DOI 10.15622/sp.58.2

J.N. GALIĆ, S.T. JOVIČIĆ, V.D. DELIĆ, B.R. MARKOVIĆ,
D.S. ŠUMARAC PAVLOVIĆ, Đ.T. GROZDIĆ

HMM-BASED WHISPER RECOGNITION USING µ-LAW FREQUENCY WARPING

Galić J.N., Jovičić S.T., Delić V.D., Marković B.R., Šumarac Pavlović D.S., Grozdić Đ.T. **HMM-based Whisper Recognition using µ-law Frequency Warping.**

Abstract. Due to the lack of sufficient amount of whisper data for training, whispered speech recognition is a serious challenge for state-of-the-art Automatic Speech Recognition (ASR) systems. Because of great acoustic mismatch between neutral and whispered speech, ASR systems are faced with significant drop of performance when applied to whisper.

In this paper, we give an analysis of neutral and whispered speech recognition based on traditional Hidden Markov Models (HMM) framework, in a Speaker Dependent (SD) and Speaker Independent (SI) cases. Special attention is paid to the neutral-trained recognition of whispered speech (N/W scenario). The ASR system is developed for recognition of isolated words from a real database (Whi-Spe) of neutral-whisper speech pairs. In the N/W scenario, a meaningful gain in robustness is achieved with the proposed frequency warping, originally developed for speech signal compression and expanding in digital telecommunication systems. Simultaneously, good performances in recognition of neutral speech are retained.

Compared to baseline recognition with Mel-frequency Cepstral Coefficients (MFCC), word recognition accuracy with cepstral coefficients using proposed frequency warping (denoted as µFCC) is improved for 7.36% (SD) and 3.44% (SI), absolute. As well, the F-measure (harmonic mean of the precission and recall) for µFCC feature vectors is increased for 6.90% (SD) and 3.59 (SI). Statistical tests confirm significance of the achieved improvement in recognition accuracy.

Keywords: automatic speech recognition, feature extraction, hidden Markov models, human voice, whisper, speech processing.

1. Introduction. Speech is the most natural and convenient form of interpersonal communication. According to the level of vocal effort, speech is classified in 5 modes: whispered, soft, normally phonated (neutral), loud and shouted speech [1]. Whisper is the most distinctive mode because of the lack of glottal vibrations and noisy excitation of the vocal tract. Humans tend to whisper or generally lower their voice for several reasons. First, it is used in situations where aloud speech is prohibited or inappropriate (e.g. in theatre or reading room); second, if some confidential information should not be heard from uninvolved parties, and third, in criminal activities for hiding their identity. In addition to conscious production of whisper, it may be phonated as a result of health issues, which appear after laryngitis or rhinitis.

State-of-the-art Automatic Speech Recognition (ASR) systems show good performances (accuracy and speed) and wide-spread commercial use. At the same time, they express high sensitivity when exposed to speech different from one used in training which is usually neutral speech recorded in controlled or even laboratory conditions. Recognition of such atypical speech with satisfactory accuracy independent from speaker is a challenging task for research community, and includes:

– speech changed in vocal effort;

– speech under different kinds of emotional states;

– various speaker dialects;

– Lombard effect speech;

– speech in adverse conditions (environment noise, reverberation, loudness, etc.).

In a range of speech modes from whisper to shouted, whispered speech has the most negative impact on the performance of the ASR system [2]. A considerable acoustic mismatch between neutral and whispered speech has dominant influence on such performance degradation. Since whisper data are not generally available (or at least not in a sufficient amount) for training of ASR systems, the greatest attention is paid to whisper recognition with ASR system trained with neutral speech only.

In this paper it is shown that using novel frequency warping for feature extraction in traditional Hidden Markov Models (HMM) framework gives accuracy in whisper recognition comparable with deep learning approach. In order to improve whisper recognition accuracy using neutral-trained ASR system, feature extraction based on μ-law frequency warping is introduced. Therefore, the improvement in whisper recognition accuracy is achieved without model adaptation, feature mapping or increase in number of cepstral coefficients. Moreover, it is shown that a filterbank resolution in high and low frequency range of speech has an influence on accuracy in mismatched train/test scenarios. This study includes recognition of isolated words in neutral and whispered phonation in both Speaker Dependent (SD) and Speaker Independent (SI) cases.

The remainder of this paper is organized in 6 sections as follows. In Section 2, the literature survey on whisper recognition and description of available speech databases are briefly discussed. Section 3 gives basic characteristics of whispered speech and comparison with neutral speech.

Explanation of proposed frequency warping is given in Section 4. Experimental preparation (speech database, feature extraction procedure and ASR system) is described in Section 5. Experimental results and discussion are given in Section 6, while concluding remarks and directions for future work are stated in Section 7.

2. Related Works. The main prerequisite for effective language-dependent use of whispered speech in modern ASR system is extensive and systematically created speech database. To the best of our knowledge, there are only few speech databases with recordings in both speech modes: English [3], Serbian [4], Mandarin [5] and Polish [6].

One of the earliest research studies in recognition of whispered speech (over a cellular phone) was conducted for Japanese at University of Nagoya [2]. The research demonstrated that using small amount of whispered speech per target speaker (10 to 50 sentences) can be effectively used for whisper recognition. Subsequent studies were focused on compensation of differences between neutral and whispered speech. Significant improvement for whisper speaker identification was obtained with frequency warping and score competition [7]. Compared with closed-set speaker ID task based on a traditional Mel-frequency Cepstral Coefficients (MFCC), an exponential based frequency warping gave absolute accuracy gain of 27%.

High accurate detection of whisper-islands embedded within continuous neutral speech was achieved with linear prediction residual and entropy-based features [8-9]. Whisper recognition based on deep neural networks and KALDI toolkit was investigated in [6].

The generation of pseudo-whisper for efficient model adaptation based on Vector Taylor Series (VTS) algorithm was demonstrated in [10-11]. Together with vocal tract length normalization and shift frequency transformation the Word Error Rate (WER) reduction from 27.7% to 17.5% (for open speaker scenario) was reported. The ASR system was speaker independent with constrained lexicon [3, 10-11]. The research studies demonstrated that WERs were considerably reduced after adapting the acoustic model toward the VTS or denoising autoencoders pseudo-whisper samples, compared to model adaptation on an available small whisper set.

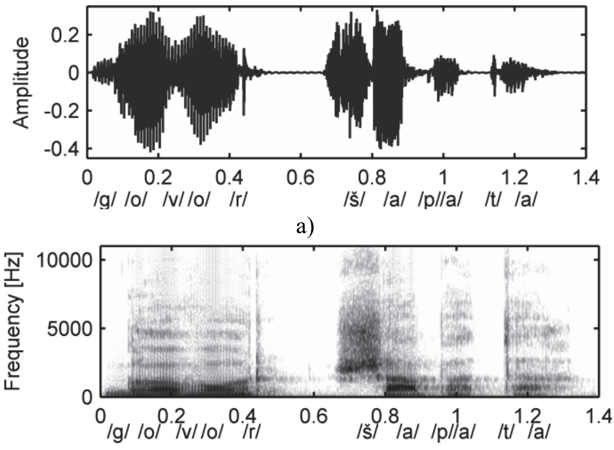
Preceding papers related to recognition of whispered speech from Whi-Spe database [4], were focused to SD case. Comparison between different normalization techniques was analyzed in [12]. The following normalization techniques were tested and compared: CMN (Cepstral Mean Normalization), CVN (Cepstral Variance Normalization), MVN (Cepstral Mean and Variance Normalization), CGN (Cepstral Gain Normalization) and quantile-based dynamic normalization techniques such as QCN and QCN-RASTA. The best results were obtained using CMN.

Recently, using Teager energy cepstral coefficients with deep denoising autoencoder and inverse filtering has brought many benefits in speaker dependent neutral-trained whisper recognition [13-14].

