ELECTRONICS AND ELECTRICAL ENGINEERING

ISSN 1392 – 1215 2012. No. 7(123)

ELEKTRONIKA IR ELEKTROTECHNIKA

SIGNAL TECHNOLOGY

SIGNALŲ TECHNOLOGIJA

Influence of the Number of Principal Components used to the Automatic Speaker Recognition Accuracy

I. Jokic, S. Jokic, M. Gnjatovic, V. Delic

University of Novi Sad-Faculty of Technical Sciences,
Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia, phones: +381 64 35 26 245, +381 21 485 25 33,
e-mails: ibahjokih@gmail.com, stevan.jokic@gmail.com, milangnjatovic@yahoo.com, vdelic@uns.ac.rs

Z. Peric

Department of Telecommunications, Faculty of Electronic Engineering, University of Niš,
Aleksandra Medvedeva 14, 18000 Niš, Serbia, phone: +381 18 529 225, e-mail: zoran.peric@elfak.ni.ac.rs

crossref http://dx.doi.org/10.5755/j01.eee.123.7.2379

Introduction

It is widely accepted that voice is a behavioral trait that may be used as a biometric characteristic. Systems intended to exploit this property of voice must be able to make a credible representation of the observed voice, and equipped with reliable procedures for automatic speaker recognition [1]. Feature extraction is one of the key components of automatic speaker recognition systems. Mel-Frequency Cepstral Coefficients (MFCCs) and their first and second derivatives are often used as a feature set. The observed speech signals are characterized by their predictive nature. MFCCs may be considered as a direct consequence of discretization of the spectrum envelope, and their derivatives as linear combinations of any two adjacent MFCCs in the infinitesimal time. Therefore, it follows that these features are mutually redundant [2].

On the other hand, the basic problem of achieving a high level of reliability of speaker recognition lies in the fact that training and testing conditions differ. Since each feature is susceptible to environmental interference, the increased complexity of a speaker model (i.e., the increased dimensionality of feature vectors) implies an increased level of noise. From the methodological point of view, to prevent such a level of noise, it is necessary to focus the recognizer on what is important in the observed recognition object. In other words, reduction of the model complexity would reduce the accumulated noise level.

These observations motivate possible efficiency improvements of a speaker recognizer. Transformation techniques that are usually applied for dimensionality reduction include Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Nonlinear Discriminant Analysis (NLDA) [3]. All these techniques are intended to train the recognizer to focus on those elements that are essential for the observed data set.

Recent work indicates that PCA is a prominent technique for dimension reduction. Conventional methods for PCA based on the full data covariance matrix require a large amount of training data [4]. In order to reduce the complexity of these methods where the eigenvector matrix of each speaker is calculated, methods for PCA that are performed on all the training data [4, 5] (and this paper), or on locally clustered data [6, 7] are introduced. In the domain of speaker recognition, PCA is often applied in Gaussian Mixture Models (GMMs) [4, 7]. In the next section, we introduce the reference environment and the modus of speaker modeling. Then, we consider the feature transformation and training of the model. Finally, we discuss the results of automatic speaker recognition.

Modus of speaker modeling

The study presented in this paper was supported by a speech corpus developed by the AlfaNum group at the Faculty of Technical Sciences, University of Novi Sad. Utterances produced by five female and five male speakers were selected from this corpus, and may be classified as follows:

(i) Digits – Each speaker produced two utterances containing only words that correspond to digits: “one two three four five” and “six seven eight nine zero”. The mean total duration of audio recordings for a speaker is 12.8s;

(ii) Words – Each speaker produced the same set of eleven preset word sequences. All the sets are disjoint. The mean total duration of audio recordings for a speaker is 70.7 s;

(iii) Names – Each speaker produced an utterance comprising his or her first name, family name, and the speaker-specific identification number. The mean duration of audio recording for a speaker is 2.5 s.

