VQ-UBM BASED SPEAKER VERIFICATION THROUGH DIMENSION REDUCTION USING LOCAL PCA

Cemal Hanilci and Figen Ertas

Department of Electronic Engineering, Uludag University
16059, Bursa, Turkey
phone: + (90) 2242942074, fax: + (90) 2242942903, email: {chanilci,fertas}@.uludag.edu.tr

ABSTRACT

The universal background model (UBM) based classifiers have recently been popular for speaker recognition. In this paper, we propose a dimension reduction method using local principal component analysis to improve the performance of speaker verification systems, where maximum a *Posteriori* (*MAP*) adapted vector quantization classifier (VQ-MAP or VQ-UBM) is employed. The proposed system first partitions the UBM data into disjoint regions (clusters) via conventional VQ algorithm and PCA is performed on the set of feature vectors in each region to obtain transformation matrix. Then, multiple speaker model is constructed using the set of transformed feature vectors closest to each cluster through MAP adaptation. Conducting experiments on NIST 2001 SRE, it is shown that transforming the data onto a lower dimensional space by the proposed method improves the recognition accuracy.

1. INTRODUCTION

Text-independent speaker verification has been a challenging area in speech technology over decades. In speaker verification the aim is to determine if the given identity claim is true or false. Gaussian mixture model with universal background model (GMM-UBM) [1] has become a standard technology in speaker recognition. It shows great performance compared to conventional systems such as vector quantization (VQ) [2] and GMM [3]. Recently Hautamaki et. al. introduced the maximum a *Posteriori* adaptation of VQ algorithm (VQ-UBM) which adapts the centroid vectors via *MAP* algorithm and showed that VQ-UBM provides recognition accuracy as good as GMM-UBM with a significant speed-up [4]. Kinnunen et. al compared the VQ-UBM with GMM-UBM and support vector machines (SVMs) on text-independent speaker verification and observed that VQ-UBM gives better recognition accuracy than GMM and SVMs on longer training and test utterances [5].

Mel-frequency cepstrum coefficients (MFCCs) are the most popular features used in speaker recognition. In speaker recognition literature, it was shown that appending the first and second order derivatives of MFCCs to the feature sets improves recognition accuracy significantly, so current speaker recognition systems use dynamic features (also known as delta and double delta coefficients) addition to MFCCs. However, this addition improves the dimensionality so the computation time. There are plenty of work in the literature which reduces the dimension of feature space while keeping the performance. In [6], a Genetic Algorithms (GA) based feature selection algorithm was presented and compared with two well-known feature transformation techniques, principal component analysis (PCA) and linear discriminant analysis (LDA). In [7], Heteroscedastic Linear Discriminant Analysis (HLDA), which provides a linear transformation that can de-correlate the features and reduce the dimensionality, was applied to GMM-UBM speaker recognition system. Seo et. al. reported that transforming training and test data using principal component analysis (PCA) with dimension reduction according to local information improves the conventional GMM based speaker identification rate while reducing the number of required parameters [8]. In [8], the training data is first divided into disjoint regions with K-means clustering algorithm and then the set of feature vectors which belong to each region is used to construct transformation matrix via PCA. Once the transformation matrix is obtained for each region then training and test data are transformed into lower dimension space and for each speaker multiple GMMs (one GMM per each region) are constructed. In [9], Lee introduced local fuzzy PCA based GMM which creates the regions using fuzzy clustering algorithm followed by PCA for each region and concluded that this technique also gives comparable speaker identification rates with reduced dimension of data. The details about the computation of required parameters in this methods can be found in [8], [9]. Our work is different from [8] and [9] in two ways: first we introduce local PCA on recently proposed *MAP* adapted VQ (VQ-UBM) which is modern and accurate on speaker recognition and second, we perform speaker verification whereas [8] and [9] use conventional GMM based speaker identification. Furthermore, we report our results on the more challenging NIST 2001 corpus which consists of conversational telephony speech that is recorded over the cellular telephone network.

The rest of the paper is organized as follows. The PCA algorithm and local PCA based VQ-UBM is described in section 2. In section 3 the experimental results are given including speaker recognition database, feature extraction and implementation of proposed system. Finally the conclusions are summarized in section 4.

2. LOCAL PCA BASED VQ-UBM

The local PCA based speaker recognition system is shown in Figure 1. In this section we briefly describe the PCA algorithm and local PCA based VQ-UBM speaker recognition system.

