PAN ARCHIVES OF ACOUSTICS
Vol. **43**, No. 1, pp. 3–9 (2018)

DOI: 10.24425/118075

Review Paper

Application of Teager Energy Operator on Linear and Mel Scales for Whispered Speech Recognition

Branko R. MARKOVIĆ, Jovan GALIĆ, Miomir MIJIĆ

Department of Acoustics
School of Electrical Engineering
Blvd. Kralja Aleksandra 73, 11000 Belgrade, Serbia
e-mail: brankomarko@yahoo.com, jgalic@etfbl.net, emijic@etf.bg.ac.rs

(*received January 28, 2017; accepted October 23, 2017*)

This paper presents experimental results on whispered speech recognition based on Teager Energy Operator for linear and mel cepstral coefficients including the Cepstral Mean Subtraction normalization technique. The feature vectors taken into consideration are Linear Frequency Cepstral Coefficients, Teager Energy based Linear Frequency Cepstral Coefficients, Mel Frequency Cepstral Coefficients and Teager Energy based Mel Frequency Cepstral Coefficients. A speaker dependent scenario is used. For the recognition process, Dynamic Time Warping and Hidden Markov Models methods are applied. Results show a respectable improvement in whispered speech recognition as achieved by using the Teager Energy Operator with Cepstral Mean Subtraction.

Keywords: Teager energy operator; cepstral mean subtraction; whispered speech recognition; linear scale; mel scale; dynamic time warping; hidden Markov models.

1. Introduction

Whisper is a specific speech mode used in situations where the communicator wishes to keep information private or discreet in public places (Ito *et al.*, 2005), when the caller hides their identity when making telephone calls, due to vocal cord problems, or for other reasons. Nowadays, when mobile telephone services have gained popularity, whisper has become a common speech mode for quiet communication that does not disturb uninvolved parties in specific settings (schools, libraries, market places, etc). Research on whispered speech is highly popular and still ongoing, focusing on signal/noise ratio, energy level (Jovičić, Šarić, 2008), vocal cord vibration (Catford, 1977), spectral slope (Zhang, Hansen, 2007), shifting the vowel formants to higher frequencies (Jovičić, 1998), formant frequency estimations (Gang, Heming, 2009), joint factor analysis for speaker verification (Gang, Heming, 2012), speaker identification (Fan, Hansen, 2014) and other areas. Therefore, whisper is still a challenge for research. The intelligibility of whispered speech is a major issue, despite its significant difference from normal speech. Hence, some studies have focused on brain activity during whisper processing, and on different areas of phonetics related to whisper (Tsunoda *et al.*, 2012).

The well-known techniques for automatic speech recognition (ASR) applied to normal speech can also be applied to whisper, after some modifications. The most popular ASRs are based on standard methods such as DTW (Dynamic Time Warping), HMM (Hidden Markov Models) (Rabiner, Juang, 1993) and ANN (Artificial Neural Networks) (Kostek, 1999). In the present research, standard DTW and HMM methods have been used as they have proved to be simple and reliable.

The application of Teager Energy Operator (TEO) for whisper has not been fully investigated. Some authors have used this operator for murmur recognition (Heracleous, 2009), others for speech under stress (normal, anger, loud, Lombard) (Hansen, Patil, 2007). The present study focuses on the use of TEO for whisper and normal speech recognition, and their match and mismatch scenarios.

This paper is structured as follows: Sec. 2 explains the use of the data recorded as the Whi-Spe (Marković *et al.*, 2013) speech corpus database. Section 3 describes preprocessing and extraction of all feature vectors: Linear Frequency Cepstral Coefficients

(LFCCs), Teager Energy based Linear Frequency Cepstral Coefficients (TELFCCs), Mel Frequency Cepstral Coefficients (MFCCs) and Teager Energy based Mel Frequency Cepstral Coefficients (TEMFCCs) with and without normalization and the first derivative. Results are presented in Sec. 4 as tables and diagrams, with the four feature vectors analyzed and compared. Final remarks and ideas for further research are given in the Conclusions.