All utterances selected from the corpus were recorded only once, and, thus, were appropriate for introducing models for text-independent speaker recognition. In our previous work [8], we introduced a speaker model, shown in Fig. 1, and a speaker recognizer, to discuss the impact of telephone channels to automatic speaker recognition. In this study, we apply the same model in order to examine impacts of feature vector dimension reduction on the accuracy of speaker recognition.

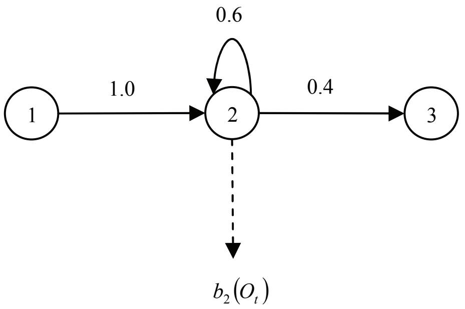

Fig. 1. A speaker model

For modeling purposes, we used the Hidden Markov Models Toolkit (HTK) [9]. For speaker modeling, we applied a Hidden Markov Model (HMM) with one emitting state only (Fig. 1). Distribution of feature vectors inside the emitting state was modeled as Gaussian Mixture Model (GMM)

$$b_2(O_t)=\sum_{k=1}^{64}\frac{1}{64}\cdot \mathrm{N}(O_t,\mu_k,\Sigma_k), \tag{1}$$

where

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

and d is the dimensionality of the used feature vectors.

The recordings from the utterance groups "*Digits*" and "*Words*" were used for training the models, while the recordings from the utterance group "*Names*" were used for testing the recognition accuracy.

From MFCC to principal components

We used a linear implementation of the PCA technique. The transformation is based on the appropriate choice of eigenvectors of the observed covariance matrix. The matrix was obtained from the whole training set *X*. Before calculating the covariance matrix, inside each dimension of the training set, a normalization is performed with the mean value of the given dimension, and the matrix X_{norm} is derived. The covariance matrix is then calculated as

$$C=\frac{1}{n-1}\cdot X_{norm}^{'}\cdot X_{norm}\,, \tag{3}$$

where *n* is the number of training vectors, i.e., the number of components in each of the observed dimensions in the training set.

The derivation of the eigenvectors, $EVect$, is based on the fact that each eigenvector is described by the appropriate eigenvalue, $EVal$, with respect to the used covariance matrix C [4], as follows

$$C\cdot EVect=EVal\cdot EVect\,. \tag{4}$$

The source feature vectors consist of zero MFCC, first 12 MFCCs and their first and second derivatives. Therefore, the covariance matrix C that corresponds to the entire training set *X* is quadratic, with dimensions 39×39. Such matrix is characterized by 39 eigenvectors and 39 corresponding eigenvalues (Fig. 2).

The eigenvectors associated with high eigenvalues represent dominant directions of variance in the observed data set. Therefore, during the formation of the transformation matrix C_{tr}, the eigenvectors with the highest eigenvalues are firstly taken into account. This means that in the one-dimensional PCA space, the transformation matrix was created only with the eigenvector that is characterized by the highest eigenvalue (in this study, this value was approximately 300.887, Fig. 2).

