

MEL-FREQUENCY CEPSTRAL COEFFICIENTS AND SPECTRUM BASED ADDITIONAL FEATURES IN AUTOMATIC SPEAKER RECOGNITION

Ivan Jokić¹, Stevan Jokić², Vlado Delić³, Zoran Perić⁴

¹Faculty of Economics and Engineering Management in Novi Sad, Srbija

²Alfa BK University, Faculty of Information Technology, Beograd, Srbija

³University of Novi Sad, Faculty of Technical Sciences, Novi Sad, Srbija

⁴University of Niš, Faculty of Electronic Engineering, Niš, Srbija


ORCID iDs: Ivan Jokić


Stevan Jokić


Vlado Delić

Zoran Perić

 <https://orcid.org/0009-0008-0083-7675>

 <https://orcid.org/0000-0003-4432-0172>

 <https://orcid.org/0000-0002-4558-9918>

 <https://orcid.org/0000-0002-8267-9541>

Abstract. *The efficiency of the proposed automatic speaker recognizer is evaluated using two speech databases. The feature vector consists of 21 mel-frequency cepstral coefficients (MFCCs), along with up to three additional features derived from the amplitude spectrum. The additional features are calculated based on the logarithm of the energy around the appropriate local maximum in the spectrum, the frequency of that maximum, and the logarithm of the energy of the maximum component in the spectrum across all frames of the observed signal. The speaker identification procedure for a closed set of speakers is tested on the Solo section of the CHAINS database and a speech database with expressed emotions, developed within the S-ADAPT project. The achieved maximum mean recognition accuracies are 97.11%, on the CHAINS database, using a feature vector of 21 MFCCs and two additional features, and 98.65% on neutral speech, as well as 98.72% on the entire database, for the S-ADAPT database, using a feature vector of 21 MFCCs.*

Key words: *accuracy, audio recording, human voice, speaker recognition, spectral analysis*

1. INTRODUCTION

Determination of speech features is one of key segments in construction of automatic speaker recognizer. The speaker features used affect the recognition accuracy of the speaker recognizer. Timbre is one of the fundamental characteristics of a speaker's,

Received December 02, 2024; revised June 30, 2025; accepted July 09, 2025

Corresponding author: Ivan Jokić

University Business Academy in NS, Faculty of Economics and Engineering Management in Novi Sad, Serbia

E-mail: ivan.jokic@fimek.edu.rs

representing the subjective perception of the listener. To achieve automatic speaker recognition, it is necessary to extricate information about the speaker's timbre.

Timbre is a consequence of the harmonic spectral content of speech, which is contained in the spectral envelope of the speech signal. Mel-frequency cepstral coefficients (MFCCs) follow the spectral envelope of the speech signal [1, 2]. This means that MFCCs contain information about the spectral envelope of the speech signal. MFCCs are short-term features of speech, they are usually calculated for speech frames of around 25 ms. Timbre, as subjective perception of listener, is long-term feature of speech. Therefore, to get the information about timbre of the speaker's speech it is necessary to apply adequate transformation to previously calculated vectors of MFCCs and represent the determined set of feature vectors in a unique way. In [3] vector quantization (VQ) is applied. Gaussian mixture model (GMM) is the type of probabilistic model that can be used for modeling of MFCC feature vectors, results in [4] show that MFCC-GMM speaker recognizer achieves higher recognition accuracy with respect to MFCC-VQ speaker recognizer. In [5] is shown that Gaussian Mixture Model-Universal background model (GMM-UBM) gives a lower equal error rate (EER) compared to GMM. Nowadays the procedures are oriented to processing of a large amount of short-term feature vectors into long-term feature, and in searching for correlations, some kind of connections or relationship, between short-term MFCC feature vectors. In [6] this is achieved by a multilayer perceptron feed-forward artificial neural network (MLPFANN) which is trained by backpropagation. Speaker recognition based on use of MFCCs from distant speech, in [7], is done with UBM i-vector (intermediate-vector) based system and with deep neural network (DNN) speaker embedding based system. This work shows the robustness of a DNN speaker embedding based speaker recognition system for distant speech data. Robust speaker recognition is done by x-vectors DNN embeddings in [8]. Emotion invariant speaker embeddings for speaker identification, i-vectors based approach, is done in [9]. Results in this paper show that when training and testing speech are in different emotions, accuracy is higher when speech in neutral emotion is used in training. Combination of MFCCs with convolutional neural network (CNN) as feature extractors and K-Nearest Neighbor (KNN) classifier is used in [10]. A supervised learning approach for speaker recognition, based on use of MFCCs and CNN, is proposed in [11]. Constrained CNN model for speaker recognition, which performs classification by processing speech spectrograms, is presented in [12]. Study the optimal design of CNN networks for speaker identification and clustering when simple spectrograms are used as input to CNN is investigated in [13]. In [14], system based on deep learning and CNN is compared with recognizer based on MFCCs and support vector machine (SVM). Introduced a point-to-point investigation of MFCCs technique in [15] emphasize easily recognize the voice with the help of MFCC techniques. Speech based security system oriented to identifying Arabic speakers, proposed in [16], uses MFCCs and radial basis function neural network (RBFNN). Combination of MFCCs, calculated for voiced frames, and inverted MFCCs (IMFCCs) calculated for unvoiced frames is described in [17].

In addition to MFCCs, other features derived from the speech spectrum are also used in the literature. These features are often combined with MFCCs to better utilize the information potential of the speaker's speech and to increase the efficiency of the speaker recognizer. Spectral subband centroids (SSC) [18-21], which represent the centroid frequency in each observed subband, are one such feature. Experiments in [18] show that MFCCs are not as good as SSCs under noisy conditions. A study of the characterization

of subband energy as a two dimensional feature, comprising Spectral Centroid Magnitude (SCM) and Spectral Centroid Frequency (SCF), was conducted in [19]. Addition of SSC features to feature vector which contains Linear Predictive Cepstral Coefficients (LPCC), in [20], is increased speaker recognition accuracy. To allow better adaptation to formant movements and other dynamic phenomena, in [21] is proposed to adapt the subband filter boundaries on a frame-by-frame basis using a globally optimal scalar quantization scheme. In [22] is proposed that SSCs are more apt for marginalization-based Missing Feature Theory (MFT). It was shown that speaker recognizer based on use of SSCs, MFT and diagonal-covariance GMM, is more robust to noise presence than speaker recognizer based on use of MFCCs and diagonal-covariance GMM. Normalized Dynamic Spectral Features (NDSF) [23] have been found to be more robust than cepstral features like MFCCs and Linear Predictive Cepstral Coefficients (LPCCs) in the case of sensor mismatch condition. The combination of MFCCs and non-linear Spectral Dimension (SD) features [24] results in better performance compared to a recognizer based solely on MFCCs. In [25-26], the combination of MFCCs with frequency modulation (FM) features increases the efficiency of speaker recognition. The combination of MFCCs and features derived by Unique Mapped Real Transform (UMRT) in [27] provides better accuracy of speaker recognition as compared to speaker recognizer based on use of MFCCs. Efficiency of MFCCs can be improved by adding additional features or by using other short-term acoustic features, alternative methods such as spectral centroids, group delay function, and integrated noise suppression can be useful for DNN-based automatic speaker verification system [28]. Research in [29] highlights the significant enhancement in computational speed and speaker recognition performance by applying adequate algorithms for feature optimization. It is done through fusion of features (MFCCs, SSCs, ...), dimension reduction employing principal component analysis (PCA) and independent component analysis (ICA), and feature optimization through genetic algorithm (GA) and marine predator algorithm (MPA).