2. Speech corpus

The Whi-Spe database created for the research on whispered speech was used in the experiment. The database contains 10 000 patterns of single words spoken in both normal and whispered mode. The vocabulary consists of 50 words of the Serbian language divided in three sub corpora: colors, numbers and phonetically balanced words. Five male and five female speakers were involved. Each pattern was recorded in the Whi-Spe database as a single wave file at a sampling rate of 22 050 Hz, 16 bits per sample. These files are inputs to the preprocessing system while the outputs include LFCC, TELFCC, MFCC and TEMFCC feature vectors and their variations. For the purpose of this research, the entire database was used.

Different parameter sets of vectors containing cepstral coefficients without normalization, cepstral coefficients with normalization and delta cepstral with cepstral coefficients with normalization were considered. These vectors were applied in two speech modes (normal and whispered) and four scenarios: normal/normal (N/N), whisper/whisper (W/W), normal/whisper (N/W) and whisper/normal (W/N). For example, in the W/N scenario, the model was trained with (W)hispered patterns, and recognition was performed using (N)ormal patterns.

3. Preprocessing and feature vector extraction

The generation of feature vectors for all these types involves the same first steps (Rabiner, Juang, 1993): pre-emphasis, blocking with overlap, windowing and Fast Fourier Transform (FFT) (Fig. 1).

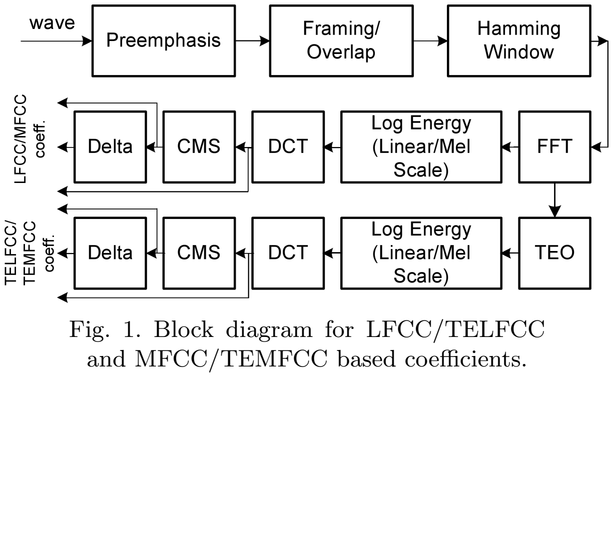

Fig. 1. Block diagram for LFCC/TELFCC and MFCC/TEMFCC based coefficients.

The pre-emphasis block produces a spectrally flattened signal and makes it less susceptible to finite precision effects later in the signal processing. In the framing/overlap block, the output signal of the pre-emphasis step is divided into N frames of 512 samples, with an overlap of 50%. Then, the frames are weighted with a Hamming window in the next block. The purpose of windowing is to taper the signal to zero at the beginning and end of each frame. The next step is the FFT, which calculates short time spectra for the signal.

Other steps to obtain cepstral coefficients for each feature vector are specific. In addition to cepstral coefficients, the first derivative is also included (delta cepstral coefficients). These coefficients are used to improve the performance of speech recognition systems. In order to calculate the first derivative, three neighboring frames are used.

The extraction of feature vectors for each of these types is explained in detail below.

3.1. LFCC and TELFCC feature vectors

Linear frequency cepstral coefficients (LFCCs) use a linear frequency scale. Some researchers have shown that, for normal speech, LFCCs are preferable for female speech because they better capture the spectral characteristics in the high frequency region (Zhou *et al.*, 2011). This can be explained by the relatively shorter vocal tract in females and the resulting higher formant frequencies of speech.

The block diagram in Fig. 1 shows the generation of LFCC/TELFCC and MFCC/TEMFCC feature vectors and their first derivative (delta coefficients).

After preprocessing blocks, log energy is calculated on a linear scale (for LFCCs). The scale is divided into 30 equal triangular filters. The filters cover a range from 0 to 11 025 Hz, and are overlapped at the central frequencies. Next, the Discrete Cosine Transform (DCT) is applied. At that point, LFCC feature vectors are generated, each with 12 coefficients.