Although a novel contribution was represented in each study, commercially available speaker independent recognition of whispered speech is an important problem that needs to be addressed in details.

3. Characteristics of Whispered Speech. Whisper is a specific style of speech which is, according to the characteristics, nature and generating mechanism, quite different from neutral speech. As already mentioned, the main characteristics of whisper are an absence of fundamental frequency and noisy excitation of the vocal tract. It was determined that the formant frequencies for whispered vowels are substantially higher than for the neutral voice [15]. Compared to normally phonated speech, whisper has lower frame energy, longer duration of speech and silence, flatter long-term spectrum and lower Sound Pressure Level (SPL) [1]. However, despite the fact that an increased effort in speech perception is needed, the intelligibility of whisper is very high. An average identification accuracy of 82% for vowels in [hVd] syllables in whisper mode has been shown in [16]. On the other hand, non-linguistic information is hardly revealed in whisper, like age, sex, emotions or identity.

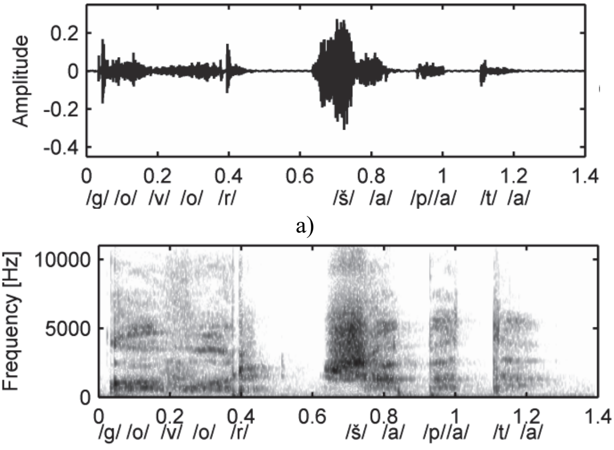
In Figures 1 and 2 the waveform and spectrogram of the short sentence in Serbian "Govor šapata." ("Whispered speech." in English), uttered in neutral and whispered speech, are depicted, respectively. The figures are supported with a phonetic transcription. Because of the lack of sonority, a difference in amplitude levels between the two modes of speech can be observed. Also, the spectrograms show that some parts of spectrum are well preserved in whisper. That is especially strong for unvoiced consonants, such as fricative /š/ (/ʃ/ in IPA notation) and plosives /p/ and /t/. A similar shape of spectrum of vibrant /r/ in Serbian is observed. Moreover, the waveform and spectrogram show that the harmonic structure of vowels is lost in whisper.



a)

b)

Fig. 1. The waveform (a) and the spectrogram (b) of a short sentence in Serbian "Govor šapata" uttered in normal phonation (neutral speech). The time in seconds is given on the abscissa



a)

b)

Fig. 2. The waveform (a) and the spectrogram (b) of a short sentence in Serbian "Govor šapata" uttered in whispered speech. The time in seconds is given on the abscissa

4. Novel Frequency Warping. Several frequency warping schemes commonly used in ASR are presented in this section, including a new one – proposed in this paper.

Mel-frequency cepstral coefficients (MFCC) are traditional and the most popular feature vectors for speech characterization in ASR systems. Their mel-warped frequency scale, which emulates human's ear sound perception, is given in the following equation:

$$f\left[\text{mel}\right] = 2595 \cdot \log_{10}\left(1 + f[\text{Hz}]/700\right). \tag{1}$$

In some special ASR tasks, Linear Frequency Cepstral Coefficients (LFCCs) have shown considerable improvement in regard to MFCCs, especially in speaker identification in whisper mode [17].

Likewise, Perceptive Linear Prediction (PLP) feature vectors [18] are frequently used, specifically in adverse conditions. They are based on bark psychoacoustical scale on which equal distances correspond with perceptually equal distances. The scale corresponds to 24 critical bands of hearing and ranges from 1 up to 24.

Frequency warping based on bark scale is given in the following equation:

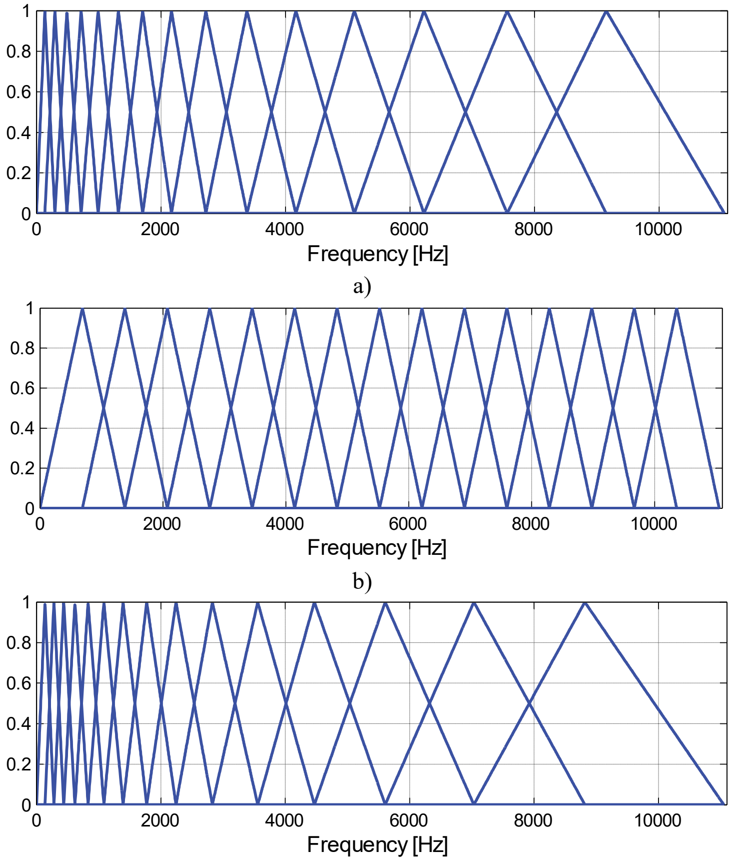
$$f\left[\text{bark}\right] = 6 \cdot \sinh^{-1}\left(f[\text{Hz}]/600\right), \tag{2}$$

where $\sinh^{-1}(x)$ denotes inverse sine hyperbolic function.

Filterbank frequency characteristics based on mel, linear, and bark frequency scale are depicted in Figure 3.

As noted in Section III, because of its unvoiced nature, compared to neutral speech, the spectrum of whispered speech tends to be more flat. As a consequence, a relatively significant portion of whispered speech information is reflected in higher range of speech frequency spectrum, wherein mel and bark scale have poor resolution (as can be seen in Figure 2). Consequently, recognizer which uses feature vectors based on mel and bark scale may neglect significant spectral details and perform poorly when applied to whisper. Compared to these scales, linear frequency scale improves resolution in higher range, but simultaneously deteriorates good resolution in lower frequency range of speech spectrum. In order to find optimal frequency warping in whisper recognition, we have considered the possibilities to combine good properties of linear and mel frequency scale; good frequency resolution in low frequency range (for mel scale) and high range (for linear scale). For that reason, we propose a novel frequency warping with non-linear mapping (originally used in speech signal compression and expanding in North America and Japan; so-called μ-law [19]), defined by the following equation:

$$warp = f_N \frac{\ln\left(1+\mu \cdot f / f_N\right)}{\ln(1+\mu)}. \qquad (3)$$



a)

b)

c)

Fig. 3. Filterbank characteristics with 15 triangular filters for: a) mel; b) linear; c) bark frequency scale

In (3), $f_N = f_s/2$ (f_s is the sampling frequency) is the Nyquist frequency, μ is a positive constant, and $\ln(x)$ refer to the natural logarithm. As evident from (3), warping functions cross the identity line for frequencies $f = 0$ and $f = f_N$. Frequency warping curves are depicted in Figure 4 for values of warping coefficient $\mu \in \{0, 1, 2\}$ and corresponding filterbank characteristics in Figure 5. From the shapes of curves depicted in Figure 4, it is evident that parameter μ determines the degree of convexity of warping functions. It can be shown (using L'Hospital's rule) that for $\mu \rightarrow 0$ warping function converges to the identity line $warp = f$, i.e., linear frequency scale.