Automatic speaker recognizer for closed set of speakers based on use of the feature vectors comprised of 21 MFCCs along with up to three additional features derived from amplitude spectrum is presented in this paper. By using covariance matrix as speaker model and training and testing on speech recordings of sentences, this speaker recognizer has small computational complexity. Impact of the proposed feature vectors to speaker identification accuracy in experiments on two speech databases is analyzed. One of the speech databases contains speaker recordings in different emotions. In addition to speaker recordings in neutral speech this speech database contains speaker recordings in four emotions: anger, joy, fear and sadness. Therefore, the impact of different emotions on speaker identification accuracy when training is performed by neutral speech also is examined in this paper. In the next parts of the paper, a description of the automatic speaker recognizer used in the experiments, a short description of the speech databases used, and the results of recognition and their discussion will be provided.

2. AUTOMATIC SPEAKER RECOGNIZER USED

Automatic speaker recognizer in experiments described in this paper work on speech signals whose frequency sampling is $f_s=44100$ Hz. The speech signal is assumed to remain stationary inside of the interval around 25 ms. This interval is usually used as the

frame duration for spectral analysis of a speech signal. Algorithm of fast Fourier transform (FFT) is used for efficient calculation of discrete Fourier transform (DFT) and amplitude spectrum of speech frames. It is necessary that frame duration N is in the form $N=2^x$, where x is a natural number, for the FFT to be applied. Therefore, feature vectors are calculated for speech frames of $N=1024=2^{10}$ samples, it is approximate 23.2 ms, mutually shifted by 368 samples or approximate 8.3 ms. Feature vector contains 21 MFCCs and one to three additional features derived from amplitude spectrum. MFCCs are basic speaker features used for experiments in this paper. Their calculation is based on cosine transform of logarithm of energy inside of fixed, $M=22$ frequency selective ranges, wide of 300 mel and mutually shifted by 150 mel. Denoting the estimation of energy inside of the m -th frequency selective range by E_m , MFCCs are determined by the equation:

$$c_n = \sum_{m=1}^{M=22} \log(E_m) \cdot \cos \left[\frac{\pi}{M} \cdot n \cdot \left(m - \frac{1}{2} \right) \right], \quad (1)$$

where $n = \{1, 2, \dots, 21\}$. Results in [30] show that maximum accuracy is achieved for maximal number of MFCCs with respect to the number of frequency selective ranges M . Here $n_{\max}=21$, since factor π is used in cosine transform $c_{22}=0$. Determination of boundaries of the selective ranges in Hz scale was done from equation:

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

The equation used between continuous frequency f and discrete frequency k in discrete spectrum calculated by DFT in $N=1024$ points is:

$$\frac{k}{N} = \frac{f}{f_s}. \quad (3)$$

The squares of the amplitude characteristics of 22 used frequency selective ranges is of sigmoid shape, this shape of square of amplitude characteristic give better results in [31] in comparison with exponential shape which is used in [30]:

$$A(k)^2 = \begin{cases} s_1(k - k_{c,m}), & k_{1,m} \leq k < k_{c,m}, \\ 1, & k = k_{c,m}, \\ s_1(k_{c,m} - k), & k_{c,m} < k \leq k_{2,m}. \end{cases} \quad (4)$$

Used sigmoid function is of shape $s_1(x) = \frac{1}{1 + e^{-0.5x}}$. Discrete frequencies $k_{1,m}$, $k_{2,m}$,

$k_{c,m} = \frac{k_{1,m} + k_{2,m}}{2}$, are the lower, upper, and the central frequency of the m -th frequency selective range. Estimation of energy inside the range of discrete frequencies $k_1 \leq k \leq k_2$ of m -th frequency selective range (4) was done by equation:

$$E_m = 2 \cdot \sum_{k=k_1}^{k_2} |X(k)|^2 \cdot A(k)^2, \quad (5)$$

$X(k)$ is the value of DFT of Hann windowed frame $x(n)$ on the discrete frequency k .

The basic feature vector comprised of 21 MFCCs is extended with one to three additional features. These features represent maximums of energy in appropriate spectral ranges of speech frame analyzed. They are determined by observing square of the amplitude of the maximum spectral component in observed spectral range and components in the nearest neighborhood of the maximum component. First additional feature is determined by observing all components of amplitude spectrum in range of discrete frequency 0 to $N/2-1$, considering the symmetry of the amplitude spectrum with respect to the discrete frequency $N/2$. Calculation of additional features is done in two steps. In first step it is calculated energy equivalent for i -th additional feature:

$$E_{\ln,i} = \ln\left(|X(k_{\max,i})|^2\right) + \sum_{\substack{j=-2 \\ j \neq 0}}^{j=2} \ln\left(|X(k_{\max,i} + j)|^2\right) \cdot s_2(-|j|), \quad (6)$$

whereby $s_2(x) = \frac{1}{1 + e^{-0.01 \cdot x}}$, $k_{\max,i}$ is discrete frequency of i -th amplitude maximum, it is frequency of amplitude maximum on the frequency range where is calculated i -th additional feature for frame $x(n)$. Finally, value of the i -th additional feature is determined in the second step:

$$e_i = \frac{E_{\ln,i} \cdot \frac{E_{\ln,i}}{\ln(|X(k)|_{\max,\text{all}})^2}}{k_{\max,i}}, \text{ if } k_{\max,i} = 0 \text{ then } k_{\max,i} = 1, \quad (7)$$

$(|X(k)|_{\max,\text{all}})^2$ represents the maximum of square of amplitude spectrum in all speech frames of signal observed. Each of additional features is determined in appropriate spectral range. First additional feature is calculated on the widest spectral range: $0 \leq k < N/2$, the second additional feature is calculated on the reduced spectral range: $0 \leq k < k_{\max,1} - 2 \cup k_{\max,1} + 10 < k < N/2$, the third additional feature is calculated on the more reduced spectral range: $0 \leq k < k_{\max,2} - 2 \cup k_{\max,2} + 10 < k < N/2$. Additional features capture information from speech spectrum that may not be fully captured by MFCCs alone. Use of these additional features increases the utilization of the information potential of amplitude spectrum. Discriminative capabilities of speaker recognizer can be increased and more comprehensive representation of speech signal is achieved. The feature extraction parameters are summarized in Table 1.

Feature vectors calculated for appropriate speech signal are grouped into matrix. The first feature vector is written into first column of matrix, the second feature vector into second column, and so on for all other feature vectors. For the matrix X of n feature vectors of dimension d , the appropriate model Σ is calculated as covariance matrix of matrix X : $\Sigma = \frac{1}{n-1} \cdot (X - \mu) \cdot (X - \mu)^T$, μ represents vector of mean values of matrix X .

Covariance matrix Σ of matrix of feature vectors X , describes appropriate recording of speaker's speech and represents long-term feature of observed speaker [32].

Table 1 Feature extraction parameters

Parameter	Description (Values)
Frame size	1024 samples, 23.2 ms, mutually shifted by 368 samples or 8.3 ms
Window type	Hann window, $w(n) = 0.5 \cdot (1 - \cos \frac{2 \cdot \pi \cdot n}{N-1})$, $0 \leq n \leq N-1$
Sampling frequency	$f_s=44100$ Hz
Frequency selective ranges	300 mel wide, mutually shifted by 150 mel, 22 ranges of sigmoidal square of amplitude characteristic
MFCCs c_n	Basic feature vector: $[c_1, c_2, \dots, c_{21}]^T$
Additional feature e_1	Range: $0 \leq k < N/2$
Additional feature e_2	Range: $0 \leq k < k_{\max 1} - 2 \cup k_{\max 1} + 10 < k < N/2$
Additional feature e_3	Range: $0 \leq k < k_{\max 2} - 2 \cup k_{\max 2} + 10 < k < N/2$

Diagonal elements of the covariance matrix, $\{\Sigma_{11}, \Sigma_{22}, \dots, \Sigma_{dd}\}$, give equivalent of measure of energy inside the first, the second, ..., d -th dimension of feature vectors. Elements outside of the main diagonal of the covariance matrix, $\{\Sigma_{i,j}\}, 1 \leq i, j \leq d, i \neq j$, give information about correlation between different dimensions, i.e. about energies in corresponding dimensional pairs. Transition from the matrix of feature vectors, matrix X , to covariance matrix Σ is consequence of the attention hope from observing only dimensions in feature vectors towards observing energies inside and between dimensions. Multiplication of matrix X of feature vectors by some other matrix represents and can be observed as some kind of attention hope from only dimensions in feature vectors towards some targeted information in observed feature vectors [33]. Appropriate matrix X of feature vectors and covariance matrix Σ as model are formed in training phase for each recording of speakers. Covariance matrices calculated for all recordings of the same speaker are named by the identity of that speaker. Automatic speaker recognizer during training phase forms its own memory, two textual files for archiving names of models and calculated models, for each speech recording used for training.

