Constrained Lexicon Speaker Dependent Recognition of Whispered Speech

Jovan Galić[1,2]
[1]Telecommunications Department
School of Electrical Engineering
Belgrade, Serbia
jgalic@etfbl.net

[2]Department of Electronics and Telecommunications
Faculty of Electrical Engineering, University of Banja Luka
Bosnia and Herzegovina

Slobodan T. Jovičić[1,3], Đorđe Grozdić[1,3]
[3]Life Activities Advancement Center, Laboratory for
Psychoacoustics and Speech Perception
Belgrade, Serbia
jovicic@etf.rs, djordjegrozdic@gmail.com

Branko Marković
Čačak Technical Colledge
Computing and Information Technology Department
Čačak, Serbia
brankomarko@yahoo.com

***Abstract*— In this paper we present results on automatic speech recognition of isolated words with part of Whi-Spe database with female speakers, in speaker dependent fashion and constrained lexicon (50 words). Word recognition rate is calculated for four train/test scenarios, with modeling of context independent monophones, context dependent triphones and whole words. As feature vectors, we used Perceptual Linear Prediction Coefficients and Mel Frequency Cepstral Coefficients. The Hidden Markov Model Toolkit was used to implement isolated word recognizer. Further improvement is achieved with reduction in number of monophone units used for modeling. Due to very high deviation in performance among different speakers, influence of Signal to Noise Ratio of tested recordings on performance of recognizer is examined in particular.**

Keywords- HTK, speech recognition, Whi-Spe database

I. INTRODUCTION

Whispering is a specific mode of speech often used in everyday life, esspecially by cellular phones. People whisper for a number of reasons, for example, in environments where normally phonated speech is not appropiate or concealment of some confidental information from the others ears. Beside conscious production of whisper, whispering may happen due to health problems which appear after rhinitis and laryngitis [1], [2]. The whisper has different characteristics compared to normally phonated speech. Due to the absence of the glottal vibrations, whispering lacks the fundamental frequency of the voice and much prosodic information. In addition, whispered speech has a significantly lower energy as compared to the normal speech [2], and the slope of the spectrum being much flatter than in the normal speech [3]. The duration of whispered speech is slightly longer [4], and the formant frequencies for whispered vowels is substantially higher than for the normal voice [4]. The amount of shift is higher for vowels with low formant frequencies [5]. Figure 1 shows waveform and Fig. 2 shows spectrogram of sentence "govor šapata" ("whispered speech" in English), uttered in normally phonated speech followed by whispered speech. Pictures are supported with phonetic transcription for both the normal (capital letters) and whispered speech (small letters). Because of the lack of sonority, difference in amplitude intensities could be observed, especially for vowels. Also, spectrogram shows that the harmonic structure of vowels in whispered speech is completely lost, which is separately presented for vowel /o/ in Fig. 3. However, spectral characteristics of unvoiced consonants (for example fricative /š/) are not significantly changed. Similar shape of spectrum of phoneme /r/ in Serbian is observed.

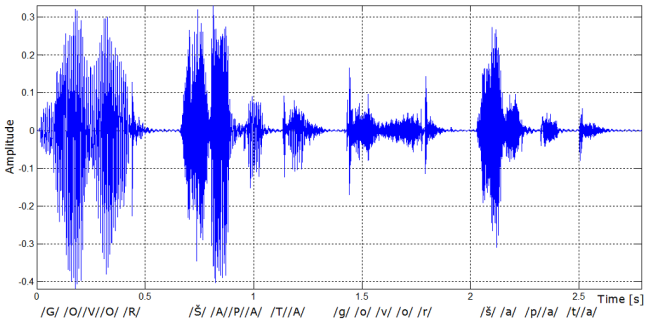

Figure 1. Waveform of sentence "Govor šapata" in normal phonation (capital letters) and whispered phonation (small letters)

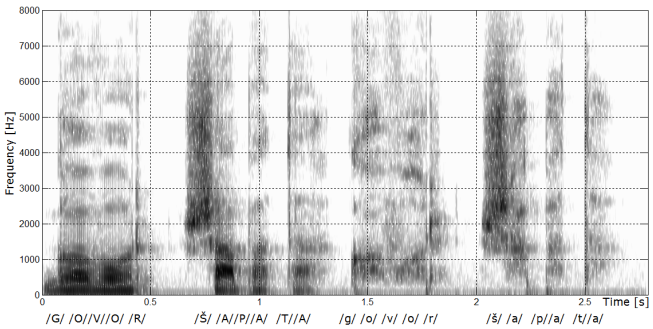

Figure 2. Spectrogram of sentence "Govor šapata" in normal phonation (capital letters) and whispered phonation (small letters)

