The Experiments in SVM-based Whispering Speaker Identification

Jovan Galić, Branko Marković and Đorđe T. Grozdić

***Abstract*—This paper presents results of automatic speaker recognition in normally phonated (neutral) and whispered speech, based on Support Vector Machines (SVM) and Whi-Spe speech database. The performance of the recognizer is examined in matched N/N (Neutral/Neutral) and W/W (Whispered/Whispered) train/test scenarios for different types of kernels (Radial basis function, Polynomial, Linear, and Sigmoid). The best accuracy is obtained with a polynomial kernel (96,12% for neutral speech and 92,16% in case of whispering). The influence of the size of training data on the performance of the recognizer is examined, as well.**

Keywords—Speaker recognition; Whispered speech; Whi-Spe database; MFCC; SVM algorithm.

I. Introduction

The technology of automatic speech and speaker recognition has made significant progress in the last two decades. Still, some disadvantages remain. The key imperfection is the considerable degradation of the performance in adverse conditions [1]. As well, speech technologies are designed for recognition of the most commonly used mode of phonation, i.e. neutral speech. Speech mode can be classified into the five main categories: whispered speech, soft speech, normally phonated speech (neutral speech), loud speech and shouted speech [2].

Nowadays, whisper is often used in a daily life, especially over the mobile phones. There are multiple reasons to use whisper: when someone doesn't like to disturb others, when the loud speech is prohibited or unpleasant, when the information to speak is secret, when someone wishes to hide identity etc. Also, whisper can be produced due to health problems: it may happen after laryngitis or rhinitis [3]. Whisper as a speech mode is characterized by a lack of glottal vibration, noisy excitation of the vocal tract and in general, the changes of the vocal tract structure.

Jovan Galić is with the Faculty of Electrical Engineering, University of Banja Luka, Department of Electronics and Telecommunications, Banja Luka, Patre 5, Bosnia and Herzegovina (e-mail: jovan.galic@etf.unibl.org)

Branko Marković is with the Faculty of Technical Science Čačak, University of Kragujevac, Department of Computer Science and Software Engineering, Čačak, Svetog Save 65, Serbia (e-mail: brankomarko@yahoo.com).

Đorđe T. Grozdić is with Grid Dynamics Holdings, Inc, Blvd. Mihajla Pupina 115, Belgrade, Serbia; and School of Electrical Engineering, University of Belgrade, Blvd. Kralja Aleksandra 73 (e-mail: djordjegrozdic@gmail.com).

There are main differences between neutral and whispered speech. It was determined that formant frequencies for whispered vowels are substantially higher than for the neutral voice [4]. Also, compared to neutral speech, whisper has less energy, longer durations of speech and silence intervals, flatter spectrum and lower sound pressure level (SPL) [2]. Despite of these "weaknesses", the intelligibility of whisper is pretty high [5]. But, non-linguistic information (like age, sex, emotions or identity), is still a big challenge for research in whispered speech.

The oscillogram and spectrogram of the short sentence "Govor šapata." ("Whispered speech" in English), uttered in neutral and whispered speech are depicted in Figures 1 and 2, respectively. Because of the lack of sonority, the difference in amplitude intensities can be observed, especially for vowels [6]. The analysis of spectrograms shows that some parts of the spectrum are well preserved in whisper, especially in the case of unvoiced consonants and plosives. Moreover, the spectrogram shows that the harmonic structure of vowels is lost in the case of a whisper [7].

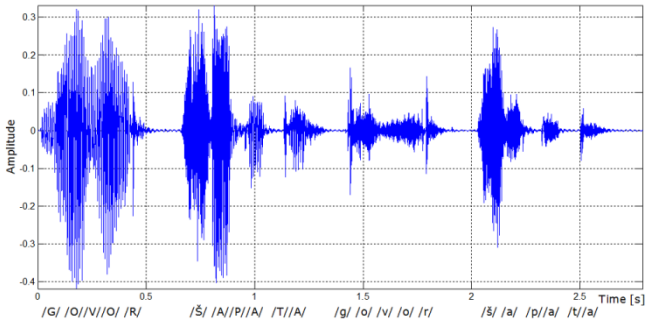

Fig. 1. The waveform of sentence "Govor šapata" in neutral phonation (capital letters) and whispered phonation (small letters).