Automatic speaker recognizer realized in C++ [34], was projected to work on closed set of speakers. Therefore, before testing of recognizer on some speech recording of observed speaker it is necessary that name and reference model of this speaker exist in memory, in two appropriate textual files, of speaker recognizer. In phase of recognizer testing, for recording of test speech also was formed appropriate matrix of feature vectors and also appropriate covariance matrix as the model of test speech recording. In the phase of decision making the first step is to define and determine the difference between two models. The difference between observed test Σ_{test} and reference Σ_{ref} model, is calculated by equation:

$$r(\Sigma_{test}, \Sigma_{ref}) = \frac{1}{d^2} \cdot \sum_{i=1}^d \sum_{j=1}^d |\Sigma_{test}(i, j) - \Sigma_{ref}(i, j)|, \quad (8)$$

where d represents the dimensionality of feature vector. Similar difference is used in [32]. During testing the model of test speech, Σ_{test} , is compared with all reference models in recognizer memory. If recognizer has in memory R reference models then identity of the most similar reference model with respect to (8) is recognized speaker, the test speech “test” has the

identity of the i -th reference model if: $r(\Sigma_{test}, \Sigma_i) < r(\Sigma_{test}, \Sigma_j)$, $j \in \{1, 2, \dots, R\} \setminus \{i\}$, where Σ_{test} represents the covariance matrix of matrix X_{test} of feature vectors of test speech recording.

3. EXPERIMENT SETUP, RESULTS AND DISCUSSION

Testing of accuracy of speaker identification was done on the part named Solo of the speech database CHAINS (CHARacterizing INDividual Speakers) [35] and on the speech database recorded in the project: Speaker/style adaptation for digital voice assistants based on image processing methods (acronym: S-ADAPT). Both speech databases are recorded in WAV format, sampling frequency of 44100 Hz and quantization resolution of 16 bit/sample. The Solo part of the CHAINS speech database contains recordings of 36 speakers and is characterized by speaking style: “subjects simply read a prepared text at a comfortable rate”. For each speaker, 33 recordings are used. These recordings represent the pronunciation of individual sentences and are characterized by duration of usually approximately 2 to 3 seconds. Sentences are of different sizes. Duration of recordings of the shortest sentences is around 1 second, of the longest sentences is greater than 3 s.

Speech database recorded in the project S-ADAPT contains 11369 speech recordings of 55 speakers. This is speech database with expressed emotions in speech. For most speakers, in addition to recordings of sentences in neutral speech, there are also recordings of the same sentences in four emotional states: anger, joy, fear and sadness. Each speaker in each of emotions or at least in neutral emotional state recorded 62 mutually different sentences, 32 longer and 30 shorter sentences, in Serbian language. Emotions were simulated, it is enough that the neutral speech was changed in pronunciation on the manner how the speaker speaks in appropriate emotion. Duration of recordings is usually around 2 to 3 seconds, but exist recordings of duration around 4, 5 or 6 seconds. During recording of speech database each speaker recorded his own voice by using his own mobile phone. Application used for recording shows the sentence that speaker should pronounce in appropriate emotional state.

Testing is arranged in ten tests. Speech databases used, the Solo part of CHAINS and S-ADAPT, are divided in ten parts. Each part was used for testing and rest 9/10 of database was used for training. Four feature vectors are used in experiments. First feature vector contains only 21 MFCCs. Next three feature vectors contain 21 MFCCs and additional features, second feature vector contains 21 MFCCs and additional feature e_1 (21 MFCCs+ e_1), third feature vector contains 21 MFCCs and additional features e_1 and e_2 (21 MFCCs+ e_1e_2), fourth feature vector contains 21 MFCCs and additional features e_1 , e_2 , and e_3 (21 MFCCs+ $e_1e_2e_3$). Results of speaker identification accuracy in the tables are given in percentage values. Mean value and standard deviation of recognition accuracy are given for each set of ten tests when the same feature vector used. For the set of ten values of recognition accuracy $\{a_1, a_2, \dots, a_{10}\}$ the mean value was calculated by:

$$\mu = (a_1 + a_2 + \dots + a_{10})/10, \text{ and standard deviation by: } \sigma = \sqrt{\sum_{i=1}^{10} (a_i - \mu)^2 / 10}.$$

Results for CHAINS database are in the next table, testing was divided in ten tests: in most cases additional features in CHAINS database increased recognition accuracy. In first column of table was indicated description (Descri) of that experiment: which recordings are used for testing (Test set) and number of reference models obtained in

appropriate training. In used part of CHAINS database each speaker is represented by 33 recordings. Therefore, in seven of ten tests testing is done on three recordings and in three tests testing is done on four recordings. Maximum value of accuracy in each experiment achieved with minimum features is shaded in lighter gray in the table. In six experiments maximum of accuracy also was repeated for feature vector with greater number of features and these values are shaded in darker gray, because that realization of recognizer is of lower efficiency.

Table 2 Results for CHAINS

Test set/Descri	21MFCCs	+e1	+e1e2	+e1e2e3
s01,s02,s03; 1080 mod.	88.89%	88.89%	89.81%	89.81%
s04,s05,s06; 1080 mod.	87.96%	90.74%	94.44%	93.52%
s07,s08,s09,s10; 1044 mod.	92.36%	95.14%	93.75%	94.44%
s11,s12,s13,s14; 1044 mod.	94.44%	95.14%	97.22%	96.53%
s15,s16,s17; 1080 mod.	99.07%	99.07%	99.07%	99.07%
s18,s19,s20; 1080 mod.	99.07%	100.00%	100.00%	100.00%
s21,s22,s23; 1080 mod.	97.22%	97.22%	98.15%	98.15%
s24,s25,s26; 1080 mod.	98.15%	99.07%	100.00%	100.00%
s27,s28,s29; 1080 mod.	100.00%	99.07%	100.00%	100.00%
s30,s31,s32,s33; 1044 mod.	97.22%	97.92%	98.61%	97.92%
mean accuracy	95.44%	96.23%	97.11%	96.94%
stan. deviation	4.13%	3.59%	3.22%	3.22%
Mean error	4.56%	3.77%	2.89%	3.06%

