http://dx.doi.org/10.5755/j01.eee.20.6.7276 ELEKTRONIKA IR ELEKTROTECHNIKA, ISSN 1392–1215, VOL. 20, NO. 6, 2014

Towards a Small Intra-Speaker Variability Models

I. D. Jokic[1], S. D. Jokic[1,2], V. D. Delic[1], Z. H. Peric[3]
[1]*Faculty of Technical Sciences, University of Novi Sad,*
Trg Dositeja Obradovica 6, 21000 Novi Sad, Serbia
[2]*DunavNET Ltd.,*
Antona Cehova 1/2, 21000 Novi Sad, Serbia
[3]*Department of Telecommunications, Faculty of Electronic Engineering, University of Nis,*
Aleksandra Medvedeva 14, 18000 Nis, Serbia
ibahjokih@gmail.com

***Abstract*—Automatic speaker recognizer used in experiments described in this paper uses vectors of mel-frequency cepstral coefficients as feature vectors and covariance matrices for speakers modelling. By comparing the models of training and test speech of the same speakers it was noticed significant differences in some model elements. Speaker models had inherent intra-speaker variability. In the observed test the distinction matrix was introduced as the measure of intra-speaker variability of all speaker models in available speech database. Based on the values of elements in distinction matrix, the ranges of validity of elements in weighting matrix were established. Each element in covariance matrix of a speaker was pondered by appropriate weighting coefficient. Application of this transformation resulted in higher accuracy of automatic speaker recognition.**

***Index Terms*—Automatic speaker recognition, mel frequency cepstral coefficients, intra-speaker variability, distinction matrix, weight matrix.**

I. INTRODUCTION

Mel Frequency Cepstral Coefficients (MFCCs) are widely used as features for automatic speaker recognition [1]. Their computation is based on the cepstrum of an observed speech segment $s(n)$ [2]

$$c_S(n) = F^{-1}\left\{\log\left|F\left\{s(n)\right\}\right|\right\}. \qquad (1)$$

Since F is Discrete Time Fourier Transform (DTFT) and F^{-1} is Inverse DTFT, it is obviously that previously defined cepstrum depends on the amplitude spectrum of observed speech segment. This definition and the fact that the speech is a complex signal produced as a result of several transformations which occur at different levels: semantic, linguistic, articulatory and acoustic [3], [4], imply an inevitable temporal variability of MFCCs. This temporal variability is undesirable and reflected in the intra-speaker variability of feature vectors. Therefore it is necessary to model feature vectors of observed speaker.

MFCCs are one of short-term speech features. Usually, their calculation is based on the spectral analysis of speech segments whose duration is about 20 ms–30 ms [1]. Each speech segment was represented by appropriate MFC feature vector. Because of feature vectors time variability, it is necessary to introduce some procedure of their temporally averaging through appropriate models. Usually it is done by one of stochastic models, Gaussian Mixture Models (GMMs) in case of text-independency or Hidden Markov Models (HMMs) for text-dependent cases [5].

Automatic speaker recognizer is based on existence of training and test model for each observed speaker. In an ideal case, estimated models for training and test data of the same speaker should be the same. Due to temporal variability of MFC feature vectors this equality can’t be realised and speaker models have undesired property of intra-speaker variability. So this is the consequence of MFCCs used as speaker features and can be solved by introducing some modifications to feature calculation or by using some new sets of features [6], [7]. The solution for reduction of intra-speaker model variability, based on analysing of models of the same speakers is presented in this paper.

In the next section the experimental setup is described: feature vectors, method of speaker modelling as well as the speech database used. Also, the method for determination of distinction matrix, necessary to describe intra-speaker variability of models, and weighting matrix determination, are presented. Finally, some experimental results of automatic speaker recognition are presented and evaluated.

II. EXPERIMENT PREPARATION

MFCCs are used as features of speaker. They are determined by [8]

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

where $n = 0,1,2,...,d-1$, $E_{\log(k)}$ represents the log-energy inside the kth auditory critical band, n is the ordinal number

Manuscript received November 3, 2013; accepted February 18, 2014.

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.

of calculated MFCC and d is the dimensionality of used feature vector i.e. the number of MFCCs used in feature vector. In this paper, for MFCCs determining we've used 20 rectangular auditory critical bands, width of the each is 300 mels and mutually shift is 150 mels.

Speaker recognizer applied in following experiments is oriented to text-independent applications. Therefore speaker modelling was based on the assumption that feature vectors are distributed in accordance with the appropriate Gaussian multidimensional distribution. The shape of this distribution is determined by the covariance matrix Σ. Each set of recordings of observed speaker was modelled by appropriate covariance matrix. For this purpose feature vectors of observed set of recordings was grouped into matrix form X_i. Then appropriate model, covariance matrix Σ_i, was determined by equality

$$\Sigma_i = \frac{1}{n-1} \cdot \left(X_i - \text{«}_i \right) \cdot \left(X_i - \text{«}_i \right)^T . \tag{3}$$

In this equality n is the number of feature vectors in modelled set of recordings and «_i is the vector of mean values.

