Automatic Speaker Recognition Dependency on Both the Shape of Auditory Critical Bands and Speaker Discriminative MFCCs

Ivan JOKIĆ[1], Vlado DELIĆ[1], Stevan JOKIĆ[1,2], Zoran PERIĆ[3]
[1]*Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6,*
21000 Novi Sad, Serbia
[2]*DunavNET Ltd., Antona Čehova 1/2, 21000 Novi Sad, Serbia*
[3]*Faculty of Electronic Engineering, University of Niš, Aleksandra Medvedeva 14, 18000 Niš, Serbia*
ivan.jokic@uns.ac.rs, vdelic@uns.ac.rs, stevan.jokic@uns.ac.rs, zoran.peric@elfak.ni.ac.rs

***Abstract*—Accuracy of an automatic speaker recognition system predominantly depends on speaker models and features that are used. An influence of the shape of auditory critical bands and a contribution of individual components of MFCC-based feature vectors are investigated in the paper and some experimental results are presented and showed their impact on the accuracy of automatic speaker recognition. The speaker-discrimination capability of the MFCCs was experimentally determined by comparing training and test models for the same speaker. The experiments are conducted with three speech databases and showed that 0th and 19th (the last one) MFCCs are non speaker discriminative. The values of MFCCs are determined by the type of applied auditory critical band. The exponential auditory critical bands based on the lower part of exponential function have outperformed the speaker recognition accuracy of other auditory critical bands such as rectangular or triangular shape.**

***Index Terms*—Automatic speaker recognition, mel-frequency cepstral coefficients, energy correction, speaker discriminative, exponential auditory critical bands.**

I. Introduction

Progress in computer technology enables development of new applications and new kinds of human-computer communications. The tendency of introducing computer based systems in everyday life results in the need of applications that might perform some recognition. A human-computer two-way communication sometimes needs automatic speaker recognition – the computer capability to recognize a person from his/her voice.

The listener decision about speaker recognition is mainly based on recognition of a timbre of the listened speaker. The spectral content of a speech signal determines the timbre. Perception of timbre is a consequence of speech signal spectral content. Therefore the listener decision is based on the spectral content of observed speech signal. This fact lead to choice of adequate feature set which represents a speaker.

Automatic recognition based on observing of speech signal, such as speech recognition, emotion or speaker recognition, as features based on speech spectral content uses mel-frequency cepstral coefficients (MFCCs). These features contain information about spectral envelope of the observed speech signal [1]. The calculation of MFCCs is based on analysis of short-term speech frames. The duration of these short-term speech frames is about 25ms [2] and they are mutually moved by about 10ms [3]. On this way only for one second of speech it is observable about 97.5 frames and the same number of the feature vectors. Speech recordings are typically longer and their duration can be of a few seconds and often more than ten seconds. Each of these speech recordings is described with a several hundred of feature vectors. The existence of such multitude of feature vectors motivates the use of the stochastic models as more adequate compared to the template models [4].

Stochastic modeling of speakers is usually performed in two ways: by the use of Hidden Markov Models (HMMs) or by using Gaussian Mixture Models (GMMs). HMMs are usually used in text-dependent applications, whereas GMMs are commonly used in text-independent applications.

The efficiency of an automatic speaker recognizer is conditioned by the discrimination capability of feature vectors that are used. Speech is a complex signal produced as a result of several transformations occurring at different levels: semantic, linguistic, articulatory and acoustic [4]. MFCCs as suitable and representative features for speech based applications, both for applications oriented to speaker recognition [3], [5], [6], as well as in the case of speech [7], [8], or emotion recognition [9-11], are determined through number of coefficients and the way of their computation. Lots of transformations responsible for the formation of speech signal cause its composite character. Therefore the timbre as the main information about speaker is hidden in the wealth of information provided by the speech signal. MFCCs describe spectral envelope of speech sample. So, they depend on energy distribution contained in speech signal. One of causes of the energy distribution in speech signal is timbre, direct characteristic of speaker. Also energy distribution of the speech depends on the text content of speech, loudness of the speech and emotional state of the speaker. As consequence, determined MFCCs beside information about timbre, as one of speaker inherent information, also carry irrelevant, non speaker discriminative information. It is obviously that the main block in automatic speaker recognizer is the block for feature extraction. Often it is necessary to do some additional transformations on determined MFCCs for better

The presented study is performed as part of the project "Development of dialogue systems for Serbian and other South Slavic languages" (TR32035), funded by the Ministry of Education, Science and Technological Development of the Republic of Serbia.

Digital Object Identifier 10.4316/AECE.2015.04004

speaker recognition. One approach is to determine the main energy directions in observed speech database, the new dimensions instead of MFCCs and their derivatives if used, for example by application of Principal Component Analysis (PCA) [12]. As consequence this can result in dimensionality reduction of feature vectors [13]. The second approach is to using some new sets of features [14] or to more detailed analyze procedure of determining MFCCs.

By definition the calculation of MFCCs is based on application of Inverse Discrete Fourier Transform on log energy in observed speech frame [1]. In accordance with this in [15] Discrete Cosine Transform was performed on log energy of whole speech frame. Differently, in most cases [16-18], the log energy was calculated inside determined auditory critical bands [19].

Very often in applications for speech or speaker recognition the shape of auditory critical bands is rectangular or triangular as it is mentioned in [16-18]. If the central component of the auditory critical band is observed as a masking component, then it can be expected that masking phenomena can be better described by auditory critical bands built by descending functions. Therefore, descending function around the central masking component should attenuate masked components as much as possible. It is possible by nonlinear functions. Exponential function has much higher slope with respect to linear function and thus we also examined the results of the automatic speaker recognition when auditory critical bands are based on exponential function. Confirmation of research direction toward exponential auditory critical bands is emphasized in [17] where auditory critical bands based on exponential function are listed as one of three lines in future development of auditory critical bands.