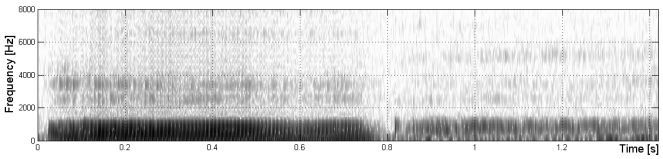

Figure 3. Spectrogram of vowel /o/ in sustained phonation (normal and whispered).

For all mentioned differences, automatic recognition of whispered speech is much more difficult than of normal speech. From [6], we know that speech can be generally classified into five categories based on vocal effort differences: whispered, soft, neutral, loud and shouted speech. The speaker identification performance significantly degrade with a change in vocal effort ranging from whisper through shouted, where whispered speech have the most serious loss in performance [7].

Nevertheless, despite of increased efforts in perception, this type of speech is perfectly understandable [8]. There are different approaches, techniques and methods of speech recognition. These techniques are usually based on algorithms of the HMM (Hidden Markov Model), the DTW (Dynamic Time Warping), the ANN (Artificial Neural Network) and their hybrid solutions [9]. This paper presents results on investigation of recognition of isolated words from part of Whi-Spe database [10] with female speakers, using a software toolkit HTK (Hidden Markov Model Toolkit). The HTK is a widely used software for ASR (Automatic Speech Recognition), that was originally developed at the Machine Intelligence Laboratory of the Cambridge University Engineering Department [11].

The reminder of this paper is organized as follows. In Section 2 the brief description of database Whi-Spe is given. In Section 3 we give the system overview used in experiments. The experimental results, as well as its discussion, are given in Section 4, while concluding remarks and further directions are stated in Section 5.

II. WHI-SPE DATABASE

The Whi-Spe database contains two parts: the first one contains speech patterns of a whispered speech, while the second one contains speech patterns of the normal speech. All patterns were collected from the five female and five male speakers. During the session of recording, each speaker read 50 isolated words of Serbian. The words were divided in three sub-corpora: basic colors (6 words), numbers (14 words) and phonetically balanced words (30 words). Balanced words were taken from the Serbian emotional speech database GEES [12], which satisfies the basic linguistic criteria of Serbian language. Sessions were repeated ten times, with a pause of a few days between recordings. Finally, the database collection grew to 10.000 utterances, half in the whispered speech and half in the normal speech. The speakers of ages between twenty and thirty were Serbian native volunteers from Čačak Technical College.

The speech was digitized by using the sampling frequency of 22.050 Hz, with 16 bits per sample, and stored in the form of Windows PCM (Pulse-Code Modulation) wave files. In this experiment, all samples from female speakers were used. Specific details about database concerning content, recording process and quality control could be find in [10].

III. SYSTEM OVERVIEW

In this work, all experiments were conducted on the latest version of HTK, 3.4.1 [13]. The toolkit was ported to Windows 7, and all experiments were done under this operating system. As a feature vectors, we used Mel Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP) coefficients, as widely used features in speech recognition [14], [15]. For obtaining a feature vector, Hamming window with preemphasis coefficient of 0.97 was used. The window size was set to 24 ms, and frame shift to 8 ms. Also, cepstral coefficient C_0, delta and acceleration coefficients were appended and cepstral mean subtraction was performed. These auxiliary features and modification techniques significantly improves recognition rate [16]. Number of filterbank channels was set to 26, and number of output cepstral coefficients per frame was set to 12. For better performance in these experiments, in filterbank analysis power was used instead of magnitude and normalization of energy was not included. The other parameters were set to default values.

