ITC 2/49

Information Technology and Control
Vol. 49 / No. 2 / 2020
pp. 224-236
DOI 10.5755/j01.itc.49.2.22258

One Solution of Extension of Mel-Frequency Cepstral Coefficients Feature Vector for Automatic Speaker Recognition

Received 2018/12/10

Accepted after revision 2020/05/06

http://dx.doi.org/10.5755/j01.itc.49.2.22258

HOW TO CITE: Jokić, I. D., Jokić, S. D., Delić, V. D., Perić, Z. H. (2020). One Solution of Extension of Mel-Frequency Cepstral Coefficients Feature Vector for Automatic Speaker Recognition. *Information Technology and Control*, 49(1), 224-236. https://doi.org/10.5755/j01.itc.49.2.22258

One Solution of Extension of Mel-Frequency Cepstral Coefficients Feature Vector for Automatic Speaker Recognition

Ivan D. Jokić

University of Novi Sad; Faculty of Technical Sciences; Trg Dositeja Obradovića 6 , 21000, Novi Sad, Serbia;
phone: +381 21 485 2521; e-mails: ivan.jokic@uns.ac.rs, vdelic@uns.ac.rs
Svezdrav rešenja; Đenerala Draže 44, 15357, Klenje, Serbia, e-mail: stevan@ecg4everybody.com

Stevan D. Jokić

Svezdrav rešenja; Đenerala Draže 44, 15357, Klenje, Serbia, e-mail: stevan@ecg4everybody.com
Faculty of Economics and Engineering Management in Novi Sad; Cvećarska 2, 21000, Novi Sad,
Serbia, e-mail: stevan.jokic@fimek.edu.rs

Vlado D. Delić

University of Novi Sad; Faculty of Technical Sciences; Trg Dositeja Obradovića 6 , 21000, Novi Sad, Serbia;
phone: +381 21 485 2521; e-mails: ivan.jokic@uns.ac.rs, vdelic@uns.ac.rs

Zoran H. Perić

University of Niš; Faculty of Electronic Engineering; Aleksandra Medvedeva 14, 18000, Niš, Serbia,
e-mail: zoran.peric@elfak.ni.ac.rs

Corresponding author: ivan.jokic@uns.ac.rs

One extension of mel-frequency cepstral feature vector for automatic speaker recognition is considered in this paper. The starting feature vector consisted of 18 mel-frequency cepstral coefficients (MFCCs). The extension was done with two additional features derived from the appropriate spectral maximums of the speech signal. The main idea behind this research is that it is possible to increase the accuracy of automatic speaker recognition which uses only MFCCs by adding additional features based on the energy maximums in the appropriate frequency ranges of observed speech frames. In the experiments, accuracy and equal error rate (EER) are compared in the case when feature vectors contain only MFCCs and in cases when additional features are used. For the case of maximum recognition accuracy achieved (92.94%), recognition accuracy increased by around 2.43%. EER values have smaller differentiation, but the results show that adding proposed additional features produced a lower decision threshold. These results indicate that tracking of proposed spectral maxima in the spectrum of the speech signal leads to more accurate automatic speaker recognizer. Determining features which track real maxima in the speech spectrum will improve the procedure of automatic speaker recognition and enable avoiding complex models.

KEYWORDS: Speaker recognition, spectrum, mel-frequency cepstral coefficients, energy, maximum.

1. Introduction

Mel-frequency cepstral coefficients (MFCCs) are introduced as features that can track the spectral envelope of the speech signal. These features are widely used as short-term speaker features [3, 21, 35, 10]. Spectral subband centroids (SSCs) are also used as features for speaker recognition [33, 24, 30, 22]. These features give the locations of local maxima of the power spectrum, the centroid frequencies of subbands. The concatenation of these features and MFCCs brings about better results in speaker recognition with respect to the case when only MFCCs used. To allow better adaptation to dynamic phenomena in speech, adaptive SSCs were proposed in [22]. The Normalized Dynamic Spectral Features (NDSF), proposed in [7], are found to be more robust than cepstral features. In addition, speaker verification combining MFCCs with the Spectral Dimension (SD) features, proposed in [5], enhances performance more than the method that is based only on MFCCs.

Speech data from the freely available CHAINS corpus [8] were used for the experiments in this paper. The speech parametrization algorithm [12], based on the AM-FM representation of the speech signal, was tested using speech data provided by the CHAINS corpus. Paper [13] presents an experimental evaluation of the effect of different speech styles on speaker identification and test of applicability of speech parameterization based on the *pyknogram frequency estimate coefficients – pykfec,* also by using CHAINS corpus.

Additional research on MFCCs features or some features derived from the spectrum is important because MFCCs are widely used features in voice applications or sound recognition, in general; MFCCs are used in application for speech recognition [14, 9, 1], emotion recognition from speech [28, 29, 2], but also for recognition of some other sounds [6, 4, 25, 27]. In addition, the exact determination of features based on the spectrum analysis can contribute to better speech synthesis or sound synthesis, in general, based on the harmonic generation [23]. The quality of this synthesizer or performance of any automatic recognizer of speech, speaker, emotion or sound depends on the quality of the input circuit of these devices; in fact, it depends on the quality of the quantizer used. Therefore, it is significant to examine the performance of the quantizer [32]. Research on the determination of speech features, or sound features, in general, based on the spectrum analysis, can also contribute to the construction of quantizers for sub-band coding of audio [36].

The determination of MFCCs is based on the application of discrete cosine transform on logarithm energies in the appropriate sub-bands of a signal, as represented in the following equation [37]:

$$c_n = \sum_{k=1}^{20} \log(E_k) \cdot \cos\left[n \cdot \left(k - \frac{1}{2} \right) \right], n = 1,2,\ldots,M. \quad (1)$$