In the continuation of this paper it was described automatic speaker recognizer used in experiments. Then through experimental results was shown the impact of the shape of auditory critical bands as well as energy inside them to accuracy of automatic speaker recognition. Since MFCCs are non strictly speaker discriminative, in the initial and final part of experimental results attention is also devoted to analysis of speaker models of the same speakers for difference speech samples.

II. Description of used Automatic Speaker Recognizer

Automatic speaker recognizer described in this paper is based on the use of MFCCs as features of voice and is set for text-independent speaker identification. He has three basic parts with the next functionalities: feature extraction, modeling and comparison of models and decision making.

Before of calculation MFCCs the observed speech signals are windowed by Hanning window function,

$$w(n)=\frac{1}{2}\cdot\left(1-\cos\frac{2\cdot\pi\cdot n}{N-1}\right),\quad 0\le n\le N-1\,, \qquad (1)$$

where N is the number of speech samples in observed frame. Frame width is about 23ms and shift between the adjacent frames is about 8.33ms. Determination of MFCCs was done by applying discrete cosine transform on log energy in observed auditory critical bands [20],

$$c_n=\sum_{k=1}^{K=20}E_{\log(k)}\cdot\cos\left[n\cdot\left(k-\frac{1}{2}\right)\right],\qquad n=0,1,...,d-1\,, \qquad (2)$$

where the k is an ordinal number of observed auditory critical band and the d, d_{max}=K=20 in accordance with the definition of discrete cosine transform, is the number of MFCCs calculated, i.e. it is the dimensionality of used feature vector. Automatic speaker recognizer was developed for sampling frequency of 22050 Hz. Width of applied auditory critical bands is approximately 300 mels and the adjacent auditory critical bands are shifted approximately by 150 mels. Used relation between frequency scale in Hz and mel frequency scale is,

$$f[mel]=2595\cdot\log_{10}\left(1+\frac{f[Hz]}{700}\right). \qquad (3)$$

Considering the sampling theorem the maximum frequency in observed speech signals is 11025 Hz, i.e. approximately 3176 mels. For covering of this range it is necessary 20 previously defined auditory critical bands, as set out in the equality 2. In initial experiments feature vector consists of all 20 MFCCs, zeroth and first 19 MFCCs.

Specificity of this paper lies in the fact that it's mainly based on the comparison of results of recognition when standard concept of auditory critical bands is changed from rectangular and triangular shape to exponential shape.

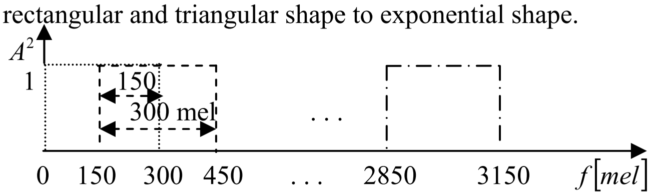


Figure 1. Arrangement of 20 applied rectangular auditory critical bands

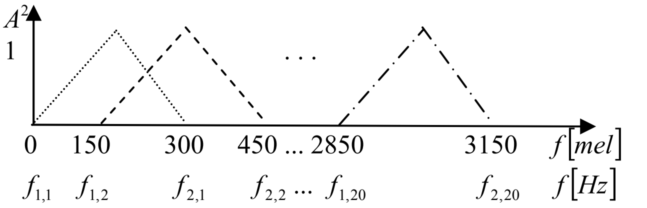


Figure 2. Arrangement of 20 applied triangular auditory critical bands

Concept of auditory critical band implies frequency range in which the listener does not different the frequency change between the two tones. This concept is related to the masking phenomena and thus the rectangular approximation of auditory critical bands is not appropriate. Representation of masking implies approximation of auditory critical bands with the help of descending functions. As descending functions often linear functions are used in the form of triangular auditory critical bands [3], [16], [18].

In this work triangular auditory critical bands (Fig. 2) are defined as:

$$A^2(k)=\begin{cases}\dfrac{2}{k_{2,n}-k_{1,n}}\cdot\left(k-k_{1,n}\right), & k_{1,n}\le k\le\dfrac{k_{1,n}+k_{2,n}}{2},\\[2ex] \dfrac{2}{k_{1,n}-k_{2,n}}\cdot\left(k-k_{2,n}\right), & \dfrac{k_{1,n}+k_{2,n}}{2}<k\le k_{2,n},\end{cases} \qquad (4)$$

where the n is the ordinal number of the observed auditory critical band from a set $n=\{1,2,...,20\}$ and the k is the discrete frequency:

$$k=N\cdot\frac{f}{f_s}\ \wedge\ 0\le k\le N-1\,, \qquad (5)$$

f_s=22050 Hz is sampling frequency and N=512 dots is length of the speech frame whose duration is about 23 ms. In later tests on speech signals whose frequency sampling is 44100 Hz, the same parameter f_s=22050 Hz in equality 5 is used but with N=1024.

MFCCs are directly dependent on log energy in observed auditory critical bands (equality 2). This energy can be artificially modified by the external influence or it may vary depending on the speaker articulation. Experiments during development of automatic speaker recognizer which is used in this paper were conducted over two speech databases: speech database 1 recorded in studio, and speech database 2 recorded in office conditions. The external influence is present in speech database 2, it is noise present in the recordings. In speech database 1 we can eventually speak about impact of the speaker articulation. Primarily due to the influence of noise in speech database 2, tests of automatic speaker recognition were conducted and when energy correction in observed auditory critical bands is applied.

