9-1-2012

Bayesian Speaker Adaptation Based on a New Hierarchical Probabilistic Model

Wen-Lin Zhang
Zhengzhou Information Science and Technology Institute

Wei-Qiang Zhang
Tsinghua University

Bi-Cheng Li
Zhengzhou Information Science and Technology Institute

Dan Qu
Zhengzhou Information Science and Technology Institute

Michael T. Johnson
Marquette University, michael.johnson@marquette.edu

Accepted version. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 20, No. 7 (September 2012): 2002-2015. DOI.

Bayesian Speaker Adaptation Based on a New Hierarchical Probabilistic Model

Wen-Lin Zhang
Department of Information Science, Zhengzhou Information Science & Technology Institute, Zhengzhou, China
Wei-Qiang Zhang
Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing, China
Bi-Cheng Li
Department of Information Science, Zhengzhou Information Science & Technology Institute, Zhengzhou, China
Dan Qu
Department of Information Science, Zhengzhou Information Science & Technology Institute, Zhengzhou, China
Michael T. Johnson
Electrical and Computer Engineering Department, Marquette University, Milwaukee, WI

Abstract: In this paper, a new hierarchical Bayesian speaker adaptation method called HMAP is proposed that combines the advantages of three conventional algorithms, maximum a posteriori (MAP), maximum-likelihood linear regression (MLLR), and eigenvoice, resulting in excellent performance across a wide range of adaptation conditions. The new method efficiently utilizes intra-speaker and inter-speaker correlation information through modeling phone and speaker subspaces in a consistent hierarchical Bayesian way. The phone variations for a specific speaker are assumed to be located in a low-dimensional subspace. The phone coordinate, which is shared among different speakers, implicitly contains the intra-speaker correlation information. For a specific speaker, the phone variation, represented by speaker-dependent eigenphones, are concatenated into a supervector. The eigenphone

IEEE Transactions on Audio, Speech, and Language Processing, Vol 20, No. 7 (September 2012): 2002-2015. DOI.

supervector space is also a low dimensional speaker subspace, which contains inter-speaker correlation information. Using principal component analysis (PCA), a new hierarchical probabilistic model for the generation of the speech observations is obtained. Speaker adaptation based on the new hierarchical model is derived using the maximum a posteriori criterion in a top-down manner. Both batch adaptation and online adaptation schemes are proposed. With tuned parameters, the new method can handle varying amounts of adaptation data automatically and efficiently. Experimental results on a Mandarin Chinese continuous speech recognition task show good performance under all testing conditions.

Section I.

Introduction

Adaptation to different speakers and environments is one of the most important functions of a modern speech recognition system. Mismatches between the training data and the testing data cannot be avoided, causing severe performance degradation even for a well-trained speech recognition system. Typical mismatches can be caused by new speakers, new speaking environments, or different transmission channels from the training data set. Adaptation techniques corresponding to these situations are referred to as speaker adaptation,[1] environment adaptation,[2] and channel compensation,[3] respectively. In this paper, we focus on the speaker adaptation of a speech recognition system based on conventional hidden Markov models (HMMs). The same adaptation techniques may also be applied to environment adaptation or channel adaptation.

The core procedure of speaker adaptation consists of maximizing the likelihood of adaptation data from a new speaker. The process can use supervised mode, where accurate transcriptions of the adaptation data are available, or unsupervised mode, where the required transcriptions must be hypothesized. Speaker adaptation can be performed in feature space or in model space. For feature space adaptation, the feature vectors of a new speaker are transformed to match the speaker independent (SI) model. Techniques of this kind include vocal tract length normalization (VTLN)[4–5,6] and feature space maximum-likelihood linear transformation (FMLLR).[7–8,9] For model space adaptation,[1,10–11,12,13,14,15] the speaker independent model is transformed to generate a speaker-dependent (SD) model for the new speaker. In this paper, only model space adaptation is considered, and both supervised and unsupervised adaptation are discussed.

Many speaker adaptation schemes have been proposed, which can be classified into three broad categories: maximum *a posteriori* (MAP),[1] maximum-likelihood linear regression (MLLR),[10] and speaker clustering.[11] In conventional MAP adaptation, a prior distribution over the SD model parameters is assumed, and the SD model parameters are

IEEE Transactions on Audio, Speech, and Language Processing, Vol 20, No. 7 (September 2012): 2002-2015. DOI.

estimated using maximum *a posteriori* criterion. The main advantage of MAP adaptation is its good asymptotic property, which means that the MAP estimate approaches the ML estimate when the adaptation data is sufficient, but it is a local update of the model parameters, in which only model parameters observed in the adaptation data can be modified from their prior values. This make it unsuitable for use with very small amounts of adaptation data. Several methods have been proposed that utilize the correlation between phones to reduce the number of parameters required by MAP methods, such as the structural Bayes method[16] and the phone-prediction method.[17]

In MLLR, however, instead of estimating the SD model directly, a set of linear transformations are estimated to transform an SI model into a new SD model. Using regression class trees, the HMM state components can be grouped into regression classes with each class having its own transformation matrix. The MLLR approach is a global adaptation scheme with lower data requirements than the MAP approach. However, its asymptotic behavior is poor, as performance improvement saturates rapidly as the adaptation data increases. The good asymptotic property of MAP adaptation is due to its Bayesian formulation, and the good performance of MLLR for smaller amounts of adaptation data can be attributed to the efficient use of correlation between different phones through regression trees. Many methods have been proposed to combine the advantages of MAP and MLLR, such as maximum *a posteriori* linear regression (MAPLR),[18] where a prior distribution of the transformation matrix is assumed, and structured maximum *a posteriori* linear regression (SMAPLR),[19] where a tree structure of the prior distributions of different transformation matrices is introduced.

Unlike MAP and MLLR, speaker clustering-based approaches deal with the speaker adaptation problem in a different way. These assume that all SD models lie in a low-dimensional manifold, so that speaker adaptation is no more than the estimation of the local or global coordinate of the new SD model. A representative of these methods is the eigenvoice method (EV).[11] where the low dimensional manifold is a linear subspace and a set of linear bases (called eigenvoices), which capture most of the variance of the SD model parameters, can be obtained by principal component analysis. During speaker adaptation, the coordinate of a new SD model is estimated using the maximum-likelihood criterion. Compared with MAP and MLLR, the eigenvoice method has fewer free parameters to be estimated, so it can yield good performance even when a few seconds of adaptation data is provided. This low data requirement is due to the explicit modeling of the correlations between different speakers through the speaker subspace. Methods combining the advantages of MAP or MLLR with eigenvoice adaptation have also been proposed, such as Bayesian speaker adaptation using probabilistic principal component analysis,[20] in which a

IEEE Transactions on Audio, Speech, and Language Processing, Vol 20, No. 7 (September 2012): 2002-2015. DOI.

probabilistic formulation of PCA is used to provide the prior of the SD models, and eigenspace-based maximum-likelihood linear regression,[21,22] where the linear subspace of SD transformation matrices is explicitly modeled.

While the explicit modeling of the speaker subspace has been widespread in many speech recognition applications,[23,24] little work has been done with subspace modeling of the phone subspace. In,[25] the "eigenphone" concept is first introduced as a set of linear bases of the phone space used in conjunction with eigenvoice. A Kullback–Leibler divergence minimization technique is introduced to estimate those phone bases and the posterior of the phone coordinates can be obtained. Experiments with a closed speaker set show good performance. However, this technique does not address the problem of how to perform speaker adaptation for a previously unseen speaker; thus, it is a multispeaker modeling technique rather than a speaker adaptation technique in the usual sense. One main contribution of the paper presented here is that we address this problem by estimating a set of speaker specific eigenphone bases for each new speaker. In our method, the same phone subspace modeling method as that of[25] is used, where the speaker specific phone variations are assumed to be in a low-dimensional linear subspace. The coordinate matrix of the whole phone set is fixed across all speakers and is estimated using the training speaker dependent models. The speaker-specific phone variation bases, which will also be called eigenphones, are estimated for each new speaker. Although the proposed method obtains better performance than the conventional ones in case of sufficient adaptation data, its performance under limited adaptation data condition (less than 10 s) is disappointing. Another contribution of this paper is that by performing eigenvoice modeling in the SD eigenphone space further a new hierarchical probabilistic model of the SD model parameters can be obtained. An efficient and flexible speaker adaptation method which yields excellent performance across a wide range of adaptation conditions can be derived under this new model. Two schemes, a batch adaptation scheme and an online adaptation scheme, are proposed. Experimental results for supervised and unsupervised speaker adaptation show good performance under all testing conditions.

This paper is organized as follows. In the next section, the construction of the phone subspace is detailed, the probabilistic generation of training speaker models using eigenphones is presented, and relationships to eigenvoice and other modeling methods are illustrated. Compact eigenvoice modeling in the eigenphone space is introduced in Section III, and the corresponding hierarchical probabilistic model is compared to that of the recent CMLLB[30] approach. In Section IV, Bayesian speaker adaptation using the new hierarchical probabilistic model is derived. Experimental results on supervised adaptation and unsupervised adaptation are presented in Section V, with conclusions in Section VI.

IEEE Transactions on Audio, Speech, and Language Processing, Vol 20, No. 7 (September 2012): 2002-2015. DOI.

Section II.

Phone Variation Subspace Modeling

Given a set of speaker independent HMMs containing a total of M mixture components across all states and models, a training speaker population comprising S speakers, and a D-dimensional speech feature vector, let $\boldsymbol{\mu}_m$ and $\boldsymbol{\Sigma}_m$ denote the speaker independent mean vector and covariance matrix, respectively, for each mixture component m, and $\boldsymbol{\mu}_m(s)$ denote the SD mean vector for a speaker s and mixture component m.

A. Eigenphones

Let $\boldsymbol{u}(m,s) = \boldsymbol{\mu}_m(s) - \boldsymbol{\mu}_m$, denoting the difference vector of mixture component m between the SD model of training speaker s and the SI model. Define a phone variation supervector $\boldsymbol{um}$ to be a supervector obtained by concatenating $\{\boldsymbol{u}(m,s)\}_{s=1}^{S}$ for some mixture component m, that is

$$\boldsymbol{u}_m = [\boldsymbol{u}(m,1)^T \boldsymbol{u}(m,2)^T \cdots \boldsymbol{u}(m,S)^T]^T. \quad (1)$$

$\boldsymbol{u}_m$ lies in an $S{\cdot}D$-dimension space, which we call the phone variation space. There are M mixture components in total, so $min(M, S \cdot D)$ bases of the phone variation space can be found using PCA. These basis vectors are called *eigenphones*,[25] denoted by $\{\boldsymbol{v}_n, n = 1,2,\cdots, min(M, S \cdot D)\}$. If we constrain all phone variation supervectors to be located in an N-dimensional subspace spanned by the first N eigenphones, an approximation for the phone variation supervectors $\{\boldsymbol{u}_m\}_{m=1}^{M}$ can be obtained as follows:

$$\begin{bmatrix} \boldsymbol{u}_1^T \\ \boldsymbol{u}_2^T \\ \vdots \\ \boldsymbol{u}_M^T \end{bmatrix} \approx \begin{bmatrix} \overline{\boldsymbol{v}}_0^T \\ \overline{\boldsymbol{v}}_0^T \\ \vdots \\ \overline{\boldsymbol{v}}_0^T \end{bmatrix} + \begin{bmatrix} l_{11} & l_{12} & \cdots & l_{1N} \\ l_{21} & l_{22} & \cdots & l_{2N} \\ \vdots & \vdots & \cdots & \vdots \\ l_{M1} & l_{M2} & \cdots & l_{MN} \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_1^T \\ \boldsymbol{v}_2^T \\ \vdots \\ \boldsymbol{v}_N^T \end{bmatrix} \quad (2)$$

where l_{mn} denotes the mth phone supervector's coordinate with respect to nth eigenphone $\boldsymbol{v}_n$, and $\overline{\boldsymbol{v}}_0 = (1/M)\sum_{m=1}^{M} \boldsymbol{u}_m$ denotes the mean of all training speaker phone variation supervectors and can be viewed as a special eigenphone determining the origin of the phone variation supervector space.

IEEE Transactions on Audio, Speech, and Language Processing, Vol 20, No. 7 (September 2012): 2002-2015. DOI.