To perform normalization, a CMS block is added. CMS is a channel normalization approach used to compensate for the acoustic channel (De Veth, Boves, 1998). When a speech signal goes through a time invariant channel, convolution distortions become multiplicative in the spectral domain and additive in the log-spectral domain. Since the cepstrum is just a linear transformation of the log-spectrum, both can be treated equally. For the speech signal, a short time analysis is performed resulting in the speech spectrum $S_t(\omega)$, and the resulting spectrum $Y_t(\omega)$. The index t indicates time dependence. The resulting spectrum is given as:

$$Y_t(\omega) = C(\omega) \cdot S_t(\omega) \tag{1}$$

and the log-spectrum (cepstrum) is:

$$y_t = c + s_t. \tag{2}$$

Since the channel is constant ($C(\omega) = \text{const}$), it can be compensated by subtracting the mean, leading to a new cepstral mean subtraction feature z_t:

$$z_t = y_t - \overline{y_t} = c + s_t - (c + \overline{s_t}) = s_t - \overline{s_t}. \tag{3}$$

Then, a block for the first derivative is applied (Delta). It produces LFCC delta coefficients. As a final result, the second branch of the block diagram in Fig. 1 generates three types of LFCC feature vectors used in the experiments.

Teager-Energy based Linear Frequency Cepstral Coefficients (TELFCCs) are produced by applying the non-linear Teager-Kaiser Energy Operator on the linear scale. This operator is proposed for tracking rapid energy changes within a glottal cycle (Kaiser, 1983), which is very attractive for whisper. For real discrete-time signals, TEO is defined as:

$$\Psi(x[n]) = x^2[n] - x[n-1] \cdot x[n+1]. \tag{4}$$

For complex discrete-time signals, TEO is the sum of the energy of real and imaginary parts of the signal (Dimitriadis *et al.*, 2005):

$$\Phi(x[n]) = \Psi(\mathrm{Re}\{x[n]\}) + \Psi(\mathrm{Im}\{x[n]\}). \tag{5}$$

The third branch in Fig. 1 shows the obtainment of TELFCC coefficients and their first derivative. After signal preprocessing (pre-emphasis, framing/overlap, windowing and FFT), the TEO of the signal is calculated, and magnitude is weighted by a linear filtered bank. Then, Log energy is computed for each subband, and finally the DCT is applied. Similarly, as for LFCC, three types of vectors are generated (depending on whether the signal goes through the CMS block or not, and whether the first derivative is calculated or not).

3.2. MFCC and TEMFCC feature vectors

Mel Frequency Cepstral Coefficients (MFCCs) represent both the model of the human auditory system and a discorrelating property of the cepstrum (Rabiner, Juang, 1993). They are the most widely used features for speech recognition. The mel scale is mapping physical frequencies to perceptual representations. The mapping between the physical frequency (in Hz) and the perceptual frequency (in mel) is given by Eq. (6):

$$mel = 2595 \cdot \log_{10}(1 + f/700). \tag{6}$$

Similarly, as with LFCC features, after the FFT, Log energy is calculated based on the mel scale. The scale is covered with 30 filters. The energy is calculated for each sub-band, and then the Discrete Cosine Transform is applied. Subsequently, blocks for normalization (CMS) and Delta are applied.

The third branch in Fig. 1 shows the obtainment of TEMFCC based coefficients. It is similar to the TELFCC explained above.

The resulting feature vectors are of three types: a) vectors of 12 cepstral coefficients without CMS; b) vectors of 12 cepstral coefficients with CMS, and c) vectors of 24 coefficients (12 cepstral plus 12 delta cepstral) with CMS. They were used in the following experiments.

4. Experimental results

The experiments were based on four feature vectors (LFCCs, TELFCCs, MFCCs, TEMFCCs) and their three types. To evaluate the suitability of these feature vectors for whispered speech, DTW (Marković *et al.*, 2013) and HMM (Galić *et al.*, 2014) methods were used for recognition. These methods are highly efficient as they use well-known techniques to compare speech patterns.

The Dynamic Time Warping method is based on dynamic programming, and focuses on finding an optimal path between the starting and ending points of two pattern representations. The speech patterns are represented by a set of vectors. The first set of patterns (50 words) is used as a reference, and the other patterns (nine sets, each consisting of 50 words) are test data. For local constraints, the type I proposed by Sakoe and Chiba (1978) is used when preference is given to a diagonal step. Global constraints are not used.

The Hidden Markov Models based ASR system is implemented using the HTK (Hidden Markov Model Toolkit, 2016). The generation of all the files needed by the HTK (i.e. script and configuration files as well as model initialization and phonetic transcription files) is fully automated using MATLAB. Also, MATLAB is used for logging ASR system performance with an evaluation using the HTK.

