New Adaptation Method Using Two-dimensional PCA for Speaker Verification

Chunyan Liang, Xiang Zhang, Lin Yang, Li Lu, Yonghong Yan
ThinkIt Speech Lab, Institute of Acoustics, CAS
Beijing, P.R. China
Email: {chyliang, xzhang, lyang, llu, yonghong.yan}@hccl.ioa.ac.cn

***Abstract*—In this paper, a new adaptation method based on two-dimensional principal component analysis is introduced into speaker recognition. In the method, mixture and dimension of mean vectors based on the Gaussian Mixture Models (GMMs) are differentiated, and the covariance matrix is computed dimension-wisely. The experiments are carried out on the core conditions of NIST 2008 speaker recognition evaluation data. The experimental results indicate that the 2DPCA-based method can achieve comparable performance to the conventional eigenvoice approach. Besides, the fusion of the two different systems can make significant performance improvement compared to the eigenvoice system alone, achieving relative reduction on EER between 7% and 25% for different test conditions.**

***Keywords*-speaker recognition; 2DPCA; eigenvoice; SVM**

I. INTRODUCTION

State-of-the-art speaker verification systems are based on statistical generative models such as Gaussian Mixture Models (GMMs). In this case, one needs to create a generative model for each client, as well as a generative model for a corresponding anti-client, often replaced by a universal background model (UBM) [1]. The support vector machines (SVMs) [2] have also proved to be effective for speaker recognition. A commonly used method for combining GMM and SVM is to concatenate GMM mean vectors as supervectors for SVM design [3].

For speaker verification, the client model is often derived by adapting the parameters of the UBM using the speaker's training speech. Some adaptation methods have been proven to be successful [4], such as eigenvoice [5]. Eigenvoice speaker adaptation has been shown to be effective for speaker recognition in recent years. The eigenvoice approach involves three steps. First, an eigenspace is established with many speaker dependent (SD) models from training speakers via principal component analysis (PCA). Each of the SD models is represented as a column vector, with the mixture and dimension treated without distinction. Then a group of eigenvoice coefficients is determined for each testing speaker. Finally, we obtain the client models which are expressed as a linear combination of bases in the eigenspace.

In this study, we adopt the speaker adaptation method based on two-dimensional PCA (2DPCA). In 2DPCA, each training SD model is represented as a matrix (the mixture and dimension of mean vectors are represented in separate directions) rather than as a vector which is the case for eigenvoice. Thus, more compact bases with lower dimension than those of eigenvoice can be obtained from 2DPCA, and the speaker adaptation formula using these bases can have a dimension-wise speaker weight. For speech recognition, the speaker adaptation method using 2DPCA has been shown to perform competitively [6]. In this paper, we introduce the new adaptation method into speaker recognition to update the GMM mean vectors of the client models and concatenate them as supervectors for SVM.

The remainder of this paper is organized as follows. In Section II, we give a brief overview of eigenvoice. Section III describes the 2DPCA method and the application of 2DPCA-based method in GMM framework for the task of speaker recognition. The details of the performed experiments and results are presented in Section IV. Finally, we conclude this paper in Section V.

II. EIGENVOICE

The underlying hypothesis of eigenvoice adaptation is that all voices represented in a space of a large dimension could in fact be well represented in a low-dimensional linear subspace [5][7]. The most commonly used tool to select the low-dimensional subspace is the well-known principal component analysis (PCA) [8].

Given a set of T speaker dependent models (SD models) already adapted by Bayesian Maximum A Posteriori (MAP), PCA is used to compute the K leading eigenvectors of the covariance matrix of the T parameter vectors. The model of a new speaker c can then be represented as a linear combination of the K eigenvectors:

$$\mu(c) = Vx + \mu(ubm). \tag{1}$$

where $\mu(ubm)$, consisting of $M \times D$ elements (M is the number of Gaussian components, and D represents the feature dimension for each Gaussian component), is the concatenated mean supervector of all the mixture component means of the UBM model, and $\mu(c)$ is the adapted supervector of the new speaker c. $V = [v_1, v_2, \cdots, v_K]$ represents the eigenspace, and it is the concatenated matrix of the K eigenvectors with the K largest eigenvalues. x is the eigenvoice coefficients of client c.