The model topology is a continuous density HMM with one Gaussian mixture component and diagonal covariance matrix. There were 5 states in total, 3 of which are emitting. As an acoustic modeling units, we used context independent (CI) monophones, context dependent (CD) triphones and whole word (WW) models. Despite more frequent use of sub-word modeling (CI and CD) in ASR systems, there are still some applications where whole word modeling presents optimal solution, especially in recognition of isolated and connected words from constrained lexicon. The WW models consisted of the same number of states as their CI and CD counterparts and followed the same transition matrix structure, that is strictly left-right, with no skips. For CI models, phonetic transcription was done manually. Stops and affricates are labeled as pairs of semi-phones that consist of occlusion and explosion parts. Phoneme /ə/ (schwa) is marked separately when phoneme /r/ is found in the consonant environment [17]. The model of silence is added at the start and the end of every utterance. Initial model parameters were estimated using the flat-start method , since training data is not time labeled. In the training phase, location of word boundaries were estimated using forced alignment. At last, in the testing phase the Viterbi algorithm was applied to determine the most likely model that best matched each test utterance.

Our goal was to compare the performance of different acoustic models in four train/test scenarios:

1. Normal/Normal (Nor/Nor) - the system is trained on normally phonated speech and tested on the speech of same mode;
2. Normal/Whisper (Nor/Whi) - the system is trained on normally phonated speech and tested against whispered speech;
3. Whisper/Normal (Whi/Nor) - the system is trained on whispered speech and tested against normally phonated speech;
4. Whisper/Whisper (Whi/Whi) - the system is trained on whispered speech and tested on the speech of same mode.

The scenarios where training and testing is in the same mode of speech are denoted as match, and in the opposite case scenarios are denoted as mismatch.

IV. RESULTS AND DISCUSSION

In match scenarios, 80% utterances were in the part for training, while the other 20% were in the part for testing. The training and test set were rotated, which gave 4 additional tests. Recognition rate is calculated as mean value of 5 tests. In mismatch scenarios, all utterances of one mode were in the part for training, while in testing part were all utterances of the other mode of speech. The results (word recognition rate - WRR) are shown in Tables I-III for modeling monophones, triphones and whole words, respectively. The speakers are labeled from G1 up to G5. The last column presents average WRR for respected scenario and feature vector. For better clarity, results are integrated, and depicted in Fig. 4 (PLP feature vector) and Fig. 5 (MFCC feature vector), where average recognition rates of all five speakers are graphically presented in dependence of scenario and modeling units. From results presented in Tables I-III, and Figures 4-5, we can observe that CD models contribute to higher scores in match scenarios, compared to CI and WW models. These results are expected, since CD models are more specialized and superior in highly matched conditions, which is the case with Whi-Spe corpus. Performance of CI and WW models could be improved (WRR above 99.5%) by increasing number of mixture components [18].

TABLE I. WORD RECOGNITION RATE FOR MONOPHONE MODELS

Speaker/ Scenario-Feature		G1	G2	G3	G4	G5	Avg
Nor/Nor	PLP	98.6	98.0	98.8	98.0	98.0	98.28
	MFCC	98.2	99.2	98.0	98.0	98.2	98.32
Nor/Whi	PLP	89.2	52.6	76.0	55.4	61.0	66.84
	MFCC	85.4	47.8	70.4	47.2	58.6	61.88
Whi/Nor	PLP	87.6	57.6	75.6	80.4	72.0	74.64
	MFCC	87.2	56.8	70.8	81.6	72.0	73.68
Whi/Whi	PLP	96.6	94.2	97.8	97.2	90.6	95.28
	MFCC	96.4	94.2	98.0	97.0	90.4	95.20

TABLE II. WORD RECOGNITION RATE FOR TRIPHONE MODELS

Speaker/ Scenario-Feature		G1	G2	G3	G4	G5	Avg
Nor/Nor	PLP	99.8	100.0	100.0	99.8	99.8	99.88
	MFCC	100.0	100.0	100.0	99.8	99.8	99.92
Nor/Whi	PLP	50.0	12.6	35.6	20.0	21.4	27.92
	MFCC	41.6	12.0	31.8	14.8	12.4	22.52
Whi/Nor	PLP	73.8	34.0	53.2	56.8	60.8	55.72
	MFCC	73.4	33.8	51.0	61.4	60.4	56.00
Whi/Whi	PLP	100.0	100.0	100.0	100.0	99.8	99.96
	MFCC	100.0	100.0	100.0	100.0	99.8	99.96

TABLE III. WORD RECOGNITION RATE FOR WHOLE WORD MODELS

Speaker/ Scenario-Feature		G1	G2	G3	G4	G5	Avg
Nor/Nor	PLP	99.2	99.2	99.8	99.8	99.2	99.44
	MFCC	98.0	99.4	99.2	100.0	99.6	99.24
Nor/Whi	PLP	57.2	22.6	41.0	35.0	37.4	38.64
	MFCC	49.2	22.8	44.6	34.2	33.2	36.80
Whi/Nor	PLP	59.8	29.8	38.0	42.0	43.4	42.60
	MFCC	57.2	25.8	40.8	49.8	41.8	43.08
Whi/Whi	PLP	98.2	98.8	98.8	97.2	93.8	97.36
	MFCC	98.2	98.6	99.0	98.0	93.2	97.40