Introducing of the energy correction is motivated by assumption that if noise or some other irregular appearance is present in frame of observed speech signal then he is most expressed in auditory critical band of minimal energy. The energy correction for each observed speech frame was applied through next three steps:

- calculation of log energy, $E_{\log(n)}$, for each auditory critical band, n={1,2,…,20},
- determination of the minimum log energy, $E_{\min\log(n)}=\min\{E_{\log(n)}\}$, n={1,2,…,20},
- calculation of the new log energy value in observed auditory critical band as,

$$E_{\log(n)new}=E_{\log(n)}-\min\{E_{\log(n)}\}. \tag{6}$$

As will be shown in chapter three of this paper, the shift from rectangular to triangular auditory critical bands increases achieved recognition accuracy. Increase of recognition accuracy by application of the triangular auditory critical bands reflects the fact that auditory critical bands should have decreasing shape around the central position. In the case of triangular auditory critical bands a decreasing shape is characterized by the slope of the linear function used. Compared to the case of the rectangular auditory critical bands, application of the triangular auditory critical bands emphasizes spectral components on their central positions. It follows that emphasis of the spectral component in central position of auditory critical band contributes to increasing of recognition accuracy. The function which better emphasizes observed value compared to values of her left side is exponential function $y=e^x$ (Fig. 3).

Applying of auditory critical bands based on the exponential function can be achieved higher suppression of components around their central positions. As is evident on Fig. 3, it exist a region on x-axis where the exponential function y is below observed linear function y_1. This is the initial part of the x-axis to the left of the point x=0. Also, on the initial part of the x-axis to the right of the point x=0, the exponential function y is below linear function y_2. Therefore, compared to the case when triangular auditory critical bands are used, these parts of the exponential function y can be used as better approximation of masking phenomena in auditory critical bands.

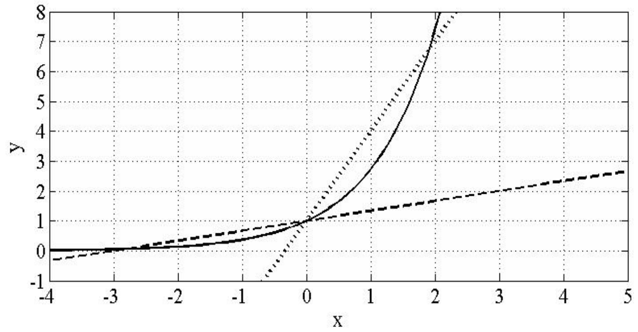


Figure 3. Example of the exponential function benefit (full: $y=e^x$) with respect to the linear functions (dashed: $y_1=\frac{1}{3}\cdot x+1$, dotted: $y_2=3\cdot x+1$) for approximation of squared amplitude of auditory critical band. In applications of these functions for approximation of auditory critical bands x-axis equally represents frequency of spectral component while y-axis is equivalent to square amplitude characteristic of observed auditory critical band

Depending on observed part of exponential function which is applied for modeling of auditory critical band in this work were used three ways of auditory critical bands modeling. As the referent point of argument is viewed point $x=0$ where the function $e^x=1$. Based on this, the three shapes of auditory critical bands (ACBs) was distinguished:

- ACBs based on the lower part of the exponential function, left of the point $x=0$ of function e^x and right of the point $x=0$ of function e^{-x},
- ACBs based on the upper part of exponential function, right of the point $x=0$ of function e^x and left of the point $x=0$ of function e^{-x},
- ACBs based on the upper – lowered part of exponential function, used parts of exponential functions as in previous case and each value is less for one.

When auditory critical bands are based on the lower part of exponential functions, their central components correspond to point $x=0$ of functions e^x and e^{-x}. Therefore points to the left and right around the central position are significantly decreased as the observed point is farther from the point $x=0$. This property is appropriate for simulating of auditory masking phenomena.

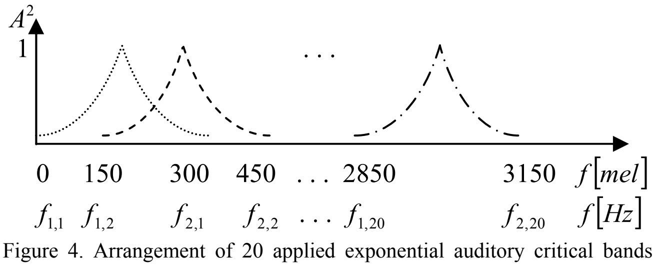


Figure 4. Arrangement of 20 applied exponential auditory critical bands based on the lower part of exponential function

Exponential auditory critical bands based on the lower part of the exponential function are defined as,

$$A^2_{\exp}(k)=\begin{cases} e^{(k-k_{c,n})s}, & k_{1,n}\le k\le k_{c,n},\\ e^{-(k-k_{c,n})s}, & k_{c,n}<k\le k_{2,n},\end{cases} \tag{7}$$

$k_{c,n}=\dfrac{k_{1,n}+k_{2,n}}{2}$ is a central discrete frequency (equality 5) of n^{th} auditory critical band and $n=\{1,2,...,20\}$ (Fig. 4). This approximation of central positions in auditory critical bands is also applied in the case of two other shapes of auditory critical bands based on the upper part of exponential functions.

On the other hand in the case of auditory critical bands based on use of the upper part of exponential function the ending components are pondered by factor 1. Exponential auditory critical bands based on the upper part of exponential function (Fig. 5) are defined as,

$$A^2_{\exp}(k)=\begin{cases} e^{(k-k_{1,n})s}, & k_{1,n}\le k\le k_{c,n}\,,\\ e^{-(k-k_{2,n})s}, & k_{c,n}<k\le k_{2,n}.\end{cases} \tag{8}$$

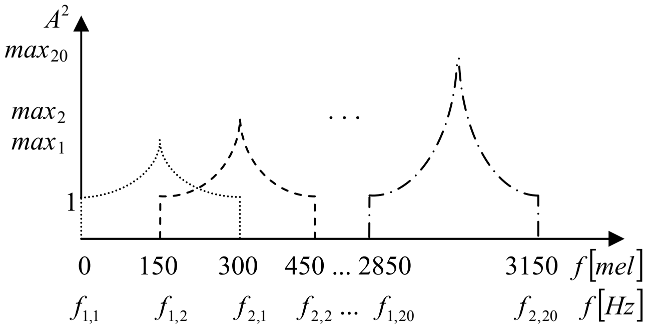


Figure 5. Arrangement of 20 applied exponential auditory critical bands based on the upper part of exponential function

Components closer to the center of auditory critical band are more emphasized related to the components on the ends of auditory critical band. So, energy of central component is emphasized by the factor *max* which is determined by the argument value of exponential function. As the *max* value is dependent of frequency and dependency between hertz and mel scale is exponential, her value increases for higher spectral components.