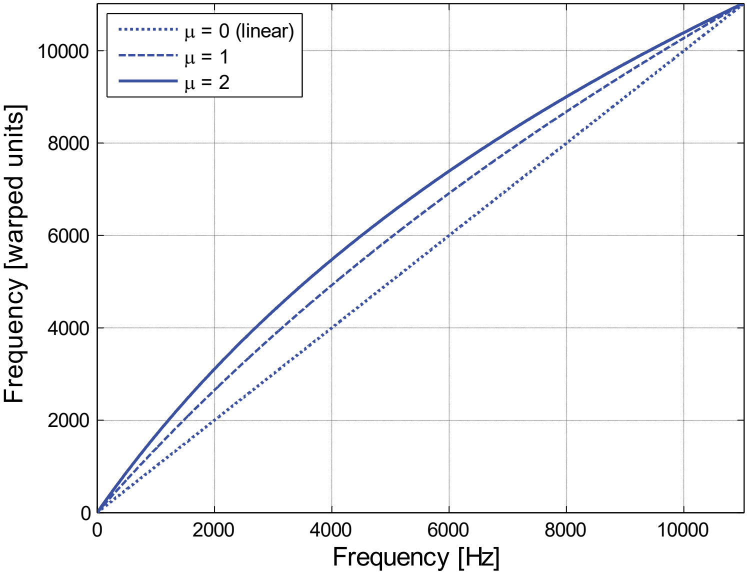
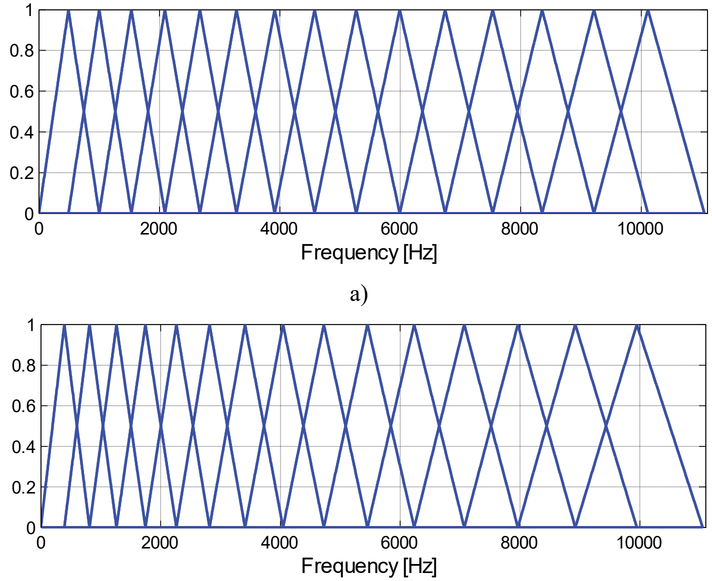


Fig. 4. Warping functions for μ-law frequency warping for three values of warping coefficient

Filterbank resolution according to μ-law frequency warping is clearly visible from Figure 5. Frequency warping using μ-law functions yields to higher frequency resolution over mel scale (in high frequency range of speech) and over linear scale (in lower part). In the research we have experimentally tested the hypothesis that using cepstral coefficients with new μ-warped frequency scale as feature vectors may provide some advantages in whisper recognition. For simplicity, these feature vectors are denoted by μFCC in the remainder of this paper.



a)

b)

Fig. 5. Filterbank characteristics with 15 triangular filters with μ-law frequency warping for values of warping coefficient: a) μ = 1; b) μ = 2

5. Experimental Preparation. This section is divided into three subsections describing speech database, feature extraction procedures, and applied ASR system.

5.A. Speech Database. For the purpose of machine recognition of neutral speech and whisper in Serbian, the speech database Whi-Spe (abbreviation of Whispered Speech) is created in the initial form [4]. The database is recorded in laboratory conditions, with a high-quality omni-directional microphone. The database was designed to have two parts: one that contains recordings of whispered words, and another one that comprises recordings of the same words uttered in neutral phonation. The corpus of 50 words is included in the database, from 10 speakers (5 female and 5 male). Each speaker read all 50 words 10 times in both speech modes. Finally, the Whi-Spe database contains 10000 recorded words, 5000 in normal speech and the same number of words recorded in whisper, or 2 hours in total. The words are divided in three sub-corpora: basic colors (6 words), numbers (14 words) and phonetically balanced words (30 words). The speech data are digitized using a sampling frequency 22050 Hz, with 16 bits per sample, in Windows linear Pulse Code Modulation (PCM) *.wav* format.

More information about the vocabulary of the Whi-Spe database, manual segmentation, the quality control and a way of labeling can be found in [4].

5.B. Feature extraction. Generation of MFCCs is performed according to procedure described in [20], using MATLAB software package. Before performing MFCC calculation, speech frames are windowed by Hamming window functions in duration of 24 ms, mutually shifted by 8 ms and pre-emphasized using a filter with coefficient 0.97. Twenty filterbank channels uniformly distributed over mel-frequency scale are used. Compared to MFCCs, the parameter configurations for both LFCCs and μFCCs are the same except for the frequency warping scale.

The PLP features are extracted according to [21], using freely available code in MATLAB. Beside conventional PLP feature vectors, the analysis of PLP feature vectors with linear frequency scale is given. These feature vectors are denoted by LPLP. The parameter configurations for both PLPs and LPLPs are the same except for the frequency warping scale.

Experiments in the SI open-speaker neutral-trained whisper recognition show the best performance for modified PLP feature vectors, with bypassed equal loudness and power-intensity processing [10]. The filterbank comprise triangular filters uniformly distributed over linear frequency scale restricted to range 0-5800 Hz. Consequently, experiments with modified LPLP vectors are appended in this study. These feature vectors are denoted by LPLP(mod). Generation of LPLP(mod) feature vectors is performed by modifying the code in MATLAB used for PLP. Recognition of whispered speech based on PLP feature vector using using Maximum Likelihood Linear Regression (MLLR) is analyzed in [22].

Speech recognition using RASTA (Relative Spectral) filtering applied on PLP features with ASR backend based on DTW algorithm is examined in [23]. The results confirmed good improvement in recognition when RASTA filtering is applied, especially in mismatch scenarios. As well, cepstral coefficients based on gammatone filterbank are analyzed in [24].

For all feature vectors in this research, each frame is represented with 39 coefficients, i.e., 13 cepstral coefficients (including the energy), along with their first and second order time derivatives. Coefficients are normalized with cepstral mean of each utterance.

5.C. HMM-GMM ASR System. In ASR systems, the conventional technology is based on HMMs with Gaussian Mixture Models (GMMs). The most commonly used modeling units in isolated words recognition are phonemes independent from their context (monophones), phonemes dependent from their context (usually triphones), and the whole words. The greatest robustness in the case of experiments with the Whi-Spe database (isolated words) is achieved for the monophone models [25]. Therefore, models of phonemes independent from their context are used in this research.

The ASR system used in this paper is completely designed by using HTK [20]. The generation of the script and configuration files, as well as the files for model initialization and phonetic transcription is automated using MATLAB. For logging the ASR system performance results MATLAB is also used.

Output probabilities are modeled with the continuous density GMMs and diagonal covariance matrices. Each monophone model is represented with strictly left-to-right topology and self loops, but without skips over states. Each word from the Whi-Spe database is transcribed manually. The number of training cycles in embedded re-estimation is fixed to 5 and the variance floor for Gaussian probability density functions is set to 1%. The number of mixture components is gradually increased and amounts to 8 (in the SD case) and 32 (in the SI case). In the testing phase, the Viterbi algorithm is applied in order to determine the most probable state sequence. The experiments are conducted in both the SD and SI cases, with 32 monophones — 30 monophones corresponding to 30 letters in the Serbian alphabet, the phoneme /ə/ (schwa) and the silence. Schwa is marked when /r/ is found in a consonant environment. The model of silence is appended at the start and the end of each utterance.