Following the phone supervector construction (1), the origin $\overline{\boldsymbol{v}}_0$ and each eigenphone $\boldsymbol{v}_n$ can also be rearranged as a partitioned block vector, where each block is a subvector corresponding to a training speaker, i.e., we can write

$$\overline{\boldsymbol{v}}_0 = [\overline{\boldsymbol{v}}(0,1)^T \overline{\boldsymbol{v}}(0,2)^T \cdots \overline{\boldsymbol{v}}(0,S)^T]^T$$

and

$$\boldsymbol{v}_n = [\boldsymbol{v}(n,1)^T \boldsymbol{v}(n,2)^T \cdots \boldsymbol{v}(n,S)^T]^T$$

where $\overline{\boldsymbol{v}}(0,s)$ and $\{\boldsymbol{v}(n,s)\}_{n=1}^{N}$ compromise the origin and the bases of the phone variation subspace of speaker *s*, respectively.

The phone supervector decomposition (2) can be written in terms of each speaker *s* as

$$\begin{aligned}
\boldsymbol{U}(s) &= \begin{bmatrix} \boldsymbol{u}(1,s)^T \\ \boldsymbol{u}(2,s)^T \\ \vdots \\ \boldsymbol{u}(M,s)^T \end{bmatrix} \\
&\approx \begin{bmatrix} \overline{\boldsymbol{v}}(0,s)^T \\ \overline{\boldsymbol{v}}(0,s)^T \\ \vdots \\ \overline{\boldsymbol{v}}(0,s)^T \end{bmatrix} + \begin{bmatrix} l_{11} & l_{12} & \dots & l_{1N} \\ l_{21} & l_{22} & \dots & l_{2N} \\ \vdots & \vdots & \dots & \vdots \\ l_{M1} & l_{M2} & \dots & l_{MN} \end{bmatrix} \begin{bmatrix} \boldsymbol{v}(1,s)^T \\ \boldsymbol{v}(2,s)^T \\ \vdots \\ \boldsymbol{v}(N,s)^T \end{bmatrix} \\
&= \begin{bmatrix} l_{11} & l_{12} & \dots & l_{1N} & 1 \\ l_{21} & l_{22} & \dots & l_{2N} & 1 \\ \vdots & \vdots & \dots & \vdots & \vdots \\ l_{M1} & l_{M2} & \dots & l_{MN} & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{v}(1,s)^T \\ \boldsymbol{v}(2,s)^T \\ \vdots \\ \boldsymbol{v}(N,s)^T \\ \overline{\boldsymbol{v}}(0,s)^T \end{bmatrix} \\
&= \hat{\boldsymbol{L}} \cdot \hat{\boldsymbol{V}}(s)
\end{aligned} \tag{3}$$

where $\hat{\boldsymbol{L}}$ is the phone coordinate matrix augmented by a column vector of 1 and $\hat{\boldsymbol{V}}(s)$ is the speaker dependent eigenphone matrix, with each row corresponding to one speaker dependent eigenphone.

IEEE Transactions on Audio, Speech, and Language Processing, Vol 20, No. 7 (September 2012): 2002-2015. DOI.

From (3), it can be observed that the augmented phone coordinate matrix $\hat{\boldsymbol{L}}$ is speaker independent and contains the relative position of each phone in the phone variation subspace, and implicitly reflects the speaker independent intra-speaker correlation information. Using the eigenphone model (3), speaker adaptation for an unknown speaker s' can be accomplished by estimating a SD eigenphone matrix $\hat{\boldsymbol{V}}(s')$ using some adaptation data. The proposed eigenphone decomposition (3) is shown graphically in Fig. 1.

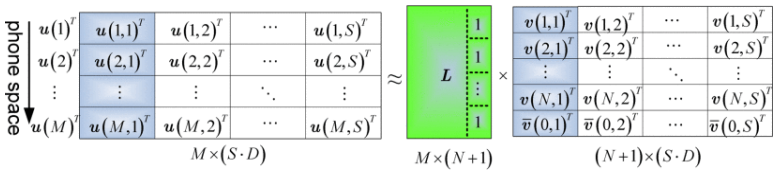

Fig. 1. Eigenphone decomposition of the training speaker phone variation supervectors. The green part shows the speaker-independent phone coordinate matrix and the blue part indicates the decomposition for the second training speaker.

B. Probabilistic Generation of the SD Models

A probabilistic formulation of PCA (probabilistic principal component analysis, PPCA) has been proposed by Tipping and Bishop.[26] Applying it to the above phone variation subspace model, we can derive a probabilistic generation model for the phone supervectors $\boldsymbol{u}_m$:

$$\boldsymbol{u}_m^T = \overline{\boldsymbol{v}}_0^T + \boldsymbol{l}_m^T \begin{bmatrix} \boldsymbol{v}_1^T \\ \boldsymbol{v}_2^T \\ \vdots \\ \boldsymbol{v}_N^T \end{bmatrix} + \boldsymbol{\varepsilon}_m^T \tag{4}$$

where $\boldsymbol{l}_m = [l_{m1} l_{m2} \ldots l_{mN}]^T$ is an Ndimensional random vector that follows a standard Gaussian distribution, and $\boldsymbol{\varepsilon}_m$ is an $S{\cdot}D$–dimensional Gaussian noise term with mean $\mathbf{0}$ and diagonal covariance matrix $\varsigma^2\boldsymbol{I}$.

Writing (4) in terms of each training speaker s, we have

$$\boldsymbol{u}(m,s)^T = \overline{\boldsymbol{v}}(0,s)^T + \boldsymbol{l}_m^T \begin{bmatrix} \boldsymbol{v}(1,s)^T \\ \boldsymbol{v}(2,s)^T \\ \vdots \\ \boldsymbol{v}(N,s)^T \end{bmatrix} + \boldsymbol{\varepsilon}(m,s)^T$$

(5)

where $\boldsymbol{\varepsilon}(m,s)$ is the Gaussian noise term corresponding to speaker s and component m.

Given the phone coordinate matrix, suppose all phone variations for speaker s are independent. Then putting together all phone variations for speaker s, we can further derive

$$\begin{aligned} \boldsymbol{U}(s) &= \begin{bmatrix} \boldsymbol{u}(1,s)^T \\ \boldsymbol{u}(2,s)^T \\ \vdots \\ \boldsymbol{u}(M,s)^T \end{bmatrix} \\ &= \begin{bmatrix} l_{11} & l_{12} & \dots & l_{1N} & 1 \\ l_{21} & l_{22} & \dots & l_{2N} & 1 \\ \vdots & \vdots & \dots & \vdots & \vdots \\ l_{M1} & l_{M2} & \dots & l_{MN} & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{v}(1,s)^T \\ \boldsymbol{v}(2,s)^T \\ \vdots \\ \boldsymbol{v}(N,s)^T \\ \overline{\boldsymbol{v}}(0,s)^T \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}(1,s)^T \\ \boldsymbol{\varepsilon}(2,s)^T \\ \vdots \\ \boldsymbol{\varepsilon}(M,s)^T \end{bmatrix} \\ &= \hat{\boldsymbol{L}} \cdot \hat{\boldsymbol{V}}(s) + \mathcal{E}(s) \end{aligned}$$

(6)

where $\mathcal{E}(s)$ is the noise matrix, with each row corresponding to one Gaussian component.

Define a speaker supervector $\boldsymbol{y}(s)$ to be a supervector obtained by concatenating the mean vectors $\boldsymbol{\mu}_m(s), m = 1,2,\cdots,M$, for a specific speaker s. Accordingly, the speaker supervector of the SI model is defined as $\boldsymbol{\mu} = [\boldsymbol{\mu}_1^T \boldsymbol{\mu}_2^T \cdots \boldsymbol{\mu}_M^T]^T$. Then the left-hand side of (6) is related to the speaker supervector $\boldsymbol{y}(s)$ via

$$\boldsymbol{y}(s) = \boldsymbol{\mu} + [\boldsymbol{u}(1,s)^T \boldsymbol{u}(2,s)^T \cdots \boldsymbol{u}(M,s)^T]^T.$$

(7)

IEEE Transactions on Audio, Speech, and Language Processing, Vol 20, No. 7 (September 2012): 2002-2015. DOI.

Substituting (6) into (7), we obtain after some manipulation the SD model for speaker s as

$$\boldsymbol{y}(s) = \boldsymbol{\mu} + \tilde{\hat{\boldsymbol{L}}}\tilde{\boldsymbol{v}}(s) + \boldsymbol{\varepsilon}(s) \tag{8}$$

where

$$\tilde{\hat{\boldsymbol{L}}} = \hat{\boldsymbol{L}} \otimes \boldsymbol{I} = \begin{bmatrix} l_{11}\boldsymbol{I} & l_{12}\boldsymbol{I} & \cdots & l_{1N}\boldsymbol{I} & \boldsymbol{I} \\ l_{21}\boldsymbol{I} & l_{22}\boldsymbol{I} & \cdots & l_{2N}\boldsymbol{I} & \boldsymbol{I} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ l_{M1}\boldsymbol{I} & l_{M2}\boldsymbol{I} & \cdots & l_{MN}\boldsymbol{I} & \boldsymbol{I} \end{bmatrix}$$

$$\tilde{\boldsymbol{v}}(s) = [\boldsymbol{v}(1,s)^T \quad \boldsymbol{v}(2,s)^T \quad \cdots \quad \boldsymbol{v}(N,s)^T \quad \overline{\boldsymbol{v}}(0,s)^T]^T \tag{9)(10}$$

is the concatenation of the SD eigenphones $\{\boldsymbol{v}(n,s)\}_{n=1}^{N}$ and the origin $\overline{\boldsymbol{v}}(0,s)$, and is called the *speaker dependent eigenphone supervector*. $\boldsymbol{\varepsilon}(s) = [\boldsymbol{\varepsilon}(1,s)^T \boldsymbol{\varepsilon}(2,s)^T \cdots \boldsymbol{\varepsilon}(M,s)^T]^T$ is an $M{\cdot}D$-dimensional Gaussian noise term with mean $\mathbf{0}$ and diagonal covariance matrix $\varsigma^2\boldsymbol{I}$.

The proof of (8) can be found in the Appendix. It reflects the probabilistic relationship between the speaker supervector and the eigenphone supervector, which will make the adaptation process similar to that of the eigenvoice method and simplify the adaptation formulation.

For a fixed phone set, $\hat{\boldsymbol{L}}$ can be viewed as a fixed matrix, or its posterior distribution can be inferred from the training data. In this paper, we fix the phone coordinate matrix with its value obtained by performing PPCA in the phone variation space. The conditional distribution of $\boldsymbol{y}(s)$ given $\tilde{\boldsymbol{L}}$ is

$$p(\boldsymbol{y}(s)|\tilde{\boldsymbol{L}}) = \mathcal{N}(\boldsymbol{y}(s)|\boldsymbol{\mu} + \tilde{\hat{\boldsymbol{L}}}\tilde{\boldsymbol{v}}(s), \varsigma^2\boldsymbol{I}). \tag{11}$$

Using (11) as a prior for the speaker supervector $\boldsymbol{y}(s)$, a Bayesian speaker adaptation method can be derived. The details will be given in Section IV.

IEEE Transactions on Audio, Speech, and Language Processing, Vol 20, No. 7 (September 2012): 2002-2015. DOI.

C. Relationship to the Eigenvoice Model

The above eigenphone decomposition scheme has a close relationship to the well-known eigenvoice modeling method. In the eigenvoice method, the decomposition is performed in the speaker space rather than the phone space. The speaker supervector $\boldsymbol{y}(s)$ is assumed to be located in a low dimensional linear subspace whose bases are called eigenvoices. Denoting the kth eigenvoice by $\boldsymbol{e}_k$ and using the probabilistic formulation of PCA, the training speaker supervectors $\{\boldsymbol{y}(s)\}_{s=1}^{S}$ can be decomposed as

$$\begin{bmatrix} \boldsymbol{y}(1)^T \\ \boldsymbol{y}(2)^T \\ \vdots \\ \boldsymbol{y}(S)^T \end{bmatrix} = \begin{bmatrix} \overline{\boldsymbol{y}}^T \\ \overline{\boldsymbol{y}}^T \\ \vdots \\ \overline{\boldsymbol{y}}^T \end{bmatrix} + \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1K} \\ w_{21} & w_{22} & \cdots & w_{2K} \\ \vdots & \vdots & \cdots & \vdots \\ w_{S1} & w_{S2} & \cdots & w_{SK} \end{bmatrix} \begin{bmatrix} \boldsymbol{e}_1^T \\ \boldsymbol{e}_2^T \\ \vdots \\ \boldsymbol{e}_K^T \end{bmatrix}$$

(12)

where $\overline{\boldsymbol{y}}$ is the mean of the training speaker supervectors $\{\boldsymbol{y}(s)\}_{s=1}^{S}$, and wsk denotes the coordinate for speaker s with respect to the kth eigenvoice $\boldsymbol{e}_k$. These K eigenvoices expand a K-dimensional subspace which implicitly contains the speaker prior information. The eigenvoice decomposition for the training speakers is shown graphically in Fig. 2.