To decrease the end points of auditory critical bands to zero value, as well as in the case of triangular auditory critical bands or approximately in the case of auditory critical bands based on the lower part of exponential function, values of previously introduced auditory critical bands based on the higher part of exponential function were reduced by one. On this idea is based introduction of auditory critical bands based on the upper – lowered part of exponential function (Fig. 6).

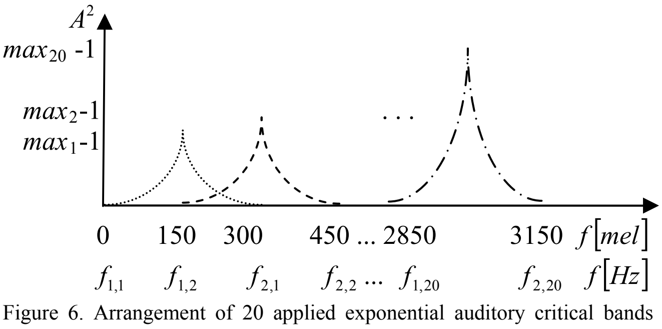


Figure 6. Arrangement of 20 applied exponential auditory critical bands based on the upper – lowered part of exponential function

According to the previously mentioned idea, exponential auditory critical bands based on the upper – lowered part of the exponential function (Fig. 6) are defined as,

$$A^2_{\exp-1}(k)=\begin{cases} e^{(k-k_{1,n})s}-1, & k_{1,n}\le k\le k_{c,n}\,,\\ e^{-(k-k_{2,n})s}-1, & k_{c,n}<k\le k_{2,n}.\end{cases} \tag{9}$$

Also by factor *s* in equalities: 7, 8 and 9, is controlled the steepness of previously introduced exponential auditory critical bands. Steepness factor *s* was varied by two values, *s*=1 and *s*=2.

Speaker recognizer is adjusted to text-independent work. Therefore modeling of speakers was based on the use of Gaussian multivariate distributions. In this case a form of Gaussian distribution is the consequence of a speech signals spectral content. For multivariate Gaussian distribution,

$$\mathrm{N}(x)=\frac{1}{\sqrt{(2\cdot\pi)^d\cdot|\Sigma|}}\cdot e^{-\frac{1}{2}\cdot(x-\mu)^T\cdot\Sigma^{-1}\cdot(x-\mu)}, \tag{10}$$

where the *x* is the *d*-dimensional feature vector, the form is determined by the covariance matrix Σ . Hence, modeling of speakers is done by appropriate covariance matrices.

Estimation of covariance matrix for each set of *n* feature vectors x_i, sorted in data matrix $X=[x_1\quad x_2\quad ...\quad x_n]$, was performed using the matrix equality,

$$\Sigma=\frac{1}{n-1}\cdot(X-\mu)\cdot(X-\mu)^T, \tag{11}$$

where μ is the vector of a mean values of matrix *X*. In training phase for each of speakers was formed appropriate data matrix and covariance matrix as the model. Testing was done on the closed set of speakers. In testing phase, for the test speech record also was formed data matrix and covariance matrix.

The measure of distinguishing between model of the i^{th} speaker and observed test model "test" was defined by,

$$m(i,test)=\frac{1}{d^2}\cdot\sum_{j=1}^{d}\sum_{k=1}^{d}\left|\Sigma_i(j,k)-\Sigma_{test}(j,k)\right|. \tag{12}$$

If a set of speakers contains of *N* speakers then the test speech "*test*" belongs to the i^{th} speaker if,

$$m(i,test)<m(j,test),\qquad \forall j\in\{1,2,...,N\}\setminus\{i\}. \tag{13}$$

III. Experiments and Results

Experiments were conducted on recordings in three speech databases. The first two speech databases are in Serbian, hereinafter referred to: speech database 1 (SD1) and speech database 2 (SD2). Both speech databases are retrieved from AlfaNum team at the Faculty of Technical Sciences in Novi Sad. Third speech database is CHAracterizing INdividuals Speakers (CHAINS) speech corpus [21]. This corpus contains speech recordings of 36 speakers, 26 of these speakers were speakers of Eastern Hiberno-English and the others are from United Kingdom and United States of America.

SD 1 contains utterances produced by 121 speakers. All utterances in this corpus were recorded only once and they are classified in three groups: "names", "digits" and "words". For each of speakers these groups contain the following contents:

- "digits" – two utterances containing words that correspond to digits: 1 – 2 – 3 – 4 – 5 and 6 – 7 – 8 – 9 – 0,
- "names" – one recording for each of speakers which contains: first name, family name, and the speaker-specific identification number,
- "words" – the same set of eleven preset word sequences.

This speech database for each of speakers contains 14 recordings. Tests of speaker recognition are performed over each of 14 speech recordings, tests 1 and 2 on recordings in “digits”, test 3 on recording in “names” and test 4 to 14 on recordings in “words”.

The recordings in SD 2 contain the utterances of four digits. Each of 44 speakers was recorded in several sessions. Most of speakers are recorded in 10 sessions and each of them contains 12 recordings. Some of speakers are recorded in a smaller number of sessions. These speakers are exempt from tests. Thus, recordings of 37 speakers were used for experiments. Tests of speaker recognition are performed over recordings in first session. Therefore over this speech database was performed 12 tests.