Speech database used in experiments described in this paper contains recordings of 121 speakers, 61 female and 60 male and for each of speakers there is 14 recordings spoken on Serbian. Recordings can be classified into 3 groups. The first group, "Names", has 1 recording for each of speakers. This recording contains the identification number of the speaker, his name and surname. Duration of these recordings is smallest in the speech database and the pronunciation in them is more spontaneous than pronunciations in two others group of recordings. The second group, "Digits", contains two recordings for each of speakers, pronunciations of isolated spoken strings of digits "1, 2, 3, 4, 5" and "6, 7, 8, 9, 0". The third group, "Words", contains 11 pronunciations of strings of words. Strings of words are the same for all of speakers. In testing phase tests are organized in 14 groups. In first group of tests as test speech file for each speaker the recording from the group "Names" is observed, second and third group of tests are devoted to tests on the recordings from the group "Digits". In remaining eleven groups of tests as test files the recordings from the group "Words" are observed.

In experiments, each speaker is characterized by training and test model. The difference between model of the i^{th} speaker and the model of the test speech is defined by equality

$$m(i, test) = \frac{1}{d^2} \cdot \sum_{i=1}^{d} \sum_{j=1}^{d} \left| \Sigma_i (i, j) - \Sigma_{test} (i, j) \right|, \tag{4}$$

where d is the dimensionality of feature vectors used. The observed test speech belongs to the i^{th} speaker if

$$m(i, test) < m(j, test), \tag{5}$$

where $\forall j \in \{1, 2, ..., N\} \setminus \{i\}$, N is the number of speakers in speech database. In experiments of automatic speaker recognition described in this paper the all speakers are observed and therefore $N = 121$.

By observing (2) it is evident that MFCCs depends on energy in observed speech signal. This dependency prevents the fulfillment of one of the main conditions appointed to speaker features, it is the property that speaker features have large between-speaker variability and small within-speaker variability [1]. So, it is the consequence of cepstrum deffinition (1) and makes the imperfection of MFCCs which is reflected in the inevitable existence of a certain degree of intra-speaker variability.

Model (3) depends of a set of feature vectors modeled. Models of training and test speech recordings of the same speaker are different. Looking at the definition (3) this is not surprising because of dependency between covariance matrix and modeled set of feature vectors. Different sets of a speaker feature vectors were produced in diferent time intervals, therefore the dependency between modeled set of feature vectors and appropriate model can be marked as time dependency of speaker model. It is idealy that model of a speaker is not time variable or eventualy very little time variable. Covariance matrix as speaker model consists of elements, inter-dimensional covariances of a set of feature vectors modeled. If element of the covariance matrix has a large time variability then he can be observed as non-speaker inherent. His non-speaker inherent ability is consequence of the method used for determining of feature set. Since MFCCs were calculated by application of short time frequency analisys, intra-speaker variability of models is unavoidable. By tracking of individual covariances in appropriate models for training and test speech of the same speakers, it is possible to more precise determine their degree of variability. By analyzing and compare training and test models of the same speakers it was observed that some elements of covariance matrices have large difference whereas some it have not. For example in our previous work [9] it was noted that $\Sigma_{0,0}$ and $\Sigma_{19,19}$ elements of covariance matrices, which modelled the zeroth and 19th MFCC, have significant differences in training and test models of the same speakers. By discarding these elements as well as complete model of zero and 19th MFCCs, all elements in the first and twentieth row and column where the observed diagonal matrix elements placed, the recognition accuracy of some tests has improved. So this is the first solution for reduction of models difference and this is a rough way to increase the recognition accuracy. This way, discarding of some elements of model or complete MFCCs, degrade the precision of speaker representation. The full covariance matrices better represents the observed set of feature vectors. Also, the feature vectors of the higher dimensionality much better describe observed frame of speech. So it is the fact that these feature vectors have much information about speaker, but not all information of the same importance, i.e. not all elements of model has the same importance.

Based on detailed analysis of time variability of speaker models, in covariance matrices can be observed some elements which have noticeable difference. Therefore it is necessary to do, before of making decision, some reduction of influence of these model elements on the final decision about recognition. Also in speaker models are noticeable the elements which do not shows significant temporal variability. These model elements may be considered as speaker inherently elements of models with respect to others elements of models. Model elements which have significant time variability, as consequence of the fact that feature vectors are determined by short time spectral analisys of speech signal, also have some information about speaker. Procentualy, these elements of models have the lesser quantity of relevant information about inherent speaker. To remain as high as possible amount of information about speaker as one solutions can be: to do some weighting of the elements of model for the feature vector of maximal dimensionality.

