A NEW EIGENVOICE APPROACH TO SPEAKER ADAPTATION

Chih-Hsien Huang[a], *Jen-Tzung Chien*[a] *and Hsin-min Wang*[b]

[a] Department of Computer Science and Information Engineering, Cheng Kung University, Tainan
[b] Institute of Information Science, Academia Sinica, Taipei
E-mail: acheron@chien.csie.ncku.edu.tw, jtchien@mail.ncku.edu.tw, whm@iis.sinica.edu.tw

ABSTRACT

In this paper, we present two approaches to improve the eigenvoice-based speaker adaptation. First, we present the maximum *a posteriori* eigen-decomposition (MAPED), where the linear combination coefficients for eigenvector decomposition are estimated according to the MAP criterion. By incorporating the prior decomposition knowledge, here we use a Gaussian distribution, the MAPED is established accordingly. MAPED is able to achieve better performance than maximum likelihood eigen-decomposition (MLED) with few adaptation data. On the other hand, we exploit the adaptation of covariance matrices of the hidden Markov model (HMM) in the eigenvoice framework. Our method is to use the principal component analysis (PCA) to project the speaker-specific HMM parameters onto a smaller orthogonal feature space. Then, we reliably calculate the HMM covariance matrices using the observations in the reduced feature space. The adapted HMM covariance matrices are estimated by transforming the covariance matrices in the reduced feature space to that in the original feature space. The experimental results show that the eigenvoice speaker adaptation using MAPED and incorporating covariance adaptation can improve the performance of the original eigenvoice adaptation in Mandarin speech recognition.

1. INTRODUCTION

It is no doubt that the performance of speech recognition is significantly degraded by mismatches between training and testing speakers/environments. To achieve the robustness of speech recognition, many speaker adaptation approaches have been proposed to adapt speaker-independent (SI) hidden Markov models (HMM) trained by a large amount of training speech, which usually covers a wide range of speakers, to a specific speaker.

Among various speaker adaptation techniques, there are three major types of model-based adaptation algorithms, namely maximum *a posteriori* (MAP) [3], maximum likelihood linear regression (MLLR) [6] and speaker clustering [8]. The eigenvoice speaker adaptation belongs to the speaker clustering family. The basic concept of eigenvoice adaptation is to apply the principal component analysis (PCA) to construct the eigenvoice space using the supervectors constructed from speaker-dependent (SD) acoustic models [4]. The principal components are then used to build the speaker-adaptive acoustic models through maximum likelihood eigen-decomposition (MLED) for a new speaker who enrolls with some adaptation data. The linear combination coefficients are estimated via the maximum likelihood criterion.

During the past few years, much research has been devoted to enhance the original eigenvoice approach. In [5], maximum likelihood eigenspace (MLES) was proposed to compact the eigenspace and MLLR was adopted to minimize the mismatches caused by noise. Chen et al. proposed an effective adaptation approach using eigenspace-based MLLR, also known as eigen-MLLR [1]. Although these approaches did enhance the original eigenvoice, the better estimation of combination coefficients and the joint adaptation of mean and covariance have not been reported yet.

In this paper, we propose the maximum *a posteriori* eigen-decomposition (MAPED) to estimate the linear combination coefficients, which is believed to be more robust than MLED, in particular with few adaptation data available. We also propose a method for the adaptation of HMM covariance matrices. The experimental results indicate that both approaches can further enhance the eigenvoice approach.

2. EIGENVOICE

Generally speaking, the eigenvoice approach is implemented with two phases, namely eigenvoice construction (the training phase) and coefficient estimation (the adaptation phase). In the training phase, a set of SD reference models from R speakers is prepared. For each SD model, we “vectorize” the model parameters

and form a "supervector". Traditionally, only the mean vectors are considered to be the elements of supervectors. Let the dimension of supervector be *D*. We can calculate a $D \times D$ covariance matrix from the *R* supervectors. PCA is then applied to this covariance matrix. The first *K* eigenvectors are selected to form a *K*-dimensional eigenspace. The selected eigenvectors possess most of the information from the training data. We have the property $K < R << D$.

