Mel-Frequency Cepstral Coefficients as Features for Automatic Speaker Recognition

Ivan D. Jokić, Stevan D. Jokić, Vlado D. Delić, *Member, IEEE*, and Zoran H. Perić, *Member, IEEE*

Abstract **— Automatic speaker recognizer can be based on the use of mel-frequency cepstral coefficients as speaker features. Mel-frequency cepstral coefficients depend on energy inside considered auditory critical bands. These auditory critical bands model masking phenomena. Application of triangular auditory critical bands results in better recognition accuracy with respect to the case when rectangular auditory critical bands are applied. Recognition accuracy when exponential auditory critical bands are applied outperforms recognition accuracy of automatic speaker recognizer when triangular or rectangular auditory critical bands are applied. Application of transformation on elements of speaker model, which target decreasing of difference between testing and training models of the same speaker, can increase recognition accuracy.**

Keywords **— Automatic speaker recognition, auditory critical bands, covariance matrix, exponential auditory critical bands, mel-frequency cepstral coefficients, multidimensional Gaussian distribution.**

I. Introduction

Automatic speaker recognition deals with the problem of speaker recognition through speech considered. The main topic in automatic speaker recognition is to solve a unique representation of voice. Speech signal is consequence of semantic, linguistic, articulatory and acoustic transformations, which occur at different levels [1]. Therefore speech signal contain information about speaker who speaking, spoken textual content and emotional state of speaker. To achieve automatic recognition of one of these informations it is necessary to determine what voice property we must consider. For automatic speaker recognition one of property is timbre.

Timbre is consequence of harmonic structure of voice. Thus information about timbre is contained in spectral envelope of voice. Usually fast spectral estimation of voice can be made by Fast Fourier Transform (FFT) over appropriate short time speech segments. Mel-Frequency Cepstral Coefficients (MFCCs) are speech features, also used for speaker recognition [2], [3], [4], whose calculation is based on short time spectral analysis of speech signal by using FFT. These features also can be used for speech recognition [5], [6], when recognizer recognizes textual content of speech, and also for emotion recognition in speech [7], [8], [9]. All characteristics of speech, spoken textual content, speaker characterization or emotion characterization, can be viewed as changes in spectral content.

Calculation of MFCCs by using short time spectral analysis implies that for every analyzed speech segment exist a lot of mel-frequency cepstral feature vectors. Modeling of speakers usually can be done by stochastic models such as Gaussian Mixture Models (GMMs) or Hidden Markov Models (HMMs). Both models are based on using of Gaussian multidimensional distribution.

In next sections of this paper description of the recognizer and results of recognition will be given.

II. Realization of Automatic Speaker Recognizer and Description of Experimental Environment

Automatic speaker recognizer used for experiments in this paper is based on the use of MFCCs as feautures of speech. Before calculation of MFCCs speech signal is windowed by Hanning window function whose duration is about 23 ms. Time shift between two adjacent Hanning windows applied on speech signal is about 8.33 ms. MFCCs are determined by application of discrete cosine transform on energy inside observed auditory critical bands using equality [10]:

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

E_k represents energy in the k^{th} auditory critical band and d is number of MFCCs calculated i.e. it is dimensionality of feature vectors. There is 20 auditory critical bands width of 300 mel and mutually shifted by 150 mel. Therefore in initial experiments feature vectors are 20-dimensional and consisted of zeroth and first 19 MFCCs.

Concept of auditory critical bands is connected with audio masking phenomena. Shape of auditory critical bands, usually used, is rectangular or triangular [11], [12], [13]. Therefore our experiments are started with rectangular shape of auditory critical bands.

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.

Ivan D. Jokić is with the Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia (phone: 381-64-3526245; e-mail: ivan.jokic@uns.ac.rs).

Stevan D. Jokić is with the Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia; (phone: 381-64-2829906, e-mail: stevan.jokic@uns.ac.rs).

Vlado D. Delić is with the Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia (phone: 381-21-4852533; e-mail: vdelic@uns.ac.rs).