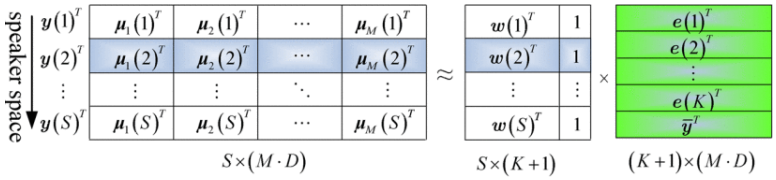

Fig. 2. Eigenvoice decomposition of the training speaker supervectors. The weighting factor for speaker $s(\boldsymbol{w}(s))$ is augmented by 1 to include the factor of the mean vector $\overline{\boldsymbol{y}}$. The green part is speaker independent and the blue part indicates the decomposition for the second training speaker.

Using the probabilistic formulation of PCA, a probabilistic model of the speaker supervector $\boldsymbol{y}(s)$ is obtained as follows:

$$\boldsymbol{y}(s) = \overline{\boldsymbol{y}} + \boldsymbol{E}w(s) + \boldsymbol{\epsilon}(s)$$

(13)

where $\boldsymbol{E} = [\boldsymbol{e}_1 \boldsymbol{e}_2 \cdots \boldsymbol{e}_K]$ and $\boldsymbol{w}(s)$ is a K dimensional random vector which follows a standard Gaussian distribution, and $\boldsymbol{\epsilon}(s)$ is a Gaussian noise term with mean $\mathbf{0}$ and covariance matrix $\sigma^2 \boldsymbol{I}$.

Although the mathematical formulation of the probabilistic eigenphone model (8) is very similar to the probabilistic eigenvoice model of (13), the intrinsic subspace decomposition methods are different, resulting in very different speaker adaptation methods. The difference can be seen graphically in Figs. 1 and 2. In fact, according to (13), the mean vector for component m of speaker s can be generated using eigenvoice by

$$\boldsymbol{\mu}_m(s) = \overline{\boldsymbol{y}}_m + \boldsymbol{E}_m \boldsymbol{w}(s) + \boldsymbol{\epsilon}(m, s) \tag{14}$$

where $\boldsymbol{E}_m$ and $\overline{\boldsymbol{y}}_m$ are the eigenvoice matrix and mean speaker supervector corresponding to component m, respectively, and $\boldsymbol{\epsilon}(m, s)$ is the corresponding Gaussian noise term.

However, using the probabilistic eigenphone model (8), we have

$$\boldsymbol{\mu}_m(s) = \boldsymbol{\mu}_m + \hat{\boldsymbol{V}}(s)^T \hat{\boldsymbol{l}}_m + \boldsymbol{\varepsilon}(m, s) \tag{15}$$

where $\hat{\boldsymbol{l}}_m = [l_{m1} l_{m2} \ldots | l_{mN} 1]^T$ and $\boldsymbol{\varepsilon}(m, s)$ is the Gaussian noise term of dimension D.

Comparing (14) and (15), it can be observed that in the eigenvoice model the basis matrix $\boldsymbol{E}_m$ of the speaker subspace is speaker independent and the speaker coordinate $\boldsymbol{w}(s)$ is unique for each speaker s, while in the eigenphone model the phone coordinate $\hat{\boldsymbol{l}}_m$ is speaker independent and the basis matrix $\hat{\boldsymbol{V}}(s)$ of the phone variation subspace is unique for each speaker s. During speaker adaptation, for a new speaker s', the eigenvoice method keeps the speaker subspace fixed and estimates the corresponding speaker coordinate $\boldsymbol{w}(s')$, while the eigenphone method keeps the relative position of each phone fixed and estimates a new set of phone variation bases. The size of the eigenphone matrix $\hat{\boldsymbol{V}}(s)$ is $(N + 1) \times D$, which has more free parameters than the eigenvoice-based method, so better adaptation performance can be expected when sufficient adaptation data is provided.

IEEE Transactions on Audio, Speech, and Language Processing, Vol 20, No. 7 (September 2012): 2002-2015. DOI.

D. Relationship to other Previous Methods

The eigenphone model also has close relationships to other previous methods, such as the structural Bayes method,[16] the phone-prediction method,[17] the conventional MLLR method[10] and the recent 2-D PCA-based method.[27]

In the structural Bayes approach, called structural MAP (SMAP),[16] a hierarchical cluster structure in the model parameter space is assumed and the probability density functions for model parameters at one level are used as priors for those of the parameters at adjacent levels. In the phone-prediction method,[17] pairwise linear regression models between sounds are built and used for prediction of unseen phones at recognition time. The effectiveness of both method can be attributed to the utilization of the correlation information between different phones. In our eigenphone model the phone space is explicitly modeled. The augmented phone coordinate matrix $\hat{\boldsymbol{L}}$ determines the relative position of each phone in the phone variation subspace and implicitly reflects the phone correlation information. Each phone variation vector $\boldsymbol{u}(m,s)$ is a linear combination of the SD eigenphones which explicitly summarizes the main phone variation patterns of speaker s.

For the conventional MLLR formulation, we can view the columns of the MLLR transform matrix as a special set of eigenphones. Consider the case in which there is a global transformation matrix. For a particular speaker s, let $\boldsymbol{A}(s)$ denote the global transformation matrix and $\boldsymbol{b}(s)$ denote the transform bias vector. The component mean $\boldsymbol{\mu}_m(s)$ is given by

$$\boldsymbol{\mu}_m(s) = \boldsymbol{\mu}_m + [\boldsymbol{A}(s) - \boldsymbol{I}\boldsymbol{b}(s)] \begin{bmatrix} \boldsymbol{\mu}_m \\ 1 \end{bmatrix}. \quad (16)$$

Comparing (16) and (15), it can be observed that if we view $\boldsymbol{b}(s)$ as the origin of the SD phone variation subspace and the columns of $\boldsymbol{A}(s) - \boldsymbol{I}$ as D eigenphones, the corresponding phone coordinate of the mth mixture is given by the SI mean vector $\boldsymbol{\mu}_m$. So the estimation of the transformation matrix and the bias vector are the same as the estimation of a $(D+1)\cdot D$-dimensional eigenphone supervector.

IEEE Transactions on Audio, Speech, and Language Processing, Vol 20, No. 7 (September 2012): 2002-2015. DOI.

The recent two-dimensional PCA-based speaker adaptation method[27] represents each training SD model as a matrix and applies 2-D PCA, resulting in a matrix decomposition of the SD component mean vector

$$\boldsymbol{\mu}_m(s) = \boldsymbol{\mu}_m + \boldsymbol{W}(s)\boldsymbol{\phi}_m^T \tag{17}$$

where $\boldsymbol{W}(s)$ is a speaker-dependent matrix of dimension $D \times K$ and $\boldsymbol{\phi}_m$ is a speaker independent vector of size K. Neglecting the noise term, the resulting decomposition (15) and (17) look the same in the mathematic form, but in the eigenphone model a subspace bias term $\overline{\boldsymbol{v}}(0, s)$ is naturally introduced and a different subspace construction method is adopted.

Section III.

Eigenvoice Modeling in the Eigenphone Space—The Compact Eigenvoice and the Hierarchical Bayesian Model

In (8), the SD eigenphone supervector $\tilde{\boldsymbol{v}}(s)$ can be estimated in an unconstrained manner using a maximum-likelihood criterion. However, when the adaptation data is limited, it cannot be estimated robustly, leading to severe overfitting problems, as will be shown in the experiments in Section V. To obtain a more robust estimation, prior information must be used. Fortunately, $\tilde{\boldsymbol{v}}(s)$ is speaker dependent and the same subspace modeling method as eigenvoice can be adopted. Applying eigenvoice analysis to the SD eigenphone supervector space results in a new hierarchical Bayesian model.

A. Modeling Method

Following the same idea as eigenvoice modeling, we decompose the SD eigenphone supervectors $\{\tilde{\boldsymbol{v}}(s)\}_{s=1}^{S}$ to be linear combinations of some common basis vectors, which we call *compact eigenvoices*. Letting $\boldsymbol{v}_p$ denote the pth compact eigenvoice, the decomposition of the S eigenphone supervectors of the training speakers can be written as

IEEE Transactions on Audio, Speech, and Language Processing, Vol 20, No. 7 (September 2012): 2002-2015. DOI.

$$\begin{bmatrix} \tilde{\boldsymbol{v}}(1)^T \\ \tilde{\boldsymbol{v}}(2)^T \\ \vdots \\ \tilde{\boldsymbol{v}}(s)^T \end{bmatrix} \approx \begin{bmatrix} \bar{\tilde{\boldsymbol{v}}}^T \\ \bar{\tilde{\boldsymbol{v}}}^T \\ \vdots \\ \bar{\tilde{\boldsymbol{v}}}^T \end{bmatrix} + \begin{bmatrix} \kappa_{11} & \kappa_{12} & \cdots & \kappa_{1P} \\ \kappa_{21} & \kappa_{22} & \cdots & \kappa_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ \kappa_{S1} & \kappa_{S2} & \cdots & \kappa_{SP} \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_1^T \\ \boldsymbol{v}_2^T \\ \vdots \\ \boldsymbol{v}_P^T \end{bmatrix}$$

(18)

where P is the number of retained bases, and $\bar{\tilde{\boldsymbol{v}}}$ is the mean of all eigenphone supervectors $\{\tilde{\boldsymbol{v}}(s)\}_{s=1}^{S}$. Define $\boldsymbol{\kappa}(s) = [\kappa_{s1}\kappa_{s2}\cdots\kappa_{sP}]^T$ the coordinate of the eigenphone supervector for speaker s, called the *compact speaker factor*. The decomposition process of the speaker supervectors can be shown graphically by Fig. 3.

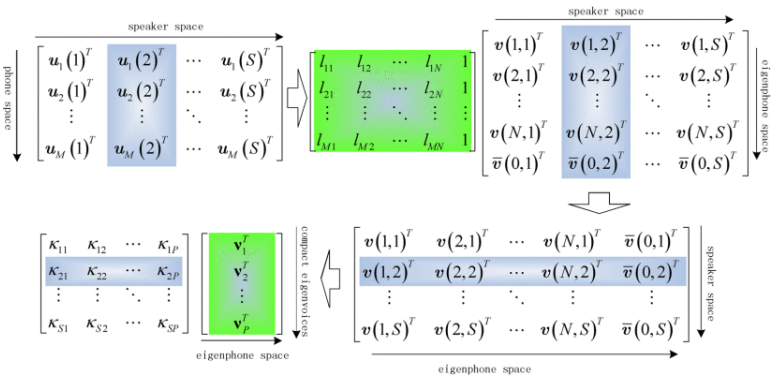

Fig. 3. Decomposition process of the SD model mean vectors. The green shaded part is speaker independent, and the blue shaded part corresponds to the decomposition of the second training speaker.

Again using PPCA, the probabilistic formulation of the eigenphone supervector can be written as

$$\tilde{\boldsymbol{v}}(s) = \bar{\tilde{\boldsymbol{v}}} + \boldsymbol{\Psi}\boldsymbol{\kappa}(s) + \boldsymbol{\zeta}$$

(19)

where $\boldsymbol{\Psi} = [\boldsymbol{v}_1\boldsymbol{v}_2\cdots\boldsymbol{v}_P]$, $\boldsymbol{\kappa}(s)$ is a P–dimensional random vector which follows a standard normal distribution, and $\boldsymbol{\zeta}$ is a Gaussian noise term with zero mean and diagonal covariance matrix $\tau^2\boldsymbol{I}$.

