Eigenspace Estimation with Missing Values and Its Application to Eigenvoice Adaptation for Speech Recognition *

Zhijian Ou, Jun Luo
Department of Electronic Engineering, Tsinghua Unversity, Beijing 100084, China
ozj@tsinghua.edu.cn

Abstract

Eigenspace estimation via principal component analysis (PCA) has been used in many applications, e.g., in eigenvoice modeling for speaker adaptation. Here the data of interest are the speaker supervectors, where each supervector is a concatenation of all the mean vectors in the speaker's speaker-dependent (SD) model. One problem is that we often do not have enough speaker-specific data to establish the individual SD models (having unseen phones). To address this issue, an approach to eigenspace estimation by expectation-maximization (EM) algorithm in situations where the training samples contain missing values is proposed, applied to eigenvoice adaptation, and experimentally evaluated in this paper.

1. Introduction

Eigenvoice modeling has been shown to be effective in speaker adaptation for speech recognition [1], [2]. Speaker adaptation uses speech data from one speaker to adjust the parameters of a speaker-independent (SI) model towards the speaker-dependent (SD) values. The eigenvoice approach was proposed in [1] for rapid speaker adaptation. The idea is that for each speaker, we can concatenate all the mean vectors in the speaker's SD model to form a supervector. The adapted supervector (representing the speaker-adapted model) is constrained to be a linear combination of a small number of (orthogonal) basis vectors, and thus greatly reduces the number of free parameters to be estimated from adaptation data. In adaptation, the combination weights are estimated based on maximum likelihood (ML) criterion [1], or maximum a posteriori (MAP) criterion [2], [3]. For both cases, an important issue is how to reliably estimate the population covariance matrix $\sum$ of the supervectors, since once we obtain $\sum$, we can perform principle component analysis (PCA), i.e., eigenvalue decomposition on $\sum$ to yield the basis vectors. The set of eigenvectors associated with the largest eigenvalues are retained as the basis vectors, namely eigenvoices, which span the eigenspace. Projection to the eigenspace gives the optimal representation of the supervector in terms of the mean square error.

Traditionally, we take the sample covariance matrix of the training speaker supervectors as an estimate of $\sum$. This assumes that we have a set of well-trained SD models, which can be separately generated for every speaker in the training data. One problem is that we often do not have enough speaker-specific data to establish the individual SD models. There are unseen phones for the training speakers, and different speakers may have different unseen phones. How to estimate the eigenvoices in situations where there are unseen phones for individual training speakers is the main issue addressed in this paper.

There have been some methods to address this issue. A practical method is to use maximum likelihood linear regression (MLLR) [4] adaptation to establish SD models [2]. By sharing of the transformation parameters across different phones, the parameters for unseen phones can also be updated. Classical MAP adaptation is not suitable here to establish SD models for supervector covariance estimation, since it only transforms the observed parameters, which leaves unseen phone still unchanged. The method by iteratively interleaving eigenvoice adaptation and eigenspace estimation in [5] is a choice. However, it is

* This work was supported by National Natural Science Foundation of China (60402029) and China 863 program (2006AA01Z149).

required that we know the desirable number of eigenvoices beforehand and keep it fixed. In practice, the number of eigenvoices may be adjusted according to the amount of adaptation data. Additionally, the resulting basis vectors after iterative estimations are not guaranteed to be orthogonal, and so they are not strictly eigenvectors. Moreover, there are no corresponding eigenvalues which measure the importance of each basis vector.

In this paper, a general approach to eigenspace estimation in situations where the training samples contain missing values is proposed. For such incomplete data, we cannot take the sample covariance matrix of the training data as an estimate of the population covariance matrix. Instead, we model the underlying population as a Gaussian distribution, and perform maximum-likelihood parameter estimation for the underlying Gaussian model by expectation maximization (EM) algorithm [6]. For eigenvoice modeling, a training sample is a speaker supervector, which may have missing subvectors for unseen phones. In this approach, the computationally intensive steps of covariance matrix estimation and PCA are carried out once, and then all eigenvectors of interest along with their corresponding eigenvalues are obtained.

After a brief outline of MAP eigenvoice adaptation in Section 3， we present the experimental results, which demonstrate the effectiveness of the new approach to eigenspace estimation.