The parameters of initial models in a flat-start training are obtained by calculating the global mean and variance. However, more accurate initial models could be achieved with the annotation of a part of database used for training, which includes labeling phoneme boundaries in utterance. In this paper, for both recognitions in the SD and SI cases, we use automatic annotation of a small database subset to bootstrap a set of HMM models. Additionally, instead of using a fixed number of states per each monophone model, a noticeable gain in robustness can be achieved with a variable number, proportional to the phoneme duration. The number of HMM states per model, proportional to the average duration of all the instances of the corresponding phoneme in the training database is proposed in [26], for all phonemes in Serbian.

The parameters of the initial monophone models are obtained by using a small part of the database (10% of utterances in neutral phonation) annotated with automatic annotation with the forced alignment implemented in the HTK.

6. Results and Discussion. This section is organized as follows. The results and discussion of initial experiments, which analyze already existing approaches for speech characterization in ASR systems, are presented in subsection A. The main objective of initial experiments is to give a baseline performance for MFCC, LFCC, PLP, LPLP and LPLP(mod) feature vectors in terms of word recognition accuracy, in 4 train/test scenarios:

– N/N and W/W — the ASR system is trained on neutral speech (N) or whispered speech (W) and tested using the speech of the same mode. These scenarios are marked as *matched*.

– N/W and W/N — the ASR system is trained on neutral speech or whispered speech and tested against the speech of the opposite mode. These scenarios are marked as mismatched.

In subsection B, the influence of μ-law warping coefficient to accuracy in neutral-trained recognition of neutral and whispered speech is examined. In addition, a comparison of the recognizer performance using μ-law frequency warping and feature vectors utilized in initial experiments, along with statistical significance of results, is given.

The experiments are conducted in both the SD and SI cases. In order to provide more reliable evaluation of the performance, cross-validation is needed. For each speaker, 1000 utterances (500 in neutral and 500 in whisper mode) are available. Word recognition accuracy is presented as metric for performance of the recognizer.

In the SD case, accuracy is calculated according to the following procedure. In matched conditions available utterances are divided in the train and test set. The train set contains 90% utterances evenly distributed between words. Remaining 50 utterances are exploited in the test set. The HTK displays the percentage of correctly recognized utterances. For example, if N denotes total number of analyzed utterances and E denotes the number of incorrectly recognized utterances, accuracy percentage is calculated in the following way:

$$accuracy = \frac{N - E}{N} \cdot 100\%. \tag{4}$$

The train and test set are rotated in 10-fold cross-validation. Accuracy for an examined speaker is calculated by averaging 10 results from cross-validation. Finally, average SD recognition accuracy is calculated as arithmetic mean of accuracies from all speakers. The procedure is the same for mismatched conditions, except the fact that test set contains all available utterances in the opposite speech mode. In train set equal number of utterances (450) is utilized in both matched and mismatched conditions.

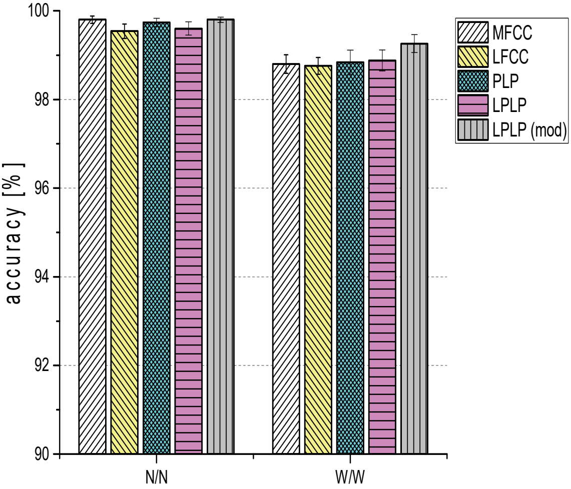
In the SI case, all 500 utterances from the examined speaker (for the respective mode) are given in the test set, whereas the utterances from the other 9 speakers (4500 for the respective mode) are given in the train set (full dataset training with leave-one-speaker-out cross-validation). Again, the accuracy is averaged across different speakers.

6.A. Initial Experiments. The results are depicted in bar graphs in Figure 6 and Figure 7, for the SD and the SI recognition, respectively. Depicted Standard Errors (SE) present standard deviation between different recognition systems divided by the square root of the sample size.

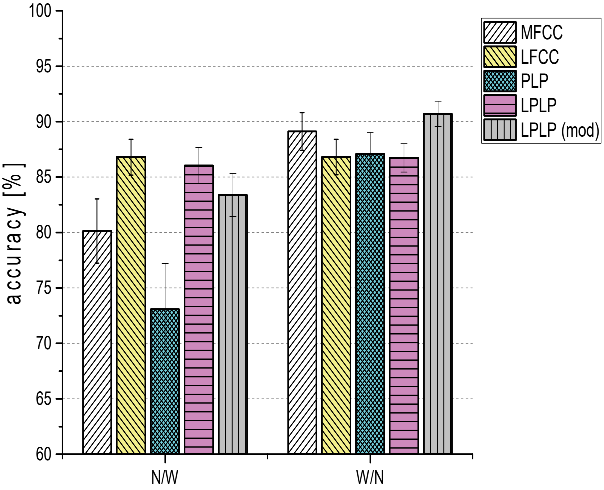
For better visual comparison of accuracies, only important part of each bar graph is shown (higher than 90% in matched and 60% mismatched scenarios). As can be seen from Figure 6(a) and Figure 7(a), the recognition of whisper is with lower success compared to the recognition of neutral speech in matched scenarios, as expected. In the SD case, recognition accuracy of neutral speech is higher than 99.5% for all feature vectors (N/N bars in Figure 6). The difference in performance between examined feature vectors is meaningless. In contrast, there is a noticeable increase in performance for modified LPLP features in recognition of whispered speech (W/W bars in Figure 6) with reached accuracy of 99.26%.

In the SI case, the performance of ASR system for LFCC and LPLP feature vectors noticeably dropped down, compared to original MFCC and

PLP features (bar graphs in Figure 7(a)). Recognition accuracy of 98.60% (neutral speech) and 96.66% (whisper) is achieved.

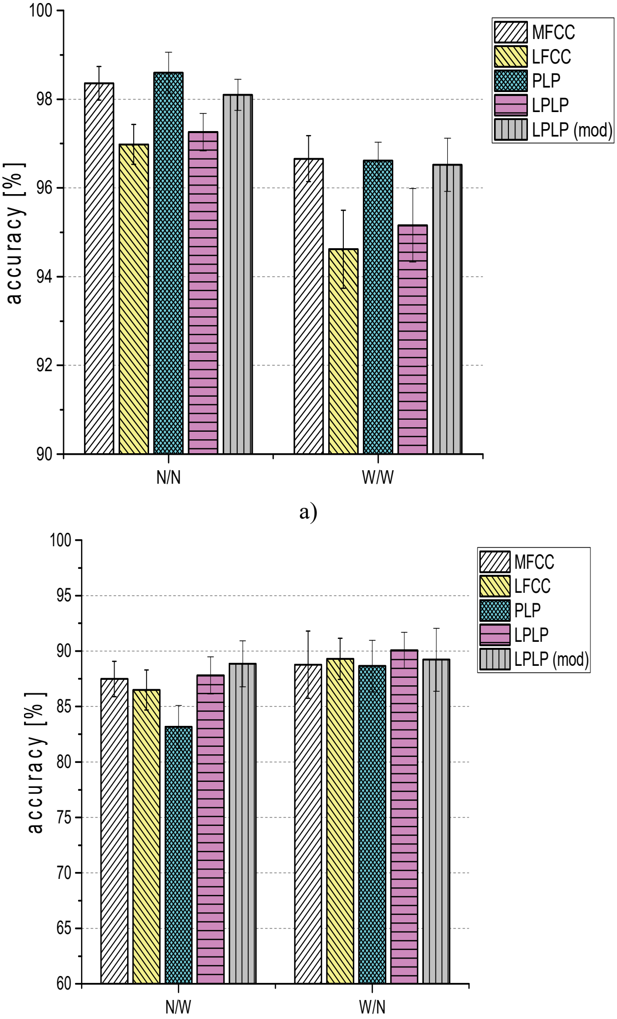


a)



b)