 ISBN: 978-1-61208-127-4

III. METHODS

A. Two-Dimensional Principal Component Analysis

In 2DPCA [6][9], the mapping from an input 'matrix' into the feature space is performed by

$$\omega = X \cdot \phi. \tag{2}$$

where $X \in \mathbf{R}^{D \times N}$ is the input matrix , ϕ is the base vector which is unitary ($\phi^T \cdot \phi = 1$), and ω is a D-dimensional feature vector (which is a scalar in PCA). A set of such bases, $\{\phi_k\}_{k=1}^K$, can be obtained by maximizing the following criterion:

$$J(\phi) = tr(E[(\omega - E[\omega])(\omega - E[\omega])^T]). \tag{3}$$

where $tr(\cdot)$ denotes the trace of a matrix. The criterion can be expressed in terms of X and ϕ as

$$J(\phi) = \phi^T \cdot \{E[(X - E(X))^T (X - E(X))]\} \cdot \phi. \tag{4}$$

The covariance matrix G is defined as

$$G = E[(X - E(X))^T (X - E(X))]. \tag{5}$$

Using a set of example matrices, $\{X_s\}_{s=1}^S$, the sample covariance matrix can be obtained by

$$G = \frac{1}{S} \sum_{s=1}^{S} (X_s - \bar{X})^T (X_s - \bar{X}). \tag{6}$$

where $\bar{X} = 1/S \sum_{s=1}^S X_s$ is the average of the example matrices. Thus, the criterion becomes

$$J(\phi) = \phi^T \cdot G \cdot \phi. \tag{7}$$

A set of orthonormal projection vectors, $\{\phi_k\}_{k=1}^K$, can be obtained as the K leading eigenvectors of the covariance matrix G to maximize the above criterion.

The set of such bases project an input matrix X_s into the feature space by

$$W(s) = (X_s - \bar{X}) \cdot \Phi. \tag{8}$$

where $\Phi = [\phi_1 \cdots \phi_k \cdots \phi_K]$ and $W(s) = [\omega_1(s) \cdots \omega_k(s) \cdots \omega_K(s)]$. Then, the low-rank approximations of the input matrix can be obtained as

$$X_s \approx \bar{X} + W(s) \cdot \Phi^T = \bar{X} + \sum_{k=1}^{K} \omega_k(s) \cdot \phi_k^T. \tag{9}$$

In (9), X_s can be exactly reconstructed when K=S.

B. Application of 2DPCA-based method in GMM framework for the task of speaker recognition

In this section, we will discuss the application of 2DPCA to speaker adaptation in the GMM framework for the task of speaker recognition. Here, let $\mu_m(s) \in R^{D \times 1}$ be the mean vector of the m-th Gaussian component of speaker s from the total S training speaker models. Adapted from the UBM model using MAP adaptation, the SD mean model of speaker s is viewed as a matrix:

$$\mu(s) = [\mu_1(s) \cdots \mu_m(s) \cdots \mu_M(s)]. \tag{10}$$

where M is the number of the Gaussian components. In the expression above, column corresponds to mixture and row corresponds to the dimension of mean vectors.

We apply 2DPCA to the training examples $\{\mu(s)\}_{s=1}^S$ as follows [6]. First, denoting $\tilde{\mu}(s) = \mu(s) - \mu(ubm)$, we obtain the covariance matrix as:

$$G = \frac{1}{S} \sum_{s=1}^{S} \tilde{\mu}(s)^T \tilde{\mu}(s). \tag{11}$$

When denoting $\tilde{\mu}_d(s) \in R^{1 \times M}$ as the d-th row vector of $\tilde{\mu}(s)$, the covariance matrix can be shown as

$$G = \frac{1}{S} \sum_{s=1}^{S} \sum_{d=1}^{D} \tilde{\mu}^d(s)^T \tilde{\mu}^d(s). \tag{12}$$

As such, 2DPCA is equivalent to line-based PCA, which partitions a matrix into lines and each line is treated as a sample data in standard PCA framework [10].