2. Maximum likelihood Gaussian parameter estimation with missing values via EM

Denote the Gaussian distribution of interest as $x \sim N(\mu,\Sigma)$ with mean μ and covariance matrix $\sum$. $x=(x_1{}^T,x_2{}^T,\cdots,x_D{}^T)^T$ denotes the random vector with D subvectors x_i, $i=1,\cdots,D$, where each x_i is in general a vector-valued random variable. Suppose that we have in hand a set of independent samples $\chi=\{x^{(1)},x^{(2)},\cdots,x^{(N)}\}$ from the above Gaussian population. And the sample data χ contain missing values, i.e., for each sample, some of its subvectors are unobserved/hidden. Let us divide the subvectors of each sample $x^{(n)}$ into two parts: the observed part $x^{(n)}_{o(n)}$ and the missing part $x^{(n)}_{h(n)}$, where $o(n)$ and $h(n)$ are the corresponding subscript sets. It is clear that we have $o(n)\cup h(n)=\{1,\cdots,D\}$ and $o(n)\cap h(n)=\varnothing$, where $\varnothing$ denotes the null set. Note that for different samples, we may have different missing part $h(n)$ varying with n.

The EM algorithm is an iterative algorithm for maximum likelihood parameter estimation from incomplete data. Denote the current parameter estimates as $\bar{\theta}=(\bar{\mu},\bar{\Sigma})$. The auxiliary function $Q(\theta,\bar{\theta})$ is defined as follows:

$$\begin{aligned} Q(\theta,\bar{\theta}) &= \sum_{n=1}^{N} E_{P\left(x^{(n)}_{h(n)}\middle|\,\bar{\theta},x^{(n)}_{o(n)}\right)} \log P\left(x^{(n)}\middle|\mu,\Sigma\right) \\ &= -\frac{1}{2}\sum_{n=1}^{N} tr\left\{\sum_{i=1}^{D}\sum_{j=1}^{D} K_{ij}\left\langle\left(x_j^{(n)}-\mu_j\right)\left(x_i^{(n)}-\mu_i\right)^T\right\rangle_n\right\} \\ &\quad -\frac{N}{2}\log|\Sigma| \end{aligned} \tag{1}$$

Here $K=\Sigma^{-1}$, and we use the notation $\left\langle f\left(x^{(n)}\right)\right\rangle_n$ to denote the conditional expectation for function $f\left(x^{(n)}\right)$ of the sample $x^{(n)}$, under the conditional distribution of the missing part $x^{(n)}_{h(n)}$ giver the observed part $x^{(n)}_{o(n)}$ and the current parameters $\bar{\theta}$:

$$\begin{aligned} \left\langle f\left(x^{(n)}\right)\right\rangle_n &\triangleq E_{P\left(x^{(n)}_{h(n)}\middle|\,\bar{\theta},x^{(n)}_{o(n)}\right)}\left[f\left(x^{(n)}\right)\right] \\ &= \int_{x^{(n)}_{h(n)}} P\left(x^{(n)}_{h(n)}\mid\bar{\theta},x^{(n)}_{o(n)}\right) f\left(x^{(n)}\right) dx^{(n)}_{h(n)} \end{aligned} \tag{2}$$

Using the above notation, we define

$$\tilde{x}^{(n)} = \left\langle x^{(n)}\right\rangle_n \tag{3}$$

$$B^{(n)} \triangleq \left\langle x^{(n)}x^{(n)T}\right\rangle_n - \left\langle x^{(n)}\right\rangle_n \left\langle x^{(n)}\right\rangle_n^T \tag{4}$$

We can regard each $\tilde{x}^{(n)}$ as a pseudo complete sample, in which the missing part is filled with the conditional expectation, namely

$$\tilde{x}_i^{(n)} = \begin{cases} E_{P\left(x^{(n)}_{h(n)}|\bar{\theta},x^{(n)}_{o(n)}\right)}\left[x_i^{(n)}\right] = \left\langle x_i^{(n)}\right\rangle_n, i\in h(n) \\ x_i^{(n)}, \qquad\qquad\qquad\qquad\qquad i\in o(n) \end{cases} \tag{5}$$

Regarding $B^{(n)}$ as a partitioned matrix, we can see that it has non-zero elements $B_{ij}{}^{(n)}$ only for $i,j\in h(n)$, and takes zero otherwise. Note that

$$\begin{aligned} & tr\left\{\sum_{i=1}^{D}\sum_{j=1}^{D} K_{ij}\left\langle\left(x_j^{(n)}-\mu_j\right)\left(x_i^{(n)}-\mu_i\right)^T\right\rangle_n\right\} \\ &= \left(\tilde{x}^{(n)}-\mu\right)^T K\left(\tilde{x}^{(n)}-\mu\right) + \sum_{i,j\in h(n)} tr\left[K_{ij}\left(B_{ij}{}^{(n)}\right)^T\right] \end{aligned} \tag{6}$$

