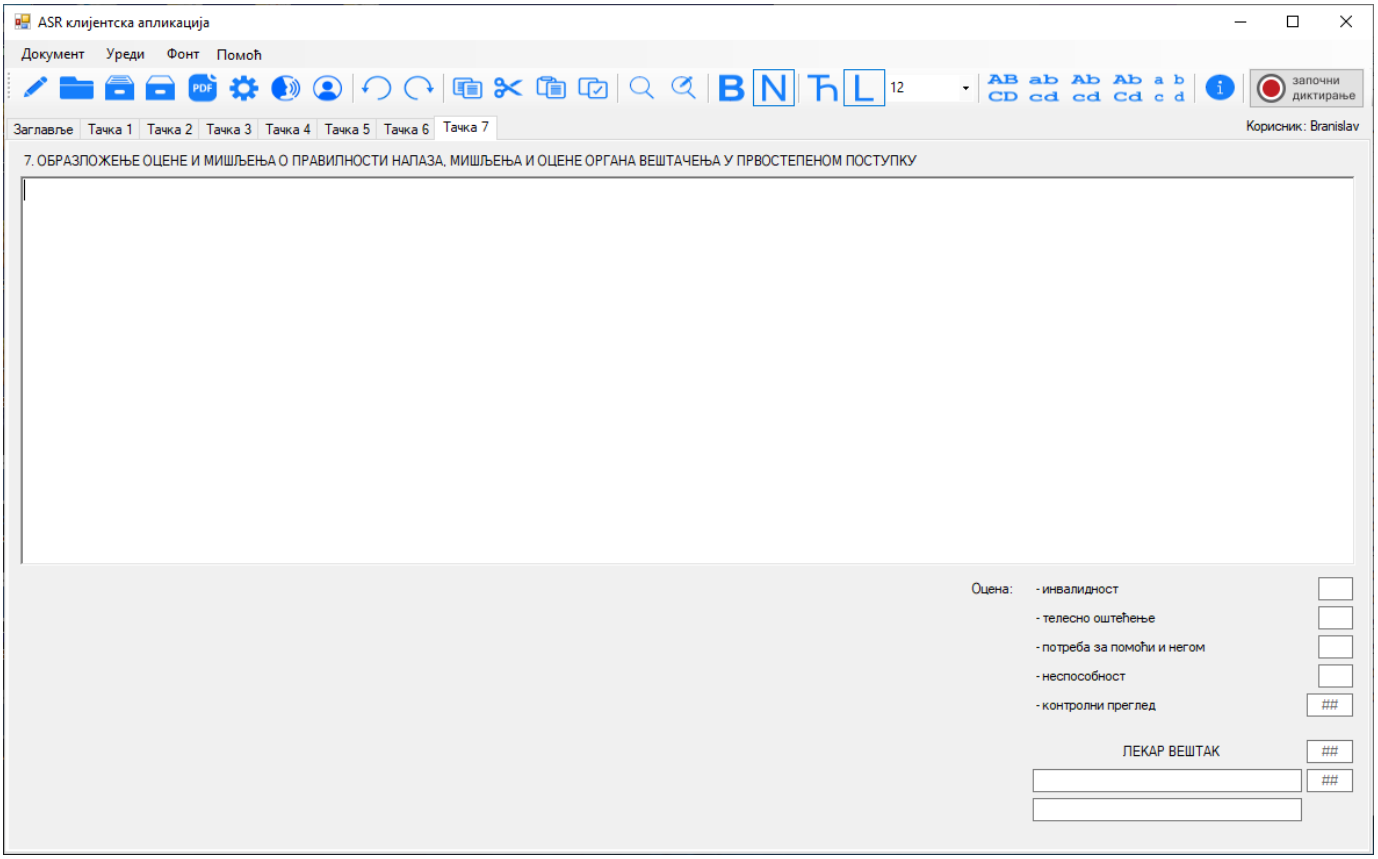

Fig. 3. Graphical user interface (diagnosis)

relation to the correct transcriptions) divided by the total number of words in the correct transcriptions. To evaluate the accuracy of the system more precisely, each digit (even in dates), punctuation mark, keyword and isolated letter were calculated as a separate word (in the same way as they are spoken). Each measurement unit is also treated as one or more separate words, depending on its transcription.