Fig. 6. The average word recognition accuracy with standard error (SE) in speaker dependent (SD) case and four train/test scenarios in: a) matched; b) mismatched scenarios for MFCC, LFCC, PLP, LPLP and modified LPLP feature vectors



a)

b)

Fig. 7. The average word recognition accuracy with standard error (SE) in speaker independent (SI) case and four train/test scenarios in: a) matched and b) mismatched scenarios for MFCC, LFCC, PLP, LPLP and modified LPLP feature vectors

In mismatched scenarios, few observations can be made. There is a significant drop of performance compared to the recognition in matched conditions, for both the SD and the SI recognitions (Figure 6*b* and Figure 7*b*).

Also, there is a pronounced asymmetry in recognition accuracy between N/W and W/N train/test scenarios for MFCC and PLP feature vectors, which is especially strong for SD recognition. On the contrary, speech parameterization based on linear frequency scale (LFCC and LPLP) contributes to the absence of asymmetric performance between mismatched scenarios. As well as in research study [10], reducing the filterbank bandwidth to range 0-5800 Hz provides further improvement in neutral-trained recognition of whispered speech (Figure 7*b*, LPLP(mod) bar in N/W scenario) while preserving good performance in recognition of neutral speech. The highest recognition accuracy in N/W scenario (as more interesting scenario) is 86.80% in the SD case (for LFCC features) and 88.86% in the SI case (for LPLP(mod) features). In order to compare the performance in N/W scenario with results from [10], Word Error Rate (WER) for UT-Vocal Effort II was 18.2%, without adaptation to whisper and with lexicon constrained to 160 words. Very high deviation of performance among different speakers is obtained. Similar observation is found in whispered speaker identification with neutral trained HMM models [27]. 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.

6.B. Recognition with Cepstral Coefficients using μ-law Frequency Warping. The experiments are done for 5 values of warping coefficient μ, from 0.5 up to 2.5 (with an increment value of 0.5).

The neutral-trained ASR system recognition accuracy is depicted in Figure 8 (for neutral speech) and Figure 9 (for whisper).

Experiments in matched scenario show the best performance (accuracy 98.56%) for value of warping coefficient μ=2 (SI bars in Figure 8). At the same time, warping coefficient has meaningless influence on recognition in the SD case. For each examined value, accuracy is higher than 99.60% (SD bars in Figure 8).

Experiments in mismatched scenario (Figure 9) show very similar tendency with regard to influence of warping coefficient to the performance. Once again, the best accuracy in the SI recognition is achieved for μ = 2, and amounts to 90.92%. Experiments in the SD case show very small change in accuracy for values of coefficient in range from μ = 1 up to μ = 2. The best obtained accuracy is 87.50% for value μ = 2. Finally, the results demonstrate that using μFCC results in increase of whisper-recognition accuracy, as compared to the neutral-trained recognition in initial experiments, while keeping very good performance in recognition of neutral speech.

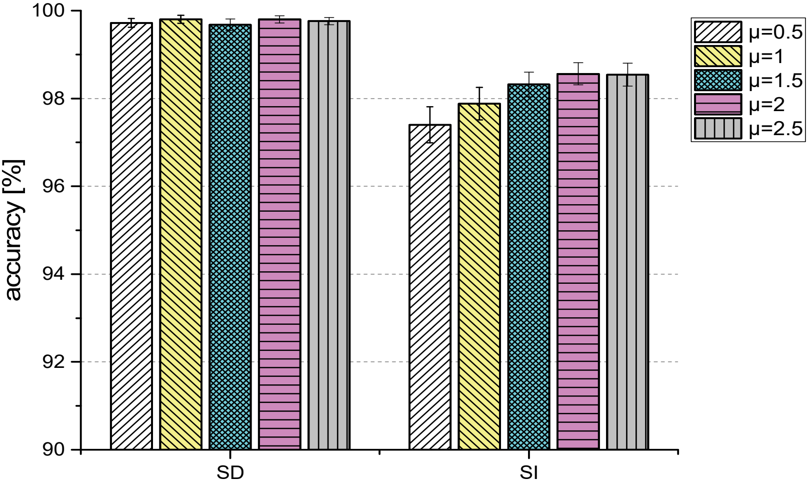


Fig. 8. The average word recognition accuracy with standard error (SE) in neutral-trained recognition of neutral speech (N/N) in speaker dependent (SD) and speaker independent (SI) case using μFCC (cepstral coefficients with μ-law frequency warping), with different values of parameter μ

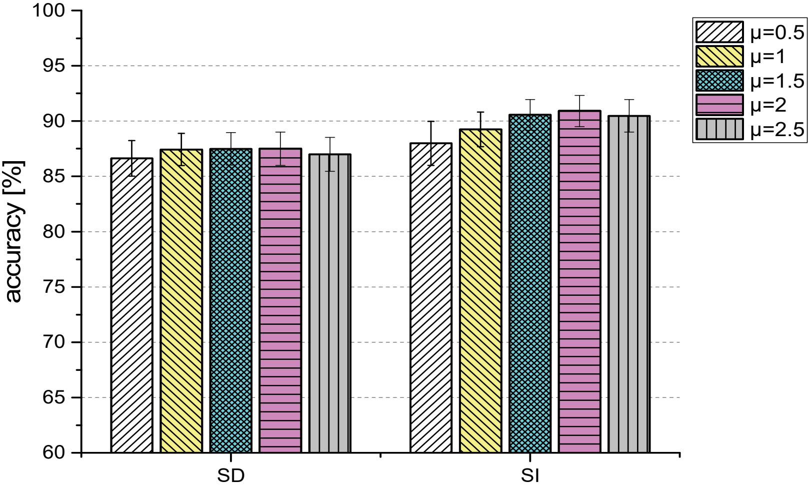


Fig. 9. The average word recognition accuracy with standard error (SE) in neutral-trained recognition of whisper (N/W) in speaker dependent (SD) and speaker independent (SI) case using μFCC (cepstral coefficients with μ-law frequency warping), with different values of parameter μ

In order to confirm the effectiveness of μFCC in mismatched scenario, statistical tests are needed. Two-tailed Wilcoxon signrank test show that improvement with proposed frequency warping is statistically significant. In Table 1, the average accuracy is given for recognition with μFCC ($\mu = 2$), as well as for feature vectors examined in the initial experiments. Belonging range of p-value is labeled with asterisks.

The tests show statistical significance of recognition using μFCC compared to recognition using all feature vectors analyzed in the initial experiments. Despite the fact that improvement in the SD case related to the accuracy for LFCC vectors is marginal (0.70%), Wilcoxon test shows high significance ($p<0.005$).