2.1 PCA Algorithm

PCA algorithm [10] (also known as Karhunen-Loeve Transform) is a powerful technique for feature extraction (especially in image processing) and dimension reduction. It projects the high dimensional data onto lower dimensional

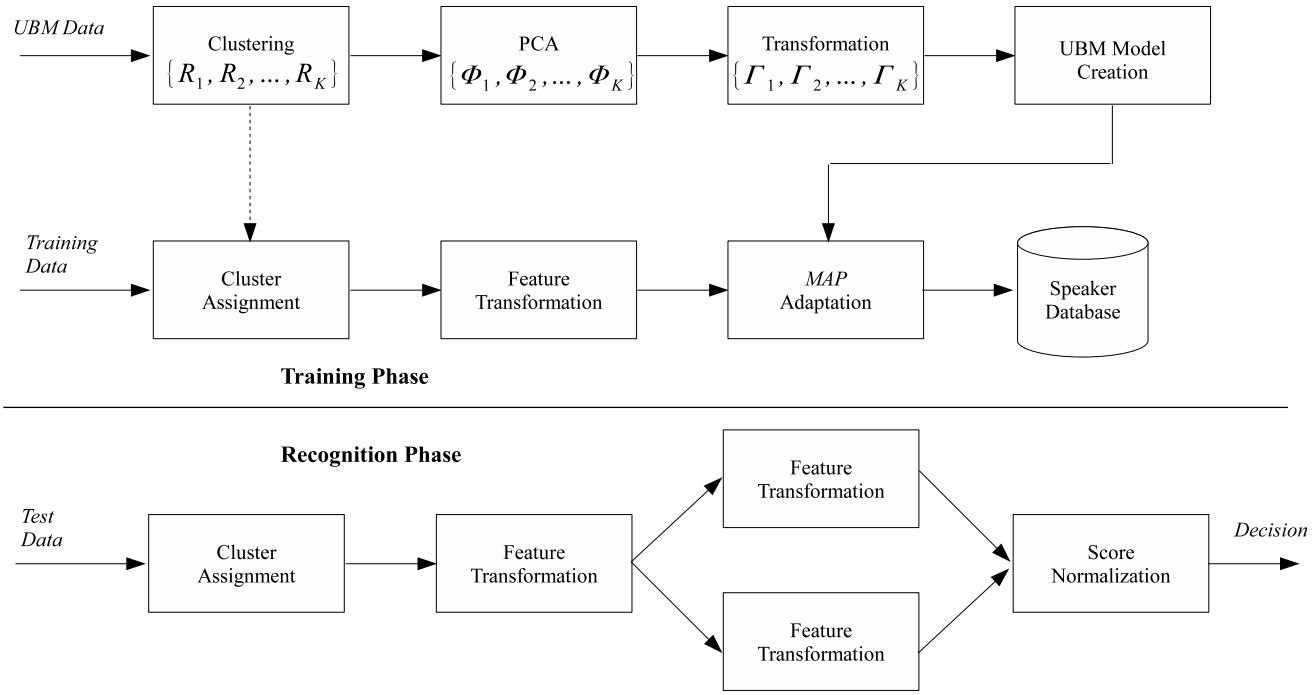

Figure 1: Local PCA based speaker recognition system

orthogonal space. In literature, PCA algorithm has been used in many speaker recognition systems. In our earlier study we proposed to use PCA as a classification method based on minimum projection error for text-independent speaker identification [11]. In [12], a new classifier presented called principal component space and with another PCA classifier, it was used as a hybrid classifier for speaker recognition. PCA algorithm can be briefly described as follows: Suppose that X is the set of n dimensional feature vectors, $X = \{x_1, x_2, ..., x_T\}$.The covariance matrix of X is calculated as:

$$\Sigma = \frac{1}{T}(X - \overline{X})(X - \overline{X})^T \quad (1)$$

where $\overline{X}$ is the sample mean of X. Let $\{\lambda_1, \lambda_2, ..., \lambda_n\}$ be the eigenvalues of covariance matrix, Σ, ordered from largest to smallest and $\Phi = \{w_1, w_2, ..., w_n\}$ be the corresponding eigenvectors. The matrix Φ is defined as the transformation matrix which projects the original data X onto orthogonal feature space. The feature vectors are transformed as:

$$Y = \Phi X \quad (2)$$

If the aim is to reduce the dimension of data, the transformation matrix Φ will consist of first L eigenvectors which is associated with largest L eigenvalues, where L is the new dimension.

2.2 Local PCA Based VQ-UBM

VQ-UBM algorithm is based on *MAP* adaptation of clustering algorithm. It first creates a codebook with desired model order (number of centroid or size of codebook) using UBM data and then each speakers feature vectors are adapted using background model through *MAP* adaptation.