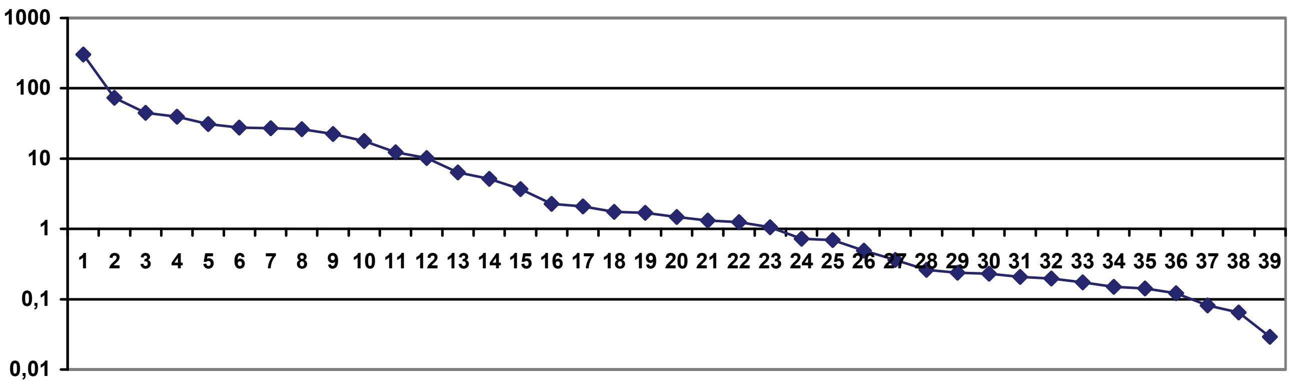

Fig. 2. Eigenvalues in descending order (1 to 39)

Generally, a *d*-dimensional transformation matrix, where $d \in \{2,3,...,39\}$, is created with *d* eigenvectors that are characterized by the *d* highest eigenvalues. For the source feature vector x, the transformed feature vector x_{tr} is derived as follows

$$x_{tr} = C_{tr} \cdot x^T . \qquad (5)$$

Training of models and testing of recognizer

The sampling frequency of the speech recordings selected from the corpus was 22050Hz. Feature extraction was performed on the speech signal segments by applying 25ms Hamming window shifted every 10ms.

The training and testing of the speaker recognizer were first conducted using the standard 39-dimensional MFCC feature vectors. In the next study phase, we applied an adequate *d*-dimensional PCA on source feature vectors, and used the obtained feature vectors for training the speaker models (Fig. 1), (1), (2). We applied the HTK function HERest to estimate parameters of the model. This function performs the re-estimation of models parameters using the Baum-Welch algorithm with embedded training. The HERest can be reapplied on the previously estimated model parameters in order to obtain a better estimation of speaker models. In this study three HERest estimations are used.

For the purpose of ensuring compliance with the procedure of training, the testing of recognition accuracy is also performed with *d*-dimensional feature vectors. Source test MFCC feature vectors are transformed with appropriate *d*-dimensional transformation matrix, as in (5).

Results

Our point of departure was to apply the PCA technique in order to reduce the dimensionality of feature vectors. Here, we discuss the recognition accuracy with respect to feature vector dimensionality. The recognition accuracy results of the introduced model are compared with the recognition results when a classical 39-dimensional MFCC feature space was used.

The testing results are summarized in Fig. 3. Previous speaker recognition tests with the classical feature space resulted in the accuracy of 100% [8]. As expected, feature vectors with extremely reduced dimensionality do not describe speakers properly. For example, considering feature vectors that are characterized by one dimension only, the eigenvalue that describes the domination of this dimension is more than four times higher than the next eigenvalue, $EVal_2 \approx 73.61$. In these cases, the speaker recognizer shows a significant decrease of recognition accuracy. The introduction of additional dimensions improves the recognition accuracy. Already for the 5-dimensional feature vectors, the recognition accuracy increases from 10 percent to 50 percent (Fig. 3). By further increasing the dimensionality of the observed feature vectors, the recognition accuracy significantly increases. Already for the 14-dimensional transformed feature vectors (Fig. 3), the recognition accuracy reaches the target value of 100 percent. In other words, it is not necessary to use the classical 39-dimensional MFCC feature vectors in order to achieve reliable speaker recognition.

In addition, the results suggest that the expectation that an increase in dimensionality implies an increase of the recognition accuracy cannot be used as a general rule. Thus, for 10-dimensional feature vectors, the accuracy is 70 percent, while for 9-dimensional feature vectors, the accuracy is – in contrast to expectations – 80 percent. This is a consequence of the linear nature of the PCA transformation. Namely, although PCA compactly groups transformed feature vectors for the observed speaker, it does not prevent overlapping or embedding of speaker models.