This is a formula for the determination of M MFCCs and E_k is the energy inside appropriate k-th filter section, i.e. k-th sub-band. These sections are fixed in the mel scale. They are 300 mel wide and mutually shifted by 150 mel. By using equality between the mel and hertz scale:

$$f[mel] = 2595 \cdot \log_{10}\left(1 + \frac{f[Hz]}{700}\right), \quad (2)$$

boundaries of the appropriate sub-band in the mel scale can be recalculated in the hertz scale. For a known sampling frequency f_s and the number of points N of the discrete Fourier transform (DFT), the discrete frequency m of the component on the continuous frequency f can be determined from the equality $\frac{m}{N} = \frac{f}{f_s}$. If k-th filter section is in the range of discrete frequencies $m_1 \le m \le m_2$ and the square of amplitude characteristic of applied filter section is $A^2(m)$ then the energy E_k is determined as:

$$E_k = \sum_{m=m_1}^{m_2} X^2(m) \cdot A^2(m), \quad (3)$$

where $X(m)$ is the amplitude of DFT of the observed signal $x(n)$. These filter sections are introduced to simulate filtering inside auditory critical bands and masking phenomena. Masking phenomena depend on masking and masked spectral components. The experiments in [17] have shown that the recognizer which uses an exponential shape of the amplitude square of applied filter sections outperforms cases when triangular or rectangular shapes are used. This can be explained by the fact that an exponential function has a higher slope with respect to a linear function, which is why the exponential critical bands better describe masking than the triangular ones. At the same time, we can mention that, since the spectral bands are fixed, real masking was not taken into account in Equation (1). Therefore, the determination of MFCCs in Equation (1) in fact does not take into consideration the real perceived spectrum of signal [19], since the maximums of applied frequency selective filters in Equation (1) are not strictly positioned at the frequencies of real maximums in the spectrum. This fact justifies research on the features which would be the picture of the real perceived spectrum in the signal, i.e. which track the real spectral maxima in the signal.

Automatic speaker recognition based on the use of short-term features, such as MFCCs, implies the determination of a model for the appropriate speaker. This model should represent a compact picture of the speaker. Covariance matrix is a compact representation of energies in the appropriate components, i.e. dimensions and between dimensions. For the set of n feature vectors grouped in the matrix X, whose vector of mean values is μ, the appropriate covariance matrix is calculated by the equation:

$$\Sigma = \frac{1}{n-1} \cdot (X-\mu) \cdot (X-\mu)^{\mathrm{T}}. \quad (4)$$

In reality, a model depends on the sample analyzed; it is the matrix of feature vectors X in Equation (4). It would be ideal if the model had the property of wholeness. Wholeness implies that for any sample of the same speaker, i.e. the matrix X, the calculated model is the same. In fact, this is a very hard request for the model. Covariance matrix as the model of speaker, as well as most other models, depends on the statistics of the source which is modeled. Application of principal component analysis (PCA) increases compactness and wholeness of the model [15]. PCA and other transformations, which are based on the additional matrix calculation, require additional calculation. In that manner, as the number of samples increases, the execution time increases as well [38], which additionally slows down the application for automatic speaker recognition. Weighting of the elements of a model can also increase compactness and wholeness of the model [16]. The algorithm of weighting of the elements of a model is simpler than PCA, but, in fact, both PCA and weighting of the elements of a model depend on the observed sample (on the training set, to be more precise). As is evident from the previous mention of PCA and weighting of the elements of a model and also from the literature [34, 31, 20, 11], it is possible to use more complex models and in such a way tend to the most perfect decision making. However, the real reason why the model is not whole is in the features used. This is the basic idea of this paper: to find features with which similar efficiency of recognition will be achieved as in the case when MFCCs and more complex procedures of modeling and decision making are used. These features, MFCCs, are oriented to tracking statistics of the speaker voice and not tracking real and accurate reasons why the observed voice has certain properties. Real spectral maximums can be attenuated because of the descending amplitude characteristic of the filters used for MFCCs determination. Therefore, the model based on the use of usually determined MFCCs, as in Equation (1), does not represent the whole picture of the observed speaker. If we can achieve compactness and wholeness of a model, then our model will be more efficient. Application of PCA or some similar transformations, for example, can increase the efficiency of an algorithm for automatic speaker recognition. Efficiency of PCA or pondering of elements of the covariance matrix show that these transformations result in a more compact model. Such models are more desirable since they better catch the essential property of the object of modeling, i.e. they provide a better differentiation between models of different speakers. In this way, their property of wholeness can be increased as well. The fact that models derived from MFCCs can be more compact indicates that the used features can also be more compact; in fact, there is a free space for achieving more compact features. From the perspective of information theory, our models also contain a certain amount of irrelevant information. If we can suppress this irrelevant information, then the resultant model will be a more suitable representation of the appropriate speaker. The model is a direct consequence of the features determined. Therefore, the existence of algorithms which, applied on the model, contribute to better performance of the automatic speaker recognizer indicates that we do not determine essential features of the speaker of interest. Assume that we have a voice recording of a speaker. For this recording we can determine vectors of square of amplitude of the discrete Fourier transform. It is obvious that if we have two different recordings of the same speaker, the spectrums will also be different although the speaker is the same. Hearing, i.e. perceiving the same timbre is the consequence of some essential features that remained unchanged. Therefore, our target is to track real voice features, through proposed spectral maxima in this paper and find its essential features in this way. Recordings of the same speaker are similar from his or her own point of view. This similarity is a feature of the speaker. If we can accurately determine this feature, we can expect a more effective performance of the automatic speaker recognizer.

In the next chapter we will describe the speaker recognizer used and the way in which additional features are determined. After that results are represented, it is compared the case when only MFCCs are used as features and the case when proposed additional features are used.

2. Automatic Speaker Recognizer Used and Experimental Setup

The used automatic speaker recognizer was organized with the aim of achieving more efficient features than in the case when only MFCCs as features were used. The feature vector of 18 MFCCs was used as a basic feature vector. MFCCs were calculated by using Equation (1). It was used 20 filter sections (Figure 1), 300 mel wide and mutually shifted by 150 mel. This arrangement of filters was taken from [37]. The arrangement covers the spectral range from 0 to 3150 mel, i.e. from 0 to 11453 Hz. The square of the amplitude of applied filter sections is of exponential shape [17]:

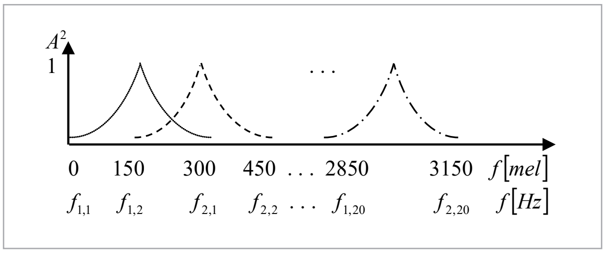
$$A(k)^2 = \begin{cases} e^{(k-k_{c,n})\cdot 2}, & k_{1,n} \le k \le k_{c,n}, \\ e^{-(k-k_{c,n})\cdot 2}, & k_{c,n} < k \le k_{2,n}. \end{cases} \quad (5)$$

$k_{c,n} = \frac{k_{1,n} + k_{2,n}}{2}$ is the central discrete frequency of n-th filter section, $k_{1,n}$ is the lower and $k_{2,n}$ is the higher discrete frequency of n-th filter section.