Zoran H. Perić is with the Faculty of Electronic Engineering, University of Niš, Aleksandra Medvedeva 14, 18000 Niš, Serbia (e-mail: zoran.peric@elfak.ni.ac.rs).

978-1-5090-0055-5/15/$31.00 ©2015 IEEE

Speaker modeling was based on the assumption that feature vectors are distributed in accordance with appropriate Gaussian multidimensional distribution. Covariance matrix of feature vectors determines the shape of this distribution, therefore speaker modeling was done by appropriate covariance matrix. Before calculation of covariance matrix feature vectors are grouped into matrix of feature vectors, $X = [x_1 \quad x_2 \quad \ldots \quad x_n]$, each of n feature vectors fills one of n columns of matrix X. Then model is calculated by matrix equality:

$$\Sigma = \frac{1}{n-1} \cdot (X - \mu) \cdot (X - \mu)^T , \qquad (2)$$

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

Measure of distinguishing between the model of the i^{th} speaker and considered test model "*test*" was defined by:

$$m(i, test) = \frac{1}{d^2} \cdot \sum_{j=1}^{d} \sum_{k=1}^{d} \left| \Sigma_i(j,k) - \Sigma_{test}(j,k) \right| . \qquad (3)$$

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

$$m(i, test) < m(j, test), \quad \forall j \in \{1,2,\ldots,N\} \setminus \{i\}. \qquad (4)$$

Algorithm for automatic speaker recognition used in this paper target to model timbre. Therefore it can be expected that its performance are independent of language in which it applies. Without loss of generality tests of recognition in this paper have been performed over two speech databases in Serbian. First speech database (SD1) was recorded in studio conditions and contains utterances produced by 121 speakers. All utterances in this speech database were recorded only once and they are classified in three groups: "Digits", "Names" and "Words". For each of speakers these groups contain following contents:

- "Digits" – two utterances containing words that correspond to two sequence of 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 from group "Digits", test 3 on recordings in "Names" and tests 4 to 14 on recordings in group "Words".

Second speech database (SD2) was recorded in office conditions. Recordings in this speech database contain utterances of four digits. In total this speech database contains utterances of 44 speakers. Each of them was recorded in multiple sessions. Most of speakers are recorded in 10 sessions and each of them contains 12 recordings. Some of speakers are recorded in smaller number of sessions and therefore 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, numbered by 1 to 12.

III. Description of Experiments and Results

In initial experiments feature vectors contain zeroth and first 19 MFCCs and auditory critical bands are of rectangular shape (Fig. 1). In that case recognition accuracy on SD1 over group "Words" is between 70% and 95%, and often has the largest values in tests on SD1. Recognition accuracy in tests on group Digits is about 68% and 80% and the smallest accuracy is on group Names, about 31%. Recognition accuracy over SD2 is more than twice, around 40%, less than recognition accuracy over groups Digits and Words in SD1. It is because the average duration of recordings in SD2 is smaller with respect to average duration of recordings in SD1. Duration of recordings in SD2 is of a few seconds, often around 5 seconds. Therefore these recordings have less number of vowels, in comparing to recordings from SD1. In tendency of tracking and if possible to ensure that test and training model of the same speakers have as less as differentiation, comparison of elements on same places in training and test matrices of same speakers is prepared. It was noted large distinguishing of the elements of model on the first positions. First element in covariance matrix represents variance of zeroth MFCC, $\Sigma_{0,0}$. Therefore this element depends on textual content of speech and present noise. Since this element is not directly dependent of considered speaker, in the next step of tests the first element of covariance matrices is set to zero and this improves accuracy of recognition.