In this paper were performed three sets of tests, 1-14 on the SD 1, 1-12 on the SD 2 and 1C, 2C on the CHAINS. In both experiments on CHAINS for each of 36 speakers training models were formed for the longest, f01 recording from “Solo” part. Average recognition accuracy over random sample of four short recordings: s01, s02, s03 and s15, in “Solo” part were monitored in test C1. “Solo” part of CHAINS for each of 36 speakers has four long recordings {f01,...,f04} and 33 short recordings {s01,...,s33}. Duration of short recordings is in range of about 2 to 4 seconds. To perform testing on longer speech recordings in duration of about 8 to 9 seconds, as well as in group “words” from SD1, three arbitrarily selected recordings, s20, s21 and s22, for each of speakers are merged in one recording. Recognition accuracy over such merged recordings for each of 36 speakers was monitored in test C2. Experiments were started for 20 rectangular auditory critical bands (Fig. 1) and feature vectors of 20 MFCCs, zeroth and first 19 MFCCs. Covariance matrix as model of observed set of MFCCs feature vectors represents the energy picture of observed feature set. Thus comparison of elements on the same places in training and test matrices of the same speakers is prepared. It was noticed that first elements show significantly differentiation. First element in covariance matrix represents variance of zeroth MFCC, i.e. the mean energy representation of zeroth MFCC. Therefore this element depends on the text of the observed speech and present noise. Thus, in the next step of tests the first element of covariance matrices is set to zero and this improves the accuracy of recognition.

The impact of the zeroth MFCC is also present in elements of first row and first column of observed covariance matrices. Because in step three of experiments the zeroth MFCC is discarded from the feature set. Experiments were repeated for feature vectors of first 19 MFCCs. In most tests recognition accuracy is increased and therefore the next experiments are devoted to investigation of the influence of auditory critical bands shape when feature vector consists of first 19 MFCCs.

Recognition accuracy has highest values in tests on group “words”, often higher than 90%. Recordings in group “names” are shortest and therefore recognition accuracy in that case is smallest, about 40%. Efficiency of the automatic speaker recognition based on MFCCs depends on the available amount of information about speech spectrum of observed speaker. Duration of test recordings is typically in ranging from a few seconds to ten seconds. Voiced segments of speech are characterized by the discrete spectral components while unvoiced have continuous spectrum. Speech spectrums of the different speakers differ in the amplitudes of the spectral components i.e. in the spectral envelope which is experienced as the timbre. Mainly, this differ is a consequence of voiced speech frames and therefore the amount of voiced speech frames impacts to the accuracy of automatic speaker recognition. In tests on SD 1, test 3 is characterized by the lowest number of voiced frames. In average in this test can be expected around 6 vowels. Remaining tests have more vowels. Pronunciation in group “names” often is faster than pronunciation in groups “digits” and “words”, name and surname often were linked spoken without of prominent break. Thus vowels in this group are shortened compared to vowels from the other two groups of recordings.

Recognition accuracy over SD 2, for 20 rectangular auditory critical bands and 20 MFCCs, zeroth and first 19 MFCCs, is more than twice, around 40%, less than recognition accuracy over SD 1. Duration of recordings is of a few seconds, often around 5 seconds. It follows that these recordings have lesser number of voiced segments compared to the recordings in SD 1. Moreover, SD 2 has a higher noise level, SD 2 was recorded in office conditions while SD 1 was recorded in the studio. So these facts were considered as justifications of a lesser recognition accuracy.

In tests on SD 1 when triangular auditory critical bands (equality 4) were used, the minimum of recognition accuracy is achieved in test 3 (Fig. 7). Recognition accuracy in test 2 and tests 4 – 14 is more than 90%. Average recognition accuracy in these tests is around 25 – 30% higher compared to recognition accuracy in test 1 and around 50% higher compared to the recognition accuracy in test 3. Application of the energy correction does not significantly affect to recognition accuracy.

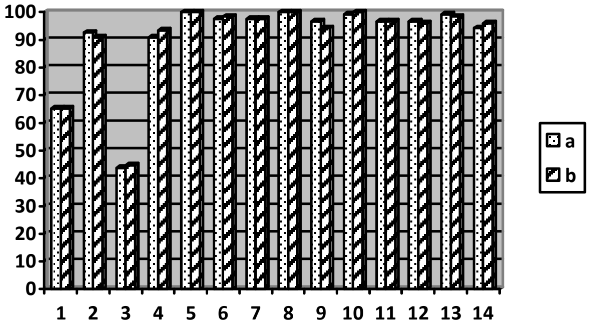


Figure 7. Percentage accuracy of recognition over SD 1 (tests: 1-14), when 20 triangular auditory critical bands are applied (Fig. 2) and feature vectors of first 19 MFCCs are used, depending on the observed energy in auditory critical bands: a – non modified, b – modified by energy correction

By analysis of training and test covariance matrices of the same speakers when feature vector of first 19 MFCCs is used, a significantly difference between $\Sigma_{19,19}$ elements is noted. The difference between $\Sigma_{19,19}$ elements is several times smaller compared to the difference between $\Sigma_{0,0}$ elements. In this phase of experiments the target is to examine the influence of the shape of auditory critical bands to recognition accuracy. Feature vector of higher dimensionality more accurately describes observed signals. MFCCs of higher order show the fine spectral details in observed speech signal [20] and therefore in further experiments was kept 19th MFCC.

Application of triangular auditory critical bands was significantly increased recognition accuracy in tests on SD 2 (Fig. 8), in the case when rectangular auditory critical bands were applied maximum of accuracy was been 70% only in test 9. Additional application of energy correction in most cases was increased recognition accuracy. Only in tests 7 and 11 application of energy correction very little reduce the recognition accuracy. In tests 7, 8 and 9 maximum values of recognition accuracy are achieved. With respect to the case when rectangular auditory critical bands and feature vectors of first 19 MFCCs were used, application of triangular auditory critical bands with energy correction was caused the significant increase of recognition accuracy, about 20%.

By application of triangular auditory critical bands a greater increase of average recognition accuracy is achieved in tests on SD 2 and CHAINS. Compared to the initial experiments, when 20 rectangular auditory critical bands are applied and feature vectors of 20 MFCCs used, average increase in tests on SD 1 is around 10% whereas on SD 2 and CHAINS this increase is around 35% and 50%.