The speaker recognizer with these characteristics, 18 MFCCs and 20 filter sections with the exponential amplitude square characteristic, proved to be efficient in our previous experiments [17, 18]. Therefore, this recognizer was the starting point for further experiments, described in this paper, towards achieving more efficient feature vector by adding additional features. The model used is determined by Equation (4). The matrix X in Equation (4) for the set of n feature vectors is formed in such a way that the first feature vector is located in the first column of the matrix X, and so on. For example, the i-th feature vector is located in the i-th column of the matrix X. The measure of the difference between models, Equation (6), and

Figure 1

Arrangement of 20 applied filter sections



decision making were not changed compared to our earlier experiments. The difference between the two models was determined with the equation:

$$d\left(\Sigma_i,\Sigma_{ref}\right)=\frac{1}{n_f^2}\cdot\sum_{m=1}^{n_f}\sum_{n=1}^{n_f}\left|\Sigma_i\left(m,n\right)-\Sigma_{ref}\left(m,n\right)\right|, \qquad (6)$$

where n_f represents the number of features used. Two tests were performed, the test of identification and the test of verification. Testing of the algorithm was conducted on the Solo part of the publicly available speech database CHAracterizing INdividual Speakers (CHAINS) [8]. The Solo part is characterized by speaking style: subjects simply read a prepared text at a comfortable rate. The used part of the CHAINS speaker database contains recordings of 36 speakers. 28 speakers speak the same dialect - 12 females and 16 males from the Eastern part of Ireland. 1 female and 2 males are from the United Kingdom, whereas 3 females and 2 males are from the USA. Each of the speakers has 37 recordings. Four of these 37 recordings for each speaker represent longer recordings of short fables, whose duration is between around half a minute to approximately one minute. The titles of these recordings contain labels: f01, f02, f03, f04. The remaining 33 recordings of 33 individual sentences, whose names contain labels s01, s02, ..., s33, are shorter in duration, around two to three seconds. These 33 recordings were used in the speaker recognition experiments described in this paper. The recordings are in wav format, their sampling rate is 44100 Hz and the quantization is 16 bit PCM.

In the initial test of identification for each of the 36 speakers, one of the recordings, marked with s15, was used for training. Each of the 33 recordings of each speaker represents the speaker's voice through the pronunciation of one predefined sentence. Speaker's voice gives information about identity of speaker, therefore in all of 33 recordings of one speaker is hidden information about identity of that speaker. Speaker's identity is a constant which is searched in the process of automatic speaker recognition. Since all of 33 recordings of one speaker contain this constant, it is sufficient and necessary to do training by using one of these 33 recordings. In the practical implementation of the automatic speaker recognizer, efficiency depends on the features used and the model determined. MFCCs based on a short-term spectral analysis depend on the sample, i.e. the recording being analyzed, which is why they do not directly point to the searched constant in the speaker's voice. MFCCs depend on the text pronounced in recording. Therefore, the efficiency of the recognizer depends on balance, i.e. congruence between test and training recordings. Having in mind this property of MFCCs, it is necessary to use phonetically richer training recordings. In accordance with this, our tendency is to develop an algorithm which will be able to recognize the constant in the speaker's voice. Being aware that our recognizer is based on the use of MFCCs, which depend on the spoken text, we used the recording s15 for training since the sentence pronounced in this recording is one of the longest. Thus, it can be expected that training based on s15 recording will result in better accuracy of the recognizer in comparison with most of the cases when one of the other 32 recordings was used for training. The sentence pronounced in the recordings marked with s15 is: "Each untimely income loss coincided with the breakdown of a heating system part". During testing, the models of other recordings that are not used in training were observed and compared with 36 reference models. Identity of the most similar reference model in terms of Equation (6) was attributed to the analyzed test recording. In the initial test of verification, one of the short recordings of every speaker, marked with s15, was chosen as a training recording for creating one reference model. Other recordings were employed for the appropriate value of the decision threshold during determinations of the probability of false rejection and false acceptance. Decision threshold was varied to get equal error rate (EER), the case when the probability of false rejection is equal to the probability of false acceptance.

It is possible to derive features for an efficient automatic speaker recognition from the speech spectrum. Therefore, the impact of two additional features whose calculation is based on the energy spectrum of the signal analyzed is examined in this paper. Since the sampling frequency in the CHAINS database is 44100 Hz, speech frames of N=1024 samples were analyzed, the duration of which was around 23.2 ms. They were mutually shifted by 368 samples, whose duration was around 8.3 ms. Experimental setup was oriented to examining how additional features based on the signal energy influence the accuracy of the previously adopted recognizer [17, 18] which is based on the use of first 18 MFCCs as features. Feature vectors were extended by two additional features. We tracked the impact of these features on recognition accuracy in the test of identification and the impact on the equal error rate in the test of verification. In what follows, the first additional feature is denoted by e_1 and the second additional feature by e_2.

The additional features were determined based on observing maximum spectral values in the appropriate spectral ranges. Since the amplitude spectrum is a symmetric function of frequency, only the first $N/2$ coefficients of the discrete Fourier transform (DFT) were analyzed. The DFT was analyzed in N=1024 points. For reasons of symmetry, the amplitude spectrum ranging from 0 to 511 was analyzed. In the initial experiments, all DFT coefficients in the range of normalized frequencies from k=0 to k=511 were observed in order to calculate the additional feature e_1. The spectral maximum was searched in that range, but this gave poor recognition accuracy. The information about the speaker identity is contained in higher spectral components, i.e. in higher harmonics [26]. Therefore, the lower boundary for calculation of the additional feature e_1 was raised. After several repeated experiments of recognition, based on the best achieved recognition accuracy, the range of normalized frequencies from k=25 to k=511 was observed for e_1 feature. The amplitude maximum of DFT was searched in this range. The natural logarithm of square of the module of the discrete Fourier transform coefficient (KMDFT) was considered and its maximum value was searched. Since fullness in the perception of a sound is the consequence of DFT components which are in the nearest neighborhood of the maximum DFT coefficient, natural logarithms of energies of two DFT coefficients in immediate surroundings of the maximum were also considered for determining the e_1 feature. The calculation of the additional feature e_1 was done in two steps. In the first step, the summation of the natural logarithm of the maximal value of KMDFT in the range of normalized frequencies from k=25 to k=511, max_1, and the natural logarithm values of KMDFT of two coefficients in immediate surroundings was determined. If the maximal value max_1 is determined for the normalized frequency k_1, then this part of the algorithm can be described by the equation:

$$E_{1\ln} = \ln(\max{}_1) + \sum_{\substack{i=-2 \\ i\neq 0}}^{2} \ln(kmdft(k_1 + i)), \qquad (7)$$

where max_1 is the square of the module of maximal DFT coefficient in the range from k=25 to k=511 in the observed frame. Finally, the additional feature e_1 was calculated by weighting of its value calculated in the previous step with respect to a maximal component of KMDFT in all frames, in the range $25 \leq k \leq 511$, and by normalization with the discrete frequency of max_1:

$$e_1 = \frac{E_{1\ln} \cdot \dfrac{E_{1\ln}}{\ln(\max KMDFT_{all})}}{k_1}, \qquad (8)$$

where E_{1ln} is the value determined in the first step, $maxKMDFT_{all}$ corresponds to maximal KMDFT in the range $25 \ \leq k \leq 511$ in all frames of the observed signal and k_1 is the discrete frequency of max_1.

The second additional feature is observed in the spectral range: $k_1 + 10 \ < k < \dfrac{N=1024}{2}$, defined in relation to a position, i.e. the discrete frequency k_1 of the first additional feature. The additional feature e_2 is determined in a similar way to the first additional feature e_1, Figure 2. First, the maximum of KMDFT, max_2, in the previously defined spectral range was determined and the following summation was calculated:

$$E_{2\ln} = \ln(\max{}_2) + \sum_{\substack{i=-2 \\ i\neq 0}}^{2} \ln(kmdft(k_2 + i)), \qquad (9)$$

where max_2=$kmdft(k_2)$. After that, the final value of the second additional feature e_2 was calculated by weighting with respect to the natural logarithm of maximal KMDFT in all frames, in the range

$25 \leq k \leq 511$, of the observed recording, ln(*maxKMDFFT$_{all}$*), and normalized by the discrete frequency of *max$_2$*, *k$_2$*:

$$e_2 = \frac{E_{2h} \cdot \dfrac{E_{2h}}{\ln(\max KMDFT_{all})}}{k_2}. \tag{10}$$

Focusing on features which will be used increases efficiency of the system which uses these features.

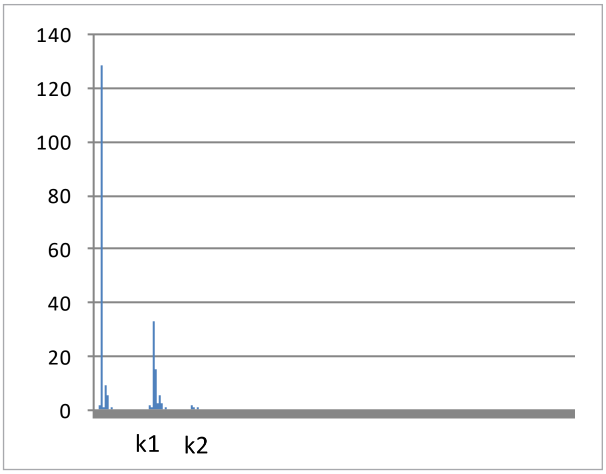


Figure 2

Illustration of typical form of the KMDFT from k=0 to k=511 with marked k_1=62 of max$_1$ and k_2=102 of max$_2$ corresponding to one speech frame, additional features good follow spectral maximums in signal. Illustration is given for the eighth frame of the signal frf01_s15_solo.wav

If we observe the equation for calculating MFCCs (Equation (1) from the introduction chapter) then we can mention that, from a geometrical point of view, the transformation used in this equation can be considered as some kind of a template with parameters. Speech signal is a dynamic signal. Spectral components are variable in time. The problem lies in the fact that the template is static. The discrete cosine transformation used in Equation (1) for the calculation of MFCCs has the appropriate parameters, indexes n and k and the parameters which describe selective filters used. However, bearing in mind possible properties of the speech spectrum, this equation acts as a template. Since the selective filters used in Equation (1) have fixed central frequencies and the fixed width, they cannot simulate real masking that really occurs in the signal. The width of selective filters is constant in the mel scale, but its changeability in frequency hertz domain and non-accurate position can lead to the wrong interpretation of spectral components. Their descending shape can lead to attenuation of wrong components which are not masked in reality. In fact, this is the consequence of the fact that we do not observe real masking and masked components in the spectrum of the observed speech. The approach in this paper, which uses a maximal component and two components in the nearest maximum environment, actually takes into account masking phenomena.

Twenty filters wide 300 mel, mutually shifted by 150 mel, cover the spectral range of 3150 mel. Using equality between the mel scale and the hertz scale stated in the introductory part of this paper (Equation (2)), the range is around 11453 Hz. Recordings in the CHAINS speech database were recorded with the sampling frequency of 44100 Hz, therefore their spectral range is from 0 to 22050 Hz, i.e. 3923 mel. Thus, before presenting the results when the additional features were added to a basic feature vector, we will report the results of recognition in the case when just one recording (s15) was used for training of each speaker and when the number of MFCCs and frequency selective filters (300 mel wide and mutually shifted by 150 mel) was varied in the available range, the number of filters can be increased to 25. We will compare more combinations by varying the number of MFCCs and the number of filters used. The best recognition accuracy of 87.41% was achieved in the configuration of 22 MFCCs and 22 filters used. Table 1 provides an overview of the best recognition results.

Table 1

Overview of comparisons of recognizer configurations in speaker identification for higher values of accuracy with respect to configuration 18 MFCCs, 20 filters

Number of MFCCs, number of filters	Accuracy [%]
18 MFCCs, 20 filters	84.03
18 MFCCs, 21 filters	85.33
20 MFCCs, 24 filters	86.28
21 MFCCs, 21 filters	87.33
21 MFCCs, 22 filters	87.33
22 MFCCs, 22 filters	87.41

It is evident that in this case a broader spectral range gives better results. For this reason, in the next chapter we will also examine the influence of additional features with respect to the configuration with the best accuracy.