Local PCA based VQ-UBM algorithm can be described as follows. First the UBM data is mapped to K disjoint regions, R_j, $j = 1, 2, ..., K$ via conventional K-means clustering algorithm. Each region is represented by its own reference vector (centroid). Then the set of feature vectors of UBM data for each region is found, $S_j = \{x \in R_j\}$. PCA algorithm is performed on the set of vectors for each region and the data of each region are transformed onto orthogonal space. The transformation matrix of each region, Φ_j, $j = 1, 2, ...K$, is stored in order to be used in training and testing phases of speaker recognition system. The background model, U, for each region is created using conventional VQ algorithm with model order M, $U = \{U_1, U_2, ..., U_K\}$. Where U_j is the codebook of j^{th} region that is created using the set of feature vectors which are closest to j^{th} region.

During training of a speaker, the feature vectors of the speaker, $X = \{x_1, x_2, ..., x_T\}$, are assigned to regions, R_j, $j = 1, 2, ..., K$, and feature vectors are transformed by using the transformation matrix of each region and a speaker model (codebook) is constructed via *MAP* adaptation with the set of transformed feature vectors for each region. For a single speaker, K models are created, $C_i = \{C_1^i, C_2^i, ..., C_K^i\}$, where C_j^i is the codebook of i^{th} speaker for j^{th} region.

In the testing phase, given a sequence of transformed feature vectors $Y = \{y_1, y_2, ..., y_N\}$ and claimed identity model, $C_i = \{C_1^i, C_2^i, ..., C_K^i\}$, a match score is computed. The match score is computed as:

$$Score = \sum_{j=1}^{K} MSE(Y_j, U_j) - \sum_{j=1}^{K} MSE(Y_j, C_j) \quad (3)$$

where Y_j, U_j and C_j are the set of vectors, background model and speaker model for the j^{th} region, respectively. *MSE* is

the mean squared error which is defined as:

$$MSE(X,Y) = \frac{1}{|X|} \sum_{x_i \in X} \min_{y_k \in Y} \|x_i - y_k\|^2 \qquad (4)$$

where $\|x_i - y_k\|^2$ is the squared Euclidean distance between the vectors x_i and y_k.

3. EXPERIMENTAL SETUP

3.1 Corpora

Experiments are conducted on NIST 2001 speaker recognition evaluation (SRE) [13]. NIST 2001 database consists of 174 speakers (74 males and 100 females) with 22,418 trials (2038 target trials and 20,380 impostors). Database contains conversational telephone speech in English. The duration of training data for each speaker is two minutes and length of test data varies from a few seconds up to one minute. We used developments set of database to create UBM models.

3.2 Feature Extraction

We use mel-frequency cepstrum coefficients (MFCC) as speaker-specific features in the experiments. First an energy based voice activity detector is applied to the speech signal to remove silence parts. MFCCs are extracted using 30 milliseconds Hamming windowed frames with 10 milliseconds overlap. 27-channel triangular filterbank is used during the extraction. Logarithmic filterbank outputs are converted into cepstral coefficients by discrete Cosine transform (DCT). The deltas (Δ) and double deltas ($\Delta\Delta$) are appended to MFCC features which yields 36 dimensional feature vectors. Delta features are obtained by convolving the MFCC feature vectors with the kernel $h = [1\ 0\ -1]$ and double deltas are computed by applying the same kernel to the delta features as in [5]. The last step is the cepstral mean and variance normalization (CMVN) to obtain features with zero mean and unit variance for removing the contribution of convolutional noises.

3.3 Experimental Results

In the experiments we used equal error rate (EER) as the performance criterion. EER is the threshold value which gives equal false acceptance rate (FAR) and false rejection rate (FRR). $M = 128$ and $M = 256$ model orders (size of codebook) were used. First we analyzed the effect of number of dimension and regions on each gender separately. 74 male speakers of database consist of 850 target and 8500 impostor trials whereas 100 female speakers of database contain 1188 target and 11,880 impostor trials. Figure 2 and Figure 3 show the relationship between the system performance, number of regions, K, and number of dimension of transformed feature vectors, L, for male and female speakers, respectively. In figures the baseline system (dotted line) corresponds to conventional VQ-UBM algorithm with model order $M = 128$. It can be seen from both figures using local information of $K = 2$ regions improves the verification performance with reduced feature dimension ($L = 26$). For female speakers $L = 25$ with $K = 2$ regions gives almost equal EER with the baseline system. For both gender using two regions ($K = 2$) is superior to the case of using four regions ($K = 4$)in local PCA.