Then, the eigenvectors corresponding to the K largest eigenvalues ($K \leq S - 1$) of G can be found as the bases $\{\phi_k^{M \times 1}\}_{k=1}^K$, corresponding to the bases $\{\phi_k^{M \times D}\}_{k=1}^K$ in eigenvoice. Thus, 2DPCA produces a more compact set of eigenvectors by the factor of D. Using the bases, we update the model for a new speaker by

$$\mu(new) = \mu(ubm) + W_{new} \cdot \Phi^T. \tag{13}$$

where $\Phi = [\phi_1 \cdots \phi_k \cdots \phi_K]$, and W_{new} is the speaker weight which can be derived in a maximum-likelihood estimation (MLE) framework as follows [6]: given the observation data $O = \{o_1 \cdots o_t \cdots o_T\}$, the auxiliary function is defined as

$$Q(\lambda, \hat{\lambda}) = -\frac{1}{2} P(O|\lambda) \times \sum_{m=1}^{M} \sum_{t=1}^{T} \gamma_m(t)(D \cdot log(2\pi) + log|C_m| + h(o_t, m)). \tag{14}$$

where λ is the current model parameter and $\hat{\lambda}$ is the re-estimated model parameter, $\gamma_m(t)$ denotes the occupation probability of being in mixture m at time t given O, and C_m is the covariance matrix for the m-th Gaussian, which

 ISBN: 978-1-61208-127-4

is diagonal in our work. The last term in (14) contains the model parameter:

$$h(o_t, m) = (o_t - (\mu_m(ubm) + W_{new} \cdot \Phi_m^T))^T \cdot C_m^{-1} \cdot (o_t - (\mu_m(ubm) + W_{new} \cdot \Phi_m^T)). \quad (15)$$

We can derive the following equation to find the weight W_{new}:

$$\sum_{m=1}^{M}\sum_{t=1}^{T}\gamma_m(t)C_m^{-1} \cdot (o_t - \mu_m(ubm)) \cdot \Phi_m = \sum_{m=1}^{M}\sum_{t=1}^{T}\gamma_m(t)C_m^{-1} \cdot W_{new} \cdot \Phi_m^T \cdot \Phi_m. \quad (16)$$

The above equation can be solved for W_{new} using the same procedure in [11].

C. Feature extraction and SVM modeling

After the speaker model are updated as (1) and (13) , the parameters from $\mu(s)$ are concatenated into a single supervector consisting of $D \times M$ elements according to Kullback-liebler divergence [3] and modeled using SVMs, where M is the number of Gaussians in UBM model and D is the dimension of mean vectors in each Gaussian component.

An SVM is trained for each target speaker by regarding the target speaker's training supervector as positive examples, and the supervectors from a background training set as negative examples. Our experiments are implemented using the SVMLight with a linear inner-product kernel function.

IV. Experiment

A. Experiment setup

The experiments for different systems based on the two kinds of speaker adaptation methods (eigenvoice and 2DPCA) are carried out on the NIST 2008 speaker recognition evaluation corpus. The NIST SRE2008 evaluation tasks are distinguished by including in the training and test conditions not only conversational telephone speech but also interview speech recorded with different microphones involving an interview scenario. We carry out the experiments on three types of trials: telephone-telephone, interview-interview and interview-telephone. The performance is measured in terms of equal error rate (EER) and DET curves [12].

The input speech utterance is first converted to a sequence of 36-dimensional feature vectors including 18 MFCC coefficients and their first order derivatives over 5 frames. To reduce channel effects, feature warping to a Gaussian distribution, CMN, CVN are performed to the feature vectors.

The gender dependent UBM models with 1024 mixture components are trained using the NIST SRE 2004 1side training corpus. The background data for SVM system are selected from the data form NIST SRE2004 and NIST SRE2005. Eigenvectors for both eigenvoice and the 2DPCA are also gender dependent. 600 eigenvoices and 600 eigenvectors for 2DPCA for both male and female are trained using the Switchboard II, Switchboard Cellular corpus as well as the data from NIST SRE2005 and NIST SRE2006. For eigenchannel compensation in feature domain, telephone and microphone data from NIST SRE2004, NIST SRE2005 and NIST SRE2006 are used.