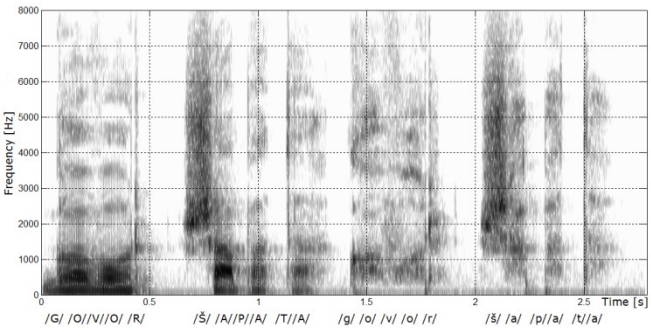

Fig. 2. The spectrogram of sentence "Govor šapata" in neutral phonation (capital letters) and whispered phonation (small letters).

Whispering speaker identification is a great challenge for state-of-the-art speaker recognition systems. In a range of speech modes from whisper to shouted, whispered speech has the most negative influence on the performance of Automatic Speech/Speaker Recognition (ASR) systems [2]. Recently, the use of Gaussian Mixture Models and K-means algorithms has been analyzed in whispering speaker ID for mel and exponential frequency scales [8]. Also, formant gap features showed higher accuracy in speaker verification compared to baseline features [9].

Automatic speaker recognition can be classified into identification and verification. Methods for speaker ID can be divided into text-independent and text-dependent [10]. For a text-independent ASR system, models for a particular speaker are irrespective of uttered speech, whereas in a text-dependent system the performance of speaker recognition depends on uttered phrases. In this research, text-dependent closed set speaker identification based on Support Vector Machines (SVM) and Mel-Frequency Cepstral Coefficients (MFCC) was analyzed.

The goal of the study presented in this paper is to analyze speaker identification accuracy for neutral and whispered speech, and classification based on SVM. This paper is organized in the following manner. In Section II the classification based on SVM is shortly discussed. The basic characteristics of the ASR system and speech database used for speaker recognition are described in Section III. The results of conducted experiments are given in Section IV whereas concluding remarks and directions for future research are stated in Section V.

II. Support Vector Machines

The SVM classifier is a relatively simple machine-learning algorithm that minimizes the structural risk [11]. Initially, the SVM classifier was introduced for linearly separable classes of objects. The separation of classes is obtained with an n-dimensional hyperplane that maximizes the margin between classes (circles and squares), as depicted in Figure 3. The margin is labeled as *M* and support vectors are hatched.

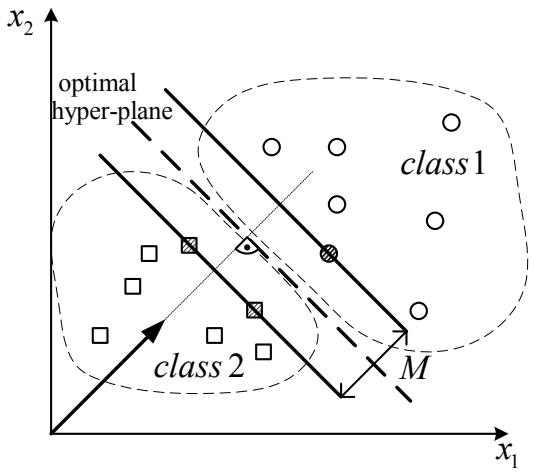

Fig. 3. Determination of hyperplane in SVM for linearly separable classes.

However, the classes are not linearly separable in most practical applications. To overcome that limitation, a non-linear transformation is performed on a feature vector. The mapping into high-dimensional feature space (in which linear separation is expected) is performed by using kernel function. Each function that satisfies necessary properties (Mercer's theorem) can be used as a kernel [12]. The most used types of kernels are:

- Radial basis function kernel (adjustable parameter γ)

$$K(x_1, x_2)=\exp\left(-\frac{\|x_1-x_2\|^2}{2\sigma^2}\right);\ \gamma=\frac{1}{2\sigma^2}. \quad (1)$$

- Polynomial kernel (adjustable parameters are the slope α, the constant term c and the polynomial degree d)

$$K(x_1, x_2)=(\alpha x_1^T x_2+c)^d. \quad (2)$$

- Linear kernel (adjustable parameter c)

$$K(x_1, x_2)=x_1^T x_2+c. \quad (3)$$

- Hyperbolic Tangent (Sigmoid) kernel (adjustable parameter are α and the constant term c)

$$K(x_1, x_2)=\tanh(\alpha x_1^T x_2+c). \quad (4)$$

Because SVM is a static classifier, meaning that it works with fixed-size input data, the application in speech/speaker recognition has some restrictions. Some hybrid solutions were developed to overcome that limitation [13]. Another issue is multiclass classification, which is commonly solved by using one of the two following techniques. The first technique includes comparison of each class against all the others (one-vs-all) and the second technique confronts each class against all the others separately (one-vs-one). In this study, a one-vs-all comparison strategy is used.