3. Results

The purpose of the experiments in this paper is to examine if it is possible to achieve better performance of a speaker recognizer by using some additional features derived from energy in the signal. The performance was compared in two different situations: when the MFCCs feature vectors were used (18 MFCCs, i.e. 22 MFCCs) and when additional features, e_1 from Equation (8) and e_2 from Equation (10), were used. Two tests were conducted, the test of speaker identification and the test of speaker verification. The speaker identification test included 1152 tests in which only the recording s15 was used for training. Therefore, the results of identification accuracy were given also relative to this summary number of tests. In the speaker identification, when only 18 MFCCs were used, accuracy was 968/1152 (84.03%), Table 2. When 18 MFCCs + e_1 were used, accuracy was 999/1152 (86.72%). Accuracy for a twenty-dimensional feature vector consisting of 18 MFCCs + e_1 + e_2 was 1007/1152 (87.41%). The same accuracy was achieved by 22 MFCCs. The best accuracy in the case when only s15 was used for training was achieved for the feature vector 22 MFCCs + e_1 + e_2 – 1018/1152 (88.37%).

The probability of false rejection (PFR) and the probability of false acceptance (PFA) were tested in the test of verification. For determining both of these probabilities for predefined training set, appropriate recordings were observed in the test. During determining PFR for each speaker, testing was done on his or her recordings which were not used for training. On the other hand, in order to determine PFA for each speaker, testing was done on the recordings from other speakers. If n_t is the number of recordings for each of the 36 speakers used for training, then we have (33-n_t)*36 tests of false rejection and 35*33*36 tests of false acceptance. PFR is determined as the ratio: *number of false rejected*/((33-n_t)*36). PFA is determined as the ratio: *number of false accepted*/(35*33*36). These probabilities depend on the threshold. The threshold value was varied to give curves of false rejection and false acceptance. The intersection point of these curves represents EER. In a practical estimate of EER, when PFR and PFA were close in value, for EER we adopted the higher of these probabilities. For example, as can be seen in Table 2, when for the feature vector 18MFCCs+e_1 and threshold τ=0.995, *PFR*≅14.41% and *PFA*≅14.17%, the higher value is taken for EER, *EER*≅14.41%. In this way we do not make a big error, with respect to the ideal case when PFR=PFA=EER, but the algorithm gives the result faster. When the feature vector contains only 18 MFCCs, equal error rate is around 14.42% for the threshold value τ=1.025. For the feature vector which contains 18 MFCCs and the first additional feature (18 MFCCs+e_1), EER is around 14.41% for τ=0.995 and when the feature vector is 18 MFCCs+e_1+e_2, EER is around 15.54% for τ=0.965. Equal error rate shows small oscillations around 15%, but the threshold values show a descending tendency. As we can see in Table 2, when the feature vector is based on 22 MFCCs, EER is somewhat lower and the threshold values also have a descending tendency.

Based on this, it follows that compactness of models was increased by adding additional features. These results prove that efficiency of an automatic speaker recognizer that uses MFCCs as features can be increased only by enhancing features which are used, by using additional features derived from the energy spectrum of speech.

Table 2

Results of speaker recognition, only s15 used in training

	Identification	Verification	
Feature vector	Accuracy[%]	EER[%]	τ
18 MFCCs	84.03	14.42	1.025
18 MFCCs+e_1	86.72	14.41	0.995
18 MFCCs+e_2	86.81	14.48	0.983
18 MFCCs+e_1+e_2	87.41	15.54	0.965
22 MFCCs	87.41	12.93	1.183
22 MFCCs+e_1	88.02	13.37	1.1545
22 MFCCs+e_2	88.19	13.23	1.1415
22 MFCCs+e_1+e_2	88.37	13.95	1.1195

Training only on one recording is too rigorous a condition for the speaker recognizer based on short-term features such as MFCCs, therefore, the accuracy is relatively low, lower than 90% (and also lower compared to our previous publications [17, 18]). In order to confirm the applicability of the procedure, to achieve results of speaker identification higher than 90%, we used an increased number of recordings for training each speaker. Due to the nature of short-term features, training must be done on a larger number of recordings. The experiments were repeated for 9 chosen recordings for each speaker used for training. This group of recordings is denoted by Γ9. Γ9 contains the following recordings: s15, s33, s01, s07, s05, s31, s18, s20, s25. In this case the number of speaker identification tests is 36*(33-9) = 864.

Table 3

Results of speaker recognition, Γ9 used in training

	Identification	Verification	
Feature vector	Accuracy[%]	EER [%]	τ
18 MFCCs	89.58	17.36	0.9526
18 MFCCs + e_1	90.51	16.78	0.9326
18 MFCCs + e_2	90.28	16.32	0.9155
18 MFCCs + e_1+e_2	92.36	16.2	0.9
22 MFCCs	90.51	13.426	1.083
22 MFCCs+e1	91.67	12.96	1.0625
22 MFCCs+e2	91.55	12.55	1.0445
22 MFCCs+e1+e2	92.94	12.96	1.0295

An increase in the training set from only s15 recording to Γ9 set of nine recordings results in increased accuracy. Depending on the feature set used, recognition accuracy varies from 774/864 (89.58%) for the feature vector of 18 MFFCs to 803/864 (92.94%) for the feature vector 22 MFCCs+e_1+e_2. The data in Table 2 and Table 3 show an obvious increase in recognition accuracy in the case when additional features are added. In the case when recognition accuracy is the greatest, 92.94% (Table 3), accuracy growth is 2.43%, between the case when 22 MFCCs are used as features and the case when the feature vector is extended as 22 MFCCs+e_1+e_2. As is evident from Table 3, the extension of the feature vector from 18 MFCCs to 18 MFCCs+e_1+e_2 increases recognition accuracy by 2.78%. The additional features e_1 and e_2 also contribute in approaching each other accuracies of speaker recognizers when the feature vectors of 18 MFCCs+e_1+e_2 and 22 MFCCs+e_1+e_2 are used. Also, by comparing accuracy values for the feature vector of 18 MFCCs (89.58%) and the feature vector of 22 MFCCs (90.51%), it can be mentioned that the benefit from the additional features e_1 and e_2, 2.43% for 22 MFCCs feature vector, i.e. 2.78% for 18 MFCCs feature vector, is higher than the benefit from 19th, 20th, 21st and 22nd MFCC (90.51% - 89.58% = 0.93%).