The Automatic Speech Recognition system backend is based on HMM models of context-independent phonemes. Output probabilities are modeled with continuous density GMMs and diagonal covariance matrices. Each monophone model has 5 states in total (3 emitting states), with a strictly left to right topology, and without skips. Each word from the Whi-Spe database is transcribed manually. The number of training cycles in embedded re-estimation is restricted to 5. To prevent variance underestimation, the variance floor for Gaussian probability density functions is set to 1% of the global variance. Initial model parameters are estimated by the flat-start method, with the models initialized using the global mean and variance (all models are initially given the same set of parameters) (Galić *et al.*, 2014). The number of mixture components is gradually increased. In the testing phase, the Viterbi algorithm is applied in order to determine the most probable state sequence.

Towards a more reliable evaluation of the performance, 5-fold cross-validation was conducted. For consistent comparison, in match scenarios (N/N and W/W), the percentage of utterances in the part for training was the same as in mismatch scenarios (N/W and W/N) i.e. 80%.

For the LFCC feature and all three types of vectors, the results expressed as the word recognition rate (WRR) for normal/normal, whisper/whisper, normal/whisper and whisper/normal scenarios are provided in Table 1. Confidence intervals were calculated for all results using a confidence level of 95%. They are described in MoE (Margin of Error) columns.

For the TELFCC feature and all three types of vectors, the results are presented in Table 2.

For MFCC and TEMFCC features, the same types of vectors as explained for LFCC and TELFCC were used. The results are given in Tables 3 and 4 for MFCC and TEMFCC, respectively.

The most interesting results were found for the type of vectors containing 24 coefficients (12 cepstral plus

Table 1. Average WRR with Margin of Error of LFCC feature vectors (in %).

Vector/Scenar.	LFCC (no CMS)				LFCC (with CMS)				LFCC+Δ (with CMS)			
	DTW		HMM		DTW		HMM		DTW		HMM	
	Avg	MoE	Avg	MoE	Avg	MoE	Avg	MoE	Avg	MoE	Avg	MoE
N/N	97.25	±0.92	97.46	±1.29	98.29	±0.78	98.50	±0.53	98.33	±0.74	99.32	±0.29
W/W	91.45	±3.50	96.18	±0.96	94.33	±2.61	97.78	±0.39	94.67	±2.58	98.74	±0.53
N/W	45.33	±6.96	28.31	±7.59	69.16	±6.54	71.19	±5.09	69.20	±6.65	77.93	±3.98
W/N	36.67	±4.53	22.76	±6.15	62.53	±5.31	61.62	±7.41	63.67	±5.14	69.66	±7.01

Table 2. Average WRR with Margin of Error of TELFCC feature vectors (in %).

Vector/Scenar.	TELFCC (no CMS)				TELFCC (with CMS)				TELFCC+Δ (with CMS)			
	DTW		HMM		DTW		HMM		DTW		HMM	
	Avg	MoE	Avg	MoE	Avg	MoE	Avg	MoE	Avg	MoE	Avg	MoE
N/N	97.22	±0.88	97.18	±1.40	98.40	±0.67	97.86	±0.86	98.36	±0.63	98.40	±0.43
W/W	91.42	±3.55	96.34	±0.95	94.64	±2.52	96.70	±0.83	94.62	±2.59	98.18	±0.50
N/W	44.08	±7.70	25.89	±5.79	69.71	±6.78	71.34	±3.67	69.53	±6.67	75.46	±3.36
W/N	37.36	±4.26	27.11	±9.51	63.29	±5.62	60.95	±6.46	64.47	±5.72	65.66	±6.23

Table 3. Average WRR with Margin of Error of MFCC feature vectors (in %).

Vector/Scenar.	MFCC (no CMS)				MFCC (with CMS)				MFCC+Δ (with CMS)			
	DTW		HMM		DTW		HMM		DTW		HMM	
	Avg	MoE	Avg	MoE	Avg	MoE	Avg	MoE	Avg	MoE	Avg	MoE
N/N	99.00	±0.24	98.16	±0.91	99.29	±0.27	98.84	±0.63	99.20	±0.27	99.28	±0.36
W/W	95.22	±2.43	95.64	±0.80	97.18	±1.60	97.14	±0.75	97.25	±1.65	98.90	±0.26
N/W	34.73	±3.84	14.41	±3.64	72.84	±5.32	67.77	±4.18	72.69	±5.61	73.03	±6.27
W/N	18.36	±2.44	20.98	±5.36	47.29	±5.12	64.34	±7.39	48.85	±5.12	77.62	±7.18

Table 4. Average WRR with Margin of Error of TEMFCC feature vectors (in %).