In the adaptation phase, we would like to adapt the existing HMM parameters to a new speaker using speaker specific adaptation data $\mathbf{X}$. We first estimate the location of a new speaker in the *K*-dimensional eigenspace. A set of weight coefficients $\{w(j), j = 1,2,\ldots,K\}$ corresponding to *K* eigenvectors $\{e(j), j = 1,2,\ldots,K\}$ should be determined. The supervector μ of a new speaker is constructed by

$$\begin{aligned}\mu &= e(0) + w(1) \times e(1) + \cdots + w(K) \times e(K) \\ &= e(0) + \sum_{k=1}^{K} w(k) e(k) = E\mathbf{w}^T,\end{aligned} \tag{1}$$

where $e(0)$ is the mean vector of *R* supervectors, $\mathbf{w} = [1, w(1), \ldots, w(K)]^T$ and $E = [e(0), \ldots, e(K)]$. The MLED can be used to estimate the weight coefficients [5] by solving

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} P(\mathbf{X} \mid \mathbf{w}). \tag{2}$$

Because the incomplete data problem is inherent in the HMM framework, we need to use the expectation and maximization (EM) algorithm [2] to solve Eq. (2). In the E-step, we calculate the expectation as

$$\begin{aligned}Q(\hat{\mathbf{w}} \mid \mathbf{w}) &= E\left[\log P(\mathbf{X}, S, M \mid \hat{\mathbf{w}}) \mid \mathbf{X}, \mathbf{w}\right] \\ &\propto \sum_s \sum_m \sum_t \gamma_m^{(s)}(t)[-n\log(2\pi) \\ &\quad - \log \mid C_m^{(s)} \mid + h(\mathbf{x}_t, s, m)]\end{aligned}, \tag{3}$$

where $\gamma_m^{(s)}(t) = P(s_t = s, m_t = m \mid \mathbf{X}, \mathbf{w})$ is the occupation probability that the observation $\mathbf{x}_t$ stays at state *s* and mixture component *m* and

$$h(\mathbf{x}_t, s, m) = (\mu_m^{(s)} - \mathbf{x}_t)^T C_m^{(s)-1} (\mu_m^{(s)} - \mathbf{x}_t). \tag{4}$$

After replacing $\mu_m^{(s)}$ with the corresponding linear combination of eigenvoices, the M-step is performed to maximize $Q(\hat{\mathbf{w}} \mid \mathbf{w})$ via $\partial Q(\hat{\mathbf{w}} \mid \mathbf{w}) / \partial w(j) = 0, j = 1, \ldots, K$. For each *j*, we have

$$\begin{aligned}&\sum_s \sum_m \sum_t \gamma_m^{(s)}(t) \left(e_m^{(s)}(j)\right)^T C_m^{(s)-1} \mathbf{x}_t \\ &= \sum_s \sum_m \sum_t \gamma_m^{(s)}(t) \times \left\{ \sum_{k=1}^{K} \hat{w}(k) \left(e_m^{(s)}(k)\right)^T C_m^{(s)-1} e_m^{(s)}(j) \right\}.\end{aligned} \tag{5}$$

There are a total of *K* equations established to solve the *K* new weights $\hat{w}(1), \ldots, \hat{w}(K)$. The current estimate is used to update the occupation probability $\gamma_m^{(s)}(t)$. The EM algorithm can guarantee the convergence of likelihood improvement.

3. MAXIMUM *A POSTERIORI* EIGEN-DECOMPOSITION

In general, the eigenvoices characterize the principal directions of speaker variations. The adapted models are determined through linear interpolation of eigenvoices. The MLED is likely to output biased estimates of combination coefficients with few adaptation data. We can restrict the variations of coefficient using a prior distribution.

To estimate the combination coefficients of eigenvoices under the MAP criterion, we define the auxiliary *R* function as follows

$$\begin{aligned}R(\hat{\mathbf{w}} \mid \mathbf{w}) &= \sum_{s=1}^{S} \sum_{m=1}^{M_s} \sum_{t=1}^{T} \gamma_m^{(s)}(t) \left\{ n\log(2\pi) + \log \left| C_m^{(s)} \right| + h(\mathbf{x}_t, s, m) \right\} \\ &\quad + \sum_{j=1}^{K} \left\{ \log(2\pi) + 2\log \sigma_{w(j)} + \frac{(\hat{w}(j) - \mu_{w(j)})^2}{\sigma_{w(j)}^2} \right\}.\end{aligned} \tag{6}$$