and substitute (6) in (1), we get:

$$\begin{aligned} Q(\theta,\bar{\theta}) &= -\frac{N}{2}\left[tr\left(\Sigma^{-1}\tilde{\Sigma}\right) + \left(\tilde{\mu}-\mu\right)^T\Sigma^{-1}\left(\tilde{\mu}-\mu\right)\right] \\ &\quad -\frac{1}{2}\sum_{n=1}^{N}\sum_{i,j\in h(n)} tr\left[K_{ij}\left(B_{ij}{}^{(n)}\right)^T\right] - \frac{N}{2}\log|\Sigma| \end{aligned} \tag{7}$$

where $\tilde{\mu}$, $\tilde{\Sigma}$ are defined as the sample mean and sample covariance matrix for the pseudo samples respectively:

$$\tilde{\mu} = \frac{1}{N}\sum_{n=1}^{N}\tilde{x}^{(n)} \tag{8}$$

$$\tilde{\Sigma} = \frac{1}{N}\sum_{n=1}^{N}\left(\tilde{x}^{(n)} - \tilde{\mu}\right)\left(\tilde{x}^{(n)} - \tilde{\mu}\right)^{T} \tag{9}$$

Let $\frac{\partial Q}{\partial \mu} = 0, \frac{\partial Q}{\partial K} = 0$,we get :

$$\hat{\mu} = \tilde{\mu} \tag{10}$$

$$\hat{\Sigma} = \tilde{\Sigma} + \frac{1}{N}\sum_{n=1}^{N}B^{(n)} \tag{11}$$

Now we obtain the re-estimation formula of the EM algorithm for ML parameter estimation of Gaussian distribution in (10) and (11). To this end, we need to compute $\left\langle x_i^{(n)} \right\rangle_n$ and $B_{ij}^{(n)}$ for $i, j \in h(n)$. It can be shown that this actually reduces to compute the conditional mean and covariance matrix of the missing part $x_{h(n)}^{(n)}$, given the observed part $x_{o(n)}^{(n)}$ and the current parameters $\bar{\theta}$. These quantities can be easily obtained, since we have a conditional Gaussian distribution $P\left(x_{h(n)}^{(n)} \mid \bar{\theta}, x_{o(n)}^{(n)}\right)$ with the conditional mean given by

$$\bar{\mu}_{h(n)} + \bar{\Sigma}_{h(n),o(n)}\left[\bar{\Sigma}_{h(n),o(n)}\right]^{-1}\left(x_{o(n)}^{(n)} - \bar{\mu}_{o(n)}\right)$$

and the conditional covariance matrix given by

$$\bar{\Sigma}_{h(n)} - \bar{\Sigma}_{h(n),o(n)}\left[\bar{\Sigma}_{o(n),o(n)}\right]^{-1}\bar{\Sigma}_{o(n),h(n)}$$

3. MAP eigenvoice adaptation

In eigenvoice modeling for speaker adaptation, we can directly apply the above EM algorithm to estimate the mean and covariance matrix of the supervectors, denoted as μ and Σ respectively. In this case, a training sample above is a speaker supervector[1], which may have missing subvectors for unseen phones. Performing PCA on Σ, we obtain the desirable eigenvoices, say, $e_1, e_2, \cdots, e_R$, and the corresponding eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_R$, where R is the dimension of the eigenspace. Suppose that every speaker supervector x can be written in the form:

$$x = \mu + \sum_{r=1}^{R}\left(w_r \times e_r\right) \tag{12}$$

where $w = \left(w_1, \cdots w_R\right)^T$ is the combination weights. The speaker adaptation problem then reduces to the estimation of the combination weights from the speaker's adaptation data.

MAP estimation is a good choice for this purpose. Denote the adaptation data as $o_1^T = o_1, \cdots, o_T$, and suppose that each frame *t*, $1 \le t \le T$, has been aligned to mixture component *m* with the occupation probability $\gamma_m(t)$. The MAP estimate of the weights *w* is:

$$\hat{w} = \arg\max_{w}\left[P\left(o_1^T \mid w\right)P_0\left(w\right)\right] \tag{13}$$

Here $P\left(o_1^T \mid w\right)$ is the likelihood function:

$$\log P\left(o_1^T \mid w\right) = \sum_m \sum_t \gamma_m(t) \log P\left(o_t \mid x, m\right) \tag{14}$$

where $P\left(o_t \mid x, m\right)$ denotes the Gaussian distribution of frame o_t associated with the mixture component *m* with mean x_m and variances C_m. $P_0(w)$ is the *a priori* distribution, which can be easily derived from the result of eigen-analysis of the supervector covariance matrix as:

$$\log P_0\left(w\right) = -\frac{1}{2}\sum_{r=1}^{R}\frac{w_r^2}{\lambda_r} \tag{15}$$