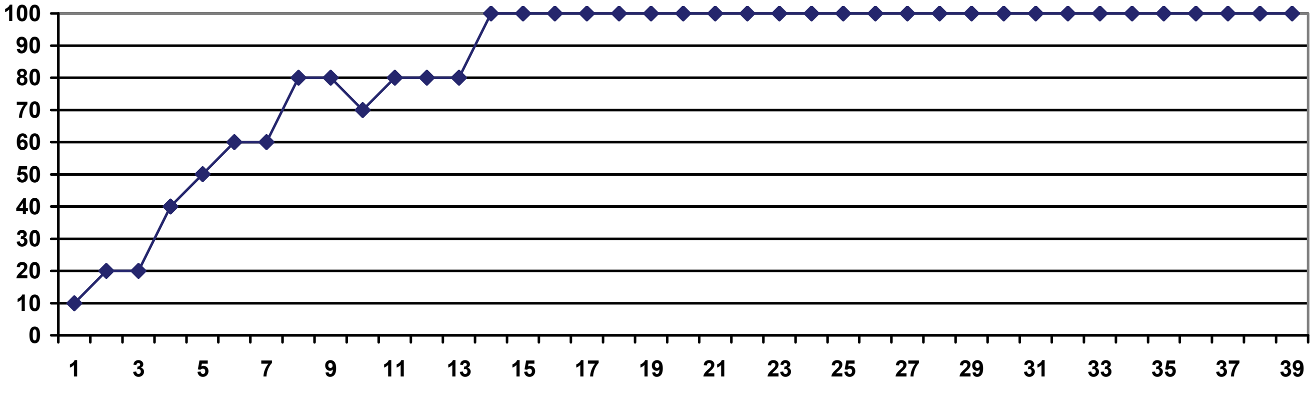

Fig. 3. Recognition accuracy with respect to used dimensionality

Consequently, when overlapping or embedding of the speaker models take place (as it happened when we tested 10-dimensional feature vectors), this may reduce the recognition accuracy.

It can be also observed that sometimes two consecutive experimental conditions provide the same recognition accuracy results. In the cases of 2-dimensional and 3-dimensional feature vectors, it is important to note that, although the accuracy result is the same, the recognized identities differ. It is evident that various transformations have different impacts on feature vectors of the observed speakers. In contrast to that, in the cases of

8-dimensional and 9-dimensional feature vectors, the recognized identities do not differ. In these cases, the applied transformations show a similar behavior with respect to the observed speakers.

Conclusions

Based on the obtained experimental results, it is evident that applying PCA on the feature vectors opens a possibility for the construction of reliable automatic speaker recognizers. It enables a dimensionality reduction that is necessary for decreasing the complexity of speaker models. In other words, it is possible to speed up decision making processes of the recognizer, and, in the same time, to preserve the recognition accuracy achieved when the standard 39-dimensional MFCC feature vectors were used. In the experiments reported in this paper, the reliable recognition was achieved already with 14-dimensional feature vectors.

The technique of PCA is applied to find the necessary and sufficient dimensionality of the observed feature vectors for reliable automatic speaker recognition. Thus, a system for automatic speaker recognition will take into account only those properties of speakers that belong to selected dimensions. Since these dimensions are determined as directions of the most prominent variance in the observed MFCC feature space, inside each of them a maximal signal-to-noise ratio in a given environment can be expected. Thus, performing the training of models within these dimensions, and focusing the recognizer on them, can be expected to increase the level of robustness of automatic speaker recognition.

Acknowledgements

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 and Science of the Republic of Serbia. The responsibility for the content of this paper lies with the authors.