To determine the quantity of time variability of elements of the speaker models it is necessary to define some rule. For detecting the quantity of changes between the same model elements of the same speakers we compare the training and test covariance matrices of the same speakers and calculate matrix of distinctions

$$D(i,j) = d\Sigma_{i,j} = \sum_{n=1}^{N} \left| \Sigma_{(i,j)(n)}^{training} - \Sigma_{(i,j)(n)}^{test} \right|, \qquad (6)$$

where i and j are the indices of the observed element in training and test covariance matrices and N = 121 is the number of speakers in speech database. In each of 14 tests appropriate matrix of distinctions was calculated. Values of appropriate weighting coefficients, as well as they regions of applying are empirically estimated during testing.

III. RESULTS

The time variability of model elements is undesirable phenomena. Elements characterized by very little time variability have the largest amount of relevant information about speaker. As the time variability of an observed element in speaker model increases, his importance and the amount of relevant information about inherent speaker decreases. The time variability of some element of model, $\Sigma(i,j)$, was determined by appropriate value in distinction matrix $D(i,j)$. Elements in matrix of distinctions which have large values marks that in these elements the covariance matrices i.e. speaker models have high intra-speaker variability and these elements are less important for efficient automatic speaker recognition. Therefore the determination of elements of weighting matrix is based on the comparison of elements in appropriate distinction matrix. The rule for determining of weighting coefficients $W(i,j)$ and $W(k,l)$ for elements of distinction matrix $D(i,j)$ and $D(k,l)$ can be expressed in the following manner

$$if \quad D(i,j) \geq D(k,l) \quad then \quad W(i,j) \leq W(k,l). \qquad (7)$$

So the weighted coefficients are the largest for elements of model characterized by the smallest values in distinction matrix. This inverse proportionality between the elements of distinction matrix $\boldsymbol{D}$ and appropriate elements of weighting matrix $\boldsymbol{W}$ in a certain way decreases the elements of models which have high degree of intra-speaker variability. On this way the importance of model elements was equalized.

Weighting coefficients are determined for each of 14 experiments separately. The main step for determining these coefficients is the determination of maximum, *max*, value in appropriate matrix of distinctions. Weight coefficients are determined according to the following rules:

$$if \quad D(i,j) > \frac{\max}{5} \quad then \quad W(i,j) = 0.05, \qquad (8)$$

$$if \quad \frac{\max}{10} < D(i,j) \leq \frac{\max}{5} \quad then \quad W(i,j) = 0.3, \qquad (9)$$

$$if \quad \frac{\max}{20} < D(i,j) \leq \frac{\max}{10} \quad then \quad W(i,j) = 0.6, \qquad (10)$$

$$if \quad \frac{\max}{40} < D(i,j) \leq \frac{\max}{20} \quad then \quad W(i,j) = 1.0, \qquad (11)$$

$$if \quad D(i,j) \leq \frac{\max}{40} \quad then \quad W(i,j) = 1.9. \qquad (12)$$

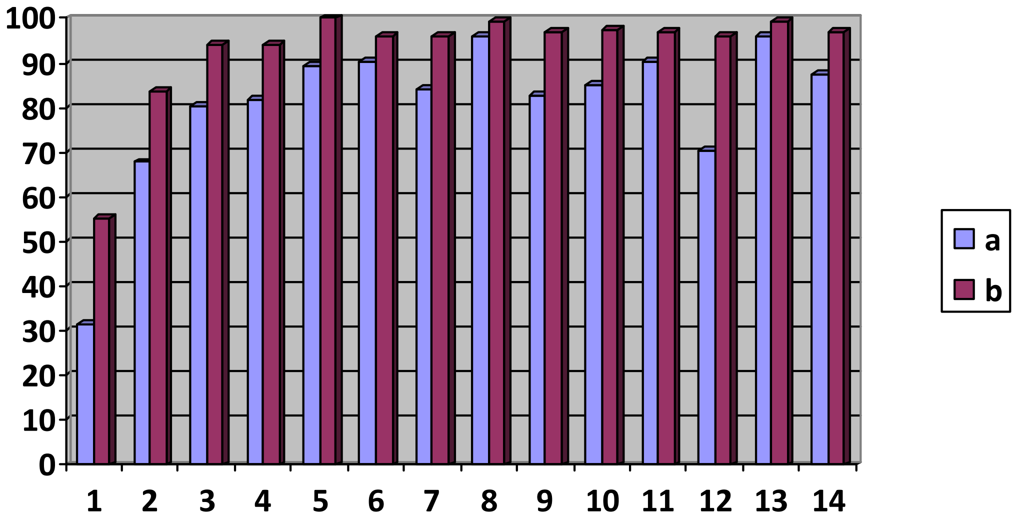

Fig. 1. Percentage accuracy of speaker recognition depending on the test file used (**1** – “Names”, **2-3** – “Digits”, **4-14** – “Words”) and the model applied: **a** – the full covariance matrix for the feature vector which contains zeroth and first 19 MFCCs, **b** – the same full covariance matrix with correction of elements as in (13).