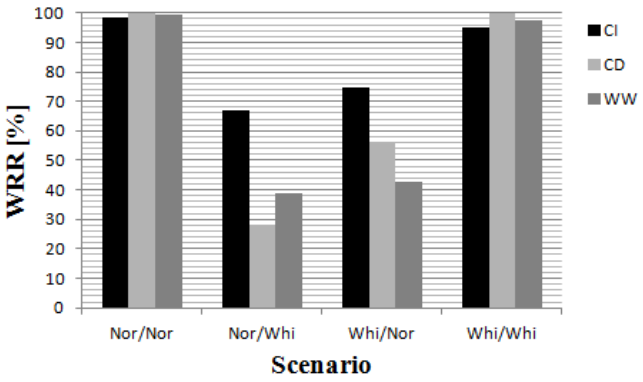

Figure 4. Average word recognition rates with PLP feature vector for context independent (CI), context dependent (CD) and whole word (WW) models and different scenario

Also, except CD modeling, in match conditions recognition of whispered speech is a few percents poorer. Both feature vectors give very similar results, and because of the "ceiling effect" it is hard to determine which is better. In mismatch scenarios, the most robust are CI models, with the average WRR of 66.84% in Nor/Whi scenario, and 74.64% in Whi/Nor scenario.

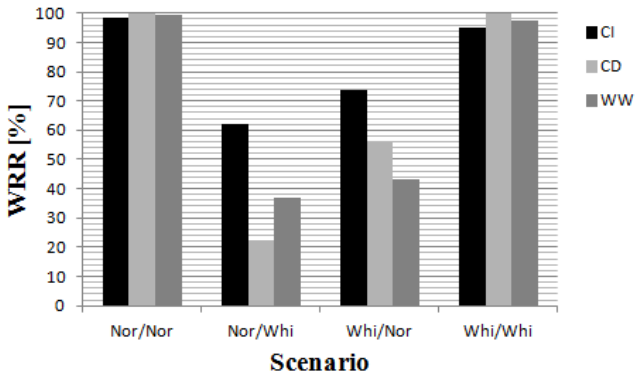

Figure 5. Average word recognition rates with MFCC feature vector for context independent (CI), context dependent (CD) and whole word (WW) models and different scenario

Utilization of PLP feature vector give an absolute improvement in average WRR of 5%, compared to recognition with MFCC feature vector. In experiments with same database and speaker independent fashion, PLP had also shown better performance, compared to MFCC feature vector [19]. The phenomenon of asymmetric performance in Nor/Whi and Whi/Nor scenario, in favor of Whi/Nor scenario, was examined in experiments with neural networks and same database [20]. Same study had shown that the sonority in speech stimuli is the main cause of difference in word recognition scores in mismatch scenarios, and that the most of whisper features are contained in normal speech, which is not the opposite case. The CD and WW give much lower recognition scores in mismatch scenarios, for both feature vectors.

Since the most of ASR systems are primarily trained on normally phonated speech, from the point of view of ASR recognition of whispered speech the greatest importance have the Nor/Whi scenario. The greatest effort in research area is devoted to maximizing performance in that scenario, because that concept does not need adaptation to whisper, of any kind. The best recognition scores for whispered speech with constrained lexicon (160 words) for recognition of English [21] are over 80% (in speaker independent fashion), so there is optimistic expectation that those scores are not far away for Whi-Spe database, for recognition with statistical ASR framework. First step is made in reduction in number of HMM models. The general approach in transcription of monophones, which includes separate modeling of inclusion/explosion parts in stops and fricative, and stressed/unstressed vowels, requires large databases for training, of several hours. That is not the case with Whi-Spe database. The unification of occlusion and explosion parts in stops and affricates, as well as stressed and unstressed vowels, has lead to reduction of HMM models to 32 monophones (30 phonemes in Serbian, schwa and silence). The results and absolute improvement with reduced number of HMM models are presented in Table IV.