III. System for Recognition

A. Speech database

One of the real problems related to the whispered speech research is a shortage of an extensive speech databases. There are some of them developed so far [14-17] for Japanese, Mandarin and English. In order to do this research, the Serbian speech database called Whi-Spe (abbreviation of *Whispered Speech*) is used [18]. The database contains two parts: the first one has recordings of whispered words, and the second one has recordings of the same words uttered with neutral speech. The vocabulary of 50 different words is divided in three groups: basic colors (6 words), numbers (14 words) and phonetically balanced (30 words). For recordings of the Whi-Spe ten volunteers (5 female and 5 male) uttered the vocabulary ten times in both speech modes, neutral and whisper. Hence, the speech database contains 10.000 represents of words in form of wave files and the total duration is 2 hours.

More details about the Whi-Spe database regarding segmentation procedure and quality control can be found in [18].

B. *ASR system*

Because SVM-based classifiers need feature vectors of fixed dimension, the variation in the duration of input speech utterances must be uniform. The two most common approaches for making a fixed number of frame windows for SVM classifier are using variable window size (with constant overlapping factor) and fixed window size (with variable overlapping factor). This causes some loss of information, especially in long speech utterances. In this paper, segmentation based on variable window size is chosen, using 13 overlapping windows, same as in the SVM-based speech recognition [19].

C. *Feature vector extraction*

The most common features used in ASR systems are Mel Frequency Cepstral Coefficients (MFCC). The diagram for obtaining the MFCC feature vector is depicted in Fig. 4.

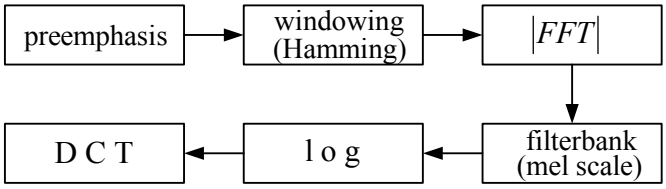

Fig. 4. The diagram for MFCC feature vector generation.

The generation of MFCC feature vectors includes the following steps: preemphasis, framing with overlap and Hamming windowing, application of the Fast Fourier Transformation (FFT), using the mel scale, calculating log energy and finally obtaining the cepstral, delta cepstral and delta-delta cepstral coefficients (based on Discrete Cosine Transformation - DCT).

The MFCC feature vector is obtained by using static features (13) along with time derivatives (delta and delta-delta) and cepstral mean normalization (39 in total). Lastly, each speech utterance from the database is represented with a vector of 507 coefficients (13 x 39) and later used as an input to the SVM classifier. The speech recognizer is developed with Python (version 3.6) using the Scikit learn package.

IV. RESULTS

In the initial experiment baseline speaker ID performance are evaluated in 2 matched train/test scenarios: N/N and W/W. In order to have a more reliable evaluation of the performance, 10-fold cross-validation was done. The average accuracy is used as a metric for performance evaluation.

Firstly, the influence of kernel selection on recognizer performance was analyzed. The experiments were done for 4 kernels: radial basis function (RBF), polynomial (with degree d=3), linear and sigmoid.

The results are graphically presented in Figure 5.

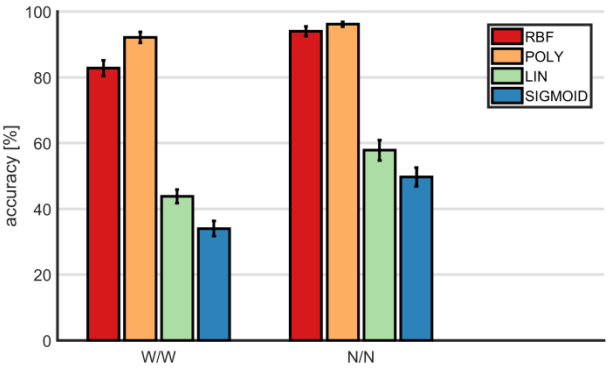

Fig. 5. Results of speaker identification (average accuracy with standard deviation) in neutral (N/N) and whispered speech (W/W) depending on type of kernel.