In cases when feature vectors of first 19 MFCCs are used, application of triangular auditory critical bands and energy correction more significantly increases recognition accuracy on SD 2 and CHAINS. Increase of average percentage recognition accuracy on SD 1 is around 2-3% whereas on SD 2 and CHAINS is around 15% and 10% respectively. SD 2 is of a poorer quality compared to the SD 1. Therefore, the positive contribution of triangular auditory critical bands and energy correction become more noticeable in tests on SD 2. Speech in CHAINS is faster pronounced with respect to recordings from "digits" and "words" groups in SD1 and therefore the results show significant difference between recognition accuracy on these two speech corpus when triangular ACBs were applied.

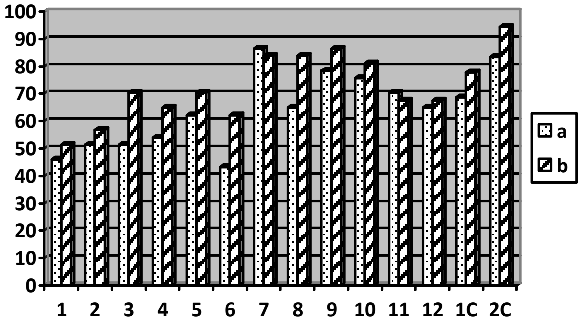
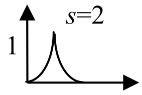
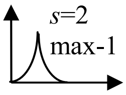
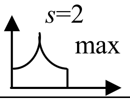
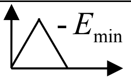
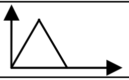
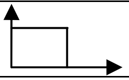


Figure 8. Percentage accuracy of recognition over SD 2 and CHAINS (tests: 1-12 and 1C, 2C), when 20 triangular auditory critical bands are applied (Fig. 2) and feature vectors of first 19 MFCCs are used, depending on the observed energy in auditory critical bands: a – non modified, b – modified by energy correction

Compared to the auditory critical bands based on the exponential functions of steepness factor *s*=1 application of auditory critical bands based on the exponential functions of steepness factor *s*=2 enables better masking of spectral components around central spectral component in observed auditory critical bands. Results of recognition were similar in the cases: *s*=1 or *s*=2. In both cases accuracy is outperformed accuracy when triangular or rectangular auditory critical bands were applied. Therefore in Table I and Table II is showed the benefit of exponential auditory critical bands of steepness factor *s*=2 with respect to triangular, with and without energy correction applied, and also with respect to rectangular auditory critical bands.

Maximum of recognition accuracy in tests: 1, 3, 4 and 14, on SD 1 when steepness factor is *s*=1, was achieved in the case of applied auditory critical bands based on the lower part of exponential function with application of energy correction. Only in test 3 in that case and absolutely maximum of about 58.68% is achieved. In test 2 application of energy correction was decreased recognition accuracy. In this test the same recognition accuracy was achieved for all three types of exponential auditory critical bands. In tests 5 – 13 recognition accuracy is fairly uniform with respect to the type of exponential auditory critical band.

TABLE I. COMPARING OF RECOGNITION ACCURACY ON SPEECH DATABASE 1, FEATURE VECTORS CONSIST OF FIRST 19 MFCCS

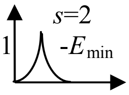
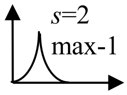
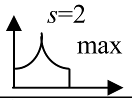
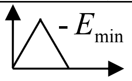
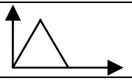
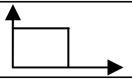
Type of crit. band	1	2	3	4	5	6	7	8	9	10	11	12	13	14
	76.9	94.2	52.9	94.2	99.2	97.5	97.5	99.2	100	100	98.3	98.3	100	99.2
	74.4	95	48.8	95	99.2	97.5	96.7	99.2	100	100	97.5	98.3	100	99.2
	76.9	94.2	52.9	94.2	99.2	97.5	97.5	99.2	100	100	98.3	98.3	100	99.2
	65.3	90.9	44.6	93.4	100	98.3	97.5	100	94.2	100	96.7	95.9	98.3	95.9
	65.3	92.6	43.8	90.9	100	97.5	97.5	100	96.7	99.2	96.7	96.7	99.2	94.2
	59.5	81	39.7	86.8	100	98.3	93.4	99.2	97.5	98.3	95	94.2	100	95.9

By comparing the recognition accuracy for steepness factor *s*=2 and steepness factor *s*=1 it is evident that achieved values of percentage recognition accuracy in most of observed 14 tests are similar. In test 1 for steepness factor *s*=2 the recognition accuracy was increased by few percent. Application of the energy correction in test 3 did not bring that improvement as in the case of steepness factor *s*=1. Also comparing to previous results when rectangular and triangular auditory critical bands are used (Table I) improvements are evident when auditory critical bands based on the lower part of exponential function were applied. The largest improvements are in tests 1 and 3, around 5% and 10-15%.

In the case of tests on SD 2 application of the energy correction significantly increases the recognition accuracy when auditory critical bands are approximated by lower part of exponential function and steepness factor is *s*=1. Also, in most tests, application of exponential auditory critical bands based on lower part of exponential function of this steepness factor and with applied energy correction is resulted in maximum of recognition accuracy. The similar case is for application of auditory critical bands with steepness factor *s*=2 (Table II). Also, in most cases by application of auditory critical bands based on lower part of exponential function maximum of recognition accuracy is achieved. Application of auditory critical bands based on the lower part of exponential function, with steepness factor *s*=2 and applied energy correction defined by equality 6, results in maximum of average recognition accuracy on SD 2 of about 78%, this value is about few percent (3-5%) higher to average accuracy when energy correction was not applied. By comparing this average recognition accuracy and the maximum values of average recognition accuracy for rectangular and triangular auditory critical bands, improvements of around 22% and 8% are evident.