When we compare results in different experiments it is evident that recognition accuracy in first four experiments is below 95%, when test recordings are from next sets: {s01, s02, s03}, {s04, s05, s06}, {s07, s08, s09, s10}, {s11, s12, s13, s14}, and when feature vector of 21 MFCCs is used. Sentences from the first four experiments are shorter with respect to sentences from next six experiments. In next six experiments, for the same feature vector, the recognition accuracy is higher than 97%. In experiments when test sets are: {s15, s16, s17} and {s18, s19, s20} recognition accuracy was greater than 99%; and when test set is {s27, s28, s29} recognition accuracy is 100%. Sentences s01, s02, ..., s33, are of different textual content and size. This indicates that recognition accuracy depends of textual content and of the size of sentences which speaker pronounces during training and testing. Accuracy is the lowest, below 90% for feature vector of maximum length (21 MFCCs + e1e2e3), in first experiment when s01, s02 and s03 recordings are used as test recordings. The sentence s01 is of medium size with respect to other sentences used, but sentences s02 and s03 are shorter compared to most other sentences [35]. The sentence s01 is the question sentence. In the set of sentences used, sentences s01, s06, s30, s31, s32, are questionable sentences. These sentences are different in intonation, mostly in the beginning of sentence, compared to declarative sentences. This intonation from the beginning of questionable sentence extends through the sentence, and makes it different from the standpoint of speaker recognition. In the second experiment recognition accuracy for the base feature vector of 21 MFCCs is the lowest, 87.96%, additional features e₁ and e₂ are increased recognition accuracy to 94.44%. It is evident that in this test is maximum increase of recognition accuracy, around 6.48%, similar increase of accuracy is achieved in [24-25]. In six experiments maximum of recognition

accuracy is achieved already after adding first two additional features to basic features. In two tests maximum of accuracy is achieved after adding one additional feature and in two tests maximum of accuracy is achieved for the basic feature vector. From the row “mean accuracy” it is evident that with additional features recognition accuracy is increased. Maximum of mean recognition accuracy, 97.11%, and minimum of standard deviation is achieved for feature vector 21 MFCCs+ele2. By comparing minimum of mean accuracy, 95.44%, for feature vector of 21 MFCCs, with respect to maximum of mean accuracy, it is evident that increasing of mean accuracy is around 1.67%. Results of identification accuracy in tests on CHAINS are similar to results in [20] and [23-25]. The mean errors vary in the range 2.89% - 4.56%. The minimum mean error is approximately 3%. This is error of speaker identification, since recognition is applied in the closed set of speakers.

Results for S-ADAPT database are in the next five tables, testing was divided in ten tests, each table shows results of two appropriate tests for two appropriate sets of test recordings (marked as underlined and double underlined). In each test approximately is used one tenth part of S-ADAPT speech database as testing set. In the name of the table is given what recordings are used for testing, for example in Table 3 these are recordings marked with 01, 02 and 03, underlined, and 04, 05, 06, double underlined. Test set 01,02,03, contains three short recordings, marked by 01, 02, 03, and three longer recordings, marked also by 01, 02, 03, in each of emotions for each of speakers. Similarly, other test sets also contain short and longer recordings. First column of each of five tables for S-ADAPT speech database contains description of the appropriate test for test set mentioned in the title of the table, the emotional states of speech recordings used for training (tra) and testing (test) and number of reference models obtained in appropriate training. First set of tests was done when training done by neutral speech, it is marked in the table by tra-n. Tests are done when: test recordings are in neutral speech also, test-n; test recordings are in anger emotional state, test-a; test recordings are in joy emotional state, test-j; test recordings are in fear emotional state, test-f; test recordings are in sadness emotional state, test-s. Sixth double row contains results when training was done in neutral speech and test recordings are in all emotional states, test-all. Results in this double row are derived from appropriate results in previous five double rows, accuracy for the appropriate feature vector is determined as ratio of sum number of correctly recognized and sum of tests done. Seventh double row give results when training was done by speech recordings in all emotional states and test recordings are also in all emotional states, tra-all, test-all. In each test, in each row, maximum results are shaded. It is evident from results that when emotional states of training speech and testing speech are different then recognition accuracy is significantly decreased with respect to the case when speech used for training and testing are in same emotional state.

Results in Table 3 for test set 01,02,03, show that when training and testing speech are in neutral state recognition accuracy is around 96%. When emotional state of testing speech is changed the recognition results are decreased, from around 10% for test speech in joy emotion to around 40% for test speech in sadness emotion. Also, it is evident significant decrease of recognition accuracy in summary results when training speech is in neutral emotion and testing speech is in all emotional states, decreasing of recognition accuracy is greater than 20%. Recognition accuracy set back to the order of magnitude when training and testing speech are in neutral speech, when in process of training is used speech of the same emotional states as in testing (tra-all, test-all). Recognition accuracy in that case is greater than 97.5% when additional features are not used i.e. feature vector

contains 21 MFCCs. Except of experiments when training speech is in neutral emotional state and testing speech is in sadness emotional state, additional features are decreased recognition accuracy.

Table 3 Results for S-ADAPT (test sets: 01,02,03; 04, 05, 06)

Description	21MFCCs	+e1	+e1e2	+e1e2e3
<u>tra-n, test-n; 3034 mod.</u>	96.75%	95.87%	96.46%	95.28%
<u>tra-n, test-n; 3046 mod.</u>	99.39%	97.55%	98.16%	97.86%
<u>tra-n, test-a; 3034 mod.</u>	81.53%	75.04%	72.48%	71.88%
<u>tra-n, test-a; 3046 mod.</u>	81.42%	75.37%	73.41%	72.37%
<u>tra-n, test-j; 3034 mod.</u>	86.39%	82.31%	79.83%	79.38%
<u>tra-n, test-j; 3046 mod.</u>	86.59%	82.31%	79.83%	79.38%
<u>tra-n, test-f; 3034 mod.</u>	75.75%	71.01%	71.90%	71.56%
<u>tra-n, test-f; 3046 mod.</u>	75.87%	70.39%	71.28%	70.95%
<u>tra-n, test-s; 3034 mod.</u>	53.73%	55.03%	59.26%	61.71%
<u>tra-n, test-s; 3046 mod.</u>	54.88%	55.76%	59.68%	61.87%
<u>tra-n, test-all; 3034 mod.</u>	75.94%	72.56%	72.45%	72.63%
<u>tra-n, test-all; 3046 mod.</u>	76.34%	72.70%	72.65%	72.71%
<u>tra-all, test-all; 10429 mod.</u>	97.77%	96.17%	95.43%	95.43%
<u>tra-all, test-all; 10233 mod.</u>	98.59%	97.97%	97.89%	97.97%

Arrangement of maximums of accuracy in rows of Table 3 for test recordings 04,05,06 is the same as for recordings 01,02,03. By comparing these results it can be mentioned significant increase in recognition accuracy when testing and training speech are both in neutral emotional state and feature vector consists of 21 MFCCs, increase of 99.39% - 96.75% = 2.64%. Similar as in results for test set 01,02,03, recognition accuracy was decreased when testing speech and training speech are in different emotional states.

Table 4 Results for S-ADAPT (test sets: 07,08,09; 10,11,12)

Description	21MFCCs	+e1	+e1e2	+e1e2e3
<u>tra-n, test-n; 3043 mod.</u>	99.70%	99.39%	99.39%	99.09%
<u>tra-n, test-n; 3043 mod.</u>	98.79%	97.57%	97.57%	97.27%
<u>tra-n, test-a; 3043 mod.</u>	81.25%	75.97%	73.19%	72.26%
<u>tra-n, test-a; 3043 mod.</u>	82.29%	75.37%	73.30%	72.26%
<u>tra-n, test-j; 3043 mod.</u>	87.12%	82.07%	80.03%	79.26%
<u>tra-n, test-j; 3043 mod.</u>	86.88%	81.74%	79.38%	78.97%
<u>tra-n, test-f; 3043 mod.</u>	75.75%	70.67%	71.95%	72.01%
<u>tra-n, test-f; 3043 mod.</u>	75.25%	69.94%	71.06%	70.84%
<u>tra-n, test-s; 3043 mod.</u>	54.46%	55.61%	60.04%	62.65%
<u>tra-n, test-s; 3043 mod.</u>	54.36%	54.41%	59.36%	61.61%
<u>tra-n, test-all; 3043 mod.</u>	76.36%	72.87%	72.95%	73.12%
<u>tra-n, test-all; 3043 mod.</u>	76.35%	72.13%	72.36%	72.47%
<u>tra-all, test-all; 10233 mod.</u>	98.59%	98.15%	98.77%	98.68%
<u>tra-all, test-all; 10235 mod.</u>	99.03%	98.24%	97.79%	97.97%

Results in Table 4 for test set 07,08,09 show increase in accuracy when training and testing are in neutral speech with respect to the results for the same conditions in Table 3 for test set 04,05,06. Accuracies in the cases when training and testing speech are in

different emotional states are similar to the results in Table 3. Results for test set 07,08,09 show increase in recognition accuracy for two additional features when training and testing speech are in all emotional states.