By taking derivative to *w*, the maximization problem in (14) can be solved as follows, for $r = 1, \cdots, R$:

$$\begin{aligned} &\sum_m \sum_t \gamma_m(t)\left(e_{r,m}\right)^T C_m^{-1}\left(o_t - \mu_m\right) \\ &= \sum_{k=1}^{R} \hat{w}_k \cdot \left[\sum_m \sum_t \gamma_m(t)\left(e_{r,m}\right)^T C_m^{-1}\left(e_{k,m}\right) + \delta_{k,r}\frac{1}{\lambda_r}\right] \end{aligned} \tag{16}$$

where $e_{r,m}$ is the subvector corresponding to mixture component *m* in the *r*-th eigenvoice e_r, and $\delta_{k,r} = 1$ iff. $k = r$, otherwise it equals to zero.

In theory, we can iterate the occupation probability computation and the weight estimation. In the first iteration, the occupation probabilities are computed using an SI model. In subsequent iterations, they are computed using the adapted model. In the experiments, we perform only one iteration, since further iterations are observed to give minor differences.

[1] The most common model for speech recognition is the hidden markov model (HMM) with guassian mixture state-output distributions. So the subvectors in a speaker supervector correspond to the means associated with the mixture components in the HMM, and will be indexed by the mixture components.

4. Experimental results

Experiments are carried out on the OGI Numbers database [7], which is an English telephone speech corpus consisting of naturally spoken numbers with 30-word vocabulary[2]. We use 6049 utterances spoken by 3059 speakers for training and 2061 utterances by 1044 speakers for testing.

The acoustic feature is 39-dimensional, formed by 12 MFCCs with normalized log-energy and their first and second order differentials. Cepstral mean subtraction is applied to the feature vector. There are 26 monophone models, a silence model, and a short-pause model. The silence and all monophones are modeled with three emitting states each, and the short-pause has only one state which is tied to the middle state of the silence model. Gaussian Mixture Density (GMD) with diagonal covariance matrices are used for state-output distributions.

We first train an SI model using all the training sentences. Then for each training speaker, the speaker's training data are viterbi aligned to state and then soft aligned to mixture components with occupation probabilities $\gamma_m(t)$ using the SI model. For the supervector sample $x^{(n)}$ from training speaker *n*, its *m*-th subvector is constructed as follows:

$$x_m^{(n)} = \frac{\sum_t \gamma_m(t) o_t}{\sum_t \gamma_m(t)} \tag{17}$$

To alleviate computation cost, we split the acoustic feature vector into 1-dimension streams. And for each stream separately, we use the algorithm in section 2 to estimate the mean and covariance matrix of the supervectors[3], and perform PCA to obtain the desirable eigenvoices. 5 eigenvoices per stream are used in the experiments. Then we conduct unsupervised MAP eigenvoice adaptation for the 1044 test speakers.

Before giving the recognition results, we show some statistics in Fig. 1, which is the average phone occupancy over all training speakers. "Phone occupancy" for a speaker is defined as the number of phones occurred in the speaker's data for each of the 26 phones. To sum up over all phones (i.e. along the horizontal axis), we can see that on average, the total number of phones observed for a speaker is 12.46, which indicates that the percentage ratio of missing values in a speaker supervector is about $13.54/26 \approx 52\%$. Using GMD-4 model, we do force alignment to obtain the state occupancy, which is displayed in Fig. 2, and again shows the data sparseness.

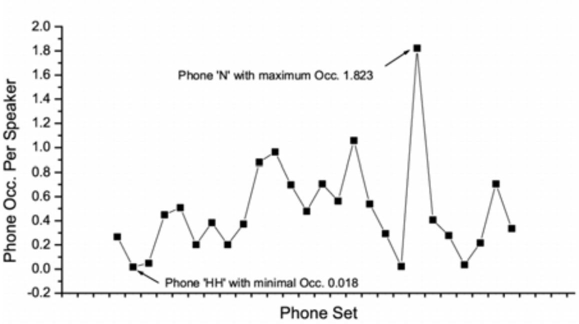

Fig. 1. Average phone occupancy over all training speakers.