Combining (8) and (19), a hierarchical probabilistic model for the speaker supervector $\boldsymbol{y}(s)$ can be constructed by

IEEE Transactions on Audio, Speech, and Language Processing, Vol 20, No. 7 (September 2012): 2002-2015. DOI.

$$p(\boldsymbol{\kappa}(s)) = \mathcal{N}(\boldsymbol{\kappa}(s)|\mathbf{0}, \boldsymbol{I}) \quad (20a)$$

$$\{p(\tilde{\boldsymbol{v}}(s)|\boldsymbol{\kappa}(s)) = \left(\mathcal{N}\tilde{\boldsymbol{v}}(s)|\bar{\tilde{\boldsymbol{v}}} + \boldsymbol{\Psi}\boldsymbol{\kappa}(s), \tau^2\boldsymbol{I}\right) \quad (20b)$$

$$p(\boldsymbol{y}(s)|\tilde{\boldsymbol{v}}(s)) = \left(\mathcal{N}\boldsymbol{y}(s)|\boldsymbol{\mu} + \tilde{\boldsymbol{L}}\tilde{\boldsymbol{v}}(s), \varsigma^2\boldsymbol{I}\right). \quad (20c)$$

The hierarchical probabilistic model (20) can be shown graphically by Fig. 4, following the convention of Bishop,[28] where random variables are denoted by open circles and deterministic parameters are shown explicitly by the smaller solid circles. Note that for a fixed phone set of a specific language, the coordinate matrix $\boldsymbol{L}$ under the phone variation subspace is deterministic in this paper, although it is presented as random variable in Fig. 4.

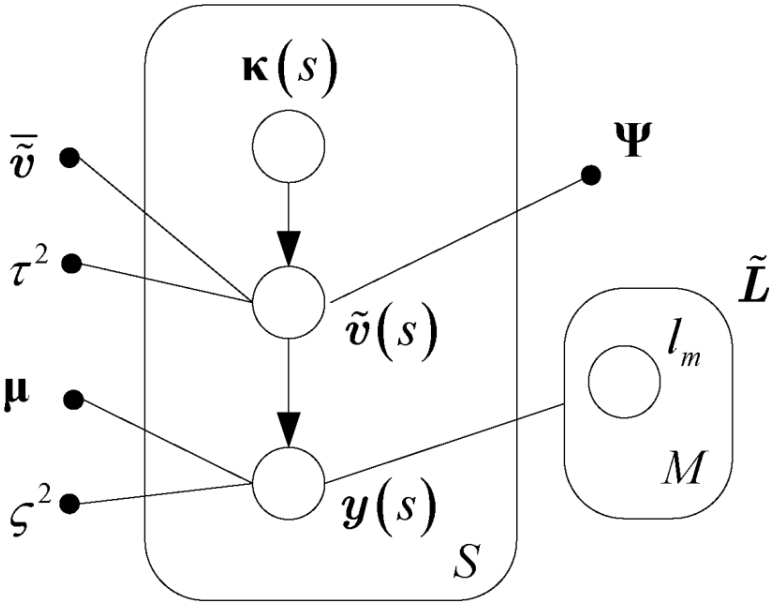

Fig. 4. Graphical representation of the hierarchical probabilistic model.

In the above hierarchical model, the phone coordinate matrix $\tilde{\boldsymbol{L}}$ is obtained by applying PPCA to the phone supervectors of the training speakers, and the compact eigenvoices matrix $\boldsymbol{\Psi}$ is calculated by performing PPCA again to the resulting eigenphone supervectors. Although maximum likelihood estimation of $\tilde{\boldsymbol{L}}$ and $\boldsymbol{\Psi}$ directly from the training data and combining the hierarchical model with the speaker adaptive training (SAT)[29] scheme are possible, we will not pursue these questions here.

IEEE Transactions on Audio, Speech, and Language Processing, Vol 20, No. 7 (September 2012): 2002-2015. DOI.

B. Relationships to Previous Methods

The compact eigenvoice approach described above is related to the clustered maximum-likelihood linear bases (CMLLB)[30] method. In CMLLB, each component mean $\boldsymbol{\mu}_m(s)$ is decomposed as

$$\boldsymbol{\mu}_m(s) = \boldsymbol{\mu}_m + \sum_{k=1}^{K} w_{sk} \boldsymbol{e}_{k,\phi(m)} \tag{21}$$

where $\phi(m)$ is a mapping function from component m to the equivalence class $\phi(m)$ and $\boldsymbol{e}_{k,\phi(m)}$ are the clustered linear bases. In the compact eigenvoice model, substituting (19) to (15) and using the equivalence between $\tilde{\boldsymbol{v}}(s)$ and $\hat{\boldsymbol{V}}(s)$, we have

$$\boldsymbol{\mu}_m(s) = \boldsymbol{\mu}_m + \overline{\hat{\boldsymbol{V}}}(s)^T \hat{\boldsymbol{l}}_m + \sum_{p=1}^{P} \kappa_{sp} \hat{\boldsymbol{V}}_p^T \hat{\boldsymbol{l}}_m \tag{22}$$

where $\overline{\hat{\boldsymbol{V}}}(s)$ and $\hat{\boldsymbol{V}}_p$ are the matrix forms of $\overline{\tilde{\boldsymbol{v}}}$ and $\boldsymbol{v}_p$, respectively, and we have neglected the noise terms. Comparing (21) and (22), it can be observed that $\hat{\boldsymbol{V}}_p^T \hat{\boldsymbol{l}}_m$ plays the same role as the clustered linear basis $\boldsymbol{e}_{k,\phi(m)}$. If we choose $\hat{\boldsymbol{l}}_m$ to be a sparse vector with only the $\phi(m)$th component equal to 1 and all other components zero, letting $\hat{\boldsymbol{V}}^T p \hat{\boldsymbol{l}} m = \boldsymbol{e} k,\phi(m)$, $\hat{\boldsymbol{V}}_p^T \hat{\boldsymbol{l}}_m = \boldsymbol{e}_{k,\phi(m)}$, $\overline{\hat{\boldsymbol{V}}}(s) = \boldsymbol{0}$ and $P = K$, the two formulations are equivalent. So the CMLLB model can be viewed as a special case of the compact eigenvoice model introduced here.

Also, the hierarchical probability model (20) has close relationships to the recent tensor based method.[31] In fact, if we set the noise terms of the eigenphone (ς^2) and compact eigenvoice (τ^2) to be zero, and let the compact speaker factor $\boldsymbol{\kappa}(s)$ be unconstrained, we obtain exactly the same tensor decomposition of the SD model as that of the multilinear decomposition in the speaker dimension ("speaking style") and the phone

IEEE Transactions on Audio, Speech, and Language Processing, Vol 20, No. 7 (September 2012): 2002-2015. DOI.

dimension ("speaking content"),[27] but our model is more intuitive and the resulting hierarchical probabilistic model (20) can be fitted to a structural Bayesian speaker adaptation framework, which is more robust and efficient.

Section IV.

Bayesian Speaker Adaptation

In this section, we will derive the Bayesian speaker adaptation method using the new hierarchical probabilistic model (20). As a first step, we reformulate the conventional MAP adaptation formula in terms of the unknown SD random variable $\boldsymbol{x}(s')$ for testing speaker s'.

A. General Framework of Bayesian Speaker Adaptation

Let $\boldsymbol{O} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, \cdots, \boldsymbol{o}_T\}$ be a sequence of feature vectors and $\boldsymbol{M} = \{m_1, m_2, \cdots, m_T\}$ represent the hypothesized mixture component sequence. Suppose the probability of observing $\boldsymbol{o}_t$ given the mixture component m and SD random variable $\boldsymbol{x}(s)$ is $p(\boldsymbol{o}(t) \mid m, \boldsymbol{x}(s))$. In Bayesian speaker adaptation, the SD random variable $\boldsymbol{x}(s')$ is assumed to follow a prior distribution $p(\boldsymbol{x}(s')|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ denotes the hyperparameters. Given the current estimate of the random variable $\hat{\boldsymbol{x}}(s)$, the auxiliary function to be optimized using the EM algorithm under the MAP criterion is given by

$$\begin{aligned} & R(\boldsymbol{x}(s), \hat{\boldsymbol{x}}(s)) \\ &= E[\log p(\boldsymbol{O}, \boldsymbol{M})|\boldsymbol{x}(s)] + E[\log p(\boldsymbol{x}(s)|\hat{\boldsymbol{x}}(s), \boldsymbol{\theta})] \\ &= \sum_{\boldsymbol{M}} p(\boldsymbol{M}|\boldsymbol{O}, \hat{\boldsymbol{x}}(s)) \log p(\boldsymbol{O}, \boldsymbol{M}|\boldsymbol{x}(s)) + \log p(\boldsymbol{x}(s)|\boldsymbol{\theta}) . \end{aligned}$$

(23)

which can be calculated to yield

IEEE Transactions on Audio, Speech, and Language Processing, Vol 20, No. 7 (September 2012): 2002-2015. DOI.

$$R(\boldsymbol{x}(s), \hat{\boldsymbol{x}}(s)) = \sum_m \sum_t \gamma_m(t) \log p(\boldsymbol{o}(t)|m, \boldsymbol{x}(s)) + \log p(\boldsymbol{x}(s)|\boldsymbol{\theta}) \quad (24)$$

where $\gamma_m(t)$ is the posterior probability of being in mixture component m at time t given the observation sequence $\boldsymbol{O}$ and $\hat{\boldsymbol{x}}(s)$. Bayesian speaker adaptation can be implemented through maximizing (24) by setting the derivatives of $\boldsymbol{x}(s)$ to zero.

B. Hierarchical MAP (HMAP) Adaptation Scheme

The probability model (20) provides a hierarchical generative model for the speaker supervector $\boldsymbol{y}(s')$. There are two levels of hyperparameters, i.e., the SD eigenphone supervector $\tilde{\boldsymbol{v}}(s')$ and the compact speaker factor $\boldsymbol{\kappa}(s')$. The MAP adaptation of each level depends on the higher level prior parameters. With decreasing adaptation data, higher level hyperparameters can be estimated more robustly than those at the lower level, as there are fewer free parameters to be estimated. A top down adaptation scheme can be performed as follows:

1. Given the adaptation data and the corresponding Gaussian level alignments for speaker s', estimate the highest level hyperparameters, i.e., the compact speaker factor $\boldsymbol{\kappa}(s')$, whose prior distribution is given by (20a).
2. Given the maximum *a posteriori* estimation of the compact speaker factor $\boldsymbol{\kappa}(s')$, estimate the second level hyperparameters, i.e., the SD eigenphone supervector $\tilde{\boldsymbol{v}}(s')$, whose prior distribution is given by (20b).
3. Given the maximum *a posteriori* estimation of the eigenphone supervector $\boldsymbol{v}(s')$, estimate the speaker supervector $\boldsymbol{y}(s')$, whose prior distribution is given by (20c).

This batch adaptation scheme can be shown graphically by Fig. 5. The detailed adaptation formula for each step will be derived in the following sections.

IEEE Transactions on Audio, Speech, and Language Processing, Vol 20, No. 7 (September 2012): 2002-2015. DOI.

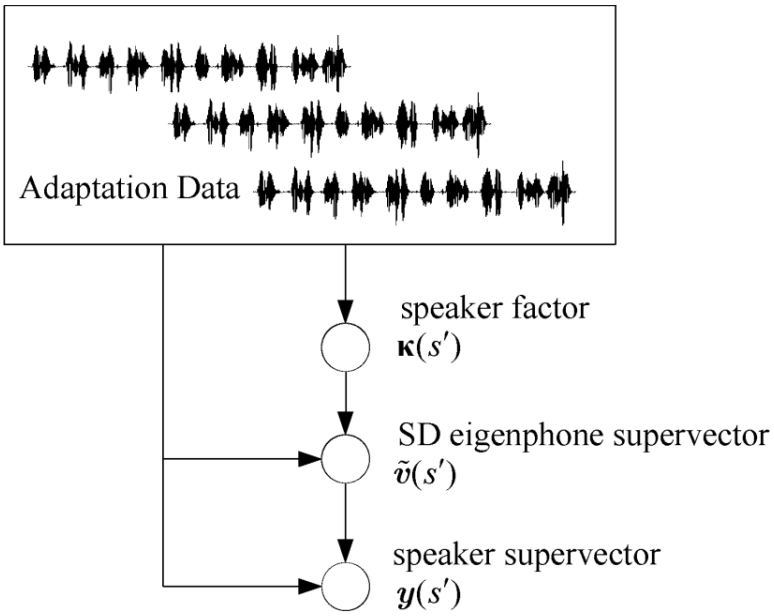

Fig. 5. Batch adaptation scheme for speaker s' using the new hierarchical probabilistic model.