As it can be seen from Figure 5, the best results in whispering speaker ID are obtained for the polynomial kernel (92,16%). As well, the best result in recognition of speakers in neutral phonation is obtained using the polynomial kernel (96,12%). The experiments showed that speaker identification performance for RBF kernel is poor compared to poly kernel. Linear and sigmoid kernels are practically useless and show poor performance, especially for speaker identification in whisper mode.

As well, the influence of the percentage of the database used for training (i.e. number of training instances) on classification performance was examined. In this experiment, the polynomial kernel was used because it showed the best recognition in the previous experiment. The results are presented in Figure 6. An important part of bar graph is emphasized (higher than 75%).

For the speaker recognition in neutral mode the recognition performance starts from 94,71% (for training 75% of full capacity database, i.e. 3750 utterances) and reaches final 96,12% (for 95%, i.e. 4750 utterances). On the other hand, whispering speaker recognition performance is in the range of 89,49% (for 75%) up to 92,16% (for 95%). As observed, the saturation effect can be seen for both neutral and whisper speech modes (saturation effect is less for whispered speech).

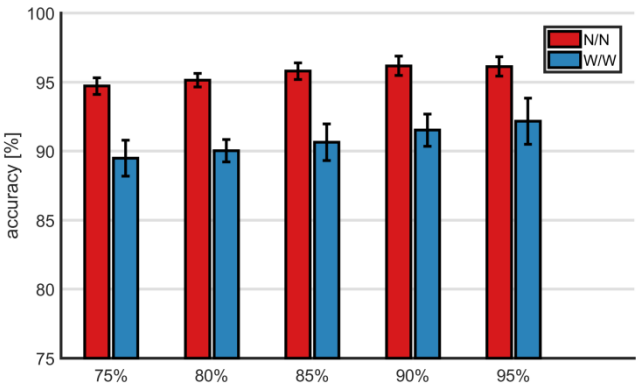

Fig. 6. Results of speaker identification (average accuracy with standard deviation) in dependence of percentage of database used for training for neutral (N/N) and whispered speech mode (W/W).

The results in N/W (neutral/whisper) train/test scenario for polynomial kernel and each speaker are presented in Table I. As can be seen, compared to N/N scenario the performance are degraded and very high difference between different speakers can be observed. Similar observation is found in whispering speaker identification with neutral trained HMM models [20].

TABLE I
ACCURACY FOR N/W SCENARIO (%)

Speaker	Accuracy
Speaker 1	65,2
Speaker 2	9,2
Speaker 3	21,4
Speaker 4	16,4
Speaker 5	62,0
Speaker 6	27,2
Speaker 7	25,4
Speaker 8	29,4
Speaker 9	33,4
Speaker 10	4,0
Average	**29,36**

V. CONCLUSION

Whispered speaker recognition is, by all means, a serious challenge for modern ASR systems.

As expected, whispering speaker ID was shown to be more difficult than recognition of a speaker that utters neutral speech. In this paper experiments on speaker identification in neutral and whisper mode for Whi-Spe speech database and SVM algorithm were conducted.

Future studies will be focused on examining more robust feature vectors in whispering speaker identification. Also, the application of the Teager energy operator [21-23] has shown improvements in robustness for whispered speech recognition, so there are reasonable expectations that it could help in speaker recognition as well.

As well, different machine learning algorithms are going to be analyzed (Gaussian mixture models and Neural networks).

ACKNOWLEDGEMENT

This research is supported by the project "Razvoj Internet of Things (IoT) aplikacija primjenom optičko-bežičnih tehnologija" of Ministry for Scientific and Technological Development, Higher Education and Information Society of Republic of Srpska.

REFERENCES

[1] J. Holms, W. Holms, *Speech Synthesis and Recognition*, Taylor & Francis, London, United Kingdom, 2001.

[2] C. Zhang, J. H. L. Hansen, "Analysis and classification of Speech Mode: Whisper through Shouted," Proceedings of Interspeech 2007, pp. 2289-2292, 2007.

[3] T. Ito, K. Takeda, F. Itakura, "Analysis and recognition of whispered speech," Speech Communication, vol. 45, no. 2, pp. 139–152, 2005.

[4] Y. Swerdlin, J. Smith, J. Wolfe, "The effect of whisper and creak vocal mechanisms on vocal tract resonances," Journal of Acoustical Society of America 127, pp. 2590–2598, 2010.