Significant improvements of recognition accuracy by application of exponential auditory critical bands are also in tests 1C and 2C on CHAINS (Table II). Recognition accuracy in test 2C is of the same order of magnitude as well as recognition accuracy on group Words from SD1 (tests 4-14).

TABLE II. COMPARING OF RECOGNITION ACCURACY ON SPEECH DATABASE 2 AND CHAINS, FEATURE VECTORS CONSIST OF FIRST 19 MFCCS

Type of crit. band	1	2	3	4	5	6	7	8	9	10	11	12	1C	2C
 *s*=2 1 -$E_{\min}$	62.2	70.3	81.1	67.6	75.7	70.3	86.5	89.2	94.6	83.8	78.4	81.1	81.2	94.4
 *s*=2 max-1	67.6	70.3	73	67.6	67.6	64.9	81.1	75.7	91.9	78.4	78.4	75.7	s=1 81.9	s=1 91.7
 *s*=2 max	62.2	64.9	70.3	64.9	67.6	62.2	78.4	75.7	91.9	81.1	83.8	73	s=1 80.6	s=1 94.4
 - $E_{\min}$	51.3	56.8	70.3	64.9	70.3	62.2	83.8	83.8	86.5	81.1	67.6	67.6	77.8	94.4
	45.9	51.3	51.3	54	62.2	43.2	86.5	64.9	78.4	75.7	70.3	64.9	68.7	83.3
	40.5	40.5	54	48.6	54	43.2	64.9	62.2	70.3	64.9	62.2	56.8	52.8	63.9

Comparing models of training and test speech of the same speakers, a large distinction is mentioned between $\Sigma_{19,19}$ elements which represent variance of the 19th MFCC. In next step of automatic speaker recognizer construction this MFCC was discarded from feature vector so that the resulting feature vector consists of first 18 MFCCs.

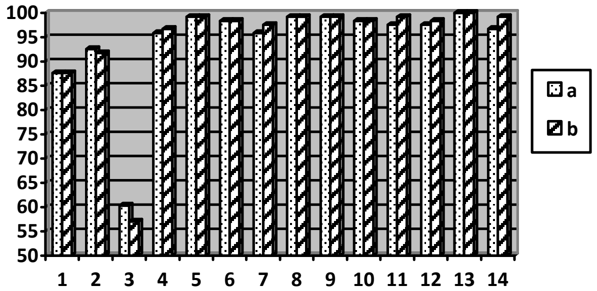


Figure 9. Percentage accuracy of recognition over SD 1 (tests 1-14) depending on the observed energy in auditory critical bands based on lower part of exponential function of steepness factor s=2 (a – non modified and b – modified by energy correction). Feature vector consist first 18 MFCCs

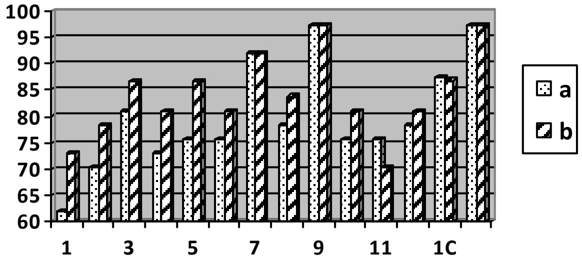


Figure 10. Percentage accuracy of recognition over SD 2 and CHAINS (tests: 1-12 and 1C, 2C) depending on the observed energy in auditory critical bands based on lower part of exponential function of steepness factor s=2 (a – non modified and b – modified by energy correction defined in equality 6). Feature vector is composed of the first 18 MFCCs

The smallest recognition accuracy on SD 1 was achieved in tests 1 and 3 (Fig. 9). Reduction of feature vector to 18 MFCCs has resulted in increase of recognition accuracy in these tests, in test 1 by 8-10% and in test 3 by few percent. Recognition accuracy in remaining tests is similar to recognition accuracy when feature vector of 19 MFCCs is used. Also recognition accuracy in tests on SD 2 and CHAINS was increased. In tests 9 and 2C recognition accuracy was increased by a few percent compared to the best achieved accuracy in previous set of tests. Recognition accuracy on SD 2 has maximum in test 9 (Fig. 10), 36 of 37 speakers were correctly recognized. With respect to the case when feature vector of first 19 MFCCs is used, maximum of average recognition accuracy is increased by few percent, from 78.38% to 82.66%. Recognition accuracy in test 1C is also increased but is smaller than accuracy in test 2C, since recordings in test 1C are much shorter.

IV. Conclusion

Based on comparing of diagonal elements in the same places of training and test covariance matrix of the same speaker, it is possible to determine which of them are significantly different. Each of diagonal elements in training or test covariance matrix, $\Sigma_{i,i}^{training}$ or $\Sigma_{i,i}^{test}$, where i represents the index or the ordinal number of the observed MFCC, $i \in \{0,1,...,d-1\}$, represents variance i.e. the expected contained energy in appropriate MFCC. Significant distinction between $\Sigma_{i,i}^{training}$ and $\Sigma_{i,i}^{test}$ reflects the fact that appropriate MFCC is dependent of speech characteristics which are non speaker specific. Therefore this MFCC can be observed as non speaker discriminative. That is why these MFCCs can be discarded from feature set. In experiments on SD 1, SD 2 and CHAINS this property was noted for zeroth and 19th MFCC. By discarding of these MFCCs the recognition accuracy was improved.

With respect to equality 2 used for calculation of MFCCs it is evident that MFCCs depend on the assumed shape of auditory critical bands. Since auditory critical band is related to the masking phenomena, its form should be descending around central position. First approximation by the use of linear functions results in triangular auditory critical bands. Related to the rectangular auditory critical bands, application of the triangular auditory critical bands results in an improvement of recognition accuracy.

The central component in auditory critical band is masking component and components around her are masked. Guided by the tendency to higher reduce masked components it is necessary to observe approximation of auditory critical bands by the nonlinear functions with higher slope compared to the linear function. As this function in this paper was tested exponential function. Application of exponential auditory critical bands based on the lower part of exponential function was resulted in higher recognition accuracy compared to the rectangular and triangular auditory critical bands.