Variation of features within one speaker is the reason for incorrect speaker recognition. The constant in the speaker's voice cannot be determined because of this variation. Therefore, we will present examples which demonstrate how the vectors MFCCs+e_1+e_2 vary within and between speakers. The determination of these variations is done for two signals of the same speaker and different speakers as well. Since for each of the speakers we have one recording of one text, we chose recordings of the same textual content for within and between speaker variability determination. These are the recordings denoted by s20 and s21. The chosen speakers are irm01 and irm16, Figure 3. Summary variation for each feature was determined as the summation of absolute values of difference for each of the adjacent frames normalized by the number of frames:

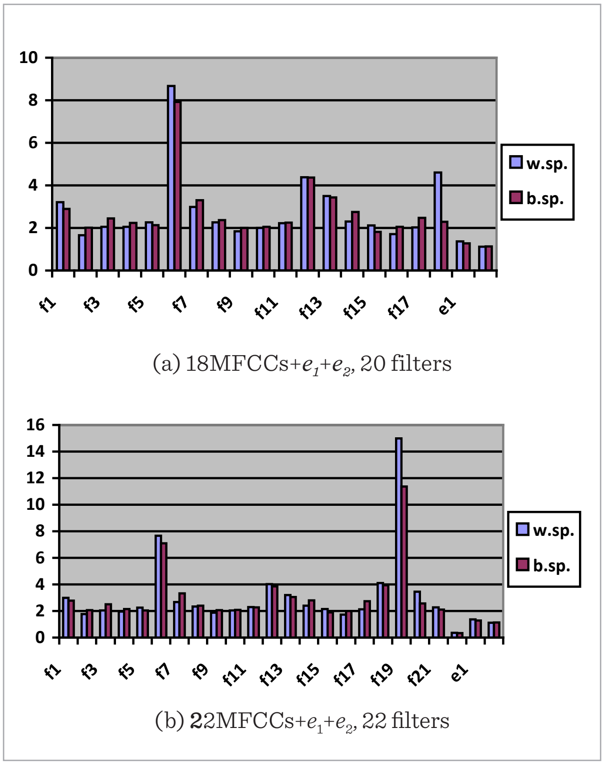
$$diff\left(f_i\right)=\frac{\sum_{n=1}^{N}\left|f_i^{(1)}\left(n\right)-f_i^{(2)}\left(n\right)\right|}{N}, \qquad (11)$$

where f_i is the observed feature, $i\in\{MFCC_1, MFCC_2,...,MFCC_{22},e_1,e_2\}$, $f_i^{(1)}(n)$ and $f_i^{(2)}(n)$ are the i^{th} features of n^{th} frame in the first and second observed signal, the observed signals are reduced to the same number of frames N, $N=\lfloor num.\_of\_frames_1, num.\_of\_frames_2\rfloor$, *num._of_frames*$_1$ and *num._of_frames*$_2$ represent the number of frames in the first and second observed signal.

The example of variation within speakers, Figure 3 (w.sp.), is determined for signals belonging to the speaker irm01. These are the recordings marked with s20: "The frightened child was gently subdued by his big brother" and s21: "The tooth fairy forgot to come when Roger's tooth fell out". The example of variations between speakers showed in Figure 3 (b.sp.) is determined for the speakers irm01 and irm16, also for the signals s20 and s21. Except for variation for $MFCC_{22}$, variations for the additional features e_1 and e_2 are the smallest. Variations of $MFCC_6$, $MFCC_{12}$, $MFCC_{18}$ and $MFCC_{19}$ have the biggest values. Variations of the additional features e_1 and e_2 are very similar for within speaker and between speaker variations.

Figure 3

Examples how the vector MFCCs+e1+e2 varies within (w.sp. – observed signals: irm01_s20_solo.wav and irm01_s21_solo.wav) and between (b.sp. – observed signals: irm01_s20_solo.wav and irm16_s21_solo.wav) speakers



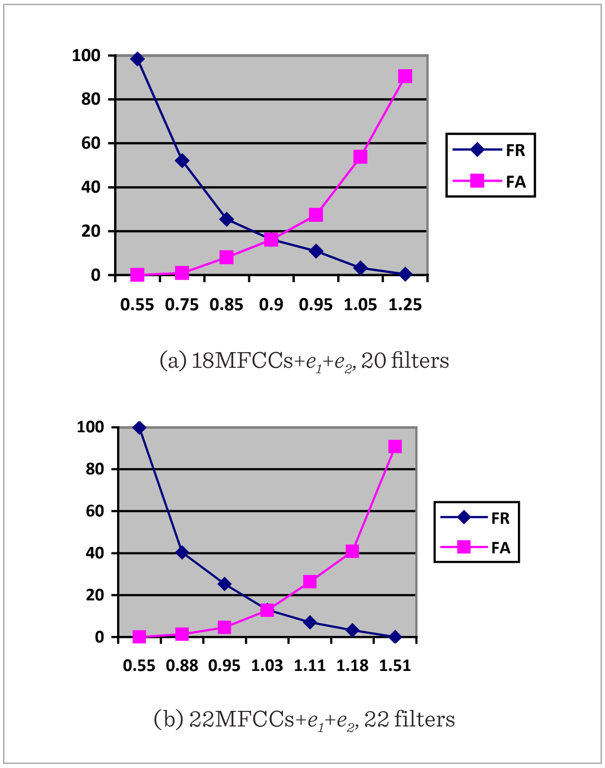
(a) 18MFCCs+e_1+e_2, 20 filters

(b) 22MFCCs+e_1+e_2, 22 filters

EER for the recognizer which uses 22 MFCCs+e_1+e_2 is below 13%, Table 3. It is also evident that when training is done with Γ9, an extension of feature vectors with the additional features e_1 and e_2 contributes to both reduction of EER and reduction of the decision threshold. Graphical examples of the curves of false rejection and false acceptance depending on the threshold value are presented in Figure 4.

Figure 4

Examples of the curves of false rejection and false acceptance depending on the threshold value, Γ9 used for training



(a) 18MFCCs+e_1+e_2, 20 filters

(b) 22MFCCs+e_1+e_2, 22 filters

A higher spectral band contains the information about the speaker [26]. This information is important for distinguishing between different speakers. By using the template in Equation (1) we do not appropriately consider higher spectral components. Thus, we lose accurate and precise information about the speaker. This problem can be viewed as analogue to the problem of uniform quantization and the reason why we use non-uniform quantization. In quantization, if we want to catch a lot of information we must introduce a larger number of quantization levels with respect to the case when we have smaller pieces of information of interest. A similar effect can be achieved by finding spectral maximums and calculating the additional features e_1 and e_2. These additional features enable us to accurately and precisely target the appropriate maximums of energy concentrations, which are in fact properties of the observed speaker's voice responsible for the timbre of his or her voice.