Next we investigated the proposed system performance of all trials (males and females). Number of regions and

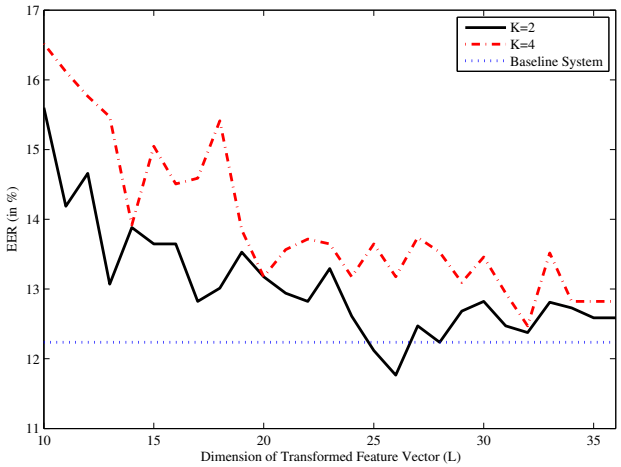

Figure 2: EERs for $K = 2$ and $K = 4$ for male speakers with model order $M = 128$

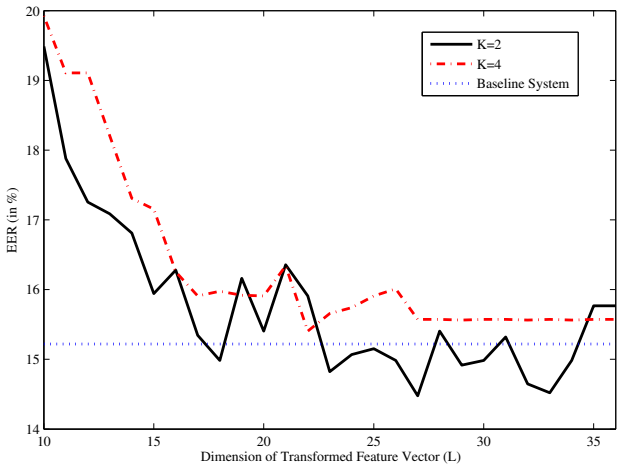

Figure 3: EERs for $K = 2$ and $K = 4$ for female speakers with model order $M = 128$

number of transformed feature dimensions are fixed to $K = 2$ and $L = 26$, respectively. The detection error trade-off (DET) curves are given in Figure 4 and Figure 5 for model orders $M = 128$ and $M = 256$, respectively. It can be seen from figures 4 and 5 that proposed system improves the recognition accuracy for both model orders. As seen from figures, the proposed method with $M = 128$ yields better speaker recognition performance than the baseline VQ-UBM system with $M = 256$ model order. If we simply define the model complexity of baseline VQ-UBM as the feature dimension (n) times model order (M), $C_b = n \times M$, and the complexity of proposed method as the number of regions (K) times number of reduced feature dimension (L) times number of model order (M), $C_p = K \times L \times M$, then it can be computed that $C_b = 9216$ ($n = 36$ and $M = 256$) and $C_p = 6656$ ($K = 2$, $L = 26$, and $M = 128$). From the results it can be concluded that, using local information yields to construct more representative speaker models and proposed system outperforms the baseline system performance with less model complexity.