TABLE IV. WORD RECOGNITION RATE IN NOR/WHI SCENARIO WITH REDUCED NUMBER OF HMM MODELS

Speaker/ Feature	G1	G2	G3	G4	G5	Avg	Improve-ment
PLP	93.6	62.8	81.4	64.0	71.0	74.56	7.72
MFCC	92.8	61.8	78.8	58.6	70.8	72.56	10.68

Results presented in Table IV clearly show greater robustness of models with reduced number of HMM models, compared to generalized case. Absolute improvement of average WRR is 7.72% for PLP feature vector, and 10.68% for MFCC feature vector. Beside better improvement of average WRR with MFCC feature vector, average WRR with PLP feature vector is again higher. It is important to note that performance of recognizer are not much degraded in match scenarios with reduced number of monophone units. The average WRR with recognition of normally phonated speech is 98.08% with PLP feature vector, and 98.04% with MFCC feature vector. Also, in recognition of whispered speech, recognizer gives the WRR score of 96.16% with PLP feature vector, and 96.24% with MFCC feature vector. The recognition of whispered speech is even higher, by comparing results in Table I.

From results in Table IV, very high difference in performance among different speakers could be observed. The recognition rate varies from poor (speakers G2 and G4) to excellent (speaker G1). Similar observation is found in whispered speaker identification with neutral trained HMM models [22], where it was stated that the degradation is concentrated for a certain number of speakers, while other speakers displayed consistent performance to that seen in neutral speech. One of the reasons for that deviation is signal to noise ratio (SNR) of tested utterances. Also, driven by that observation we examined correlation between SNR and WRR of all five speakers. In Table V are shown average SNR of all recordings in whispered speech, for all female speakers. Because of the way the manual segmentation of recordings is done, last 100 samples are taken into account for calculating power of noise.

TABLE V. AVERAGE SNR FOR RECORDINGS OF FEMALE SPEAKERS

Speaker	G1	G2	G3	G4	G5
SNR [dB]	15.3	8.0	13.0	11.4	9.7

From results in Tables IV-V, in this experiment the highest SNR leads to highest performance, and vice versa, the lowest SNR leads to lowest WRR (for PLP feature vector). In order to quantify the degree of correlation between average SNR of tested utterances and corresponding WRR, we determined the coefficient of correlation. For two random variables X (with mean μ_X and standard deviation σ_X) and Y (with mean μ_Y and standard deviation σ_Y), the correlation coefficient and covariance are defined according equations 1 and 2, respectively. In equation 2, E denotes expectation of random variable.

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \quad (1)$$

$$Cov(X,Y) = E\left[(X-\mu_X)(Y-\mu_Y)\right] \quad (2)$$

Also, corresponding p-value is calculated for testing the hypothesis of no correlation. For SNR values given in Table V and WRR in Table IV (for PLP feature vector), correlation coefficient of 0.89 and corresponding p-value of 0.04 are obtained. Besides relatively small number of tested speakers, obtained values unambiguously show high and significant correlation.

V. CONCLUSSION AND FUTURE WORK

Whispered speech, as an alternative mode of speech production is not seldom used in everyday life. Inspired by the fact that whisper is well understandable in human to human communication, performance of statistical ASR approach based on HMM is examined in this paper, for whispered speech recognition and speaker dependent fashion. A part of Whi-Spe database with female speakers is used in this study. In match train/test conditions, recognizer had the best performance with context dependent models, where recognition of normally phonated speech, as well as whispered speech was nearly 100%. In mismatch scenarios, the monophone models had shown best results with average word recognition rate of 66.84% in Nor/Whi scenario, and 74.64% in Whi/Nor scenario (both with PLP feature vector). The greatest attention was paid to performance improvement of whispered speech recognition with models trained on normal speech. Reducing numbers of HMM models had lead to significant absolute improvement of word recognition rate of 7.72% for PLP, and 10.68% for MFCC feature vector. Due to high difference in performance among speakers, the hypothesis of correlation between tested average SNR and obtained WRR is tested. The results of hypothesis induce significant correlation.