As a benefit of the full Bayesian formulation, the adaptation scheme can be adjusted to perform online speaker adaptation, where the prior distribution of the compact speaker factor $\boldsymbol{\kappa}(s')$ in the current adaptation epoch is set to be the posterior distribution of the previous adaptation epoch. This online adaptation scheme can be shown graphically by Fig. 6.

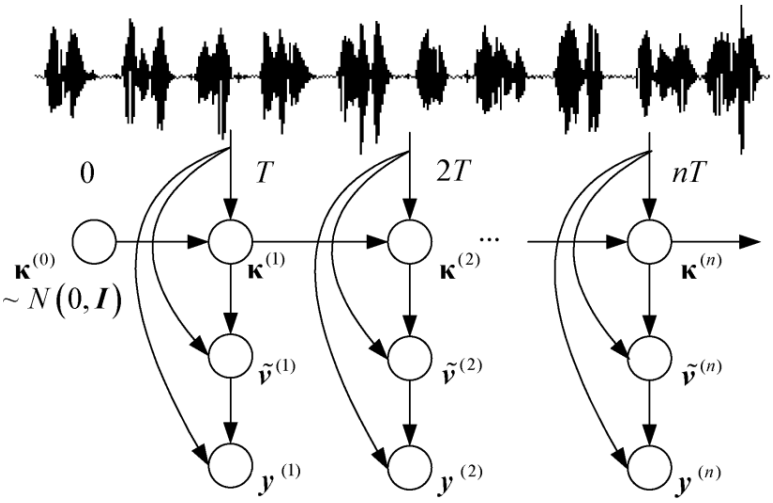

Fig. 6. Online adaptation scheme using the new hierarchical probabilistic model. $\boldsymbol{\kappa}^{(n)}$, $\tilde{\boldsymbol{v}}^{(n)}$, and $\boldsymbol{y}^{(n)}$ represent the compact speaker factor, the SD eigenphone supervector and the speaker supervector of the nth adaptation epoch, respectively. T is the updating epoch.

IEEE Transactions on Audio, Speech, and Language Processing, Vol 20, No. 7 (September 2012): 2002-2015. DOI.

C. MAP Adaptation of the Compact Speaker Factor

In this section, we consider MAP adaptation of the compact speaker factor $\boldsymbol{\kappa}(s')$ given the adaptation data. Let $\tilde{\boldsymbol{L}}_m$ denote the part of (9) corresponding to the mth mixture:

$$\tilde{\boldsymbol{L}}_m = [l_{m1}\boldsymbol{I} l_{m2}\boldsymbol{I} \ldots l_{mN}\boldsymbol{I}\boldsymbol{I}]. \tag{25}$$

Then the hierarchical model of mixture m generating observation $\boldsymbol{o}(t)$ is as follows:

$$\begin{cases} p(\boldsymbol{\kappa}(s')) = \mathcal{N}(\boldsymbol{\kappa}(s')|\boldsymbol{0}, \boldsymbol{I}) & \text{(26a)} \\ p(\tilde{\boldsymbol{v}}(s')|\boldsymbol{\kappa}(s')) = \mathcal{N}(\tilde{\boldsymbol{v}}(s')|\dot{\tilde{\boldsymbol{v}}}(s'), \tau^2\boldsymbol{I}) & \text{(26b)} \\ p(\boldsymbol{\mu}_m(s')|\tilde{\boldsymbol{v}}(s')) = \mathcal{N}(\boldsymbol{\mu}_m(s')|\dot{\boldsymbol{\mu}}_m(s'), \varsigma^2\boldsymbol{I}) & \text{(26c)} \\ p(\boldsymbol{o}(t)|m, \boldsymbol{\mu}_m(s')) = \mathcal{N}(\boldsymbol{o}(t)|\boldsymbol{\mu}_m(s'), \boldsymbol{\Sigma}_m) & \text{(26d)} \end{cases}$$

where $\dot{\tilde{\boldsymbol{v}}}(s') = \bar{\tilde{\boldsymbol{v}}} + \boldsymbol{\Psi}\kappa(s')$ and $\dot{\boldsymbol{\mu}}_m(s') = \boldsymbol{\mu}_m + \tilde{\boldsymbol{L}}_m\tilde{\boldsymbol{v}}(s')$ denote the prior mean of the SD eigenphone supervector $\tilde{\boldsymbol{v}}(s')$ and the SD component mean $\boldsymbol{\mu}_m(s')$, respectively.

In order to estimate $\boldsymbol{\kappa}(s')$ from the given observations, we must integrate across the unknown random variables, i.e., the eigenphones supervector $\tilde{\boldsymbol{v}}(s')$ and the SD mean $\boldsymbol{\mu}_m(s')$ from (26), to get the conditional distribution of observation $\boldsymbol{o}(t)$ given the compact speaker factor $\boldsymbol{\kappa}(s')$. Note that the hierarchical model (26) is a linear Gaussian model and the marginal distribution of each random variable is also a Gaussian. Applying the linear Gaussian model,[28] we arrive at

$$\begin{cases} p(\boldsymbol{\kappa}(s')) = \mathcal{N}(\boldsymbol{\kappa}(s')|\boldsymbol{0}, \boldsymbol{I}) & \text{(27a)} \\ p(\boldsymbol{o}(t)|m, \boldsymbol{\kappa}(s')) = \mathcal{N}\left(\boldsymbol{o}(t)|\boldsymbol{\mu}_m + \tilde{\boldsymbol{L}}_m\dot{\tilde{\boldsymbol{v}}}(s'), \boldsymbol{\Sigma}_m + \sigma_m^2\boldsymbol{I}\right) & \text{(27b)} \end{cases}$$

where $\sigma_m^2 = \varsigma^2 + \tau^2\left(\sum_{n=1}^{N} l_{mn}^2 + 1\right)$.

IEEE Transactions on Audio, Speech, and Language Processing, Vol 20, No. 7 (September 2012): 2002-2015. DOI.

Substituting (27) into (24), and setting the derivative of the auxiliary function with respect to $\boldsymbol{\kappa}(s')$ to zero, the estimation formula for the compact speaker factor is

$$\boldsymbol{\kappa}(s') = (\boldsymbol{A}_\kappa + \boldsymbol{I})^{-1}\boldsymbol{b}_\kappa. \tag{28}$$

where

$$\boldsymbol{A}_\kappa = \sum_m s_0(m)(\tilde{\boldsymbol{L}}_m\boldsymbol{\Psi})^T(\boldsymbol{\Sigma}_m + \sigma_m^2\boldsymbol{I})^{-1}(\tilde{\boldsymbol{L}}_m\boldsymbol{\Psi})$$

$$\boldsymbol{b}_\kappa = \sum_m (\tilde{\boldsymbol{L}}_m\boldsymbol{\Psi})^T(\boldsymbol{\Sigma}_m + \sigma_m^2\boldsymbol{I})^{-1} \times \left(\boldsymbol{s}_1(m) - s_0(m)(\boldsymbol{\mu}_m + \tilde{\boldsymbol{L}}_m\bar{\tilde{\boldsymbol{v}}})\right) \tag{29)(30}$$

and $s_0(m) = \sum_t \gamma_m(t)$ and $\boldsymbol{s}_1(m) = \sum_t \gamma_m(t)\boldsymbol{o}(t)$ are the zeroth-order and first-order statistics of the observations, respectively.

D. MAP Adaptation of the Speaker Dependent Eigenphones

Given the maximum *a posteriori* estimation of the compact speaker factor $\boldsymbol{\kappa}(s')$ by (28), the prior distribution of the SD eigenphone supervector is then obtained using (26b). In order to estimate the eigenphone supervector $\tilde{\boldsymbol{v}}(s')$, integrating across the unknown variable $\boldsymbol{\mu}_m(s')$ from (26) yields

$$\begin{cases} p(\tilde{\boldsymbol{v}}(s')|\boldsymbol{\kappa}(s')) = \mathcal{N}\left(\tilde{\boldsymbol{v}}(s')|\dot{\tilde{\boldsymbol{v}}}(s'), \tau^2\boldsymbol{I}\right) & \text{(31a)} \\ p(\boldsymbol{o}(t)|m, \tilde{\boldsymbol{v}}(s')) = \mathcal{N}\left(\boldsymbol{o}(t)|\boldsymbol{\mu}_m + \tilde{\boldsymbol{L}}_m\tilde{\boldsymbol{v}}(s'), \boldsymbol{\Sigma}_m + \varsigma^2\boldsymbol{I}\right). & \text{(31b)} \end{cases}$$

Substituting (31) into (24), and setting the derivative of the auxiliary function with respect to $\tilde{\boldsymbol{v}}(s')$ to zero, the eigenphone supervector solution is

IEEE Transactions on Audio, Speech, and Language Processing, Vol 20, No. 7 (September 2012): 2002-2015. DOI.

$$\tilde{\boldsymbol{v}}(s') = (\boldsymbol{A}_{\tilde{v}} + \tau^{-2}\boldsymbol{I})^{-1}\left[\boldsymbol{b}_{\tilde{v}} + \tau^{-2}\dot{\tilde{\boldsymbol{v}}}(s')\right] \tag{32}$$

where $\boldsymbol{A}_{\tilde{v}} = \sum_m s_0(m)\tilde{\boldsymbol{L}}_m^T(\boldsymbol{\Sigma}_m + \varsigma^2\boldsymbol{I})^{-1}\tilde{\boldsymbol{L}}_m$ and $\boldsymbol{b}_{\tilde{v}} = \sum_m \tilde{\boldsymbol{L}}_m^T(\boldsymbol{\Sigma}_m + \varsigma^2\boldsymbol{I})^{-1}(\boldsymbol{s}_1(m) - s_0(m)\boldsymbol{\mu}_m)$.

From (32), it can be observed that the inverse variance term τ^{-2} determines the tradeoff between the prior information introduced by the compact speaker factor, i.e., the prior mean $\dot{\tilde{\boldsymbol{v}}}(s')$, and the direct maximum-likelihood estimation of the eigenphone supervector. When τ^{-2} is large, more relative weight will be put on the prior information, while for small values of τ^{-2}, (32) will approach the maximum-likelihood estimated eigenphone supervector.

E. MAP Adaptation of the Mixture Means

Given the maximum *a posteriori* estimate of the SD eigenphone supervector $\tilde{\boldsymbol{v}}(s')$ from (32), the maximum *a posteriori* estimate of the SD mixture means can be derived using (26c). Substituting (26c) and (26d) into (24), and setting the derivative with respect to $\boldsymbol{\mu}_m(s')(m = 1,2,\cdots,M)$ to zero, the estimation formula for the SD mixture mean vectors becomes

$$\boldsymbol{\mu}_m(s') = (\boldsymbol{A}_{\mu_m} + \varsigma^{-2}\boldsymbol{I})^{-1}[\boldsymbol{b}_{\mu_m} + \varsigma^{-2}\dot{\boldsymbol{\mu}}_m(s')] \tag{33}$$

where $\boldsymbol{A}_{\mu_m} = s_0(m)\boldsymbol{\Sigma}_m^{-1}$ and $\boldsymbol{b}_{\mu_m} = \boldsymbol{\Sigma}_m^{-1}\boldsymbol{s}_1(m)$.

Formula (33) is very similar to that of the conventional MAP method. It can be observed that the inverse variance ς^{-2} plays the role of balancing the prior information introduced by the eigenphone supervector, i.e., the prior mean $\dot{\boldsymbol{\mu}}_m(s')$, with respect to the maximum-likelihood estimated speaker supervector. When ς^{-2} is large, more weight will be put on the prior information, while smaller values of ς^{-2} give more emphasis to the maximum likelihood estimate of the speaker supervector. Note that for mixture

IEEE Transactions on Audio, Speech, and Language Processing, Vol 20, No. 7 (September 2012): 2002-2015. DOI.

components that are not observed in the adaptation data, the total occupation $s_0(m) = 0$, so that the update formula (33) is reduced to $\boldsymbol{\mu}_m(s') = \boldsymbol{\mu}_m(s')$.