[5] V. C. Tartter, "Identifiability of vowels and speakers from whispered syllables", Perception & Psychophysics, vol. 49, no. 4, pp. 365–372, 1991.

[6] J. Galić, B. Popović, D. Šumarac Pavlović, "Whispered Speech Recognition using Hidden Markov Models and Support Vector Machines," Acta Politechnica Hungarica, vol. 15, no. 5, pp. 11-29, 2018.

[7] J. Galić, S.T. Jovičić, Đ. Grozdić, B. Marković, "Constrained Lexicon Speaker Dependent Recognition of Whispered Speech," Proceedings of International Symposium on Industrial Electronics and Applications (INDEL), pp. 180-184, 2014.

[8] A. Singh, A. M. Joshi, "Speaker Identification Through Natural and Whisper Speech Signal," Optical and Wireless Technologies, Lecture Notes in Electrical Engineering, vol. 546, pp. 223-231, 2020.

[9] A. R. Naini, A. Rao, P. K. Ghosh, "Formant-gaps features for speaker verification using whispered speech," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 6231–6235, 2019.

[10] S. Furui, "50 Years of progress in speech and speaker recognition," Proceeding of the International Conference Speech and Computer SPECOM, Patras, Greece, pp. 1-9, 2005.

[11] P. Clarkson, P. J. Moreno, "On the use of support vector machines for phonetic classification," IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP 99, vol. 2, pp. 585-588, 1999.

[12] Z. Ćirković, Z. Banjac, "Jedna primena SVM klasifikatora u verifikaciji govornika nezavisno od teksta," Proceedings of the International Symposium Infoteh Jahorina, vol. 11, pp. 833-837, 2013.

[13] Z. Qu, Y. Liu, L. Zhang, M. Shao, "A Speech Recognition System Based on a Hybrid HMM/SVM Architecture," First International Conference on Innovative Computing, Information and Control (ICICIC) Beijing, China, pp. 100-104, 2006.

[14] N. Kawaguchi, K. Takeda, S. Matsubara, I. Yokoo, T. Ito, K. Tatara, T. Shinde, F. Itakura, "Ciair speech corpus for real world applications," International Joint Conference of the 5th Symposium on Natural Language Processing, pp. 288–295, 2002.

[15] P. X. Lee, D. Wee, H. S. Yin Toh, B. P. Lim, N. Chen, B. Ma, "Whispered Mandarin Corpus for Speech Technology Applications," Proceedings of Interspeech 2014, pp. 1598–1602, 2014.

[16] T. Tran, S. Mariooryad, C. Busso, "Audiovisual corpus to analyze whisper speech," presented at the International Conference on Acoustics, Speech and Signal Processing, pp. 8101–8105, 2013.

[17] F. Cummins, M. Grimaldi, T. Leonard, J. Simko, "The chains corpus: Characterizing individual speakers," Proceedings of International Conference on Speech and Computer SPECOM, St. Petersburg, Russia, pp. 421–435, 2006.

[18] B. Marković, S. T. Jovičić, J. Galić, Đ. Grozdić, "Whispered speech database: design, processing and application," In: Habernal, I., Matousek, V. (eds.), TSD 2013, LNAI 8082, Springer-Verlag Berlin Heidelberg, pp. 591–598, 2013.

[19] J. M. Garcia-Cabellos, C. Peleaz-Moreno, A. Gallardo-Antolin, F. Perez-Cruz, F. Diaz-de-Maria, "SVM classifiers for ASR: A discussion about paramterization," Proceedings of 12th European Signal Processing Conference, pp. 2067-2070, 2004.

[20] X. Fan, J. H. L. Hansen, "Speaker identification within whispered speech audio stream," IEEE Transactions on Audio, Speech and Language Processing, vol. 19, no. 5, 1408-1421, 2011.

[21] Đ. Grozdić,, "Application of Neural Networks in Whispered Speech Recognition", PhD dissertation, University of Belgrade, Belgrade, Serbia, 2017.

[22] Đ. Grozdić, S. Jovičić, M. Subotić, "Whispered speech recognition using deep denoising autoencoder", Engineering Applications of Artificial Intelligence, vol. 59, pp 15-22, 2017.

[23] Đ. Grozdić, S. Jovičić, "Whispered speech recognition using deep denoising autoencoder and inverse filtering", IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 12, pp. 2313-2322, 2017.