For determined weight matrix in the appropriate k^{th} test, $W^{(k)}$ *and* $1 \le k \le 14$, the elements of final speaker model, $\Sigma_i^{final}(i,j)$, were calculated as product of element in original model, i.e. in covariance matrix which models speaker (3), $\Sigma_i(i,j)$, and appropriate element in weight matrix $W^{(k)}(i,j)$

$$\Sigma_i^{final}(i,j) = \Sigma_i(i,j) \cdot W^{(k)}(i,j). \tag{13}$$

As is evident on Fig. 1 by application of these transformations the recognition accuracy was increased (case b). These results are better with respect to results when elements of models on positions by the largest values in distinction matrix, $\Sigma_{0,0}$ and $\Sigma_{19,19}$, are discarded or when complete zeroth and 19th MFCCs are discarded [9]. The relative difference between maximal and minimal values in distinction matrices of observed tests is of the order of 10^3. Because of this large range in which are the elements of distinction matrices it is assumed that one fifth of maximum in distinction matrices is the limit value for the least significant elements in covariance matrices of speaker models. These elements were weighted by very small value of $W(i,j) = 0.05$. On this way the expected model elements of high intra-speaker variability are significantly decreased. The next boundaries are referred to the first, limit boundary, in order to define additional parameters for pondering of elements of original speaker model. In order to achieve better recognition it is evident that it is not enough to decrease only the impact of elements of models for which distinction matrix have the biggest values or values which close to the biggest values. Therefore the width of first range is the largest. For this range the weighting coefficients are the smallest. The test recordings in the test 1 are the shortest, therefore recognition accuracy in this test is the smallest also after application of transformation on models of speakers.

IV. CONCLUSIONS

MFCCs depend on the energy in observed speech frame. Therefore they are also text-dependent. This dependence is mapped to covariance matrices, which are used as speaker models, and manifested in difference between training and test models of the same speakers, i.e. in intra-speaker variability of speaker models. Therefore, some procedure for decreasing of model intra-speaker variability can be introduced. One approach presented in this paper provides improvements in accuracy of automatic speaker recognition. By multiplying appropriate model element and the weighting coefficient text dependency of that element as well as complete model is decreased. Results show that it is necessary to determine the measure of the model elements variability. In real applications when identity of the test model is unknown this can be done by analysis and comparing of different training models of the same speakers.

ACKNOWLEDGMENT

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

REFERENCES

[1] T. Kinnunen, H. Li, “An overview of text-independent speaker recognition: From features to supervectors”, *Speech Communication 52*, pp. 12–40, 2010. [Online]. Available: http://dx.doi.org/10.1016/j.specom.2009.08.009

[2] L. L Molgaard, K. W Jorgensen, “Speaker recognition – special course”, *IMM-DTU*, 2005, p. 28. [Online]. Available: http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/4414/pdf/imm4414.pdf

[3] J. P. Campbell, Jr., “Speaker recognition: a tutorial”, in *Proc. IEEE*, vol. 85, no. 9, 1997, pp. 1437–1462. [Online]. Available: http://dx.doi.org/10.1109/5.628714

[4] V. D. Delic, M. S. Secujski, N. M. Jakovljevic, “Action model of human-machine speech communication”, in *Proc. 16th Telecommunications forum (TELFOR 2008)*, Serbia, Belgrade, 2008, pp. 680–683. (in Serbian)

[5] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacretaz, D. A. Reynolds, “A tutorial on text-independent speaker verification”, *EURASIP Journal on Applied Signal Processing*, vol. 4, pp. 430–451, 2004. [Online]. Available: http://dx.doi.org/10.1155/S1110865704310024

[6] B. Salna, J. Kamarauskas, “Evaluation of effectiveness of different methods in speaker recognition”, *Elektronika ir Elektrotechnika*, no. 2, pp. 67–70, 2010.

[7] I. Jokic, S. Jokic, Z. Peric, M. Gnjatovic, V. Delic, “Influence of the number of principal components used to the automatic speaker recognition accuracy”, *Elektronika ir Elektrotechnika*, no. 7, pp. 83–86, 2012.

[8] B. R. Widermoth, “Text-independent speaker recognition using source based features”, M.S. thesis, Griffith University, Brisbane, Australia, 2001, p. 101.

[9] I. Jokic, V. Delic, S. Jokic, Z. Peric, “Influence of the discarding non-speaker specific model parameters and features to accuracy of automatic speaker recognition”, in *Proc. Second Int. Conf. (TAKTONS 2013)*, Novi Sad, Serbia, 2013, pp. 96–99.