F. Online Bayesian Adaptation

For the online Bayesian adaptation scheme (Fig. 6), because of the conditional independence between the eigenphone supervector $\tilde{\boldsymbol{v}}(s')$ and the historical observations given the current compact speaker factor $\boldsymbol{\kappa}(s')$, the updates of the SD eigenphone supervector $\tilde{\boldsymbol{v}}(s')$ and speaker supervector $\boldsymbol{y}(s')$ are the same as in (32) and (33). The only difference between the batch adaptation mode and online adaptation mode for our hierarchical model lies in the update of the compact speaker factor. In online adaptation mode, the posterior distribution of the compact speaker factor summarizes all speaker information contained in the observation history, and can be used as prior distribution of the compact speaker factor for the current adaptation epoch. Recall that in a linear Gaussian model, the posterior of each random variable is also Gaussian. Suppose the posterior of the compact speaker factor in the last adaptation epoch is $\mathcal{N}(\boldsymbol{\kappa}(s')|\boldsymbol{\kappa}(s')^{(n-1)}, \boldsymbol{\Sigma}_{s'}^{(n-1)})$. Then given new adaptation data $\boldsymbol{O}(T) = \{\boldsymbol{o}(t)\}_{t=1}^{T}$ of the current epoch, the log likelihood of the joint distribution of the compact speaker factor and adaptation data is given by

$$\begin{aligned}
&\log p(\boldsymbol{O}(T), \boldsymbol{\kappa}(s')) \\
&\quad = \log p(\boldsymbol{O}(T)|\boldsymbol{\kappa}(s')) + \log p\left(\boldsymbol{\kappa}(s')|\boldsymbol{\kappa}(s')^{(n-1)}\right) \\
&\quad = \sum_t \sum_m \gamma_m(t) \log p(\boldsymbol{o}(t)|m, \boldsymbol{\kappa}(s')) \\
&\qquad + \log p\left(\boldsymbol{\kappa}(s')|\boldsymbol{\kappa}(s')^{(n-1)}\right) \\
&\quad = -\frac{1}{2}\boldsymbol{\kappa}(s')^T \left[\boldsymbol{A}_\kappa + (\boldsymbol{\Sigma}_{s'}^{(n-1)})^{-1}\right] \boldsymbol{\kappa}(s') \\
&\qquad + \boldsymbol{\kappa}(s')^T \left[\boldsymbol{b}_\kappa + (\boldsymbol{\Sigma}_{s'}^{(n-1)})^{-1} \boldsymbol{\kappa}(s')^{(n-1)}\right] \\
&\qquad + \text{Constant}.
\end{aligned} \tag{34}$$

Using the "completing the square" technique,[28] the posterior mean and variance of the current compact speaker factor $\boldsymbol{\kappa}(s')$ can be derived as

IEEE Transactions on Audio, Speech, and Language Processing, Vol 20, No. 7 (September 2012): 2002-2015. DOI.

$$\begin{cases} \left(\boldsymbol{\Sigma}_{s'}^{(n)}\right)^{-1}\boldsymbol{\kappa}(s')^{(n)} = \left(\boldsymbol{\Sigma}_{s'}^{(n-1)}\right)^{-1}\boldsymbol{\kappa}(s')^{(n-1)} + \boldsymbol{b}_{\kappa} & \text{(35a)} \\ \left(\boldsymbol{\Sigma}_{s'}^{(n)}\right)^{-1} = \left(\boldsymbol{\Sigma}_{s'}^{(n-1)}\right)^{-1} + \boldsymbol{A}_{\kappa}. & \text{(35b)} \end{cases}$$

During online speaker adaptation, we use (35) to update the mean and variance of the compact speaker factor. The initial mean ($\boldsymbol{\kappa}(s')^{(0)}$) and variance ($\boldsymbol{\Sigma}_{s'}^{(0)}$) are set to $\mathbf{0}$ and $\boldsymbol{I}$, respectively.

Section V.

Experiments

Performance of the proposed method was evaluated with speaker-independent Mandarin Chinese continuous speech recognition experiments on the Microsoft speech database.[32] Utterances from 100 male speakers were used for training data, and those from the other 25 male speakers were used for evaluation. Each training speaker contributed 200 sentences for training (about 33 hours total) and each test speaker had 20 sentences available for testing (each testing sentence lasts for about 5 seconds). All experiments were based on the standard HTK (v 3.4.1)[33] tool set. The frame length and frame step size were set as 25 ms and 10 ms, respectively. Each speech frame was parameterized by a 39-dimensional feature vector consisting of 13 Mel-frequency cepstral coefficients and their first-order and second-order time derivatives. Each Mandarin tonal syllable was modeled by a three-state left-to-right HMM without skips. After state clustering, there were 19 136 different Gaussian components in the SI model. We used a standard regression class tree based MLLR method to obtain the 100 training speakers' SD HMM models. In the recognition experiments, we drew 1, 2, 4, 6, 8, and 10 sentences from each testing speaker for adaptation, and tonal syllable recognition rate was averaged among all the remaining sentences.

A. Existence of Phone Subspace and Speaker Subspace

Initially, in order to demonstrate the existence of phone subspace, standard principal component analysis was performed on the training speakers' phone variation supervectors. The cumulative contribution of the first 100 largest eigenvalues is plotted in Fig. 7(a). Most of the variance is represented by the top 40 eigenvalues (about 81%), suggesting a low-dimension phone subspace does exist.

IEEE Transactions on Audio, Speech, and Language Processing, Vol 20, No. 7 (September 2012): 2002-2015. DOI.

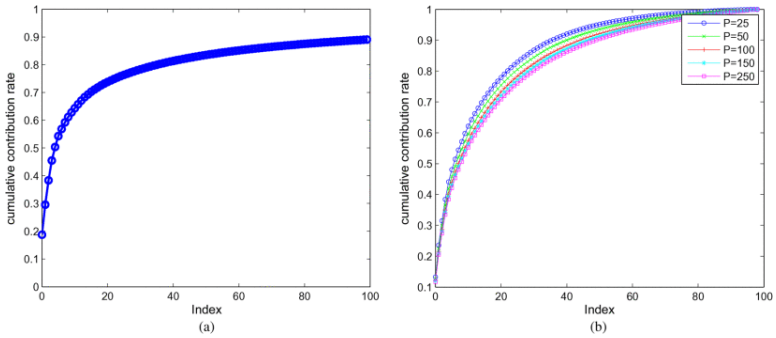

Fig. 7. (a) Cumulative contribution rate of the largest 100 eigenvalues of the phone supervector matrix. (b) Cumulative contribution rate of the 100 eigenvalues of the speaker supervector matrix in different dimensional phone subspaces.

We then constructed compact speaker supervectors for each training speaker by concatenating the corresponding SD eigenphones and performed standard principal component analysis on all the training speakers' eigenphone supervectors. The cumulative contribution of all eigenvalues for varying phone subspace dimension are plotted in Fig. 7(b). Results again support the existence of a compact speaker subspace.

B. Supervised Adaptation Experiments

For the purpose of comparison, we carried out five experiments using conventional MAP, MLLR, MLLR+SAT, MLLR+MAP, and eigenvoice adaptation methods. For MAP adaption, the weighting factor α of the prior means was varied between 10 and 20. For MLLR, the transformation matrix is 3-block-diagonal and the number of regression classes (RC=16, 32 and 64) was varied. For eigenvoice adaptation, between 10 and 100 eigenvoices were obtained from the 100 training speaker supervectors using PCA, and the maximum-likelihood eigen decomposition (MLED) formula[11] was implemented for adaptation. Adaptation experiment results of the five conventional methods are summarized in Table I. The baseline recognition accuracy of the SI model is 53.04%.

IEEE Transactions on Audio, Speech, and Language Processing, Vol 20, No. 7 (September 2012): 2002-2015. DOI.

Table I Average Tonal Syllable Recognition rate (%) Using Supervised Speaker Adaptation Based on Three Conventional Methods

Methods	Settings	Number of adaptation sentences					
		1	2	4	6	8	10
MAP	$\alpha = 10$	53.27	**53.34**	53.21	53.71	**54.78**	54.66
	$\alpha = 15$	**53.32**	53.32	53.40	**53.80**	54.49	**54.83**
	$\alpha = 20$	53.29	53.29	**53.44**	53.69	54.20	54.26
MLLR	$RC = 16$	53.04	54.68	57.41	57.81	58.81	58.92
	$RC = 32$	**53.04**	**54.68**	**57.41**	**57.93**	**58.83**	**58.98**
	$RC = 64$	53.04	54.68	57.41	57.81	58.83	58.92
MLLR+SAT	$RC = 32$	**51.03**	**53.32**	**58.06**	**58.18**	**59.04**	**59.44**
MLLR+MAP ($RC = 32$)	$\alpha = 10$	53.29	54.83	**57.85**	58.43	59.48	59.99
	$\alpha = 15$	**53.32**	**54.93**	57.83	**58.50**	**59.65**	**60.16**
	$\alpha = 20$	53.29	54.83	57.85	58.43	59.48	59.99
Eigenvoice	$K = 10$	54.93	55.50	55.46	55.73	55.76	55.77
	$K = 20$	55.73	56.38	56.61	56.90	57.11	57.05
	$K = 30$	**55.90**	56.71	56.92	57.39	57.34	57.47
	$K = 40$	55.67	56.59	57.03	57.26	57.62	57.45
	$K = 50$	55.73	56.63	57.41	57.76	58.04	57.78
	$K = 60$	55.80	**57.01**	57.15	57.36	57.87	57.95
	$K = 70$	55.41	56.95	**57.60**	**57.76**	57.93	57.99
	$K = 80$	55.37	56.97	57.39	57.45	**58.14**	58.18
	$K = 90$	55.27	56.86	56.97	57.30	57.91	58.29
	$K = 100$	55.20	56.90	57.15	57.36	57.91	**58.39**

From Table I, it can be seen that recognition results for conventional MAP adaptation method show limited improvement over the SI model for the limited adaptation data available. For the MLLR method, best results are obtained when 3-block-diagonal transformation matrix is used with 32 regression classes, and the performance consistently improves when more adaptation data is available. The MLLR+speaker adaptive training (SAT) method gives better results than MLLR when the adaptation data is sufficient. The MLLR+MAP method further improves the recognition rate. When the adaptation data is more than six sentences (about 30 s), best results are obtained when using MLLR+MAP with the prior weight α=15. Speaker adaptation using the eigenvoice method yields the best recognition results by a significant margin when the adaptation data is limited to two sentences (about 10 s) or less.

1. Speaker Adaptation Based on Maximum-Likelihood Speaker-Dependent Eigenphone Estimation

In order to determine the best number of eigenphones for our system, speaker adaptation experiments were conducted based on the maximum-likelihood eigenphone estimation described in Section IV-D. We used (32) for adaptation with $\tau^{-2} = 0$ and calculated the speaker supervector $\boldsymbol{y}(s')$ according to (8). Experimental results on different phone subspace dimensions are summarized in Table II.

IEEE Transactions on Audio, Speech, and Language Processing, Vol 20, No. 7 (September 2012): 2002-2015. DOI.

Table II Average Tonal Syllable Recognition rate (%) Using Supervised Speaker Adaptation Based on ML Eigenphone Estimation

N	Number of adaptation sentences					
	1	2	4	6	8	10
10	**51.45**	**56.71**	56.95	57.41	57.87	58.12
25	47.25	55.73	57.99	**59.36**	59.34	59.57
50	33.74	51.38	**58.16**	59.00	**59.84**	**60.62**
100	19.14	41.46	54.30	57.91	59.44	60.13

From Table II, it can be observed that when the adaptation data is limited, a small phone subspace yields better performance, and as the adaptation data increases, a larger phone subspace is preferred. The reason for this increase is that a larger phone subspace requires more free parameters be estimated, thus demanding more adaptation data. When the adaptation data is severely limited, such as with 1 sentence available (equivalent to about 5 s), the performance is worse than that of the baseline SI model. When the adaptation data is sufficient the best result is consistently obtained with 50 eigenphones, so in the following experiments, we set the dimension of the phone subspace to be 50. Notice that with N=50, the amount of adaptation data must be greater than four sentences, i.e., at least 20 seconds in order to obtain a reliable eigenphone estimation.