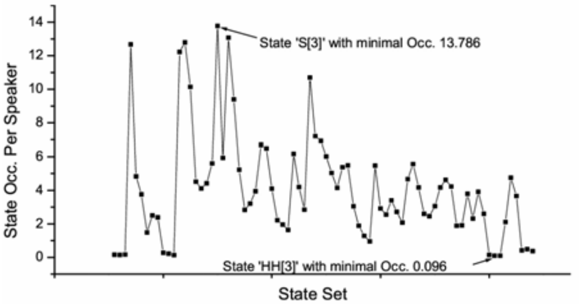

Fig. 2. Average state occupancy over all training speakers.

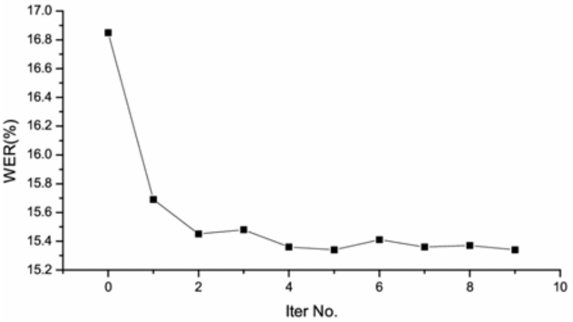

Fig. 3. WER results as a function of the iteration steps (using GMD-2).

[2] This is a relatively simple task. It will be shown below how sparse the data is in this task, and thus is suitable for investigating the problem of eigenspace estimation with missing values.

[3] For EM initialization, we copy the means and variances of the Gaussian components in the SI model to construct μ and (diagonal) $\sum$ respectively.

Table 1. WER results for the baseline, MLLR, MAP, MLLR+EV, EM+EV.

	Baseline	MLLR	MAP	MLLR+EV	EM+EV
GMD-1	20.86%	20.71%	20.75%	20.79%	19.82%
GMD-2	16.85%	16.79%	16.83%	16.27%	15.34%
GMD-4	13.34%	13.25%	13.32%	12.59%	12.20%

The word error rate (WER) results for the baseline and various adaptation methods are listed in Table 1. “MLLR+EV” refers to using MLLR adaptation to establish SD models for eigenspace estimation. “EM+EV” refers to using the proposed EM algorithm for eigenspace estimation. For both cases, MAP eigenvoice speaker adaptation is used. We experiment with varying number of Gaussian mixture components for HMM states. The results show that the EM+EV system outperform the MAP, MLLR and MLLR+EV systems consistently. Fig. 3 plots the WER results as a function of the iteration steps for the proposed EM+EV method (using GMD-2). That is, we conduct a test for the eigenvoices obtained after each EM iteration. It can be seen that the WER is gradually reduced with the iterations, which demonstrate the benefit of the iterative EM algorithm for supervector covariance estimation.

5. Conclusions

In this paper, a general approach to eigenspace estimation in situations where the training samples contain missing values is proposed and applied to eigenvoice adaptation for speech recognition. We model the underlying population as a Gaussian distribution, and perform maximum-likelihood parameter estimation for the underlying Gaussian model via EM algorithm. Experimental results show that the eigenvoice adaptation using the eigenvoices estimated by the proposed approach consistently outperforms that using MLLR adaptation to establish SD models for eigenspace estimation, and also performs better than the MLLR and MAP adaptation.

References

[1] R. Kuhn, J. C. Junqua, P. Nguyen, and et al, “Rapid speaker adaptation in eigenvoice space,” *IEEE Trans. Speech and Audio Proc.*, no. 6, pp. 695–707, 2000.

[2] H. Botterweck, “Very fast adaptation for large vocabulary continuous speech recognition using eigenvoices,” in *Proc. ICSLP*, 2000, pp. 354– 357.

[3] J. Luo, Z. Ou, and Z. Wang, “Fast eigenspace-based map adaptation within correlation subspace,” *Tsinghua Science and Technology (in chinese)*, pp. 829–832, 2004.

[4] C.J.Leggetter and P.C.Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models,” *Comput. Speech Lang.*, pp. 171–185, 1995.

[5] G. B. Patrick Kenny and P. Dumouchel, “Eigenvoice modeling with sparse training data,” *IEEE Trans. Acoustics, Speech and Signal Proc.*, no. 3, pp. 345–354, MAY 2005.

[6] D. R. AP Dempster, NM Laird, “Maximum likelihood from incomplete data via the em algorithm,” *Journal of the Royal Statistical Society*, pp. 1–38, 1977.

[7] T. L. RM Fanty, M Noel, “Telephone speech corpus development at cslu,” in *Proc. ICSLP*, 1994.