In addition, the previously mentioned research studies [12-13], in which the experiments were also performed over the CHAINS speech database, justify the view that it is possible to develop speech features from the spectrum which will be more efficient than commonly used MFCCs. Therefore, the results of these studies, as well as the findings presented in this paper, show that there is a space where we can find more efficient features derived from the spectrum.

4. Conclusion

The results of this paper prove the fact: As the object becomes clearer, i.e. described with more characteristic details, its recognition becomes greater. Introducing the additional features e_1 and e_2, the characteristic details in energy spectrum which describe local energy maximums in higher spectral components, it is increased clearness of the used models and thus improved accuracy of the used speaker recognizer. MFCCs contain information about the speaker, but the results presented in this paper show that this information about the speaker identity can be clearer if we add the additional features e_1 and e_2. The calculation of MFCCs does not take into account real masking in speech signal. Therefore, we lose the clarity of the used features. Based on the results of this study, it can be concluded that tracking of local maximums in spectrum can additionally improve accuracy of the used speaker recognizer. The explanation for this lies in the fact that local maximums in the spectrum are real places of masking in the spectrum. In that manner we can only keep on determining as accurate features as possible and the model can stay covariance matrix.

To achieve clearness of the features and consequently of the applied model we must track real characteristic properties in the observed speech signal. Each of the features in the appropriate feature vector should be determined depending on property i.e. characteristic of the speaker we want to catch. Since the information about speaker identity is contained in the spectrum of his or her voice, it is necessary to calculate features based on a detailed analysis of the energy spectrum. One solution for the features can be tracking spectral maximums and concentration of energy around these maximums in the appropriate spectral ranges. In this way we can be sure that we will catch real features from speech. The first step would be determination of the global spectral maximum in the observed speech recording. In the next step we would determine other local spectral maximums. Apart from tracking real masking in surroundings of the appropriate local maximum, we can also track the real spectral envelope and timbre as well.

Acknowledgement

Research done in this paper received funding from TR32035, TR32032, DANSPLAT and CABUNS projects.

References

1. Arora, S. V. Effect of Time Derivatives of MFCC Features on HMM Based Speech Recognition System. International Journal on Signal and Image Processing, 2013, 4(3), 50-55.
2. Attabi, Y., Alam, M. J., Dumouchel, P., Kenny, P., O'Shaughnessy, D. Multiple Windowed Spectral Features for Emotion Recognition. Published in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, BC, May 26-31, 2013, 7527-7531. https://doi.org/10.1109/ICASSP.2013.6639126
3. Bimbot, F., Bonastre, J.-F., Fredouille, C., Gravier, G., Margin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-Garcia, J., Petrovska-Delacrétaz, D., Reynolds, D. A. A Tutorial on Text-Independent Speaker Verification. EURASIP Journal on Applied Signal Processing, 2004, 2004(4), 430-451. https://doi.org/10.1155/S1110865704310024
4. Cai, J., Ee, D., Pham, B., Roe, P., Zhang, J. Sensor Network for the Monitoring of Ecosystem: Bird Species Recognition. Proc. 2007 3rd International Conference on Intelligent Sensors, Sensor Networks and Information, Melbourne, Queensland, Australia, December 3-6, 2007, 293-298. URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4496859&isnumber=4496790 https://doi.org/10.1109/ISSNIP.2007.4496859
5. Chen, W.-S., Huang, J.-F. Speaker Recognition with Spectral Dimension Features of Human Voices for Personal Authentication. Journal of Network Communications and Emerging Technologies (JNCET), 2015, 5(3), 6-11.

6. Chou, C.-H., Ko, H.-Y. Automatic Birdsong Recognition with MFCC Based Syllable Feature Extraction. In Hsu CH., Yang L.T., Ma J., Zhu C. (Eds.), Ubiquitous Intelligence and Computing. UIC 2011. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2011, 6905, 185-196. https://doi.org/10.1007/978-3-642-23641-9_17

7. Chougule, S. V., Chavan, M. S. Robust Spectral Features for Automatic Speaker Recognition in Mismatch Condition. Procedia Computer Science, 2015, 58, 272-279. https://doi.org/10.1016/j.procs.2015.08.021

8. Cummins, F., Grimaldi, M., Leonard, T., Simko, J. The CHAINS Corpus: CHAracterizing INdividual Speakers. In Proceedings of the 11th International Conference «Speech and Computer» SPECOM'2006, St. Petersburg, Russia, June 25-29, 2006, 431-435.

9. Dhingra, S. D., Nijhawan, G., Pandit, P. Isolated Speech Recognition Using MFCC and DTW. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 2013, 2(8), 4085-4092.

10. Dobrović, M. M., Delić, V. D., Jakovljević, N. M., Jokić, I. D. Comparison of the Automatic Speaker Recognition Performance Over Standard Features. In Proceedings of the 2012 IEEE 10th Jubilee International Symposium on Intelligent Systems and Informatics (SISY 2012), Subotica, Serbia, September 20-22, 2012, 341-344. https://doi.org/10.1109/SISY.2012.6339541

11. Ghahabi, O., Hernando, J. Deep Belief Networks for I-Vector based Speaker Recognition. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, May 4-9, 2014, 1700-1704. https://doi.org/10.1109/ICASSP.2014.6853888

12. Grimaldi, M., Cummins, F. Speaker Identification Using Instantaneous Frequencies. IEEE Transactions on Audio, Speech, and Language Processing, 2008, 16(6), 1097-1111. https://doi.org/10.1109/TASL.2008.2001109

13. Grimaldi, M., Cummins, F. Speech Style and Speaker Recognition: A Case Study. Proceedings of the 10th Annual Conference of the International Speech Communication Association, INTERSPEECH 2009, Brighton, United Kingdom, September 6-10, 2009, 920-923.

14. Ittichaichareon, C., Suksri, S., Yingthawornsuk, T. Speech Recognition Using MFCC. In Proceedings of International Conference on Computer Graphics, Simulation and Modeling (ICGSM'2012), Pattaya, Thailand, July 28-29, 2012, 135-138.