The raw score are speaker-normalized by means of gender-dependent ZTnorm. For Znorm and Tnorm, telephone and interview utterances are drawn from the NIST SRE2006 corpus.

We use the linear fusion for the two systems, with the weight of 0.5 for each system.

Table I
SVM SYSTEMS BASED ON DIFFERENT ADAPTATION METHODS ACROSS ALL MALE SPEAKERS IN THE TEST CORPUS. THE VALUE IN EACH TABLE CELL IS THE EER (%) .

task	eigenvoice	2DPCA	fusion
telephone-telephone	5.62	5.54	4.98
interview-interview	2.38	2.97	2.04
interview-telephone	4.74	5.21	3.59

B. Experiment results

In this subsection, we list the results of the systems based on both eigenvoice and the 2DPCA-based method as well as the fusion of them on the three test conditions in NIST SRE 2008. The DET curves are also given below.

Table I lists the performance of the SVM system based on eigenvoice and 2DPCA-based method on the three trial conditions across all male speakers. From Table I, we can see that the system based on 2DPCA can achieve comparable performance to the conventional eigenvoice system for male speakers. As well, the fusion of the two systems makes significant performance improvement compared to the eigenvoice system alone, yielding 11.4% improvement on EER for the telephone-telephone condition, 14.3% for the interview-interview condition and 24.3% for the interview-telephone condition.

Table II shows the results of the SVM system based on the two different adaptation methods across all female speakers. It can also be seen that the performance of the 2DPCA-based system is comparable to the eigenvoice system. Compared to the single system based on eigenvoice, the fusion of the two systems achieve relative reduction of 7.14% on EER for the telephone-telephone condition, 11.9% for the interview-interview condition and 7.3% for interview-telephone.

Figure 1 and Figure 2 show the DET curves of the systems based on different speaker adaptation methods for male and female speakers respectively.

Table III summarizes the approximate average training time per file. The training time mainly consists of two parts, the time of estimating the parameters of feature vector and the time of SVM training. As we can see, our method uses

 ISBN: 978-1-61208-127-4

Table II
SVM SYSTEMS BASED ON DIFFERENT ADAPTATION METHODS ACROSS ALL FEMALE SPEAKERS IN THE TEST CORPUS. THE VALUE IN EACH TABLE CELL IS THE EER (%) .

task	eigenvoice	2DPCA	fusion
telephone-telephone	7.14	8.11	6.63
interview-interview	4.53	6.22	3.99
interview-telephone	6.46	11.12	5.99

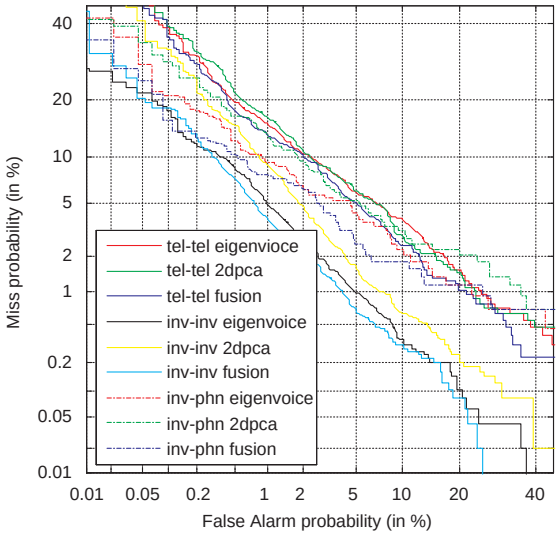

Figure 1. DET curves comparing systems based on eigenvoice and 2DPCA as well as the fusion system for male speakers

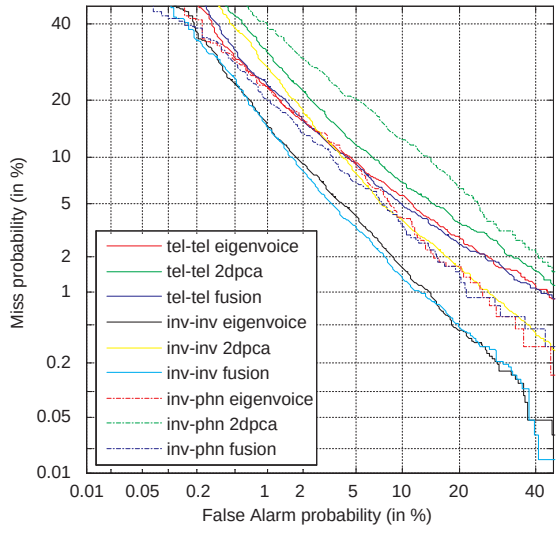

Figure 2. DET curves comparing systems based on the eigenvoice and 2DPCA as well as the fusion system for female speakers

more time cost than the eigenvoice system, which implies its limits in real time work.