Results of recognition accuracy for test set 07,08,09 for training and testing in neutral speech are very similar to each other. Little variability also is evident in results when training and testing speech are in all emotional states. Variability is greater in experiments when training speech is neutral and test speech is in some other emotion. Results of recognition accuracy for test set 10,11,12 are similar to results of tests for previous test sets on S-ADAPT database. Recognition accuracy for testing and training speech in all emotional states is approximately 99% for feature vector of 21 MFCCs.

Table 5 Results for S-ADAPT (test sets: 13,14,15; 16,17,18)

Description	21MFCCs	+e1	+e1e2	+e1e2e3
<u>tra-n, test-n; 3045 mod.</u>	98.17%	98.17%	98.47%	97.56%
<u>tra-n, test-n; 3050 mod.</u>	97.52%	98.76%	98.14%	98.45%
<u>tra-n, test-a; 3045 mod.</u>	81.20%	74.82%	72.64%	71.33%
<u>tra-n, test-a; 3050 mod.</u>	81.03%	74.77%	73.30%	72.26%
<u>tra-n, test-j; 3045 mod.</u>	86.10%	81.87%	79.46%	78.97%
<u>tra-n, test-j; 3050 mod.</u>	86.71%	81.87%	79.38%	79.22%
<u>tra-n, test-f; 3045 mod.</u>	75.53%	69.27%	69.78%	69.89%
<u>tra-n, test-f; 3050 mod.</u>	76.26%	70.78%	71.62%	71.90%
<u>tra-n, test-s; 3045 mod.</u>	55.19%	55.40%	59.78%	62.13%
<u>tra-n, test-s; 3050 mod.</u>	54.46%	56.18%	60.67%	63.33%
<u>tra-n, test-all; 3045 mod.</u>	76.10%	72.15%	72.09%	72.19%
<u>tra-n, test-all; 3050 mod.</u>	76.20%	72.65%	72.78%	73.19%
<u>tra-all, test-all; 10235 mod.</u>	98.41%	97.97%	98.24%	97.79%
<u>tra-all, test-all; 10248 mod.</u>	98.66%	98.66%	97.95%	98.13%

Results of recognition accuracy for test set containing recordings numbered 13,14,15, when training and testing speech are in neutral emotional state, show maximum for feature vector of 21 MFCCs + e1e2, Table 5. Variability of recognition accuracy for different test sets, shows that these features depend of the text content of recordings used for training and testing of the speaker recognizer. For test set 16,17,18, (Table 5) maximum of recognition accuracy in neutral speech is achieved for feature vector 21 MFCCs + e1. Maximum of the recognition accuracy when training and testing speech are in all emotional states is for the feature vector of 21 MFCCs but the same accuracy was achieved also for feature vector 21 MFCCs + e1. Recognition accuracy when training and testing speech are in different emotional states is similar to results in previous tables.

For test set 19,20,21, (Table 6) when training and testing recordings are both in the same emotional states, neutral or all emotional states, maximum of recognition accuracy around 99% is achieved for feature vector 21 MFCCs + e1e2.

Table 6 Results for S-ADAPT (test sets: 19,20,21; 22,23,24)

Description	21MFCCs	+e1	+e1e2	+e1e2e3
<u>tra-n, test-n; 3049 mod.</u>	98.46%	99.07%	99.38%	98.15%
<u>tra-n, test-n; 3049 mod.</u>	99.38%	99.07%	99.07%	98.76%
<u>tra-n, test-a; 3049 mod.</u>	82.18%	75.64%	73.73%	72.48%
<u>tra-n, test-a; 3049 mod.</u>	81.25%	75.75%	73.46%	72.48%
<u>tra-n, test-j; 3049 mod.</u>	86.63%	82.44%	79.87%	79.50%
<u>tra-n, test-j; 3049 mod.</u>	86.80%	82.11%	79.75%	79.22%
<u>tra-n, test-f; 3049 mod.</u>	76.26%	71.12%	72.12%	71.84%
<u>tra-n, test-f; 3049 mod.</u>	75.14%	69.94%	70.95%	70.45%
<u>tra-n, test-s; 3049 mod.</u>	55.24%	55.82%	60.25%	62.65%
<u>tra-n, test-s; 3049 mod.</u>	54.15%	55.71%	60.09%	62.75%
<u>tra-n, test-all; 3049 mod.</u>	76.65%	73.02%	73.09%	73.15%
<u>tra-n, test-all; 3049 mod.</u>	76.03%	72.67%	72.69%	72.81%
<u>tra-all, test-all; 10250 mod.</u>	98.57%	98.48%	98.93%	98.39%
<u>tra-all, test-all; 10257 mod.</u>	99.28%	98.56%	98.47%	98.56%

When set of testing recordings numbered by 22,23,24, (Table 6), arrangement of maximum values of recognition accuracy is similar to results in Table 3, but in the case of training and testing recordings in neutral speech, the results of recognition accuracy for feature vector 21 MFCCs, when is achieved maximum of accuracy, and for feature vectors 21 MFCCs + e₁, 21 MFCCs + e₁e₂, are very similar each other.

For test set 25,26,27,28 (Table 7 - underlined), maximum of recognition accuracy when training and testing speech are in neutral emotional states is achieved for feature vector of 21 MFCCs, the same accuracy is achieved for feature vector 21 MFCCs + e₁e₂e₃. Accuracies for two other feature vectors are very similar. For test set 29,30,31,32 (Table 7 - double underlined) maximum of accuracy when training and testing speech are in all emotional states is achieved for feature vector 21 MFCCs + e₁e₂e₃.

Table 7 Results for S-ADAPT (test sets: 25,26,27,28; 29,30,31,32)

Description	21MFCCs	+e1	+e1e2	+e1e2e3
<u>tra-n, test-n; 2940 mod.</u>	98.61%	98.38%	98.38%	98.61%
<u>tra-n, test-n; 3058 mod.</u>	99.68%	98.41%	99.05%	98.73%
<u>tra-n, test-a; 2940 mod.</u>	81.09%	75.20%	72.91%	71.72%
<u>tra-n, test-a; 3058 mod.</u>	81.47%	75.15%	73.02%	72.32%
<u>tra-n, test-j; 2940 mod.</u>	86.35%	81.09%	78.93%	78.81%
<u>tra-n, test-j; 3058 mod.</u>	86.51%	82.23%	79.71%	79.34%
<u>tra-n, test-f; 2940 mod.</u>	75.53%	70.39%	71.90%	71.95%
<u>tra-n, test-f; 3058 mod.</u>	76.70%	71.28%	71.45%	71.23%
<u>tra-n, test-s; 2940 mod.</u>	54.88%	56.23%	59.99%	62.54%
<u>tra-n, test-s; 3058 mod.</u>	55.14%	55.87%	60.20%	62.65%
<u>tra-n, test-all; 2940 mod.</u>	76.38%	72.77%	72.82%	73.13%
<u>tra-n, test-all; 3058 mod.</u>	76.55%	72.84%	72.69%	72.93%
<u>tra-all, test-all; 9886 mod.</u>	99.12%	98.52%	98.31%	98.65%
<u>tra-all, test-all; 10315 mod.</u>	99.15%	98.96%	98.86%	99.43%

As is evident from results on S-ADAPT, in some experiments when training and test speech are in the same emotional state additional features are increased recognition accuracy or not decreased (Table 4 – test set 07,08,09, in all emotional states; Table 5 – test set

13,14,15, in neutral state; Table 5 – test set 16,17,18 and Table 6 – test set 19,20,21, in neutral and in all emotional states; Table 7: for test set 25,26,27,28, in neutral state, for test set 29,30,31,32, in all emotional states). The consequence is that the results in those cases are very slowly variable.