Vector/Scenar.	TEMFCC (no CMS)				TEMFCC (with CMS)				TEMFCC+Δ (with CMS)			
	DTW		HMM		DTW		HMM		DTW		HMM	
	Avg	MoE	Avg	MoE	Avg	MoE	Avg	MoE	Avg	MoE	Avg	MoE
N/N	98.87	±0.29	98.16	±1.04	99.24	±0.31	98.98	±0.67	99.18	±0.28	99.38	±0.37
W/W	95.33	±2.56	97.20	±0.88	97.18	±1.70	97.36	±0.64	97.11	±1.76	99.20	±0.27
N/W	34.71	±4.12	11.07	±2.27	73.15	±5.30	66.27	±4.61	73.22	±4.94	75.76	±4.57
W/N	18.56	±2.56	12.02	±2.45	48.60	±4.81	60.45	±7.41	49.56	±5.05	77.38	±6.74

12 delta cepstral coefficients). Figures 2 and 3 present word recognition rates (with confidence intervals) for match scenarios (normal/normal, whisper/whisper, respectively), and the type of vectors consisting of 24 coefficients, for both recognition methods.

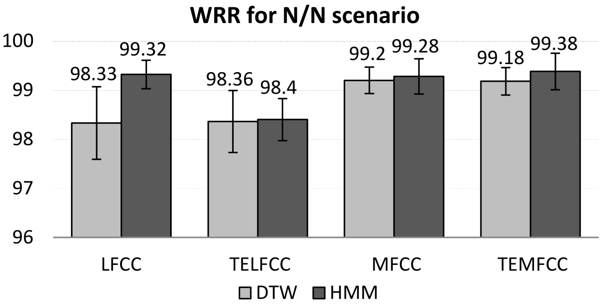

Fig. 2. Average word recognition rate for normal/normal scenario.

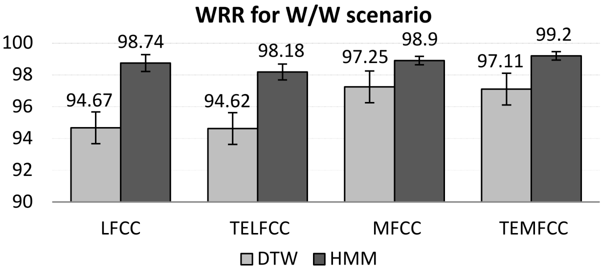

Fig. 3. Average word recognition rate for whisper/whisper scenario.

Figures 4 and 5 present word recognition rates for mismatch scenarios (normal/whisper and whisper/normal, respectively).

An additional statistical analysis was performed using a p-value and H_0/H_1 hypothesis (Neyman, Pearson, 1933). Two pairs of hypothesis were analyzed:

When CMS is used or not: Hypothesis H_{0A}: "Both algorithms (with no CMS and with CMS) produce the same WRR". Hypothesis H_{1A}: "The algorithm with CMS gives better results".

When TEO is used or not: Hypothesis H_{0B}: "Both algorithms (with no TEO and with TEO) produce the same WRR". Hypothesis H_{1B}: "The algorithm with TEO gives better results".

Results for H_{0A}/H_{1A} hypotheses are provided in Table 5, where the Hypothesized Mean Difference is set to zero, and the Alpha value is set to 0.05.

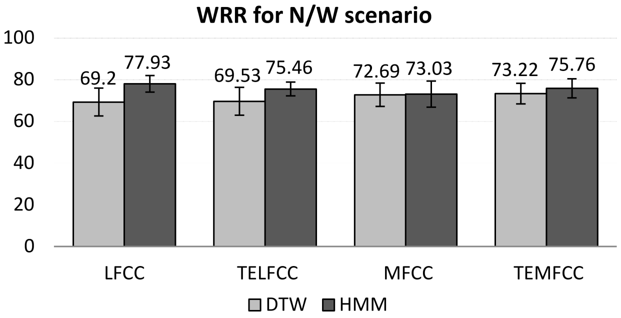

Fig. 4. Average word recognition rate for normal/whisper scenario.