Table 1. Average Recognition Accuracy for Different Feature Vectors in Neutral-trained Recognition of Whispered Speech (N/W scenario)

Feature vector	Accuracy [%]	
	SD	SI
μFCC	87.50	90.92
MFCC	80.14^{**}	87.48^{**}
LFCC	86.80^{**}	86.50^{**}
PLP	73.07^{**}	83.16^{**}
LPLP	86.05^{**}	87.80^{**}
LPLP(mod)	83.38^{**}	88.86^{*}

(p<0.05 *; p<0.005 **; Confidence interval = 95%)

In order to evaluate how well particular ASR system performs, precision (P) and recall (R) are sometimes used as measures. The F-measure of the system, which is defined as weighted harmonic mean of its precision and recall, is given in the following [28]:

$$F = \left(\alpha \frac{1}{P} + (1-\alpha)\frac{1}{R} \right)^{-1}. \quad (5)$$

The balanced F-measure (commonly denoted as F_1) equally weights precision and recall (i.e., $\alpha = 0{,}5$), that is,

$$F_1 = \frac{2PR}{P+R}. \quad (6)$$

In N/W scenario, for each individual word, precision and recall are determined by analyzing HTK recognition output file. The balanced F-measure is averaged across words and calculated for each speaker.

The results are presented for MFCC and µFCC feature vectors in Table 2, in both the SD and SI cases. The speakers denoted as Speaker 1 up to Speaker 5 are female speakers whereas speakers denoted as Speaker 6 up to Speaker 10 are male speakers. Obtained results suggest that the ASR system which exploits µFCC feature vectors can achieve higher *F*-measure in whisper recognition, compared to conventional MFCCs. Average balanced *F*-measure for µFCC feature vectors is higher for 7% in the SD case and 3.5% in the SI case, approximately (the last raw in Table 2). As well as for accuracy, high deviation of *F*-measure between speakers is obtained.

Table 2. Average Balanced F-measure for all Speakers in Neutral-trained Recognition of Whispered Speech (N/W scenario)

Speaker	SD		SI	
	MFCC	µFCC	MFCC	µFCC
Speaker 1	0.9763	0.9695	0.9053	0.9453
Speaker 2	0.7629	0.8667	0.8630	0.8899
Speaker 3	0.8631	0.9093	0.8592	0.9186
Speaker 4	0.7324	0.8495	0.8449	0.9290
Speaker 5	0.7703	0.8227	0.7750	0.8199
Speaker 6	0.7274	0.8352	0.9444	0.9596
Speaker 7	0.8636	0.9119	0.9311	0.9438
Speaker 8	0.8273	0.8871	0.9130	0.9352
Speaker 9	0.8108	0.9152	0.9054	0.9497
Speaker 10	0.8560	0.9130	0.8895	0.8992
Average	0.8190	0.8880	0.8831	0.9190

7. Conclusion. This study has been motivated by insight in whispered speech recognition that modern ASR systems are not capable to handle when tested with whisper, due to high acoustic mismatch with neutral speech. Speech usually contains more energy at lower frequencies due to formant structure of vocals, while the whisper has relatively strong higher frequencies. Since traditional mel and bark frequency scales do not have good frequency resolution in high frequency range of speech, this study has investigated frequency warping schemes and compared performances for MFCC and LFCC as well as PLP and LPLP feature vectors, in both speaker dependent and speaker independent cases. The results in the initial experiments confirmed that filterbank resolution affects the recognizer performance in mismatched scenarios.

In order to find better frequency warping for neutral-trained whisper recognition, frequency warping based on μ-law compression mapping has been proposed. Conducted experiments have shown effectiveness of proposed warping: the best performances for both the SD and SI recognition have been obtained for the value of warping coefficients $\mu = 2$. This new approach in generation of feature vectors has some advantages in whisper recognition: (i) notably higher recognition accuracy is observed in mismatched scenario compared to the traditional speech parameterization; (ii) whisper data in training needed for model adaptation or multi-condition training is not prerequisite; and, (iii) feature vectors are easily obtained without increase in feature dimensionality. Compared to MFCC based recognition, robustness of recognizer with μFCC has been improved for 7.36% (SD case) and 3.44% (SI case).

Our current and future work aims to find more robust speech parameterization for neutral-trained whispered speech recognition. Because Teager energy cepstal coefficients show superiority over MFCC, combined effect of Teager operator and μ-law warping on HMM-based recognition of bimodal speech will be examined.

References

1. Zhang C., Hansen J.H.L. Analysis and classification of speech mode: whispered through shouted. Eighth Annual Conference of the International Speech Communication Association. 2007. pp. 2289–2292.
2. Ito T., Takeda K., Itakura F. Analysis and recognition of whispered speech. *Speech Communication*. 2005. vol. 45. no. 2. pp. 129–152.
3. Ghaffarzadegan S., Boril H., Hansen J.H.L. UT-VOCAL EFFORT II: Analysis and constrained-lexicon recognition of whispered speech. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2014. pp. 2544–2548.
4. Marković B., Jovičić S.T., Galić J., Grozdić Đ. Whispered speech database: Design, processing and application. International Conference on Text, Speech and Dialogue. 2013. pp. 591–598.
5. Lee P.X. et al. A whispered Mandarin corpus for speech technology applications. Fifteenth Annual Conference of the International Speech Communication Association. 2014. pp. 1598–1602.
6. Kozierski P. et al. Kaldi toolkit in Polish whispery speech recognition. *Przeglad Elektrotechniczny*. 2016. vol. 92. pp. 301–304.
7. Fan X., Hansen J.H.L. Speaker identification for whispered speech based on frequency warping and score competition. Ninth Annual Conference of the International Speech Communication Association. 2008. vol. 1. pp. 1313–1316.
8. Zhang C., Hansen J.H.L. Advancements in whisper-island detection using the linear predictive residual. 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP). 2010. pp. 5170–5173.

9. Zhang C., Hansen J.H.L. Whisper-island detection based on unsupervised segmentation with entropy-based speech feature processing. *IEEE Transactions on Audio Speech and Language Processing*. 2011. vol. 19. no. 4. pp. 883–894.
10. Ghaffarzadegan S., Bořil H., Hansen J.H.L. Model and feature based compensation for whispered speech recognition. Fifteenth Annual Conference of the International Speech Communication Association. 2014. pp. 2420–2424.
11. Ghaffarzadegan S., Bořil H., Hansen J.H.L. Generative modeling of pseudo-whisper for robust whispered speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processin*g. 2016. vol. 24. no. 10. pp. 1705–1720.
12. Grozdić Đ. et al. Comparison of cepstral normalization techniques in whispered speech recognition. *Advances in Electrical and Computer Engineering*. 2017. vol. 17. no. 1. pp. 21–26.
13. Grozdić Đ., Jovičić S.T. Whispered Speech Recognition Using Deep Denoising Autoencoder and Inverse Filtering. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 2017. vol. 25. no. 12. pp. 2313–2322.
14. Marković B., Galić J., Mijić M. Application of Teager Energy Operator on Linear and Mel Scales for Whispered Speech Recognition. *Archives of Acoustics*. 2018. vol. 43. no. 1. pp. 3–9.
15. Swerdlin Y., Smith J., Wolfe J. The effect of whisper and creak vocal mechanisms on vocal tract resonances. *The Journal of the Acoustical Society of America*. 2010. vol. 127. no. 4. pp. 2590–2598.
16. Tartter V.C. Identifiability of vowels and speakers from whispered syllables. *Perception & psychophysics*. 1991. vol. 49. no. 4. pp. 365–372.
17. Fan X., Hansen J.H.L. Speaker identification with whispered speech based on modified LFCC parameters and feature mapping. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2009). 2009. pp. 4553–4556.
18. Hermansky H. Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America*. 1990. vol. 87. no. 4. pp. 1738–1752.
19. Sklar B. Digital Communications: Fundamentals and Applications: 2nd edition. Prentice-Hall. 1988. 776 p.
20. Young S. et al. The HTK Book (for HTK Version 3.2). Cambridge University Engineering Department. 2006. 355 p. Available at: http://speech.ee.ntu.edu.tw/homework/DSP_HW2-1/htkbook.pdf (accessed: 17.04.2018).
21. Hermansky H., Morgan N. RASTA processing of speech. IEEE transactions on speech and audio processing. 1994. vol. 2. no. 4. pp. 578–589. Available at: https://labrosa.ee.columbia.edu/matlab/rastamat/ (accessed: 17.04.2018).
22. Galić J. et al. Speaker dependent recognition of whispered speech based on MLLR adaptation. Proc. of 11th Conference Digital Speech and Image Processing DOGS. 2017. pp. 29–32.
23. Marković B. et al. Recognition of Normal and Whispered Speech Based on RASTA Filtering and DTW Algorithm. Proceedings of the Int. Conf. IcETRAN-2017. 2017. pp. AK1.8.2–4.
24. Marković B., Jovičić S., Galić J., Grozdić Đ. Recognition of the Multimodal Speech Based on the GFCC features. Proceedings of the Int. Conf. IcETRAN-2015. 2015. pp. AK1 1.3 1–5.
25. Galić J., Jovičić S., Grozdić Đ., Marković B. HTK-Based Recognition of Whispered Speech. International Conference on Speech and Computer (SPECOM-2014). 2014. pp. 251–258.