In next table will be mean accuracy, standard deviation and mean error for each of experiments on S-ADAPT. The difference between maximum and minimum mean recognition accuracy, Table 8, when training and testing speech both are in neutral emotional state is approximately $98.65\% - 97.98\% = 0.67\%$, and when training and testing speech both are in all emotional states is approximately $98.72\% - 98.06\% = 0.66\%$. Mean recognition accuracy in these cases for the base feature vector of 21 MFCCs is greater than 98.5% and to increase recognition accuracy closer to 100% it is necessary to use additional features of maximum efficiency. In [20] feature vector of 38 coefficients is used, in future work number of additional features can be carefully increased to achieve higher increase in accuracy.

Table 8 Mean accuracy, standard deviation and mean error for S-ADAPT

Description	21MFCCs	+e1	+e1e2	+e1e2e3
tra-n, test-n	98.65%	98.23%	98.41%	97.98%
	0.91%	0.98%	0.86%	1.05%
	1.35%	1.77%	1.59%	2.02%
tra-n, test-a	81.47%	75.31%	73.14%	72.14%
	0.41%	0.37%	0.36%	0.35%
	18.53%	24.69%	26.86%	27.86%
tra-n, test-j	86.61%	82.00%	79.62%	79.20%
	0.28%	0.37%	0.31%	0.21%
	13.39%	18.00%	20.38%	20.80%
tra-n, test-f	75.80%	70.48%	71.40%	71.26%
	0.46%	0.59%	0.66%	0.69%
	24.20%	29.12%	28.60%	28.74%
tra-n, test-s	54.65%	55.60%	59.93%	62.39%
	0.47%	0.52%	0.40%	0.51%
	45.35%	44.40%	40.07%	37.61%
tra-n, test-all	76.29%	72.64%	72.66%	72.83%
	0.21%	0.27%	0.28%	0.32%
	23.71%	27.36%	27.34%	27.17%
tra-al, test-al	98.72%	98.17%	98.06%	98.10%
	0.43%	0.73%	0.96%	1.00%
	1.28%	1.83%	1.94%	1.90%

Mean recognition accuracy is around 98% when training and test done on neutral speech. When test is done in some other emotion, mean accuracy is decreased. In test of anger decrease is around 20%, in test of joy mean accuracy is decreased around 17%, in test of fear decrease is around 25% and in test of sadness decrease is around 40%. Mean recognition accuracy is returned on the value as well as in test for neutral speech when training is done in all emotional states. Standard deviation of recognition accuracy in experiments is below 1.1%. This indicates good compliance of recognition accuracy in appropriate tests. Accuracy on S-ADAPT when training and test speech are in same emotion is around 98% or often higher than 97%. These results are comparable with results in [20], [23-24].

Errors of speaker identification are errors in experiments described on the speech database S-ADAPT, since recognition is applied in the closed set of speakers. Mean error is in the range 1.28% - 2.02%, when training and test speech both are in neutral emotional state or in all emotional states. This error increases when training and test speech are in different emotions. In experiments when test speech is in anger, joy or fear emotion, additional features are increased mean error. The largest value of mean error is in the case when test speech is in sadness emotion. In that case additional features decreases mean error, maximum of accuracy when test speech is in sadness emotion is achieved for feature vector 21 MFCCs + e1e2e3.

The range of mean recognition accuracy for different feature vectors depends on emotions of training and test speech (Table 8). This can be used in applications on groups of speakers where can be expected changes in emotion of speech, and where it is necessary to estimate the emotion of observed test speech of speakers. In all experiments additional feature e3 is increased accuracy when test speech is in sadness emotion. This finding additionally can be used for detecting sadness in speakers. Suppose that we have the training models for observed group of speakers in neutral speech. If difference between accuracies for test speech in neutral emotion and test speech in observed emotion is around 40%, if this difference decreases when additional features are added so that for the feature vector 21 MFCCs + e1e2e3 this difference is the smallest, then can be expected that test speech is in sadness emotion.

This paragraph presents results from the literature for the purpose of comparison with results of this paper. Gaussian mixture model (GMM) in combination with MFCCs on dataset of 15 speakers, 10 male and 5 female, in [4], give accuracy of 86.27% for text independent case and 94.12% for text dependent case. Gaussian Mixture Model – Universal Background Model (GMM-UBM) applied in the verification of speakers with the Constraint of Limited data (<15 sec) in [5], gives equal error rate (EER) around 37% on the NIST-SRE-2003 database. The proposed method in [6] on the database of Indian scenario named as IITG Multivariability Speaker Recognition Database, Part-III phase-IV (the subject was asked to read some text), which performs classification by using a Multilayer perceptron feed-forward neural network trained by backpropagation and MFCCs feature vectors, gives an accuracy of 94.44%. In experiments in [8], on Speakers in the Wild and NIST SRE 2016 Cantonese data sets, EER vary from 9.68% for i-vector based speaker recognizer to 4.16% for speaker recognizer based on x-vectors. Speaker recognizer based on use of i-vectors in [9] give recognition accuracy of 93.8% when test speech is in neutral emotion, IEMOCAP database is used. In emotion invariant system accuracy is increased when test speech is in happiness, anger or sadness emotion, accuracies in those cases are 91.3%, 89.3% and 89.6%. The speaker recognition system developed in [10], which combines the CNN method with MFCC, achieved a high accuracy of 96% on live speaker recordings, and 94.66% on used TIMIT dataset. Research was conducted on 50 speakers from the TIMIT dataset, which contained eight utterances for each speaker and 60 speakers from live recording using a smartphone. CNN using MFCC in [11], on the dataset VoxForge Speech Corpus achieves an accuracy of 96.95%. Unconstrained and constrained CNN model applied on speech spectrograms of 1 second of speech in [12] was tested in emotional speech when training was done on neutral speech, on the closed set of speakers of the Serbian emotional amateur corpus (SEAC) recorded by amateur speakers using mobile phones. The proposed unconstrained CNN speaker model in [12] achieves high average recognition accuracy of 99.248% on neutral