2. Speaker Adaptation Based on Compact Eigenvoices

In order to determine the best dimension of the speaker subspace, speaker adaptation experiments were performed based on using compact eigenvoices, i.e., the eigenvoices estimated in the SD eigenphone space. We used (28) to estimate the compact speaker factor with $\sigma_m^2 = 0$, and obtained the speaker supervector by $\boldsymbol{y}(s') = \boldsymbol{\mu} + \tilde{\overline{\boldsymbol{L}}}\overline{\boldsymbol{v}} + \tilde{\boldsymbol{L}}\boldsymbol{\Psi}\boldsymbol{\kappa}(s')$. The dimension of the phone subspace was fixed to 50. Experimental results with different speaker subspace dimensions are shown in Table III.

Table III Average Tonal Syllable Recognition rate (%) Using Supervised Speaker Adaptation Based on Maximum-Likelihood Compact Eigenvoice (N=50)

P	Number of adaptation sentences					
	1	2	4	6	8	10
10	54.28	54.24	54.49	54.76	54.64	54.43
20	54.58	55.14	55.69	55.67	55.87	55.79
40	54.78	56.27	56.46	56.65	56.57	56.74
60	54.85	56.23	56.61	56.90	57.01	57.24
80	54.85	56.25	56.86	57.26	57.30	57.39
90	**55.01**	**56.63**	56.90	57.30	**57.64**	57.66
100	54.66	56.61	**57.09**	**57.32**	57.41	**57.76**

IEEE Transactions on Audio, Speech, and Language Processing, Vol 20, No. 7 (September 2012): 2002-2015. DOI.

From Table III, it can be seen that performance is improved compared to the baseline SI model. More compact eigenvoices are required to achieve comparable performance improvement than with the conventional eigenvoice method (see Table I). The benefit of using compact eigenvoices is that the storage demands are significantly less than that of the conventional eigenvoice method. For example, in our system, to use 20 conventional eigenvoices we have to store $20 \times 19{,}136 \times 39 = 14{,}926{,}080$ float parameters. For 90 compact eigenvoices, the storage requirement is reduced to $90 \times 50 \times 39 = 175{,}500$. The only additional cost is the storage of the phone coordinate matrix, with size $19{,}136 \times 50 = 956{,}800$, giving a total storage requirement of $175{,}500 + 956{,}800 = 1{,}132{,}300$, about 7.5% of that of the conventional eigenvoice method. In order to obtain best adaptation performance with limited adaptation data, the dimension of the speaker subspace was set to 90 in the following experiments.

3. Speaker Adaptation Based on the new Hierarchical Bayesian Model

From the above experiments, we can conclude that when the adaptation data is sufficient, maximum-likelihood eigenphone adaptation provides the best speaker adaptation performance, and when the adaptation data is limited, maximum *a posteriori* compact eigenvoice adaptation performs better, giving comparable performance to that of the conventional eigenvoice based method. In this section, we investigate the adaptation performance of our proposed method of Section IV, that is, the hierarchical Bayesian model (20) based speaker adaptation method combining the two in a consistent Bayes probabilistic way.

Initially, MAP estimation of the SD eigenphone supervector using compact eigenvoice as the prior mean is tested based on adaptation formulae (28) and (32). Currently, MAP adaptation of the speaker supervector discussed in Section IV-E is not performed, so we call this approach partial-HMAP in the following sections. The performance of new method greatly depends on the variance terms ς^2 and τ^2. In order to investigate the influence of the two parameters, we fix one and vary the other around the value obtained by PPCA, denoted by $\varsigma^2_{\mathrm{PPCA}}$ and τ^2_{PPCA}, respectively.

Initially, we set ς^2 to be zero and let τ^2 be equal to τ^2_{PPCA}. The speaker adaptation results under different speaker subspace settings for this case are presented in Table IV.

IEEE Transactions on Audio, Speech, and Language Processing, Vol 20, No. 7 (September 2012): 2002-2015. DOI.

Table IV Average Tonal Syllable Recognition Rate (%) Using Speaker Adaptation Based on the Partial-HMAP Method With $N = 50, \varsigma^2 = 0, \tau^2 = \tau^2_{\text{PPCA}}$

P	τ^2_{PPCA}	Number of adaptation sentences					
		1	2	4	6	8	10
30	0.007451	48.24	54.80	58.37	**59.42**	**59.67**	**59.92**
40	0.004797	49.54	55.06	**58.52**	59.21	59.38	59.88
50	0.003107	50.31	55.33	58.18	59.09	59.40	59.76
60	0.001997	50.86	55.46	58.31	58.83	59.42	59.76
70	0.001202	51.49	55.79	58.14	59.02	59.50	59.78
80	0.000637	52.58	56.15	58.46	58.60	58.94	59.38
90	0.000231	**52.94**	**56.55**	58.22	58.40	58.75	59.17

From Table IV, it can be observed that when the adaptation data is limited, a larger speaker subspace is preferred. When the adaptation data is sufficient, smaller speaker subspace yields better performance. At a first glance, this contradicts our intuition and the previous experimental results, where a larger speaker subspace outperforms a smaller one when the adaptation data is sufficient. In fact, this phenomenon is due to the different value of the variance term τ^2_{PPCA}, which decreases quickly as P increases. From Section V, we have seen that a larger τ^2 will give more weight to the directly maximum-likelihood estimated eigenphones, so for sufficient adaptation data, larger τ^2 is preferred; while for insufficient adaptation data, smaller τ^2 is required. In addition, because of the small speaker population, the variance term τ^2 tends to be underestimated using PPCA, so we set τ^2 to a range of larger values. The experimental results are summarized in Table V.

Table V Average Tonal Syllable Recognition Rate (%) Using Speaker Adaptation Based on the Partial-HMAP Method With $N = 50, P = 90, \varsigma^2 = 0$

τ^2	Number of adaptation sentences					
	1	2	4	6	8	10
0.5	37.35	52.01	58.29	59.13	59.88	**60.62**
0.05	43.10	54.28	58.96	59.57	60.03	60.34
0.04	44.04	54.51	59.00	59.57	**60.18**	60.28
0.03	44.65	54.85	59.09	59.53	60.13	60.30
0.02	45.95	54.93	**59.23**	59.50	59.97	60.55
0.01	47.82	54.97	59.09	**59.69**	59.80	60.47
0.001	51.83	55.92	58.39	59.02	59.30	59.95
0.0001	**53.11**	**56.53**	57.76	58.29	58.35	58.77

From Table V, we can see that the adaptation performance is improved significantly when the adaptation data is sufficient (more than 20 s). As more and more data is available, a larger τ^2 is required. The reason for this is that with more adaptation data, the maximum-likelihood estimation of the eigenphone supervector becomes more robust, so its prior

IEEE Transactions on Audio, Speech, and Language Processing, Vol 20, No. 7 (September 2012): 2002-2015. DOI.

constraints should be relaxed to allow the estimated value deviate from the prior mean introduced by the compact eigenvoice, requiring a larger prior variance. However, when the adaptation data is very limited, the performance is still not as good as the compact eigenvoice based adaptation method even with a very small τ^2 (see Table III).

The performance under the limited adaptation data condition can be improved with appropriate setting of the variance term ς^2. To investigate this, we set $\tau^2 = \tau^2_{\text{PPCA}}$ and let ς^2 vary around $\varsigma^2_{\text{PPCA}}$. The results are given in Table VI.

Table VI Average Tonal Syllable Recognition rate (%) Using Speaker Adaptation Based on the Partial-HMAP Method With $N = 50, P = 90, \tau^2 = \tau^2_{\text{PPCA}} = 0.00023$

ς^2	Number of adaptation sentences					
	1	2	4	6	8	10
0.5	55.27	56.88	57.68	58.10	**58.46**	**59.00**
$0.4(\varsigma^2_{\text{PPCA}})$	55.14	56.90	57.76	**58.14**	58.43	58.83
0.2	55.18	57.36	57.87	57.95	58.14	58.81
0.1	55.56	**57.49**	**57.91**	57.89	58.08	58.73
0.08	**55.64**	57.39	57.89	57.91	58.12	58.79

From Table VI, it can be observed that with ς^2 set to a small nonzero value, the speaker adaptation results of the limited adaptation data case can be improved greatly. When the adaptation data is 1 sentence, the result is very close to the that of the conventional eigenvoice method. A more significant result is obtained with two sentences for adaptation, where the recognition rate is 57.49% with ς^2=0.1. The reason for the performance improvement may be that, when the adaptation data is insufficient, the estimation of the eigenphones is unreliable even using MAP estimation based on compact eigenvoices. So the variance term ς^2 cannot be neglected.

Based on these results, we can see that fixed variance terms ς^2 and τ^2 are not suitable for all adaptation data conditions, and that they should instead be changed dynamically according to the amount of the adaptation data. According to the results in Tables V and VI, one robust choice could be

$$\begin{cases} \varsigma^2 = 0.1, \tau^2 = \tau^2_{\text{PPCA}} = 0.00023 & \text{if } n \le 3 \\ \varsigma^2 = 0, \tau^2 = 0.01 \times (n-2) & \text{if } n > 3 \end{cases} \quad \text{(36a)(36b)}$$

where $n = \sum_t \sum_m \gamma_m(t)/500$ is proportional to the amount of the available adaptation data (measured in 5-s units). Ideally, the formula should be obtained on development data, independently from the test set, but we did not have separate development data, so (36) is obtained using a simple piecewise linear function for robustness. If the simple piecewise function (36) yields better performance than other tuned methods, the new method should give even more improvement with a well-tuned parameter function.

From (36), it can be observed that when the adaptation data is insufficient (less than 15 s), τ^2 is set to a small value ($= \tau^2_{\text{PPCA}}$), providing a tight prior constraint for the eigenphone estimation, and ς^2 is fixed to 0.08 according to Table VI. When the adaptation data is sufficient (more than 15 s), τ^2 is increased linearly as the adaptation data increases, putting more weight on the eigenphone estimation results, and ς^2 is set to 0 which means that the MAP estimation of the eigenphones can be trusted.

The partial-HMAP speaker adaptation results, using the dynamic linear parameter formulas from (36) without MAP estimation of the speaker supervector, are presented in Table VII. The best results of MLLR, MLLR+SAT, MLLR+MAP, and eigenvoice are also shown in Table VII for comparison. The dynamic linear settings improve performance greatly. Under all adaptation data conditions the recognition rates are consistently higher than those of the conventional methods when the adaptation data is more than two sentences (about 10 s), and the performance is very close to the best result of the conventional eigenvoice method with 1 sentence. Note that the result of the partial-HMAP method is not as good as that of the ML-based eigenphone method in Table II when the number of adaptation sentences is ten. The reason for this may be that the partial-HMAP method estimates the SD eigenphone supervector $\boldsymbol{v}(s')$ in a constrained manner and the adaptation data is still not enough for the MAP estimate to deviate from the prior mean $\boldsymbol{v}(s')$. Theoretically, as more data become available, the SD eigenphone supervector $\boldsymbol{v}(s')$ obtained by the partial-HMAP method should approaches those obtained by the ML-based eigenphone method.

Table VII Average Tonal Syllable Recognition rate (%) Using Supervised Speaker Adaptation With N=50,P=90 and Variance Parameters From (36)

Methods	Number of adaptation sentences					
	1	2	4	6	8	10
MLLR	53.04	54.68	57.41	57.93	58.83	58.98
MLLR+SAT	51.03	53.32	58.06	58.18	59.04	59.44
MLLR+MAP	53.32	54.93	57.85	58.50	59.65	60.16
Eigenvoice	**55.90**	57.01	57.60	57.76	58.14	58.39
partial-HMAP	55.56	57.49	59.09	59.74	60.13	60.53
full-HMAP	55.80	**57.78**	**59.40**	**60.41**	**60.93**	**61.81**

IEEE Transactions on Audio, Speech, and Language Processing, Vol 20, No. 7 (September 2012): 2002-2015. DOI.

Finally, we perform MAP estimation of the speaker supervector based on the full HMAP model. Partial-HMAP adaptation is performed using formulae (28) and (32) with parameter setting (36), then MAP adaptation of the speaker supervector is performed using formula (33) with $\varsigma^2 = 0.1$. The results are presented in Table VII as full-HMAP. It can be observed that combined with the MAP adaptation of the speaker supervector, performances is further improved. The improvement is more significant as more adaptation data become available, showing good asymptotic behavior. Compared to the best baseline method, MLLR+MAP, about 1.5% absolute improvement is achieved when the number of adaptation sentences is highest.