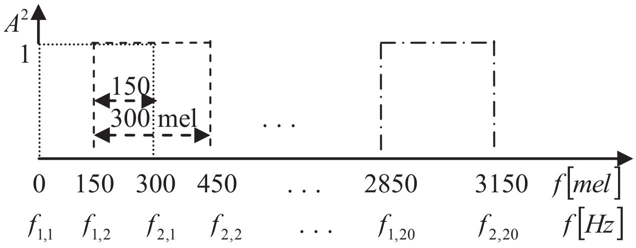


Fig. 1. Arrangement of 20 applied rectangular auditory critical bands.

By discarding first element from covariance matrix recognition accuracy is increased, in average by about few percent, in all 14 experiments on SD1. Also recognition accuracy in tests on SD2 is increased. Impact of the zeroth MFCC also is present on elements of first row and first column of covariance matrices. Because the zeroth MFCC is discarded from the feature set. Experiments were repeated for feature vectors of first 19 MFCCs. In most tests the recognition accuracy is increased and therefore 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. In some tests from this group accuracy is reached values of 100%. Recordings in group Names are shortest and therefore recognition accuracy in that case is smallest, about 40%. By discarding the first element from covariance matrix average recognition accuracy on SD2 is increased from about 37% to about 46%, and when zeroth MFCCs is discarded from feature vector average accuracy is additionally increased on about 55%.

Representation of masking implies approximation of auditory critical bands with help of descending functions. As descending functions often linear functions are used in the form of triangular auditory critical bands [2], [11], [13]. In this work triangular auditory critical bands are defined by:

$$A^2(k)=\begin{cases}\dfrac{2}{k_{2,n}-k_{1,n}}\cdot\left(k-k_{1,n}\right), & k_{1,n}\le k\le\dfrac{k_{1,n}+k_{2,n}}{2},\\[2ex] \dfrac{2}{k_{1,n}-k_{2,n}}\cdot\left(k-k_{2,n}\right), & \dfrac{k_{1,n}+k_{2,n}}{2}<k\le k_{2,n},\end{cases}\qquad(5)$$

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

$$k=N\cdot\frac{f}{f_s}\quad\wedge\quad 0\le k\le N-1,\qquad(6)$$

f_s=22050 Hz is sampling frequency and N=512 dots is length of speech frame whose duration is about 23 ms.

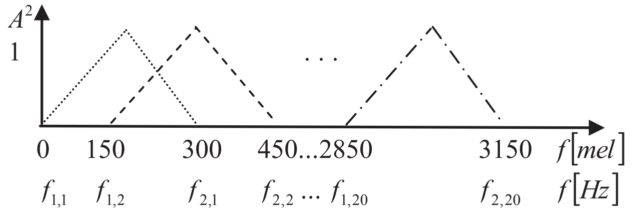


Fig. 2. Arrangement of 20 applied triangular auditory critical bands.

MFCCs are directly dependent on log energy in observed auditory critical bands (equality 1). SD2 is recorded in office conditions and have some amount of noise. Primarily due to the influence of noise in SD2 tests of automatic speaker recognition were conducted and when energy correction in observed auditory critical bands is applied. In SD1 we can eventually speak about impact of the speaker articulation. Therefore introducing of energy correction is motivated by the assumption, if noise or some other irregular appearance is present in frame of speech signal then he is most expressed in auditory critical band with minimal energy. Energy correction for each observed speech frame was applied through next three steps:

- calculation of log energy, log(E_k), for each auditory critical band, k={1,2,...,20}, of observed speech frame,
- determination of the minimum log energy, log(E_k)$_{min}$=min{log(E_k)}, k={1,2,...,20},
- calculation of the new log energy value in observed auditory critical band,

$$\log\left(E_k\right)_{new}=\log\left(E_k\right)-\min\left\{\log\left(E_k\right)\right\}.\qquad(7)$$

Application of triangular auditory critical bands and energy correction was increased recognition accuracy especially in the case of tests on SD2 (Table II), because in SD2 has present noise.

Shift from rectangular to triangular auditory critical bands increases recognition accuracy (Table I and Table II). Increase of recognition accuracy by application of triangular auditory critical bands reflects fact that auditory critical bands should have decreasing shape around central position. In the case of triangular auditory critical bands decreasing shape is characterized by slope of linear function used. Achievement of higher slope is possible by approximation of auditory critical bands by exponential function of type $y=e^x$.