test speech, around 85% in the case of test speech in anger, fear or joy emotion, and 79.84% when test speech is in sadness emotion. Constrained 8-bit representation of weights results in negligible difference in achieved recognition accuracy with respect to full-precision unconstrained 32-bit model. The difference in the case of the ternary quantization model is up to 3.3%, degradation in the case of the binary quantization model is up to 10.55%. Accuracies of 97.22% and 94.01% are achieved in the cases of ternary and binary quantization, when training and test speech are in neutral emotional state. Speaker recognition accuracy of 97% by using simple spectrograms as input to CNN, in [13], is achieved in experiments on the TIMIT dataset. The average speaker identification accuracy of experiments on the speech from five speakers speak in Thai language of which voices are extracted from YouTube, in [14], achieved for the system using MFCCs and SVM is 91.26% and for CNN applied on spectrograms of 2 seconds of speech is 95.83%. MFCCs and radial basis function neural network (RBFNN) in [16], tested on the database of Arabic speakers who pronounce “Thank you” in Arabic, achieves accuracy of 97.5%. Combination of MFCCs, calculated for voiced frames, and inverted MFCCs (IMFCCs) calculated for unvoiced frames, increased speaker recognition accuracy in [17] from 80% for only MFCCs used to 90% for MFCCs and IMFCCs used. Addition of 26 SSCs to feature vector of 12 LPCC and modeling by GMM in experiments on TIMIT speech database [20], results in increase of speaker recognition accuracy by 2.9% and reached accuracy is 99.1%. Graphically presented results of experiments in [22] on the TIMIT database, where diagonal-covariance GMMs with 32 mixtures are used as speaker models, show that for signal to noise ratio (SNR) around 25 dB speaker identification accuracy for SSCs and marginalization-based MFT features is around or higher than 90% while for MFCCs accuracy is smaller than 80%. NDSF features with included spectral subtraction in combination with cepstral mean normalization [23], in experiments on Hindi and IITG datasets, increase accuracy to almost 100% with respect to accuracy achieved by using MFCCs or LPCCs features. By addition of SD features to MFCCs and using GMM_128 model in experiments on AURORA2.0 data set in [24], increase of identification accuracy is higher than 5% and achieved accuracy is approximately 95%. An overall 5% relative improvement in accuracy over the conventional MFCC-based front-end was obtained in [25] by addition of FM features, in experiments on the NIST 2001 Evaluation database by using GMMs. Combinations of MFCC and UMRT based features and use of Multi-layer perceptron (MLP), in [27], in experiments on the data set of ten samples of words each spoken by 15 persons (8 female and 7 male) is resulted in increase of average accuracy by 3%. The accuracy achieved is around 97.91% for speech dependent system and around 94.44% for speech independent system. A comparative re-assessment of feature extractors for deep speaker embeddings conducted on x-vector system, in [28] show that: minimal EER of 3.89% on Voxceleb1-E data set is achieved for MFCC+SCMC (Spectral centroid magnitude coefficients)+Multi-taper features, minimal EER of 6.08% on SITW-DEV data set is achieved for Power-normalized cepstral coefficients (PNCCs) features. The research in [29] yields exceptional results across different datasets and classifiers. For instance: on the TIMIT babble noise dataset (120 speakers), feature fusion achieves speaker identification accuracy of 92.7%; speaker identification accuracy of 93.5% on the TIMIT babble noise dataset (630 speakers) using a KNN classifier with feature optimization; on the TIMIT white noise dataset (120 and 630 speakers) – speaker identification accuracies of 93.3% and 83.5%, respectively, by using PCA dimension reduction and feature optimization techniques (PCA-MPA) with KNN

classifiers; on the Voxceleb1 dataset PCA-MPA feature optimization with KNN classifiers achieves a speaker identification accuracy of 95.2%.

Since the proposed additional features are short-term speech features, as well as already good defined MFCCs, and since the results on S-ADAPT show that with feature vector of 21 MFCCs can be achieved mean accuracy greater than 98.5%, therefore one of directions of future work can be in investigation for efficient long-term features of speakers i.e. for more efficient models of speakers. Covariance matrix of matrix of feature vectors is only one possibility for speaker modeling, it is one of a lot of manners to transform matrix of feature vectors in a more compact shape. These transformations are possibilities or potential possibilities which help in forming a more compact representation of speaker from information potential contained in matrix of feature vectors and in finding relations between feature vectors, i.e. embedded specific information about speaker. To achieve better recognition accuracy it is necessary to improve and as fully as possible take advantage of information potential of the matrix of feature vectors X . Also, the impact of amount and of text content of speech material used for training and testing on accuracy of speaker recognition will be interesting to detailed examine.

4. CONCLUSION

Results of recognition show that in training it is necessary to use recordings in all emotions which are expected in test conditions. By this way the robustness of speaker recognizer to changes of emotions in speech is ensured. Mean recognition accuracy for feature vector of 21 MFCCs in experiments on CHAINS speech database is 95.44%. It is more than 3% smaller with respect to mean accuracy on S-ADAPT speech database when feature vector of 21 MFCCs used and training and testing speech both are in the same emotional state. Also, a smaller amount of speech is used in training and testing on the CHAINS database. Impact of additional features on increase of recognition accuracy is more expressed in experiments on CHAINS speech database. Mean recognition accuracy is increased by adding additional features in experiments on CHAINS approximately 1.67% and decreased on S-ADAPT less than 0.67% when training and testing speech are neutral or in all emotions.

Additional features used in this paper are efficient in recognition on smaller closed sets of speakers when lesser number of recordings is used for training and testing. When in experiments on S-ADAPT, for training and test speech in same emotion, a larger number of recordings are used in training and testing, then in most experiments feature vector of 21 MFCCs is sufficient for efficient recognition. However, these additional features improve accuracy of recognition in some cases in experiments on S-ADAPT when accuracy higher than 98% is achieved for feature vector of 21 MFCCs. In experiments on CHAINS database: when s18, s19, s20, are used as test recordings recognition accuracy is increased from 99.07% to 100%; and when s24, s25, s26, are used as test recordings additional features are increased recognition accuracy from 98.15% towards 100%. These results on CHAINS and S-ADAPT databases indicate that proposed additional features can in some experiments improve recognition accuracy even though with feature vector of 21 MFCCs achieved accuracy is 98% or 99%. Achieved maximums of mean recognition accuracy on both speech databases are higher than 97%. These results are comparable with results in presented literature. Used speaker recognizer has a small number of parameters. The complexity of the proposed method is not higher than complexity of the methods in presented literature.

Acknowledgement: *This research has been supported by the Ministry of Education, Science and Technological Development through the project no. 451-03-68/2020-14/200156: "Inovative scientific and artistic research from the FTS activity domain". Thanks to the project team of the AI projects S-ADAPT and AI-SPEAK, who made the newly developed speech database available to us, as well as 426 by the European Union's Horizon 2023 research and innovation programme through the 427 AIDA4Edge Twinning project grant ID 101160293.*