Acknowledgment

We thank AlfaNum Ltd. from Novi Sad on the speech databases used in experiments described in this paper.

References

[1] F. de Leon, K. Martinez, “Enhancing timbre model using MFCC and its time derivatives for music similarity estimation,” in Proc. 20th European Signal Processing Conference (EUSIPCO 2012), Bucharest, Romania, August 27 – 31, 2012, pp. 2005-2009.

[2] T. Kinnunen, H. Li, “An overview of text-independent speaker recognition: From features to supervectors,” Speech Communication, vol. 52, no. 1, pp. 12-40, 2010.

[3] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacrétaz, and D. A. Reynolds, “A Tutorial on Text-Independent Speaker Verification,” EURASIP Journal on Applied Signal Processing 2004:4, pp. 430-451, 2004.

[4] J. P. Campbell, Jr., “Speaker recognition: a tutorial,” Proceedings of the IEEE, Vol. 85, No. 9, pp. 1437-1462, 1997. [Online]. Available: http://dx.doi.org/10.1109/5.628714

[5] M. M. Dobrović, V. D. Delić, N. M. Jakovljević, I. D. Jokić, “Comparison of the Automatic Speaker Recognition Performance over Standard Features,” in Proc. of the 2012 IEEE 10th Jubilee International Symposium on Intelligent Systems and Informatics (SISY 2012), Subotica, Serbia, 20 – 22 September 2012, pp. 341 – 344. [Online]. Available: http://dx.doi.org/10.1109/SISY.2012.6339541

[6] V. Tiwari, “MFCC and its applications in speaker recognition,” International Journal on Emerging Technologies, vol. 1(1), pp. 19-22, 2010.

[7] C. Ittichaichareon, S. Suksri, and T. Yingthawornsuk, “Speech Recognition using MFCC,” in Proc. International Conference on Computer Graphics, Simulation and Modeling (ICGSM'2012), July 28-29, 2012 Pattaya (Thailand), pp. 135-138.

[8] S. D. Dhingra, G. Nijhawan, P. Pandit, “Isolated speech recognition using MFCC and DTW,” International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol. 2, Issue 8, August 2013, pp. 4085-4092.

[9] D. Neiberg, K. Elenius and K. Laskowski, “Emotion Recognition in Spontaneous Speech Using GMMs,” in INTERSPEECH 2006 – ICSLP, 17-21 September 2006, Pittsburg, Pennsylvania, pp. 809-812.

[10] B. Panda, D. Padhi, K. Dash, Prof. S. Mohanty, “Use of SVM Classifier & MFCC in Speech Emotion Recognition System,” International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 3, March 2012, pp. 225-230.

[11] Y. Attabi, M. J. Alam, P. Dumouchel, P. Kenny, D. O'Shaughnessy, “Multiple windowed spectral features for emotion recognition,” Published in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 26-31 May 2013, Vancouver, BC, pp. 7527-7531.

[12] D. Wu, B. Li, and H. Jiang, “Normalization and Transformation Techniques for Robust Speaker Recognition,” Source: Speech Recognition, Technologies and Applications, Book edited by: France Mihelič and Janez Žibert, ISBN 987-953-7619-29-9, pp. 550, 311-330, November 2008, I-Tech, Vienna, Austria.

[13] I. Jokić, S. Jokić, Z. Perić, M. Gnjatović, V. Delić, “Influence of the Number of Principal Components used to the Automatic Speaker Recognition Accuracy,” Electronics and Electrical Engineering, Kaunas: Technologija, 2012, No. 7(123), pp. 83-86.

[14] B. Šalna, J. Kamarauskas, “Evaluation of Effectiveness of Different Methods in Speaker Recognition,” Electronics and Electrical Engineering, Kaunas: Technologija, 2010, No. 2(98), pp. 67-70.

[15] S. Molau, M. Pitz, R. Schlüter, and H. Ney, “Computing Mel-Frequency Cepstral Coefficients on the Power Spectrum,” in Proc. International Conference on Acoustic, Speech and Signal Processing, Salt Lake City, UT, June 2001, Vol. 1, pp. 73-76.

[16] C. Lee, D. Hyun, E. Choi, J. Go, and C. Lee, “Optimizing Feature Extraction for Speech Recognition,” IEEE Transactions on Speech and Audio Processing, Vol. 11, No. 1, January 2003, pp. 80-87. [Online]. Available: http://dx.doi.org/10.1109/TSA.2002.805644

[17] R. F. Lyon, A. G. Katsiamis, E. M. Drakakis, “History and Future of Auditory Filter Models,” Proceedings of 2010 IEEE International Symposium on Circuits and Systems (ISCAS 2010), May 30 – June 2 2010, Paris, France, pp. 3809-3812. [Online]. Available: http://dx.doi.org/10.1109/ISCAS.2010.5537724

[18] M. Siafarikas, T. Ganchev, N. Fakotakis, G. Kokkinakis, “Wavelet Packet Approximation of Critical Bands for Speaker Verification,” International Journal of Speech Technology, ISSN 1381 – 2416, vol.10, no.4, 2007, Springer, pp. 197-218.

[19] A. C. den Brinker, “An interpretation of the auditory critical bands using a local Kautz transformation,” in Proc. ProRISC 8th anual Workshop on Circuits, Systems and Signal Processing, Mierlo, The Netherlands, 27-28 Nov. 1997, pp. 83-88.

[20] B. R. Wildermoth, “Text-Independent Speaker Recognition Using Source Based Features,” pp. 19-20, M. Phil. Thesis, Griffith University, Brisbane, Australia, Janury 2001.

[21] F. Cummins, M. Grimaldi, T. Leonard, and J. Simko, “The CHAINS speech corpus: CHaracterizing INdividual Speakers,” in Proc. of SPECOM, 2006, pp. 1-6.