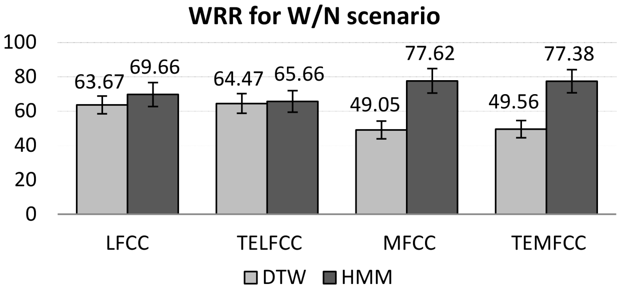

Fig. 5. Average word recognition rate for whisper/normal scenario.

These results show that differences for N/W and W/N scenarios were statistically significant ($p < 0.05$) and those for N/N and W/W scenarios close to being statistically significant.

Results for H_{0B}/H_{1B} hypotheses are shown in Table 6, where vectors of 24 coefficients are used.

Table 6. p-values for H_{0B}/H_{1B} hypotheses.

Vector/Scenario	LFCC/TELFCC		MFCC/TEMFCC	
	DTW	HMM	DTW	HMM
N/N	0.482	0.002	0.454	0.353
W/W	0.490	0.075	0.457	0.074
N/W	0.473	0.182	0.445	0.249
W/N	0.420	0.207	0.424	0.481

These results show no significant difference (in most cases $p > 0.05$). Also, the results based on the HMM method are closer to a low p-value than those based on the DTW method.

Table 5. p-values for H_{0A}/H_{1A} hypotheses.

Vector/Scenario	LFCC		TELFCC		MFCC		TEMFCC	
	DTW	HMM	DTW	HMM	DTW	HMM	DTW	HMM
N/N	0.054	0.081	0.025	0.214	0.063	0.123	0.051	0.105
W/W	0.110	0.004	0.082	0.291	0.102	0.007	0.127	0.389
N/W	5.87 E–05	1.6 E–08	5.8 E–05	6.96 E–11	5.89 E–10	1.31 E–13	7.38 E–10	1.99 E–14
W/N	4.6 E–07	1.43 E–07	5.26 E–07	9.12 E–06	4.45 E–09	1.32 E–08	1.34 E–09	2.05 E–10

5. Conclusions

The results suggest the following conclusions:

1. Cepstral Mean Subtraction normalization provided a huge improvement in speech recognition in all scenarios, especially when mismatch scenarios were used.
2. TEO with CMS improved recognition in most cases when applied on a mel scale.
3. Hidden Markov Models as a recognition method outperformed DTW for all scenarios when CMS normalization was used.
4. The highest WRR for match scenarios was 99.38% (for N/N scenario) and 99.20% (for W/W scenario) when the TEMFCC feature vector and the HMM method were used.
5. The highest WRR for mismatch scenarios was around 77% (for N/W and W/N scenarios), and was obtained using TEMFCC, MFCC and LFCC feature vectors.
6. The use of the first derivative of cepstral coefficients led to an important improvement in recognition (up to 17%), especially when the HMM method was applied.

In general, the TEMFCC feature vector ensures respectable results, especially when HMM is used as a recognition tool. Using a normalization technique such as CMS is necessary for good results. Its impact is essential for N/W and W/N scenarios.

Further research can focus on ASR using these feature vectors, different recognition tools, such as Kaldi Toolkit (Kozierski *et al.*, 2016), new methods i.e. ANN (Kostek, 1999), and an enlarged whisper database.