15. Jokic, I., Jokic, S., Peric, Z., Gnjatovic, M., Delic, V. Influence of the Number of Principal Components used to the Automatic Speaker Recognition Accuracy. Elektronika ir Elektrotechnika, 2012, 123(7), 83-86.https://doi.org/10.5755/j01.eee.123.7.2379

16. Jokic, I. D., Jokic, S. D., Delic, V. D., Peric, Z. H. Towards a Small Intra-Speaker Variability Models. Elektronika ir Elektrotechnika, 2014, 20(6), 100-103. https://doi.org/10.5755/j01.eee.20.6.7276

17. Jokić, I., Delić, V., Jokić, S., Perić, Z. Automatic Speaker Recognition Dependency on Both the Shape of Auditory Critical Bands and Speaker Discriminative MFCCs. Advances in Electrical and Computer Engineering, 2015, 15(4), 25-32. https://doi.org/10.4316/AECE.2015.04004

18. Jokić, I. D., Jokić, S. D., Delić, V. D., Perić, Z. H. Mel-Frequency Cepstral Coefficients as Features for Automatic Speaker Recognition. In Proceedings of the 23rd Telecommunications Forum TELFOR 2015, Belgrade, Serbia, November 24-26, 2015, 419-424. https://doi.org/10.1109/TELFOR.2015.7377497

19. Jokic, I., Jokic, S., Delic, V., Peric, Z. About the Need for Sound Features Independent of the Sound Pattern. Proceedings of 62nd Conference on Electronics, Telecommunications, Computing, Automation and Nuclear Technology (ETRAN 2018), Palić, Serbia, June 11-14, 2018, 108-111 (in Serbian).

20. Kanagasundaram, A., Vogt, R., Dean, D., Sridharan, S., Mason, M. I-Vector Based Speaker Recognition on Short Utterances. INTERSPEECH 2011, Florence, Italy, August 28-31, 2011, 2341-2344.

21. Kinnunen, T., Li, H. An Overview of Text-Independent Speaker Recognition: From Features to Supervectors. Speech Communication, 2010, 52(1), 12-40. https://doi.org/10.1016/j.specom.2009.08.009

22. Kinnunen, T., Zhang, B., Zhu, J., Wang, Y. Speaker Verification with Adaptive Spectral Subband Centroids. In: Proceedings of the International Conference on Biometrics (ICB 2007), Seoul, Korea, August 2007, 58-66. https://doi.org/10.1007/978-3-540-74549-5_7

23. Korvel, G., Šimonytė, V., Slivinskas, V. A Phoneme Harmonic Generator. Information Technology and Control, 2016, 45(1), 7-12. https://doi.org/10.5755/j01.itc.45.1.7657

24. Kua, J. M. K., Thiruvaran, T., Nosratighods, M., Ambikairajah, E., Epps, J. Investigation of Spectral Centroid Magnitude and Frequency for Speaker Recognition. In Odyssey-2010: The Speaker and Language Recognition Workshop, Brno, Czech Republic, June 28 - July 1, 2010, paper 007, 34-39.

25. Ma, L., Milner, B., Smith, D. Acoustic Environment Classification. ACM Transactions on Speech and Language Processing (TSLP), 2006, 3(2), 1-22. https://doi.org/10.1145/1149290.1149292

26. Monson, B. B., Hunter, E. J., Lotto, A. J., Story, B. H. The Perceptual Significance of High-Frequency Energy in the Human Voice. Frontiers in Psychology, 2014, 5(587), 1-11. https://doi.org/10.3389/fpsyg.2014.00587

27. Muhammad, G., Alghathbar, K. Environment Recognition for Digital Audio Forensics Using MPEG-7 and Mel Cepstral Features. The International Arab Journal of Information Technology, 2013, 10(1), 43-50.

28. Neiberg, D., Elenius, K., Laskowski, K. Emotion Recognition in Spontaneous Speech Using GMMs. In INTERSPEECH 2006 - ICSLP, Pittsburg, Pennsylvania, USA, September 17-21, 2006, 809-812.

29. Panda, B., Padhi, D., Dash, K., Mohanty, S. Use of SVM Classifier & MFCC in Speech Emotion Recognition System. International Journal of Advanced Research in Computer Science and Software Engineering, 2012, 2(3), 225-230.

30. Qarachorloo, M., Farahani, G. New Features to Improve Speaker Recognition Efficiency with Using LPCC and SSC Features. International Journal of Signal Processing Systems, 2016, 4(4), 295-299. https://doi.org/10.18178/ijsps.4.4.295-299

31. Salehghaffari, H. Speaker Verification using Convolutional Neural Networks. ArXiv, 2018, abs/1803.05427.

32. Tančić, M. Ž., Perić, Z. H., Simić, N., Tomić, S. S. Performance of Quasi-Logarithmic Quantizer for Discrete Input Signal. Information Technology and Control, 2017, 46(3), 395-402. https://doi.org/10.5755/j01.itc.46.3.16197

33. Thian, N. P. H., Sanderson, C., Bengio, S. Spectral Subband Centroids as Complementary Features for Speaker Authentication. In: Proceedings of the First International Conference on Biometric Authentication (ICBA 2004), Hong Kong, China, July 15-17, 2004, 631-639. https://doi.org/10.1007/978-3-540-25948-0_86

34. Tirumala, S. S., Shahamiri, S. R. A Review on Deep Learning Approaches in Speaker Identification. Proceedings of the 8th International Conference on Signal Processing Systems ICSPS 2016, Auckland, New Zeland, November 21-24, 2016, 142-147. https://doi.org/10.1145/3015166.3015210

35. Tiwari, V. MFCC and Its Applications in Speaker Recognition. International Journal on Emerging Technologies, 2010, 1(1), 19-22.

36. Tomić, S., Perić, Z., Tančić, M., Nikolić, J. Backward Adaptive and Quasi-Logarithmic Quantizer for Sub-Band Coding of Audio. Information Technology and Control, 2018, 47(1), 131-139. https://doi.org/10.5755/j01.itc.47.1.16190

37. Wildermoth, R. B. Text-Independent Speaker Recognition Using Source Based Features. M. Phil. Thesis, Griffith University, Brisbane, Australia, January 2001, 19-20.

38. Zubova, J., Kurasova, O., Liutvinavičius, M. Dimensionality Reduction Methods: The Comparison of Speed and Accuracy. Information Technology and Control, 2018, 47(1), 151-160. https://doi.org/10.5755/j01.itc.47.1.18813