References

1. **Šalna B., Kamarauskas J.** Evaluation of Effectiveness of Different Methods in Speaker Recognition // Electronics and Electrical Engineering. – Kaunas: Technologija, 2010. – No. 2(98). – P. 67–70.
2. **Wang X., O’ Shaugnessy D.** Improving the efficiency of Automatic Speech Recognition by Feature Transformation And Dimensionality Reduction // In Proc. of Interspeech 2003. – Geneva, Switzerland, 2003. – P. 1025–1028.
3. **Wu D., Li B., Jiang H.** Normalization and Transformation Techniques for Robust Speaker recognition // Source: Speech Recognition, Technologies and Applications. – I-Tech, Vienna, Austria, 2008. – 550 p.
4. **Kim M., Kim E., Seo C., Jeon S.** Speaker Verification and Identification Using Principal Component Analysis Based on Global Eigenvector Matrix // Hybryd Artificial Intelligence Systems, Lecture Notes in Computer Science. – Springer–Verlag Berlin Heidelberg. 2010. – Vol. 6076. – P. 278–285.
5. **Hatch O. A., Kajarekar S., Stolcke A.** Within–Class Covariance Normalization for SVM–based Speaker Recognition // In: Proc. Interspeech 2006 (ICSLP). – Pittsburgh, Pennsylvania, USA, 2006. – P. 1471–1474.
6. **Hanilci C., Ertas F.** VQ–UBM Based Speaker Verification Through Dimension Reduction Using Local PCA // 19th European Signal Processing Conference (EUSIPCO’2011), Barcelona, Spain, 2011. – P. 1303–1306.
7. **Lee J., Rheem Y. J., Lee Y. K.** GMM Based on Local Fuzzy PCA for Speaker Identification // IDEAL’2003, LNCS 2690. – Springer–Verlag Berlin Heidelberg, 2003. – P. 1000–1007.
8. **Jokić I., Jokić S., Gnjatović M., Sečujski M., Delić V.** The Impact of Telephone Channels on the Accuracy of Automatic Speaker Recognition // Telfor Journal. – Belgrade: Telecommunications Society, Academic Mind, 2011. – Vol. 3. – No. 2. – P. 100–104.
9. **Young S., Evermann G., Gales M., Hain T., Kershav D., Liu X., Moore G., Odell J., Ollason D., Povey D., Valtchev V., Woodland P.** The HTK Book (for HTK Version 3.4). – Microsoft Corporation, Cambridge University Engineering Department, 2009. – 384 p.

Received 2011 12 13
Accepted after revision 2012 01 14

I. Jokic, S. Jokic, M. Gnjatovic, V. Delic, Z. Peric. Influence of the Number of Principal Components used to the Automatic Speaker Recognition Accuracy // Electronics and Electrical Engineering. – Kaunas: Technologija, 2012. – No. 7(123). – P. 83–86.

This paper discusses possibilities to reduce dimensionality of the standard MFCC feature vectors by applying the technique of Principal Component Analysis (PCA). The reported experimental results suggest that PCA is an appropriate technique to reduce dimensionality without reducing the accuracy of recognition. The applied automatic speaker recognizer shows that already for a 14-dimensional PCA feature space, the recognition accuracy reaches the target value as in the 39-dimensional MFCC feature space. This gives motivation for further research towards more efficient speaker recognizers. Ill. 3, bibl. 9 (in English; abstracts in English and Lithuanian).

I. Jokic, S. Jokic, M. Gnjatovic, V. Delic, Z. Peric. Pagrindinių komponentų kiekio įtaka automatinio kalbančiojo atpažinimo tikslumui // Elektronika ir elektrotechnika. – Kaunas: Technologija, 2012. – Nr. 7(123). – P. 83–86.

Nagrinėjama galimybė sumažinti standartinių MFCC požymių vektorių dimensijas pritaikant pagrindinių komponentų analizę (PKA). Pateikti eksperimentiniai rezultatai rodo, kad PKA yra tinkamas metodas dimensijoms sumažinti nesumažinant atpažinimo tikslumo. Naudojant automatinį kalbančiojo atpažinimo įrenginį parodyta, kad net esant 14-matei PKA požymių erdvei atpažinimo tikslumas pasiekia norimą lygį kaip ir 39-matėje MFCC požymių erdvėje. Tai skatina tęsti tyrimus kuriant efektyvesnius kalbančiojo atpažinimo įrenginius. Il. 3, bibl. 9 (anglų kalba; santraukos anglų ir lietuvių k.).