With respect to the point x=0 where exponential function $y=e^x=1$ it is possible to distinguish between the two parts of exponential function, to the left or to the right from the point x=0 i.e. lower or upper part with respect to the value of y=1 [14]. In some tests application of exponential auditory critical bands based on lower part of exponential function give better recognition accuracy with respect to application of auditory critical bands based on upper part of exponential function [14], but in some cases recognition accuracy is similar. Exponential auditory critical bands based on the lower part of the exponential function are defined as:

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

$k_{c,n}=\dfrac{k_{1,n}+k_{2,n}}{2}$ is central discrete frequency (equality 6) of n^{th} auditory critical band and n={1,2,...,20} (Fig. 3), s is steepness factor in experiments varied through vales s=1 and s=2. Exponential auditory critical bands with higher steepness factor better mask components in surrounding of central component and therefore in table II are results only for s=2. As is expected, application of exponential auditory critical bands gives better recognition accuracy with respect to recognition accuracy when triangular auditory critical bands were applied (Table I and Table II).

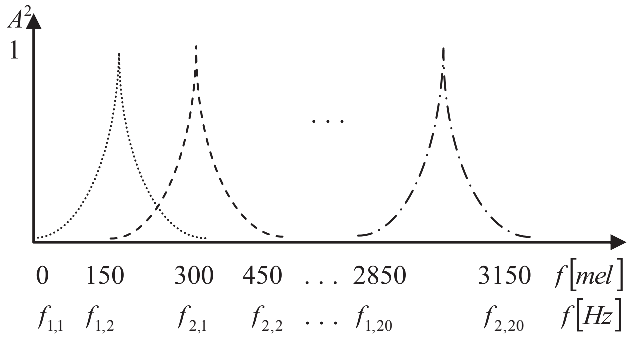
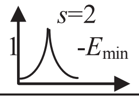
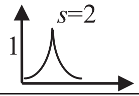
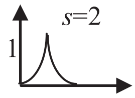
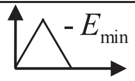
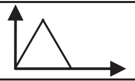
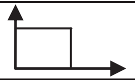


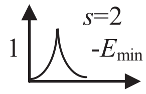
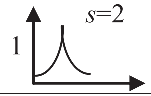
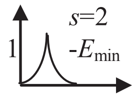
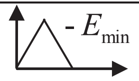
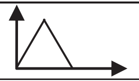
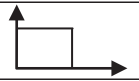
Fig. 3. Arrangement of 20 applied exponential auditory critical bands based on the lower part of exponential function.

Table I. Comparing of percentage recognition accuracy on SD1 depending on: test (1-14), number of MFCCs, type of auditory critical bands and energy inside them.

Type of crit. band	1	2	3	4	5	6	7	8	9	10	11	12	13	14
18 MFCCs														
	87.6	91.7	57.0	96.7	99.2	98.3	97.5	99.2	99.2	98.3	99.2	98.3	100	99.2
	87.6	92.6	60.3	95.9	99.2	98.3	95.9	99.2	99.2	98.3	97.5	97.5	100	96.7
19 MFCCs														
	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

Application of exponential auditory critical bands on SD2 also increases recognition accuracy with respect to accuracy when triangular auditory critical bands were applied, and these are higher on SD2.

Table II. Comparing of percentage recognition accuracy on SD2 depending on: test (1-12), number of MFCCs, type of auditory critical bands and energy inside them.