References

1. Catford J.C. (1977), *Fundamental problems in phonetics*, Edinburgh: Edinburgh University Press.
2. De Veth J., Boves L. (1998), *Channel normalization techniques for automatic speech recognition over the telephone*, Speech Communication, **25**, 149–164.
3. Dimitriadis D., Maragos P., Potamianos A. (2005), *Auditory Teager energy cepstrum coefficients for robust speech recognition*, Proc. of European Conf. on Speech Communication and Technology – Interspeech 2005, Lisbon, Portugal, 3013–3016.
4. Fan X., Hansen J.H.L. (2014), *Speaker identification with whispered speech based on modified LFCC parameters and feature mapping*, Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, May 2014, pp. 4553–4556.
5. Galić J., Jovičić S.T., Grozdić Đ., Marković B. (2014), *HTK-based recognition of whispered speech*, A. Ronzhin *et al.* [Eds.]: SPECOM 2014, LNAI 8773, Springer International Publishing Switzerland 2014, 251–259.
6. Gang L., Heming Z. (2009), *Formant frequency estimations of whispered speech in Chinese*, Archives of Acoustics, **34**, 2, 127–135.
7. Gang L., Heming Z. (2012), *Joint factor analysis of channel mismatch in whispering speaker verification*, Archives of Acoustics, **37**, 4, 555–559.
8. Hansen J.H.L., Patil S. (2007), *Speech under stress: analysis, modeling and recognition*, [in:] Müller C. [Ed.], *Speaker Classification I: Fundamentals, Features, and Methods*, Springer, Berlin–Heidelberg, pp. 108–137.
9. Heracleous P. (2009), *Using teager energy cepstrum and HMM distances in automatic speech recognition and analysis of unvoiced speech*, International Journal of Information and Communication Engineering, **5**, 1, 31–37.
10. Hidden Markov Model Toolkit (2016), http://htk.eng.cam.ac.uk/ (retrieved June 15, 2016).
11. Ito T., Takeda K., Itakura F. (2005), *Analysis and recognition of whispered speech*, Speech Communication, **45**, 139–152.
12. Jovičić S.T. (1998), *Formant feature differences between whispered and voiced sustained vowels*, Acustica united with Acta Acoustica, **84**, 4, 739–743.
13. Jovičić S.T., Šarić Z.M. (2008), *Acoustic analysis of consonants in whispered speech*, Journal of Voice, **22**, 3, 263–274.
14. Kaiser J.F. (1983), *Some observations on vocal tract operation from a fluid flow point of view*, in: Vocal Fold Physiology: Biomechanics, Acoustics and Phonatory Control, Titze I.R., Scherer R.C. [Eds.], Denver Center for the Performing Arts, Denver, CO, pp. 358–386.
15. Kostek B. (1999), *Soft computing in acoustics, applications of neural networks, fuzzy logic and rough sets to musical acoustics*, Springer-Verlag, Berlin.
16. Kozierski P., Sadalla T., Drags S., Dobrowski A., Horla D. (2016), *Kaldi toolkit in Polish whispery speech recognition*, Przeglad Elektrotechniczny, **R.92**, 11, 301–304.
17. Marković B., Galić J., Grozdić Đ., Jovičić S.T. (2013), *Application of DTW method for whispered speech recognition*, Proc. of 4th International Conference on Fundamental and Applied Aspects of Speech and Language, Belgrade, 308–315.
18. Marković B., Jovičić S.T., Galić J., Grozdić Đ. (2013), *Whispered speech database: design, processing and application*, Proc. of 16th International Conference, TSD 2013, I. Habernal and V. Matousek [Eds.]: TSD 2013, LNAI 8082, Springer-Verlag Berlin Heidelberg, pp. 591–598.
19. Neyman J., Pearson E. (1933), *On the problem of the most efficient tests of statistical hypotheses*, Philosophical Transactions of the Royal Society of London. Series A, **231**, 289–337.
20. Rabiner L., Juang B-H. (1993), *Fundamentals of speech recognition*, Prentice Hall, New Jersey.

21. Sakoe H., Chiba S. (1978), *Dynamic programming optimization for spoken word recognition*, IEEE Transactions on Acoustics, Speech, and Signal Processing, **26**, 1, 43–49.

22. Tsunoda K., Sekimoto S., Baer T. (2012), *Brain activity in aphonia after a coughing episode: different brain activity in healthy whispering and pathological aphonic conditions*, Journal of Voice, **26**, 5, 668.e11–668.e13.

23. Zhang C., Hansen J.H.L. (2007), *Analysis and classification of speech mode: whisper through shouted*, Proc. of Interspeech 2007, pp. 2289–2292.

24. Zhou X., Garcia-Romero D., Duraiswami R., Espy-Wilson C., Shamma S. (2011), *Linear versus mel frequency cepstral coefficients for speaker recognition*, 2011 IEEE Workshop on Automatic Speech Recognition & Understanding, ASRU 2011, Waikoloa, HI, USA, December 11–15, pp. 559–564.