Future work will examine performance of recognizer in multi-condition training, where training corpus is composed of normal and whispered utterances. It would be interesting to examine if the performance could reach separately trained conditions and the amount of whisper data added to training corpus for a satisfactory recognition. Since any part of data in training process had not been labeled, flat-start method in initialization of HMM models was used. Thus, to bootstrap a set of HMM models, part of utterances in training process is to be manually labeled. That activity is in a progress and preliminary results for completed speakers show noticeable improvement in recognition rate. Using alternative feature vectors will be examined, especially those which are reported to be very robust in highly mismatch condition. From the perspective of application of whispered speech in ASR, the greatest challenge will be optimizing performance without adaptation to whisper, in speaker independent fashion. Due to very high influence of amount of data in training to performance, speaker independent ASR recognizer is expected to have better performance, compared to speaker dependent recognizer. Also, the future research should include comparative analysis of word recognition efficiency using different algorithms such as DTW, HMM and ANN, using the same feature set and each speaker from database.

REFERENCES

[1] T. Ito, K. Takeda, F. Itakura, "Analysis and recognition of whispered speech," Speech Communication 45, 2005, pp. 139–152.

[2] S.T. Jovičić, Z. M. Šarić, "Acoustic analysis of consonants in whispered speech," Journal of Voice, Vol. 22, No. 3, 2008, pp. 263-274.

[3] S.T. Jovičić, "Formant feature differences between whispered and voiced sustained vowels," Acta Acustica, 84(4), 1998, pp.739-743.

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

[5] X. Fan, J.H.L. Hansen, "Speaker identification with whispered speech based on modified LFCC parameters and feature mapping," ICASSP 2009, pp. 4553-4556, 2009.

[6] C. Zhang, J.H.L. Hansen, "Advancements in whisper-island detection using the linear predictive residual," ICASSP 2010, pp. 5170-5173, 2010.

[7] C. Zhang, J.H.L. Hansen, "Analysis and classification of speech mode: whisper through shouted," Interspeech 2007, pp. 2289–2292, 2007.

[8] Đ.T. Grozdić, B. Marković, J. Galić, S. T. Jovičić, "Application of neural networks in whispered speech recognition," Telfor Journal, Vol. 5, No. 2, pp. 103-106 , 2013.

[9] J. Holms, W. Holms, Speech synthesis and recognition. Taylor & Francis, London, 2001.

[10] B. Marković, S.T. Jovičić, J. Galić, Đ.T. Grozdić, "Whispered speech database design, processing and application," 16th International Conference TSD 2013, pp. 591-598, Pilsen, Czech Republic , 2013.

[11] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason,D. Povey,V. Valtchev, P. Woodland, The HTK Book v.3.2.1, Cambridge University Engineering Department, 2002

[12] S. T. Jovičić, Z. Kašić, M. Đorđević, M. Rajković, "Serbian emotional speech database: design, processing and evaluation," Proceedings of SPECOM-2004, St. Petersburg, Russia, pp. 77-81 , 2004.

[13] The Hidden Markov Model Toolkit, http://htk.eng.cam.ac.uk/

[14] S.B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, no. 4, pp. 357–366, 1980.

[15] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," The Journal of the Acoustical Society of America, vol. 87, no. 4, pp. 1738–1752, 1990.

[16] J. Kacur, G. Rozinaj, "Practical issues of building robust HMM models using HTK and SPHINX systems," Speech Recognition, France Mihelic and Janez Zibert (Ed.), ISBN: 978-953-7619-29-9

[17] S. Sovilj-Nikić, V. Delić, I. Sovilj-Nikić, M. Marković, "Tree-based phone duration modeling of the serbian language," Electronics and Electrical Engineering (Elektronika ir Elektrotechnika), Kaunas University of Technology, Vol. 20, No. 3, pp. 1392-1215, 2014.

[18] J. Galić, S.T. Jovičić, Đ. Grozdić, B. Marković,"HTK-based recognition of whispered speech," SPECOM 2014, pp. 251-258, in press.

[19] J. Galić, S.T. Jovičić, Đ. Grozdić, B. Marković, "The influence of feature vector selection on performance on performance of automatic recognition of whispered speech," Speech and Language 2013, pp. 258-264 , 2013.

[20] Đ.T. Grozdić, S.T. Jovičić, D. Šumarac-Pavlović, B. Marković, J. Galić, "Whispered speech recognition with neural networks," unpublished.

[21] S. Ghaffarzadegan, H. Boril, J. H. L. Hansen, "UT-VOCAL EFFORT II: Analysis and constrained-lexicon recognition of whispered speech," ICASSP 2014, pp. 2544-2548, 2014.

[22] Fan, X., Hansen, J. H. L, "Speaker identification within whispered speech audio stream," IEEE Transactions on Audio, Speech and Language Processing, 19(5), 1408-1421, 2011.