The coefficient, $w(j)$, is modeled by a Gaussian distribution with mean, $\mu_{w(j)}$, and variance, $\sigma_{w(j)}^2$, i.e. $P(w(j)) = N(\mu_{w(j)}, \sigma_{w(j)}^2)$. The combination coefficients can be obtained through maximizing the *R* function

$$\hat{\mathbf{w}}_{MAP} = \arg\max_{\hat{\mathbf{w}}} R(\hat{\mathbf{w}} \mid \mathbf{w}). \tag{7}$$

By taking the partial derivative with respect to $\hat{w}(j)$, we obtain

$$\begin{aligned}&\frac{\partial}{\partial \hat{w}(j)} R(\hat{\mathbf{w}} \mid \mathbf{w}) \\ &= \frac{\partial}{\partial \hat{w}(j)} \sum_{s=1}^{S} \sum_{m=1}^{M_s} \sum_{t=1}^{T} \gamma_m^{(s)}(t) \left\{ n\log(2\pi) + \log \left| C_m^{(s)} \right| + h(\mathbf{x}_t, s, m) \right\} \\ &\quad + \frac{\partial}{\partial \hat{w}(j)} \sum_{i=1}^{K} \left\{ \log(2\pi) + 2\log \sigma_{w(i)} + \frac{(\hat{w}(i) - \mu_{w(i)})^2}{\sigma_{w(i)}^2} \right\} \\ &= \sum_{s=1}^{S} \sum_{m=1}^{M_s} \sum_{t=1}^{T} \gamma_m^{(s)}(t) \left\{ 2 \left[\left(e_m^{(s)}(j)\right)^T C_m^{(s)-1} \mathbf{x}_t + \sum_{k=1}^{K} \hat{w}(k) \left(e_m^{(s)}(k)\right)^T \right. \right. \\ &\quad \left. \left. \times C_m^{(s)-1} e_m^{(s)}(j) \right] \right\} + 2 \times \frac{\hat{w}(j) - \mu_{w(j)}}{\sigma_{w(i)}^2}.\end{aligned} \tag{8}$$

After setting Eq. (8) to zero, the combination coefficients are estimated by the following *K* linear equations:

$$\begin{aligned}&\frac{\mu_{w(j)}}{\sigma_{w(j)}^2} + \sum_{s=1}^{S} \sum_{m=1}^{M_s} \sum_{t=1}^{T} \gamma_m^{(s)}(t) \left(e_m^{(s)}(j)\right)^T C_m^{(s)-1} \mathbf{x}_t \\ &\quad = \sum_{k=1}^{K} \hat{w}(k) \left\{ \sum_{s=1}^{S} \sum_{m=1}^{M_s} \sum_{t=1}^{T} \gamma_m^{(s)}(t) \left(e_m^{(s)}(k)\right)^T C_m^{(s)-1} e_m^{(s)}(j) \right. \\ &\quad \left. + \delta(k-j) \frac{\hat{w}(j)}{\sigma_{w(j)}^2} \right\}.\end{aligned} \tag{9}$$

Finally, the new MAP estimates $\{\hat{w}(1),\ldots,\hat{w}(K)\}$ are found by solving the $K \times K$ linear system.

4. EIGENVOICE-BASED COVARIANCE ADAPTATION

In [9], the covariance matrices were shown to be effective to serve as the evaluation measurement for speaker recognition. However, the eigenvoice approach based on PCA was developed only for the adaptation of HMM mean vectors [5]. Herein, we present the PCA approach to the adaptation of HMM mean vectors as well as covariance matrices.

The covariance matrices can be adapted using the approach similar to the adaptation of mean vectors. We hope that the adapted model can precisely characterize a specific speaker and, thus, improve the speech recognition performance. We first vectorize the covariance matrices of each reference speaker's model. Here, we assume that all the covariance matrices are diagonal. The diagonal elements of the covariance matrices are concatenated to form the supervector for each speaker. The supervector of covariance matrices of a new speaker's model can be expressed in a form of linear combination of eigenvectors $\{C_m^{(s)}(j)\}$ of covariance matrices,

$$\hat{C}_m^{(s)} = \sum_{j=0}^{K} w(j) C_m^{(s)}(j) \cdot \tag{10}$$

Like the adaptation of mean vectors, we can perform MLED to find $\hat{C}_m^{(s)}$ by setting the differentiation of the Q function with respect to $\hat{w}(j)$ to zero,

$$\frac{\partial Q(\hat{\mathbf{w}} \mid \mathbf{w})}{\partial \hat{w}(j)} = 0, \quad j = 1 \cdots K \cdot \tag{11}$$