Table III
AVERAGE TRAINING TIME PER FILE FOR EIGENVOICE AND 2DPCA.

Systems	Training time cost(sec)
eigenvoice	2.88
2DPCA	7.86

V. CONCLUSION

In this paper, we have introduced the new adaptation method using 2DPCA into speaker recognition. The 2DPCA of training models produces the more compact bases whose dimension is lower than that of eigenvoice, and the speaker weight consists of dimensional elements. Experiments show that the system based on 2DPCA can achieve comparable performance to the conventional eigenvoice system and the fusion of the two systems can further improve the performance, yielding 7%-25% improvement on EER for different tasks, which indicates that the 2DPCA-based method and eigenvoice are complementary to each other to some extent when used in speaker recognition. Future work include generalizing this approach to other PCA-based modeling methods such as eigenspace-based MLLR [13].

ACKNOWLEDGEMENT

This work is partially supported by The National Science and Technology Pillar Program (2008BAI50B03), National Natural Science Foundation of China (No. 10925419, 90920302, 10874203, 60875014).

REFERENCES

[1] D. A. Reynolds, T. F. Quatieri and R. B. Dunn, “Speaker verification using adapted Gaussian mixture models”, Digital Signal Processing, vol. 10, pp. 19-41, 2000.

[2] N. Cristianini and J. Shawe-Taylor, “ Support vector machines”, Cambridge University Press, Cambridge, UK, 2000.

[3] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, “SVM based speaker verification using a GMM supervector kernel and NAP variability compensation”, IEEE ICASSP 2006, vol. 1, 2006.

[4] J. Mariethoz and S. Bengio, “A comparative study of adaptiation methods for speaker verification”, In Seventh International Conference on Spoken Language Processing. 2002.

[5] B. Mak, J. Kwok, and S. Ho, “Kernel eigenvoice speaker adaptation”, IEEE Transactions on Speech And Audio Processing, 13(5):984, 2005.

[6] Y. Jeong and H.S. Kim, “New speaker adaptation method using 2-D PCA”, IEEE Signal Process. Lett., vol. 17, no. 2, pp. 193-196, 2010.

[7] H. Wang, Q. Zhao and Y. Yan, “Using Eigenvoice Coefficients as Features in Speaker Recognition”, 2009 International Conference on Electronic Computer Technology, pp. 262-266. 2009.

[8] I. Jolliffe, “Principal component analysis”, Springer New York, 2002.

 ISBN: 978-1-61208-127-4

[9] J. Yang, D. Zhang, A.F. Frangi, and J.Y. Yang, “Two-dimensional PCA: A new approach to appearance-based face representation and recognition”, IEEE Trans. Pattern Anal. Mach. Intell., vol. 26, no. 1, pp. 131-137, Jan. 2004

[10] L. Wang, X. Wang, X. Zhang, and J. Feng, “The equivalence of two-dimensional PCA to line-based PCA”, Pattern Recognition. Lett., vol. 26, no. 1, pp. 57-60, Jan. 2005.

[11] C.J. Leggetter and P.C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models”, Comput. Speech Lang., vol. 9, no. 2, pp. 171-185, Apr. 1995.

[12] “The NIST Year 2008 Speaker Recognition Evaluation Plan”, http://www.nist.gov/speech/tests/spk/2008/index.html, 2008.

[13] K.T. Chen, W.W. Liau, H.M. Wang, and L.S. Lee, “Fast speaker adaptation using eigenspace-based maximum likelihood linear regression”, ICSLP 2000, vol. 3, pp. 742-745, Oct. 2000.

 ISBN: 978-1-61208-127-4