REFERENCES

- [1] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Margin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacrétaz and D. A. Reynolds, "A Tutorial on Text-Independent Speaker Verification", *EURASIP J. Appl. Signal Process.*, vol. 2004, no. 4, pp. 430-451, Apr. 2004.
- [2] T. Kinnunen and H. Li, "An Overview of Text-Independent Speaker Recognition: From Features to Supervectors", *Speech Commun.*, vol. 52, no. 1, pp. 12-40, Jan. 2010.
- [3] V. Tiwari, "MFCC and Its Applications in Speaker Recognition", *Int. J. Emerg. Technol.*, vol. 1, no. 1, pp. 19-22, 2010.
- [4] A. Maurya, D. Kumar and R. K. Agarwal, "Speaker Recognition for Hindi Speech Signal using MFCC-GMM Approach", In Proceedings of the 6th International Conference on Smart Computing and Communications (ICSCC), 2017, Kurukshetra, India, in *Procedia Comput. Sci.*, vol. 125, pp. 880-887, 2018.
- [5] T. R. Jayanthi Kumari, R. Anita, and T. Suraj Duncan, "Speaker Verification Comparison between GMM and GMM-UBM Under Limited Data Condition", *J. Electr. Syst.*, vol. 20, no. 11s, pp. 1345-1350, 2024.
- [6] K. J. Devi, A. A. Devi, and K. Thongam, "Automatic Speaker Recognition using MFCC and Artificial Neural Network", *Int. J. Innov. Technol. Explor. Eng. (IJITEE)*, vol. 9, no. 1S, pp. 39-42, Nov. 2019.
- [7] M. K. Nandwana, J. van Hout, M. McLaren, A. Stauffer, C. Richey, A. Lawson and M. Graciarena, "Robust Speaker Recognition from Distant Speech under Real Reverberant Environments Using Speaker Embeddings", In Proceedings of the Interspeech, Hyderabad, India, 2018, pp. 1106-1110.
- [8] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition", In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 2018, pp. 5329-5333.
- [9] B. D. Sarma and R. K. Das, "Emotion Invariant Speaker Embeddings for Speaker Identification with Emotional Speech", In Proceedings of the APSIPA Annual Summit and Conference, Auckland, New Zealand, 2020, pp. 610-615.
- [10] A. Wirdiani, S. N. Machetho, I. K. G. D. Putra, M. Sudarma, R. S. Hartati and H. A. Ferdian, "Improvement Model for Speaker Recognition using MFCC-CNN and Online Triplet Mining", *Int. J. Adv. Sci., Eng. Inform. Technol.*, vol. 14, no. 2, pp. 420-427, 2024.
- [11] S. Srivastava, G. Chaudhary and C. Shukla, "Text-Independent Speaker Recognition Using Deep Learning", In: S. Srivastava, M. Khari, R. Gonzales Crespo, G. Chaudhary, P. Arora (eds) *Concepts and Real-Time Applications of Deep Learning*. EAI/Springer Innovations in Communication and Computing. Cham: Springer, pp. 41-51, 2021.
- [12] N. Simić, S. Suzić, T. Nosek, M. Vujović, Z. Perić, M. Savić and V. Delić, "Speaker Recognition Using Constrained Convolutional Neural Networks in Emotional Speech", *Entropy*, vol. 24, no. 3, p. 414, 2022.
- [13] Y. Lukic, C. Vogt, O. Dürr, T. Stadelmann, "Speaker Identification and Clustering Using Convolutional Neural Networks", In Proceedings of the 2016 IEEE International Workshop on Machine Learning for Signal Processing, Salerno, Italy, 2016, pp. 1-6.
- [14] S. Bunrit, T. Inkian, N. Kerdprasop and K. Kerdprasop, "Text-Independent Speaker Identification Using Deep Learning Model of Convolution Neural Network", *Int. J. Mach. Learn. Comput.*, vol. 9, no. 2, pp. 143-148, Apr. 2019.
- [15] G. D. Saxena, N. A. Farooqui and S. Ali, "Extricate Features Utilizing Mel Frequency Cepstral Coefficient in Automatic Speech Recognition System", *Int. J. Eng. Manuf.*, vol. 12, no. 6, pp. 14-21, 2022.
- [16] A. Al-Qaisi, "Arabic Word Dependent Speaker Identification System Using Artificial Neural Network", *Int. J. Circuits, Syst. Signal Process.*, vol. 14, pp. 290-295, 2020.
- [17] Latha, "Robust Speaker Identification Incorporating High Frequency Features", In Proceedings of the Twelfth International Multi-Conference on Information Processing (IMCIP-2016), in *Procedia Computer Science*, vol. 89, pp. 804-811, 2016.
- [18] N. P. H. Thian, C. Sanderson and S. Bengio, "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, in: D. Zhang, A. K. Jain (eds)

- Biometric Authentication*, ICBA 2004, Lecture Notes in Computer Science, vol. 3072. Heidelberg, Berlin: Springer, 2004, pp. 631-639.
- [19] J. M. K. Kua, T. Thiruvaran, M. Nosratighods, E. Ambikairajah and J. Epps, "Investigation of Spectral Centroid Magnitude and Frequency for Speaker Recognition", in *Proceedings of the Odyssey-2010: The Speaker and Language Recognition Workshop*, Brno, Czech Republic, 2010, pp. 34-39.
 - [20] M. Qarachorloo and G. Farahani, "New Features to Improve Speaker Recognition Efficiency with Using LPCC and SSC Features", *Int. J. Signal Process. Syst.*, vol. 4, no. 4, pp. 295-299, Aug. 2016.
 - [21] T. Kinnunen, B. Zhang, J. Zhu and Y. Wang, "Speaker Verification with Adaptive Spectral Subband Centroids", In *Proceedings of the International Conference on Biometrics (ICB 2007)*, Seoul, Korea, 2007, in: SW. Lee, S.Z. Li (eds) *Advances in Biometrics*, ICB 2007, Lecture Notes in Computer Science, vol. 4642. Heidelberg, Berlin: Springer, 2007, pp. 58-66.
 - [22] A. Nicolson, J. Hanson, J. Lyons and K. Paliwal, "Spectral Subband Centroids for Robust Speaker Identification Using Marginalization-based Missing Feature Theory", *Int. J. Signal Process. Syst.*, vol. 6, no. 1, pp. 12-16, Mar. 2018.
 - [23] S. V Chougule and M. S Chavan, "Robust Spectral Features for Automatic Speaker Recognition in Mismatch Condition", In *Proceedings of the Second International Symposium on Computer Vision and the Internet (VisionNet'15)*, in *Procedia Computer Science*, vol. 58, pp. 272-279, 2015.
 - [24] W.-S. Chen and J.-F. Huang, "Speaker Recognition with Spectral Dimension Features of Human Voices for Personal Authentication", *J. Netw. Commun. Emerg. Technol. (JNCET)*, vol. 5, no 3, pp. 6-11, 2015.
 - [25] T. Thiruvaran, E. Ambikairajah and J. Epps, "Speaker Identification Using FM Features", In *Proceedings of the 11th Australian International Conference on Speech Science & Technology*, ed. P. Warren & C. I. Watson, ISBN 0 9581946 2 9, University of Auckland, New Zealand, 2006, pp. 148-152.
 - [26] T. Thiruvaran, E. Ambikairajah and J. Epps, "FM Features for Automatic Forensic Speaker Recognition", In *Proceedings of the Interspeech 2008, Interspeech 2008 Special Session: Forensic Speaker Recognition – Traditional and Automatic Approaches*, Brisbane, Queensland, Australia, 2008, pp. 1497-1500.
 - [27] A. Antony and R. Gopikakumari, "Speaker Identification Based on Combination of MFCC and UMRT Based Features", In *Proceedings of the 8th International Conference on Advances in Computing and Communication (ICACC-2018)*, in *Procedia Computer Science*, vol. 143, pp. 250-257, 2018.
 - [28] X. Liu, M. Sahidullah and T. Kinnunen, "A Comparative Re-Assessment of Feature Extractors for Deep Speaker Embeddings", In *Proceedings of the Interspeech 2020*, 2020, Shanghai, China, pp. 3221-3225.
 - [29] N. Chauhan, T. Isshiki and D. Li, "Enhancing Speaker Recognition Models with Noise-Resilient Feature Optimization Strategies", *Acoustics*, vol. 6, pp. 439-469, 2024.
 - [30] I. D. Jokić, S. D. Jokić, V. D. Delić and Z. H. Perić, "One Solution of Extension of Mel-Frequency Cepstral Coefficients Feature Vector for Automatic Speaker Recognition", *Inf. Technol. Control*, vol. 49, no. 2, pp. 224-236, 2020.
 - [31] I. Jokić, V. Delić and Z. Perić, "Application of Mel-Frequency Cepstral Coefficients in Automatic Speaker Recognition as Part of IoT Solutions for Security and Optimization in Smart Cities", *Alfatech J.*, no. 1, pp. 5-10, 2025.
 - [32] M. Sigmund, "Speaker Discrimination Using Long-Term Spectrum of Speech", *J. Inf. Technol. Control*, vol. 48, no. 3, pp. 446-453, 2019.
 - [33] N. N. An, N. Q. Thanh and Y. Liu, "Deep CNNs with Self-Attention for Speaker Identification", *IEEE Access*, vol.7, pp. 85327-85337, May 2019.
 - [34] [Online]. Available: https://github.com/stevanjokic/speaker_identification
 - [35] F. Cummins, M. Grimaldi, T. Leonard, J. Simko, "The CHAINS Corpus: CHAracterizing INdividual Speakers", In *Proceedings of the 11th International Conference "Speech and Computer" SPECOM'2006*, St. Petersburg, Russia, 2006, pp. 431-435.