The differentiation consists of two parts, i.e. $\frac{\partial \log(\left|\hat{C}_m^{(s)}\right|)}{\partial \hat{w}(j)}$ and $\frac{\partial (\mathbf{x}_t - u_m^{(s)}(t))^T \hat{C}_m^{(s)-1} (\mathbf{x}_t - u_m^{(s)}(t))}{\partial \hat{w}(j)}$, which can be, respectively, written as

$$\frac{\partial \log(\left|\hat{C}_m^{(s)}\right|)}{\partial \hat{w}(j)} = \sum_{d=1}^{D} \frac{C_m^{(s)}(j)_d}{\sum_{k=0}^{K} \hat{w}(k) C_m^{(s)}(k)_d} \tag{12}$$

and

$$\begin{aligned} &\frac{\partial (\mathbf{x}_t - u_m^{(s)}(t))^T \hat{C}_m^{(s)-1} (\mathbf{x}_t - u_m^{(s)}(t))}{\partial \hat{w}(j)} \\ &= \frac{\partial}{\partial \hat{w}(j)} \sum_{d=1}^{D} \frac{(\mathbf{x}_t - u_m^{(s)}(t))^T (\mathbf{x}_t - u_m^{(s)}(t))}{\sum_{k=0}^{K} \hat{w}(k) C_m^{(s)}(k)_d} \\ &= -\sum_{d=1}^{D} (\mathbf{x}_t - u_m^{(s)}(t))^T (\mathbf{x}_t - u_m^{(s)}(t)) \left[\sum_{k=0}^{K} \hat{w}(k) C_m^{(s)}(k)_d \right]^{-2} C_m^{(s)}(j)_d, \end{aligned} \tag{13}$$

where d is the index of parameter dimension. The coefficients are calculated according to the linear equations as follows

$$\begin{aligned} &\sum_{m=1}^{M} \sum_{\tau=1}^{T} \gamma_m(\tau) \sum_{d=1}^{D} \left\{ \frac{C_m^{(s)}(j)_d}{\sum_{k=0}^{K} \hat{w}(k) C_m^{(s)}(k)_d} + (\mathbf{x}_t - u_m^{(s)}(t))^T (\mathbf{x}_t - u_m^{(s)}(t)) \right. \\ &\left. \left[\sum_{k=0}^{K} \hat{w}(k) C_m^{(s)}(k)_d \right]^{-2} C_m^{(s)}(j)_d \right\} = 0. \end{aligned} \tag{14}$$

Unfortunately, there is no unique solution to this system. Therefore, we look for an alternative approach using the transformation scheme. First, the training observations and the corresponding mean vectors are used to calculate the full covariance matrix $C_{full}^{(s)}$. We then apply PCA to obtain the eigenvector set $E^{(s)}$. During adaptation, we use the adaptation data of a new speaker to perform feature transformation by

$$\mathbf{x}_t^r = E^{(s)T} \mathbf{x}_t. \tag{15}$$

Subsequently, the MLED/MAPED is applied to obtain the adapted mean vectors using

$$\hat{\mu}_m^{r(s)} = E^{(s)T} \hat{\mu}_m^{(s)}. \tag{16}$$

The covariance matrix of mixture component m in this reduced feature subspace is obtained by

$$C_m^{r(s)} = \frac{\sum_t \gamma_m^{(s)}(t) (\mathbf{x}_t^r - \mu_m^{r(s)})(\mathbf{x}_t^r - \mu_m^{r(s)})^T}{\sum_t \gamma_m^{(s)}(t)}. \tag{17}$$

After extracting the diagonal elements $\mathrm{diag}(C_m^{r(s)})$ of $C_m^{r(s)}$, the adapted covariance matrix can be computed using the inverse transformation

$$\mathbf{C}_m^{(s)} = E^{(s)} \mathrm{diag}(C_m^{r(s)}) E^{(s)T}. \tag{18}$$

In this way, we can adapt the covariance matrices associated with the seen data, while keeping the covariance matrices associated with the unseen data unchanged.