Type of crit. band	1	2	3	4	5	6	7	8	9	10	11	12
18 MFCCs												
	72.97	78.38	86.49	81.08	86.49	81.08	91.89	83.78	97.30	81.08	70.27	81.08
	62.16	70.27	81.08	72.97	75.68	75.68	91.89	78.37	97.30	75.68	75.68	78.38
19 MFCCs												
	62.16	70.27	81.08	67.57	75.68	70.27	86.49	89.19	94.59	83.78	78.38	81.08
	51.35	56.76	70.27	64.86	70.27	62.16	83.78	83.78	86.49	81.08	67.57	67.57
	45.95	51.35	51.35	54.05	62.16	43.24	86.49	64.86	78.38	75.68	70.27	64.86
	40.54	40.54	54.05	48.65	54.05	43.24	64.86	62.16	70.27	64.86	62.16	56.76

By comparing test and training models of the same speakers when feature vector of first 19 MFCCs is used, significant difference between $\Sigma_{19,19}$ elements is noted. Difference between $\Sigma_{19,19}$ elements is several times smaller compared to difference between $\Sigma_{0,0}$ elements.

MFCCs depend on energy in speech signal (equality 1). As consequence MFCCs depend on text and emotion. If we assume that emotion in speech is constant, then MFCCs depend on textual content of speech. In an ideal case it is desirable that test and training model of the same speaker do not have or have very small differences. This can be achieved by discarding elements of model where training and test models of the same speakers have large differences. Previous results prove that by discarding $\Sigma_{0,0}$ and $\Sigma_{19,19}$ elements from speaker models or by discarding all elements of speaker models that represent covariance with zeroth or 19th MFCCs, recognition accuracy was increased. Discarding of some elements of speaker model is equivalent to multiplication of these elements with zero. In this approach elements in covariance matrices are multiplied by one or by zero. On that way is reduced the resolution i.e. the number of elements of model. For more precisely speaker model it is desirable to keep all elements of model but with assumption that not all of elements with the same importance.

Determination of the measure of time variability of elements of model in some test can be done by definition of matrix of differences:

$$D_t(i,j)=\sum_{n=1}^{N}\left|\Sigma^{training}_{(i,j)(n)}-\Sigma^{test}_{(i,j)(n)}\right|, \qquad (9)$$

where t indices the ordinal number of test considered in speech database, for example $t\in\{1,2,\ldots,14\}$ for tests on SD1 or $t\in\{1,2,\ldots,12\}$ for test on SD2, i and j represents indices of elements in training and test covariance matrices and N is number of speakers in speech database. Based on values in matrix of differences it is possible to determine weights for elements of model by next rule:

$$D_t(i,j)\geq D_t(k,l) \;\Rightarrow\; W(i,j)\leq W(k,l), \qquad (10)$$

and resultant element of speaker model is calculated as multiplication:

$$\Sigma^{resul\tan t}(i,j)=\Sigma(i,j)\cdot W(i,j). \qquad (11)$$

It is done one experiment for intuitively determined weighted factors and ranges of their application, for case of 20 rectangular auditory critical bands (Fig. 1) and feature vector consisted of zeroth and first 19 MFCCs (Fig. 4):

$$D_t(i,j)>\frac{\max}{5} \;\Rightarrow\; W(i,j)=0.05, \qquad (12.1)$$

$$\frac{\max}{10}<D_t(i,j)\leq\frac{\max}{5} \;\Rightarrow\; W(i,j)=0.3, \qquad (12.2)$$

$$\frac{\max}{20}<D_t(i,j)\leq\frac{\max}{10} \;\Rightarrow\; W(i,j)=0.6, \qquad (12.3)$$

$$\frac{\max}{40}<D_t(i,j)\leq\frac{\max}{20} \;\Rightarrow\; W(i,j)=1.0, \qquad (12.4)$$

$$D_t(i,j)\leq\frac{\max}{40} \;\Rightarrow\; W(i,j)=1.9. \qquad (12.5)$$

Borders are defined with respect to the largest value, *max*, in appropriate matrix of differences.

As is evident from Fig. 4 recognition accuracy has been significantly improved by application of weights. Test 1 was repeated for the case when 20 exponential critical bands based on the lower part of exponential function were applied and energy correction (equality 7) is also applied. By application of transformation (equalities 12.1 – 12.5) recognition accuracy was increased from about 68% to about 90%. Resultant accuracy in this case is greater than recognition accuracy when feature vector contain 18 MFFCs (Table I).