C. Unsupervised Adaptation Experiments

In this section, unsupervised speaker adaptation using conventional methods and the new hierarchical Bayesian method are compared. For each adaptation data condition, the corresponding 1-best recognition result is used as the hypothesized transcription. Recognition results are summarized in Table VIII. For our new methods, the parameter settings are the same as those for Table VII. For other comparing methods, best results for each experiment are given. Note that MLLR+SAT seems to perform worse than the MLLR method. This may be due to the limited size of the adaptation data.

Table VIII Average Tonal Syllable Recognition Rate (%) Using Unsupervised Speaker Adaptation

Methods	Number of adaptation sentences					
	1	2	4	6	8	10
MLLR	53.04	54.34	56.27	57.39	57.76	58.02
MLLR+SAT	51.03	52.79	56.46	57.30	57.41	57.93
MLLR+MAP	53.02	54.66	56.76	57.15	58.43	58.56
Eigenvoice	54.66	55.29	55.69	56.67	56.71	56.77
partial-HMAP	54.78	56.13	56.67	**58.16**	58.50	59.06
full-HMAP	**54.82**	**56.21**	**56.82**	58.12	**59.06**	**59.48**

Again, better performance is obtained compared with the conventional methods under all conditions. Compared with results of the partial-HMAP method, the relative improvement of the full-HMAP method is small because the hypotheses are not reliable under the unsupervised condition. When the adaptation data is ten sentences, about an absolute 1.0% improvement is achieved over the MLLR+MAP method. Note that because of the inaccurate alignment under the unsupervised condition, compared with the partial-HMAP method, no improvement is obtained for the full-HMAP method when the number of adaptation sentences is six.

IEEE Transactions on Audio, Speech, and Language Processing, Vol 20, No. 7 (September 2012): 2002-2015. DOI.

D. Unsupervised Online Adaptation Experiments

The online adaptation scheme from Section IV-F is tested in this section. The partial-HMAP scheme is adopted in this experiment. The HMAP adaptation parameters are again linear dynamic using (36). The test set contains all 20 sentences of each test speaker. Speaker adaptation is performed every 1, 2, and 5 sentences in unsupervised mode. A two pass recognition scheme is adopted in which the adaptation data used for adaptation in the current epoch is re-recognized after the current adaptation epoch is completed. Online adaptation using conventional MLLR method with 3-block-diagonal transformation matrix and 32 regression classes is also evaluated. The recognition results are given by Table IX. Compared with the MLLR method, an absolute 1.0% improvement is obtained when the updating epoch is five sentences.

Table IX Average Tonal Syllable Recognition Rate (%) Using Unsupervised Online Speaker Adaptation

Methods	Adaptation epoch (in sentences)		
	1	2	5
MLLR	57.02	57.29	57.56
partial-HMAP	**57.65**	**58.03**	**58.61**

Section VI.

Conclusion

In this paper, a new hierarchical probabilistic model for speaker adaptation called HMAP is proposed. The intra-speaker correlation and the inter-speaker correlation information of the SD model parameters are modeled simultaneously in a consistent and robust way. When the adaptation data is limited, the method focuses on the compact speaker factor level, yielding comparable performance with the conventional eigenvoice method. As the adaptation data increases, robust estimation at the SD eigenphone level can be obtained, giving consistently better performance than the conventional MLLR method. Combining the advantages of other methods through a hierarchical probabilistic formulation, HMAP gives excellent performance across a wide range of adaptation data, with experimental results showing improvement over all baseline methods.

IEEE Transactions on Audio, Speech, and Language Processing, Vol 20, No. 7 (September 2012): 2002-2015. DOI.

Appendix Proof of (8)

According to (6) and (7), we have

$$\begin{aligned}\boldsymbol{y}(s) = &\quad \boldsymbol{\mu} + [\boldsymbol{u}(1,s)^T \quad \boldsymbol{u}(2,s)^T \quad \cdots \quad \boldsymbol{u}(M,s)^T]^T \\ = &\quad \boldsymbol{\mu} + \mathrm{rvec}\left(\begin{bmatrix}\boldsymbol{u}(1,s)^T \\ \boldsymbol{u}(2,s)^T \\ \vdots \\ \boldsymbol{u}(M,s)^T\end{bmatrix}\right) = \boldsymbol{\mu} + \mathrm{rvec}(\boldsymbol{U}(s))\end{aligned}$$

(37)

where rvec(·) is a row vectorization operator by which

$$\mathrm{rvec}(\boldsymbol{U}(s)) = [\boldsymbol{u}(1,s)^T \boldsymbol{u}(2,s)^T \cdots \boldsymbol{u}(M,s)^T]^T.$$

(38)

Substituting (3) to (37) yields

$$\begin{aligned}\boldsymbol{y}(s) = &\quad \boldsymbol{\mu} + \mathrm{rvec}(\boldsymbol{U}(s)) = \boldsymbol{\mu} + \mathrm{rvec}\left(\hat{\boldsymbol{L}} \cdot \hat{\boldsymbol{V}}(s) + \mathcal{E}(s)\right) \\ = &\quad \boldsymbol{\mu} + \left(\hat{\boldsymbol{L}} \otimes \boldsymbol{I}\right) \cdot \mathrm{rvec}\left(\hat{\boldsymbol{V}}(s)\right) + \mathrm{rvec}(\mathcal{E}(s))\end{aligned}$$

(39)

where $\otimes$ is the Kronecker product operator.

Define $\tilde{\boldsymbol{L}} = \hat{\boldsymbol{L}} \otimes \boldsymbol{I}$, $\tilde{\boldsymbol{v}}(s) = \mathrm{rvec}(\hat{\boldsymbol{V}}(s))$ and $\boldsymbol{\varepsilon}(s) = \mathrm{rvec}(\mathcal{E}(s))$, we get (8).

References

[1]C.-H. Lee, C.-H. Lin, B.-H. Juang, "A study on speaker adaptation of the parameters of continuous density hidden Markov models", *IEEE Trans. Signal Process.*, vol. 39, no. 4, pp. 806-814, Apr. 1991.

[2]A. Acero, R. M. Stern, "Environmental robustness in automatic speech recognition", *Proc. ICSLP*, vol. 2, pp. 849-852, 1990-Apr.

[3]A. Solomonoff, W. Campbell, I. BoardmanCampbell, "Advances in channel compenstation for SVM speaker recognition", *Proc. ICASSP*, vol. I, pp. 629-632, 2005.

[4]T. Glaes, I. Dologlou, L. ten Bosch, D. V. Compernolle, "A novel feature transformation for vocal tract length normalization in automatic speech recognition", *IEEE Trans. Speech Audio Process.*, vol. 6, no. 6, pp. 549-557, Nov. 1998.

[5]L. Lee, R. C. Rose, "A freqency warping approach to speaker normalization", *IEEE Trans. Speech Audio Process.*, vol. 6, no. 1, pp. 49-60, Jan. 1998.

[6]S. P. Rath, S. Umesh, "Acoustic class specific VTLN-warping using regression class trees", *Proc. Interspeech*, pp. 556-559, 2009.

[7]M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition", *Comput. Speech Lang.*, vol. 12, no. 2, pp. 75-98, Apr. 1998.

[8]B. Varadarajan, D. Povey, S. M. Chu, "Quick FMLLR for speaker adaptation in speech recognition", *Proc. ICASSP*, pp. 4297-4300, 2008.

[9]A. Ghoshal, D. Povey, M. Agarwal, "A novel estimation of feature-space MLLR for full-covariance models", *Proc. ICASSP*, pp. 4310-4313, 2010.

[10]C. J. Leggetter, P. C. Woodland, "Flexible speaker adaptation using maximum likelihood linear regression", *Proc. ARPA SLS Technol. Workshop*, pp. 110-115, 1995.

[11]R. Kuhn, J.-C. Junqua, P. Nguyen, N. Niedzielski, "Rapid speaker adaptation in eigenvoice space", *IEEE Trans. Speech Audio Process.*, vol. 8, no. 6, pp. 695-707, Nov. 2000.

[12]D. K. Kim, N. S. Kim, "Rapid online adaptation using speaker space model evolution", *Comput. Speech Lang.*, vol. 42, no. 3-4, pp. 467-478, 2004.

[13]B. Gowtham Krishna, T. V. Sreenivas, "A comparative study of speaker adaptation methods", *Proc. TENCON-IEEE Region 10 Conf.*, pp. 2508-2511, 2008-Nov.

[14]W. X. Teng, G. Gravier, F. Bimbot, F. Soufflet, "Speaker adaptation by variable reference model subspace and application to large vocabulary speech recognition", *Proc. ICASSP*, pp. 4381-4384, 2009.

[15]K. Yu, M. Gales, P. C. Woodland, "Unsupervised adaptation with discriminative mapping transforms", *IEEE Trans. Audio Speech Lang. Process.*, vol. 17, no. 4, pp. 714-723, May 2009.

[16]K. Shinoda, C.-H. Lee, "A structural Bayes approaches to speaker adaptation", *IEEE Trans. Speech Audio Process.*, vol. 9, no. 3, pp. 276-287, Mar. 2001.

[17]S. Cox, "Predictive speaker adaptation in speech recognition", *Comput. Speech Lang.*, vol. 9, pp. 1-17, 1995.

[18]O. Siohan, C. Chesta, C.-H. Lee, "Hidden Markov model adaptation using maximum a posteriori linear regression", *Proc. Workshop Robust Methods Speech Recognition Adverse Conditions*, 1999.

[19]O. Siohan, T. A. Myrvoll, C. H. Lee, "Structural maximum a posteriori linear regression for fast HMM adaptation", *Comput. Speech Lang.*, vol. 16, no. 1, pp. 5-24, Jan. 2002.

[20]D. K. Kim, N. S. Kim, "Bayesian speaker adaptation based on principal component analysis", *Proc. ICSLP*, vol. 3, pp. 734-737, 2000-Oct.

[21]K.-T. Chen, W.-W. Liau, H.-M. Wang, "Fast speaker adaptation using eigenspace-based maximum likelihood linear regression", *Proc. ICSLP*, vol. 3, pp. 742-745, 2000-Oct.

[22]X. Cui, J. Xue, B. Zhou, "Improving online incremental speaker adaptation with eigen feature space MLLR", *Proc. ASRU*, pp. 136-140, 2009.

[23]P. Kenny, P. Ouellet, N. Dehak, V. Gupta, P. Dumouchel, "A study of interspeaker variability in speaker verification", *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 5, pp. 980-988, Jul. 2008.

[24]F. Castaldo, S. Cumani, P. Laface, D. Colibro, "Language recognition using language factors", *Proc. Interspeech*, pp. 176-179, 2009.

IEEE Transactions on Audio, Speech, and Language Processing, Vol 20, No. 7 (September 2012): 2002-2015. DOI.

[25]P. Kenny, G. Boulianne, P. Ouellet, "Speaker adaptation using an eigenphone basis", *IEEE Trans. Speech Audio Process.*, vol. 12, no. 6, pp. 579-589, Nov. 2004.

[26]M. Tipping, C. M. Bishop, "Probabilistic principal component analysis", *J. R. Statist. Soc.: Series B (Statist. Methodol.)*, vol. 3, no. 61, pp. 611-622, 1999.

[27]Y. Jeong, H. S. Kim, "New speaker adaptation method using 2-D PCA", *IEEE Signal Process. Lett.*, vol. 17, no. 2, pp. 193-196, Feb. 2010.

[28]C. M. Bishop, Pattern Recognition and Machine Learning, New York:Springer, Oct. 2007.

[29]T. Anastasakos, J. McDonough, R. Schwartz, J. Makhoul, "A compact model for speaker-adaptive training", *Proc. ICSLP*, pp. 1137-1140, 1996.

[30]Y. Tang, R. Rose, "Rapid speaker adaptation using clustered maximum-likelihood linear basis with sparse training data", *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 3, pp. 607-616, Mar. 2008.

[31]Y. Jeong, "Speaker adaptation based on the multilinear decomposition of training speaker models", *Proc. ICASSP*, pp. 4870-4873, 2010-Mar.

[32]E. Chang, Y. Shi, J. Zhou, "Speech lab in a box: A Mandarin speech toolbox to jumpstart speech related research", *Proc. Eurospeech*, pp. 2799-2802, 2001.

[33]S. Young, G. Evermann, M. Gales, *The HTK Book (for HTK Version 3.4)*, 2009.