5. EXPERIMENTS

The proposed approach was evaluated on two tasks, namely the large vocabulary continuous Mandarin speech recognition and the connected Mandarin digit recognition. For the task of continuous speech recognition, we used the benchmark TCC300 microphone speech database. The speech of 100 speakers was used for training, and the speech of the other 20 speakers was used for testing. First of all, the SI model was trained from all the training data. For each training speaker, the SD model was generated via the MAP adaptation. The Initial/Final HMM's were adopted here. Each Initial HMM has three states, while each Final HMM has six states. The number of Gaussian mixture components was 32 at most. The acoustic feature vector contained 26 dimensions, including twelve MFCCs, one log energy, and their derivatives. The dimension D of

supervector is 137000. For the task of connected digit recognition, 1000 clean utterances spoken by 50 males and 50 females were used to train the SI model. A noisy speech database was adopted to evaluate the performance of speaker/noise adaptation. The speech was recorded in the 50km/hour car driving environment. There are ten speakers, and each speaker had 15 sentences, in which five were used for adaptation and ten for recognition. Each digit was modeled using a HMM with six states. Each state has sixteen mixture components. The dimension of supervector is 24960($6\times10\times16\times26$).

The first experiment was conducted to evaluate the proposed MAPED and covariance adaptation. The first 20 eigenvoices were used as the bases of eigenspace The experimental results, as shown in Table 1, were averaged over all testing speakers and reported in terms of syllable recognition rates. From Table 1, we find that the performance of MAPED mean adaptation is higher than that of MLED mean adaptation in both the continuous speech recognition and connected digit recognition tasks. If both the covariance matrices and mean vectors were adapted, the performance was further improved, though the improvement was moderate. Different from the isolated word recognition task conducted in [4], the number of parameters for the continuous speech recognition task here is very large so that the amount of speaker specific training utterances is relatively insufficient. In the case of connected digit recognition, the improvement is relatively significant because the training data are relatively sufficient.

	Syllable Recognition Rate (%)	
	TCC300	connected digits
Baseline	64.3	75.7
MLED mean adapt.	66.1	78.1
MAPED mean adapt.	66.8	78.6
MLED mean adapt. + covariance adapt.	67.0	79.1
MAPED mean adapt. + covariance adapt.	67.3	79.4

Table 1. Performance comparison between different methods

Number of eigenvoices (K)	Syllable Recognition Rate (%)	
	TCC300	Connected digits
K=20	66.1	78.1
K=40	67.5	79.4
K=60	68.4	80.4

Table 2. Recognition rates using MLED with different number of eigenvoices

We have also tested the recognition performance using 20, 40 and 60 eigenvoices. The experimental results are summarized in Table 2. It is clear that the more the eigenvoices kept, the better the recognition rates achieved for both tasks. The best performance in this experiment is achieved when we keep as many as 60 eigenvoices.

6. CONCLUSION

We proposed the MAPED and covariance adaptation to improve the eigenvoice speaker adaptation. MAPED constrained the estimation of combination coefficients of eigenvoices by a prior distribution, and accordingly improved the robustness of speaker adaptation. Covariance adaptation provided better modeling of speaker variations so that the adaptation performance could be further improved. It is desirable that both the MAPED and covariance adaptation can improve the recognition accuracy according to the experimental results.

7. REFERENCES

[1] K. T. Chen, W. W. Liau, H. M. Wang and L. S. Lee, "Fast speaker adaptation using eigenspace-based maximum likelihood linear regression," in *Proc. ICSLP*, pp. 742-745, 2000.

[2] P. Dempster, N. M. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society (B)*, vol. 39, pp. 1-38, 1977.

[3] J. L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 291-298, 1994.

[4] R. Kuhn et al., "Rapid speaker adaptation in eigenvoice space," *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 695-707, 2000.

[5] P. Nguyen et al., "Maximum likelihood eigenspace and MLLR for speech recognition in noisy environments," in *Proc. Eurospeech*, pp. 2519-2522, 1999.

[6] C. J. Leggetter and P. C. Woodland, "Maximum likelihood regression for speaker adaptation of continuous density hidden Markov models," in *Computer Speech and Language*, vol. 9, pp. 171-185, 1995.

[7] A. Ljolje, "The importance of cepstral parameter correlations in speech recognition," *Computer Speech Language*, vol. 8, pp. 223-232, 1994.

[8] P. C. Woodland, "Speaker adaptation: Techniques and challenges," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, pp.85-90, 1999.

[9] R. D. Zilca, "Text-independent speaker verification using utterance level scoring and covariance modeling," *IEEE Trans. Speech Audio Processing*, vol.10, no.6, pp.363-370, 2002.