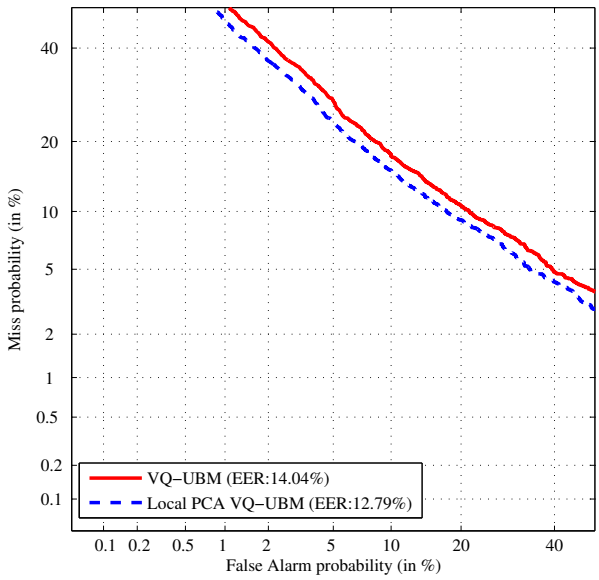

Figure 4: DET curves for all trials with model order $M = 128$

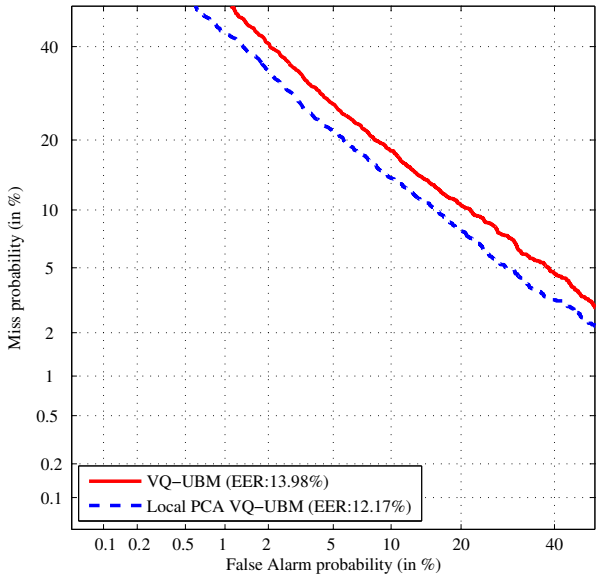

Figure 5: DET curves for all trials with model order $M = 256$

4. CONCLUSIONS AND FUTURE WORK

We have introduced the local PCA algorithm on the recently proposed *MA*P adapted VQ (VQ-UBM) on speaker verification. The experiments show that the using local information and reducing the dimension of feature vectors via PCA algorithm improves the speaker verification performance. It is observed that using $K = 2$ local regions and $L = 26$ (the dimension of transformed vector) are the best choices of parameters for the proposed system. For future work, applying the proposed system to the GMM-UBM and comparisons of results with baseline GMM-UBM performance will be interesting. It would also be interesting to analyze the proposed method for the case of fuzzy clustering algorithm employed instead of conventional *K*-means algorithm.

REFERENCES

[1] D. A. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, Jan. 2000.

[2] F. K. Soong, A. E. Rosenberg, L. R. Rabiner, and B. H. Juang, "A vector quantization approach to speaker recognition," in *Proc. ICASSP 1985*, April 1985, pp. 387–390.

[3] D. A. Reynolds, and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech and Audio Processing*, vol. 3, pp. 72–83, Jan. 1995.

[4] V. Hautamaki, T. Kinnunen, I. Karkkainen, J. Saastamoinen, M. Tuononen, and P. Franti, "Maximum a Posteriori adaptation of the centroid model for speaker verification," *IEEE Signal Processing Leters*, vol. 15, pp. 162–165, 2008.

[5] T. Kinnunen, J. Saastamoinen, V. Hautamaki, M. Vinni, and P. Franti, "Comparative evaluation of maximum a Posteriori vector quantization and Gaussian mixture models in speaker verification," *Pattern Recogition Leters*, vol. 30, pp. 341–347, March 2009.

[6] M. Zamalloa, L. J. Rodriguez-Fuentes, M. Penagarikano, G. Bordel, and J. P. Uribe, "Feature dimensionality reduction through genetic algorithms for faster speaker recognition," in *Proc. EUSIPCO 2008*, Lausanne, Switzerland, August 25-29. 2008.

[7] L. Burget, P. Matejka, P. Schwarz, O. Glembek, and J. H. Cernocky, "Analysis of feature extraction and channel compensation in a GMM speaker recognition system," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, pp. 1979–1986, Sep. 2007.

[8] C. Seo, K. Y. Lee, and J. Lee, "GMM based on local PCA for speaker identification," *Electronics Leters*, vol. 37, pp. 1486–1488, Nov. 2001.

[9] K. Y. Lee, "Local fuzzy PCA based GMM with dimension reduction on speaker identification," *Pattern Recognition Leters*, vol. 25, pp. 1811–1817, Dec. 2004.

[10] I. T. Jolliffe, *Principal component analysis*. Springer Verlag, 2002.

[11] C. Hanilci, and F. Ertas, "Principal component based classification for text-independent speaker identification," in *Proc. ICSCCW 2009*, Famagusta, North Cyprus, September 2-4. 2009, pp. 1–4.

[12] W. Zhang, Y. Yang, and Z. Wu, "Exploiting PCA classifiers to speaker recognition," in *Proc. Neural Networks 2003*, July 2003, pp. 820–823.

[13] NIST 2001 Speaker Recognition Evaluation website "*http://www.itl.nist.gov/iad/mig//tests/spk/2001/*"