26. Jakovljević N. An application of sparse representation in Gaussian mixture models used in speech recognition task. Ph.D. thesis. University of Novi Sad. 2013.
27. Fan X., Hansen J.H.L. Speaker identification within whispered speech audio stream. *IEEE Transactions on Audio, Speech and Language Processing*. 2011. vol. 19. no. 5. pp. 1408–1421.
28. Zhang E., Zhang Y. F-Measure. Encyclopedia of Database Systems. 2009. pp. 1147.

Galić Jovan Neđo — Ph.D. student of School of Electrical Engineering, University of Belgrade, assistant of the Faculty of Electrical Engineering, University of Banja Luka. Research interests: speech processing, speech enhancement, robust automatic speech recognition and compression of audio signals. The number of publications — 30. jovan.galic@etf.unibl.org, http://www.etf.unibl.org; 5, Patre, 78000, Banja Luka, Republic of Srpska, Bosnia and Herzegovina; office phone: +387-51-221-876.

Jovičić Slobodan Toma — Ph.D., Dr. Sci., professor of chair of Telecommunications of School of Electrical Engineering, University of Belgrade, head of laboratory for forensic acoustics and phonetics, Life Advancement Activities Center (Belgrade), scientific adviser for the speech signal processing and forensic speaker identification, Life Advancement Activities Center (Belgrade). Research interests: speech communications, man-machine communications, natural language processing, cognition and psychology of speech, speech enhancement, speech technologies. The number of publications — 300. jovicic@etf.rs; 73, Bul. Kralja Aleksandra, 11120, Belgrade, Serbia; office phone: +381-11-3218-361.

Delić Vlado Dragomir — Ph.D., Dr. Sci., professor, head of the Chair of telecommunications and signal processing of Department of power, electronic and telecommunications engineering of Faculty of technical sciences, University of Novi Sad, visiting professor of the Faculty of Electrical Engineering, University of Banja Luka. Research interests: speech technologies, audio signal processing. The number of publications — 300. vdelic@uns.ac.rs; 6, Trg Dositeja Obradovića, 21000, Novi Sad, Serbia; office phone: +381-21-485-2533.

Marković Branko Rade — Ph.D. student of School of Electrical Engineering, University of Belgrade, lecturer, Čačak Technical College. Research interests: speech recognition, multimodal speech, pattern's matching technology, microphones arrays, Internet technology and networking. The number of publications — 50. branko333@mts.rs, http://www.etf.bg.ac.rs; 65, Svetog Save, 32000, Čačak, Serbia; office phone: +381-32-322-321.

Šumarac Pavlović Dragana Staniša — Ph.D., Dr. Sci., professor of chair of Telecommunications of School of Electrical Engineering, University of Belgrade. Research interests: processing audio and speech signals, room acoustic design and sound filed modeling, building acoustics and noise control systems. The number of publications — 150. dsumarac@etf.rs, http://www.etf.bg.ac.rs; 73, Bul. Kralja Aleksandra, 11120, Belgrade, Serbia; office phone: +381-11-3218-361.

Grozdić Đorđe Tomislav — Ph.D., Dr. Sci., software engineer – data scientist, Fincore Ltd. Research interests: speech signal processing, automatic speech recognition, speaker identification. The number of publications — 50. djordje.grozdic@fincore.com; 7, Mutapova, 11000, Belgrade, Serbia; office phone: +381-62-8081-921.

Acknowledgements. This research is partially supported by the Ministry od Education, Science and Technological Development of Republic of Serbia (grants OI178027, TR32032 and TR32035, as well as E!9944).

УДК 004.5 **DOI** 10.15622/sp.58.2

Й.Н. Галич, С.Т. Йовичич, В.Д. Делич, Б.Р. Маркович,
Д.С. Шумарац Павлович, Г.Т. Гроздич

РАСПОЗНАВАНИЕ ШЕПОТНОЙ РЕЧИ С ИСПОЛЬЗОВАНИЕМ СММ И ЧАСТОТНОГО ПРЕОБРАЗОВАНИЯ ПО μ-ЗАКОНУ

Галич Й.Н., Йовичич С.Т., Делич В.Д., Маркович Б.Р., Шумарац Павлович Д.С., Гроздич Г.Т. **Распознавание шепотной речи с использованием СММ и частотного преобразования по μ-закону.**

Аннотация. Отсутствие достаточного количества данных шепотной речи для обучения является серьезной проблемой современных систем автоматического распознавания речи (АРР). Из-за большого акустического различия между обычной и шепотной речью АРР системы значительно снижают производительность при обработке шепота.

В статье приведен анализ подходов к распознаванию нейтральной и шепотной речи на основе традиционных скрытых марковских моделей (СММ) для дикторозависимых (SD) и дикторонезависимых (SI) случаев. Особое внимание уделяется распознаванию шепота с использованием нейтральной речи на этапе обучения (сценарий N/W). Система АРР разработана для распознавания изолированных слов из базы данных (Whi-Spe), включающей пары слов реально произнесенной речи нейтрально и шепотом. В сценарии N/W увеличение надежности достигается с применением предлагаемого частотного преобразования, изначально разработанного для сжатия и декомпрессии речевого сигнала в цифровых телекоммуникационных системах. Вместе с тем сохраняются хорошие показатели в распознавании нейтральной речи.

По сравнению с базовой моделью распознавания с применением Мел-частотных кепстральных коэффициентов (MFCC) точность распознавания слов с использованием кепстральных коэффициентов, полученных с помощью предложенного частотного деформирования (обозначаемого как μFCC), улучшена на 7,36% (SD) и 3,44% (SI) в абсолютных значениях. Кроме того, F-мера (гармоническое среднее значение точности и полноты) для векторов признаков μFCC увеличивается на 6,90% (SD) и 3,59 %(SI). Статистические тесты подтверждают значимость достигнутого улучшения точности распознавания.

Ключевые слова: автоматическое распознавание речи, извлечение признаков, скрытые марковские модели, человеческий голос, шепот, обработка речи.

Галич Йован Недьо — аспирант электротехнического факультета, Белградский университет, ассистент электротехнического факультета, Университет Баня-Лука. Область научных интересов: обработка речи, шумоочистка речи, робастное автоматическое распознавание речи, сжатие аудиосигнала. Число научных публикаций — 30. jovan.galic@etf.unibl.org, http://www.etf.unibl.org; Патре, 5, 78000, Баня-Лука, Республика Сербская, Босния и Герцеговина; р.т.: +387-51-221-876.

Йовичич Слободан Тома — д-р техн. наук, профессор кафедры телекоммуникаций электротехнического факультета, Белградский университет, заведующий лабораторией судебной акустики и фонетики, Центр улучшения жизни (Белград), научный консультант по обработке речевого сигнала и идентификации судебных носителей, Центр улучшения жизни (Белград). Область научных интересов: речевые коммуникации, человекомашинные коммуникации, обработка естественного языка, познание и психология речи, улучшение речи, речевые технологии. Число научных публикаций — 300. jovicic@etf.rs; Король Александар Бульвар, 73, 11120, Белград, Сербия; р.т.: +381-11-3218-361.