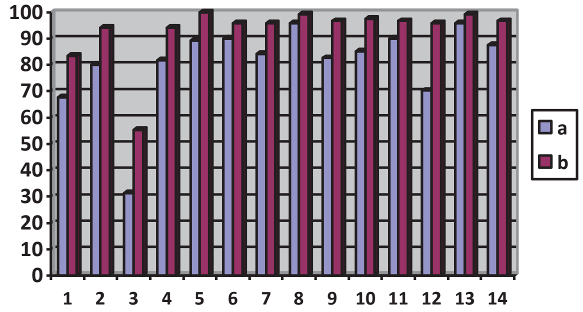


Fig. 4. Percentage accuracy of speaker recognition over SD1 depending on the test file used and model applied (20 rectangular auditory critical bands are used): a – full covariance matrix for feature vector which contains zeroth and first 19 MFCCs, b – same full covariance matrix with applied weights by rules 12.1 to 12.5.

If test set is unknown then it is necessary to apply weights on the basis of elements of model. If elements of model have higher values then it can be expected that difference between test and training model on that place have high value. To reduce difference it is necessary to apply some nonlinear function on elements of model. This function must decrease or limit high values in speaker model. Sigmoid function, $f(x)=\dfrac{1}{1+e^{-x}}$, seems suitable for this purpose. Therefore test 3 in SD 1 for feature vector of 18 MFCCs was repeated for the case when sigmoid function was applied on elements of model, and recognition accuracy has been increased (Fig. 5).

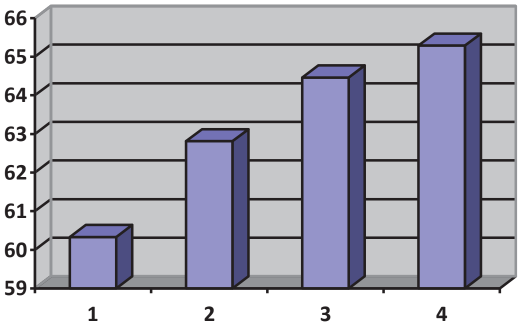


Fig. 5. Percentage recognition acccuracy in test 3 on SD1 depending on the setting parameters of the applied model based on feature vectors of 18 MFCCs and 20 exponential auditory critical bands based on the lower part of exponential function with steepness factor *s*=2 (Basic model): 1 – Basic model, 2 – Basic model and sigmoid applied on all elements of model, 3 – Basic model and sigmoid applied on elements of model which are higher of *max*/2, 4 – Basic model and sigmoid applied on elements of model which are higher of *max*/1.7. *Max* represents the maximum value in observed Basic model.

IV. Conclusion

During calculation of MFCCs it is possible apply some modifications to achieve higher recognition accuracy of automatic speaker recognizer. To achieve reliable automatic speaker recognition it is necessary to decrease difference between test and training models of the same speakers. All methods and transformations described in this paper, the use of exponential auditory critical bands and weighting of elements of model, on some way discard some irrelevant information from speech signal, irrelevant from the viewpoint of automatic speaker recognition. These results encourages research towards the very little time variable models of the same speakers in same emotional state.

ACKNOWLEDGMENT

Thanks to AlfaNum Ltd. from Novi Sad for the speech databases used in experiments described in this paper.

REFERENCES

[1] J. P. Campbell, Jr., "Speaker recognition: a tutorial," *Proceedings of the IEEE*, Vol. 85, No. 9, 1997, pp. 1437-1462.

[2] 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*, 2004, pp. 430-451.

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

[14] I. Jokić, "Analysis of mel-frequency cepstral coefficients as features used for automatic speaker recognition," *PhD thesis (Doctoral dissertation)*, mentor: PhD Vlado Delić, Faculty of Technical Sciences, University of Novi Sad, 2014, pp. 54-64. (In Serbian)