Fifteen different medical workers, both male and female, have evaluated the system accuracy, as well as its speed of returning recognized words (i.e., the real-time factor). All the speakers performed adaptation to their particular voices first. The speed of the system depends heavily on the used hardware, and the general consensus is that about 1 CPU core is needed per channel (simultaneous recognition) with a mid-range CPU (or better) to diminish the delay in obtaining the results to a manageable small value. On the other hand, the reported accuracy of 97% varied a bit from speaker to speaker, without major outliers. The typical recognition errors included out-of-vocabulary (OOV) words, as the set of possible words is naturally not a fixed or even a determinable set, so future supplementation of the language model for the purpose of covering an even larger number of words and contexts is planned. Other than that, the spelling of abbreviations produced most errors (modelling of abbreviations is a known weakness of the used acoustic model). Regardless of the few mentioned issues, the testers have rated this ASR system as a potentially very useful tool.

V. Conclusion

The system for automatic medical speech recognition in Serbian, implemented upon request of the Pension and Disability Insurance Fund of the Republic of Serbia, enables easier and faster textual input using dictation, which can simplify the work of medical workers, increase their productivity and at the same time prevent some of the typical spelling errors. The system enables detection of terms spoken in Serbian or Latin depending on the selected context, while achieving high recognition accuracy and reliable operation under specified conditions. Further improvements of accuracy can be achieved by processing a larger set of documents for the training of language models, which is a hard, highly expensive and time-consuming process, but on the other hand, significantly contributes to the robustness of recognition as well as end-user satisfaction.

Acknowledgment

The research described in this paper has been supported in part by the Serbian Ministry of Education, Science and Technological Development through the project no. 451-03-68/2020-14/200156: "Innovative Scientific and Artistic Research from the Faculty of Technical Sciences Activity Domain".

REFERENCES

[1] E. Pakoci, B. Popović and D. Pekar, "Using morphological data in language modeling for Serbian large vocabulary speech recognition," in Computational Intelligence and Neuroscience, Special Issue on Advanced Signal Processing and Adaptive Learning Methods, vol. 2019, 8 pages, 2019.

[2] B. Popović, E. Pakoci and D. Pekar, "A comparison of language model training techniques in a continuous speech recognition system for Serbian," in Proceedings of the 20th International Conference on Speech and Computer (SPECOM) – Lecture Notes in Artificial Intelligence, vol. 11096, pp. 522-531, Leipzig, Germany, September 2018.

[3] E. Pakoci, B. Popović and D. Pekar, "Improvements in Serbian speech recognition using sequence-trained deep neural networks," in SPIIRAS Proceedings, vol. 3, no. 58, pp. 53-76, 2018.

[4] V. Peddinti, D. Povey and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in Proceedings of the 16th Annual Conference of the International Speech Communication Association (INTERSPEECH), pp. 3214-3218, Dresden, Germany, September 2015.

[5] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, "Front-end factor analysis for speaker verification," in IEEE Transactions on Audio, Speech and Language Processing, vol. 19, no. 4, pp. 788-798, 2011.

[6] D. Rosenthal, F. Chew, D. Dupuy, S. Kattapuram, W. Palmer, R. Yap and L. Levine, "Computer-based speech recognition as a replacement for medical transcription," in American Journal of Roentgenology, vol. 170, no. 1, pp. 23-25, 1998.

[7] L. Zhou, S.V. Blackley, L. Kowalski, R. Doan, W.W. Acker, A.B. Landman, E. Kontrient, D. Mack, M. Meteer, D.W. Bates and F.R. Goss, "Analysis of errors in dictated clinical documents assisted by speech recognition software and professional transcriptionists," JAMA Network Open, vol. 1, no. 3, 13 pages, 2018.

[8] C-C. Chiu, A. Tripathi, K. Chou, C. Co, N. Jaitly, D. Jaunzeikare, A. Kannan, P. Nguyen, H. Sak, A. Sankar, J. Tansuwan, N. Wan, Y. Wu and X. Zhang, "Speech recognition for medical conversations," in Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH), pp. 2972-2976, Hyderabad, India, September 2018.

[9] I. Hammana, L. Lepanto, T.G. Poder, C. Bellemare and M-S. Ly, "Speech recognition in the radiology department: A systematic review," in Health Information Management Journal, vol. 44, no. 2, pp. 4-10, 2015.

[10] T.G. Poder, J. Fisette and V. Déry, "Speech recognition for medical dictation: Overview in Quebec and systematic review," in Journal of Medical Systems, vol. 42, no. 5, pp. 89, 2018.

[11] J.P. Lyons, S.A. Sanders, D.F. Cesene, C. Palmer, V.L. Mihalik and T. Weigel, "Speech recognition acceptance by physicians: A temporal replication of a survey of expectations and experiences," in Health Informatics Journal, vol. 22, no. 3, pp. 768-778, 2016.

[12] E. Pakoci, "Influence of morphological features on language modeling with neural networks in speech recognition systems," Ph.D. thesis, Dept. Power, Electronic and Telecommunication Engineering, University of Novi Sad, Serbia, 2019.