Делич Владо Драгомир — д-р техн. наук, профессор, заведующий кафедрой телекоммуникаций и обработки сигналов департамента энергетики, электроники и телекоммуникационного инжиниринга факультета технических наук, Нови-Садский университет, приглашенный профессор электротехнического факультета, Университет Баня-Лука. Область научных интересов: речевые технологии, обработка звуковой сигнал. Число научных публикаций — 300. vdelic@uns.ac.rs; Трг Доситейа Обрадовича, 6, 21000, Нови Сад, Сербия; р.т.: +381-21-485-2533.

Маркович Бранко Раде — аспирант электротехнического факультета, Белградский университет, преподаватель, Высшая техническая школа Чачак. Область научных интересов: распознавание речи, мультимодальная речь, распознавание образов, массив микрофонов, Интернет-технологии и сети. Число научных публикаций — 50. branko333@mts.rs, http://www.etf.bg.ac.rs; Светог Саве, 65, 32000, Чачак, Сербия; р.т.: +381-32-322-321.

Шумарац Павлович Драгана Станиша — д-р техн. наук, профессор кафедры телекоммуникаций электротехнического факультета, Белградский университет. Область научных интересов: обработка аудио и речевых сигналов, проектирование акустики помещений и моделирование распространения звука, построение систем управления акустикой и шумом. Число научных публикаций — 150. dsumarac@etf.rs, http://www.etf.bg.ac.rs; Король Александар Бульвар, 73, 11120, Белград, Сербия; р.т.: +381-11-3218-361.

Гроздич Георгий Томислав — д-р техн. наук, разработчик программного обеспечения – специалист по обработке и анализу данных, Fincore Ltd. Область научных интересов: обработка речевого сигнала, автоматическое распознавание речи, идентификация диктора. Число научных публикаций — 50. djordje.grozdic@fincore.com; Мутапова, 7, 11000, Белград Белград, Сербия; р.т.: +381-62-8081-921.

Поддержка исследования. Работа выполнена при частичной финансовой поддержке Министерства образования, науки и технологического развития Республики Сербия (гранты OI178027, TR32032 и TR32035, E!9944).

Литература

1. *Zhang C., Hansen J.H.L.* Analysis and classification of speech mode: whispered through shouted // Eighth Annual Conference of the International Speech Communication Association. 2007. pp. 2289–2292.
2. *Ito T., Takeda K., Itakura F.* Analysis and recognition of whispered speech // Speech Communication. 2005. vol. 45. no. 2. pp. 129–152.
3. *Ghaffarzadegan S., Boril H., Hansen J.H.L.* UT-VOCAL EFFORT II: Analysis and constrained-lexicon recognition of whispered speech // 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2014. pp. 2544–2548.

4. *Marković B., Jovičić S.T., Galić J., Grozdić Đ.* Whispered speech database: Design, processing and application // International Conference on Text, Speech and Dialogue. 2013. pp. 591–598.
5. *Lee P.X. et al.* A whispered Mandarin corpus for speech technology applications // Fifteenth Annual Conference of the International Speech Communication Association. 2014. pp. 1598–1602.
6. *Kozierski P. et al.* Kaldi toolkit in Polish whispery speech recognition // Przeglad Elektrotechniczny. 2016. vol. 92. pp. 301–304.
7. *Fan X., Hansen J.H.L.* Speaker identification for whispered speech based on frequency warping and score competition // Ninth Annual Conference of the International Speech Communication Association. 2008. vol. 1. pp. 1313–1316.
8. *Zhang C., Hansen J.H.L.* Advancements in whisper-island detection using the linear predictive residual // 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP). 2010. pp. 5170–5173.
9. *Zhang C., Hansen J.H.L.* Whisper-island detection based on unsupervised segmentation with entropy-based speech feature processing // IEEE Transactions on Audio Speech and Language Processing. 2011. vol. 19. no. 4. pp. 883–894.
10. *Ghaffarzadegan S., Bořil H., Hansen J.H.L.* Model and feature based compensation for whispered speech recognition // Fifteenth Annual Conference of the International Speech Communication Association. 2014. pp. 2420–2424.
11. *Ghaffarzadegan S., Bořil H., Hansen J.H.L.* Generative modeling of pseudo-whisper for robust whispered speech recognition // IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2016. vol. 24. no. 10. pp. 1705–1720.
12. *Grozdić Đ. et al.* Comparison of cepstral normalization techniques in whispered speech recognition // Advances in Electrical and Computer Engineering. 2017. vol. 17. no. 1. pp. 21–26.
13. *Grozdić Đ., Jovičić S.T.* Whispered Speech Recognition Using Deep Denoising Autoencoder and Inverse Filtering // IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2017. vol. 25. no. 12. pp. 2313–2322.
14. Marković B., Galić J., Mijić M. Application of Teager Energy Operator on Linear and Mel Scales for Whispered Speech Recognition // Archives of Acoustics. 2018. vol. 43. no. 1. pp. 3–9.
15. *Swerdlin Y., Smith J., Wolfe J.* The effect of whisper and creak vocal mechanisms on vocal tract resonances // The Journal of the Acoustical Society of America. 2010. vol. 127. no. 4. pp. 2590–2598.
16. *Tartter V.C.* Identifiability of vowels and speakers from whispered syllables // Perception & psychophysics. 1991. vol. 49. no. 4. pp. 365–372.
17. *Fan X., Hansen J.H.L.* Speaker identification with whispered speech based on modified LFCC parameters and feature mapping // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2009). 2009. pp. 4553–4556.
18. *Hermansky H.* Perceptual linear predictive (PLP) analysis of speech // The Journal of the Acoustical Society of America. 1990. vol. 87. no. 4. pp. 1738–1752.
19. *Sklar B.* Digital Communications: Fundamentals and Applications: 2nd edition // Prentice-Hall. 1988. 776 p.
20. *Young S. et al.* The HTK Book (for HTK Version 3.2). Cambridge University Engineering Department. 2006. 355 p. URL: http://speech.ee.ntu.edu.tw/homework/DSP_HW2-1/htkbook.pdf (accessed: 17.04.2018).

21. *Hermansky H., Morgan N.* RASTA processing of speech. IEEE transactions on speech and audio processing. 1994. vol. 2. no. 4. pp. 578–589. URL: https://labrosa.ee.columbia.edu/matlab/rastamat/ (дата обращения: 17.04.2018).
22. *Galić J. et al.* Speaker dependent recognition of whispered speech based on MLLR adaptation // Proc. of 11th Conference Digital Speech and Image Processing DOGS. 2017. pp. 29–32.
23. *Marković B. et al.* Recognition of Normal and Whispered Speech Based on RASTA Filtering and DTW Algorithm // Proceedings of the Int. Conf. IcETRAN-2017. 2017. pp. AK1.8.2–4.
24. *Marković B., Jovičić S., Galić J., Grozdić Đ.* Recognition of the Multimodal Speech Based on the GFCC features // Proceedings of the Int. Conf. IcETRAN-2015. 2015. pp. AK1 1.3 1–5.
25. *Galić J., Jovičić S., Grozdić Đ., Marković B.* HTK-Based Recognition of Whispered Speech // International Conference on Speech and Computer (SPECOM-2014). 2014. pp. 251–258.
26. *Jakovljević N.* An application of sparse representation in Gaussian mixture models used in speech recognition task // Ph.D. thesis. University of Novi Sad. 2013.
27. *Fan X., Hansen J.H.L.* Speaker identification within whispered speech audio stream // IEEE Transactions on Audio, Speech and Language Processing. 2011. vol. 19. no. 5. pp. 1408–1421.
28. *Zhang E., Zhang Y.* F-Measure // Encyclopedia of Database Systems. 2009. pp. 1147.