Speaker Adaptation Based on Speaker-Dependent Eigenphone Estimation

Wen-Lin Zhang [#1], Wei-Qiang Zhang [*2], Bi-Cheng Li [#3]

[#] *Department of Information Science, Zhengzhou Information Science and Technology Institute*
Zhengzhou, China
[1] zwlin_2004@163.com
[3] lbclm@163.com

[*] *Department of Electronic Engineering, Tsinghua University*
Beijing, China
[2] wqzhang@tsinghua.edu.cn

***Abstract*—Based on speaker dependent eigenphone estimation, a novel speaker adaptation technique is proposed in this paper. Different from conventional speaker adaptation approaches, the proposed method explicitly models the phone variations for each speaker through subspace modeling in the phone space. The phone coordinate, which is shared by all speakers, contains correlation information between different phones. During speaker adaptation, two schemes for estimation of the new speaker specific phone variation bases (namely eigenphones) are derived under maximum likelihood (ML) criterion and maximum a posteriori (MAP) criterion respectively. Supervised speaker adaptation experiments on a Mandarin Chinese continuous speech recognition task show that the new method outperforms both eigenvoice and maximum likelihood linear regression (MLLR) methods when sufficient adaptation data is available.**

I. Introduction

In conventional speech recognition system, a set of speaker-independent (SI) Hidden Markov Models (HMMs) are trained on a multi-style database which contains different speech variation examples. Speaker adaptation is usually performed to obtain speaker dependent (SD) models from the SI models with small amount of adaptation data. To deal with the sparseness of the adaptation data, some prior or correlation information must be used to constrain the SD model parameters. There are two categories of correlation information presented in human speech: the inter-speaker correlation which is the correlation between different speakers (SD models), and the intra-speaker correlation which is the correlation between different phone pronunciations of a specific speaker.

Current speaker adaptation methods can generally be classified into three major categories: maximum a posteriori (MAP) [1] , maximum likelihood linear regression (MLLR) [2] and speaker clustering [3]. In conventional MAP adaptation method, a prior distribution over the SD model parameters is assumed, and the SD model parameters is estimated using maximum a posteriori criterion. Besides its limitation of large data requirement, 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. Different from MAP, MLLR obtain the SD model through estimating a set of linear transformation matrices. A regression class tree is usually built to capture the intra-speaker phone correlations, thus phones belonging to the same class can share the same transformation matrix. While in speaker clustering method, the inter-speaker correlation information is modeled explicitly by speaker clustering or by finding a speaker subspace for the SD model parameters. During speaker adaptation, the new SD model is obtained through a linear combination of some reference speakers (reference speaker weighting, RSW [4]) or by estimating the new speaker's coordinate under the speaker subspace using the maximum likelihood criterion (eigenvoice adaptation [3]).

In this paper, a novel speaker adaptation method through subspace modeling of the intra-speaker information is proposed. The phone variations for a particular speaker are assumed to be in a low dimensional subspace, which is called *phone variation subspace*. A set of phone variation bases, called *eigenphones*, which best represent the phone variation patterns for each speaker, can be obtained by principal component analysis (PCA). Different from conventional subspace-based adaptation methods, the new method keeps the coordinate of each phone fixed, and estimates the bases for each speaker. The coordinate matrix of the whole phone set implicitly contains the intra-speaker phone correlation information, which can be used as constraint for new SD models. During adaptation, the eigenphones for a new speaker can be estimated using maximum likelihood (ML) criterion. With a Gaussian prior assumption, an adaptation scheme based on maximum a posteriori (MAP) criterion can also be derived.

This paper is organized as follows. In the next section, subspace modeling of the phone variations is decribed, and the concept of eigenphones and their relations to conventional methods are presented. Two schemes for speaker adaptation based on eigenphone estimation are derived in Section III. Experimental results on supervised adaptation are presented in Section IV. In Section V the conclusions are given.

II. Subspace modeling in phone variation space

Suppose there are a set of speaker independent HMMs containing a total of M mixture components and a training speaker population comprising S speakers using D dimensional feature

vector. For mixture component m, let $\boldsymbol{\mu}_m$ and $\boldsymbol{\Sigma}_m$ denote the speaker independent mean vector and covariance matrix respectively. For speaker s and mixture component m, let $\boldsymbol{\mu}(m,s)$ denote the speaker dependent mean vector. In this paper, we only discuss the adaptation of the mean vectors.

A. Eigenphones

Let $\boldsymbol{u}(m,s) = \boldsymbol{\mu}(m,s) - \boldsymbol{\mu}_m$, which denotes the difference vector of mixture component m between the SI model and the SD model of training speaker s. In order to find the speaker independent correlation between different phone variations, all speaker specific vectors $\{\boldsymbol{u}(m,s)\}_{s=1}^{S}$ are concatenated to form a supervector, called the *phone supervector* of mixture component m, which is denoted by

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

It can be observed that $\boldsymbol{u}(m)$ lies in an $S \cdot D$ dimensional space and there are M mixture components, so $\min(M, S \cdot D)$ orthogonal bases of the phone supervector space can be found using PCA. As a dual of eigenvoice [3], these basis vectors (denoted by $\boldsymbol{v}_n, n = 1, 2, \cdots, \min(M, S \cdot D)$) are called *eigenphones* in this paper. If we confine the phone supervectors to be located in an N-dimensional subspace which is spanned by the first N eigenphones, an approximation of the phone supervectors $\{\boldsymbol{u}(m)\}_{m=1}^{M}$ can be obtained by

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

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

Similar to (1), $\overline{\boldsymbol{v}}_0$ and $\boldsymbol{v}_n$ can be rearranged as partitioned block vectors, where each block is a subvector corresponding to a specific training speaker, i.e. we can write

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

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

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.

Then (2) can be written in terms of each speaker as

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

where matrix $\boldsymbol{L}$ is speaker independent and contains relative position of each phone in the phone variation subspace of all speakers, and implicitly reflects the speaker independent intra-speaker correlation information.

The left hand side of (3) is related to the conventional speaker supervector $\boldsymbol{y}(s)$, which is the concatenation of the mean vectors $\{\boldsymbol{\mu}(m,s)\}_{m=1}^{M}$ for speaker s, by

$$\boldsymbol{y}(s) = \boldsymbol{\mu} + \mathrm{rvec}\left(\boldsymbol{U}(s)\right), \quad (4)$$

where $\boldsymbol{\mu}$ denotes the concatenation of the SI mean vectors, and $\mathrm{rvec}(\cdot)$ is a row vectorization operator by which

$$\mathrm{rvec}\left(\boldsymbol{U}(s)\right) = [\boldsymbol{u}(1,s)^T, \boldsymbol{u}(2,s)^T, \cdots, \boldsymbol{u}(M,s)^T]^T. \quad (5)$$

Substituting (3) to (4) yields

$$\boldsymbol{y}(s) = \boldsymbol{\mu} + \mathrm{rvec}\left(\boldsymbol{L}\boldsymbol{V}(s)\right) = \boldsymbol{\mu} + \left(\boldsymbol{L} \bigotimes \boldsymbol{I}\right) \cdot \mathrm{rvec}\left(\boldsymbol{V}(s)\right). \quad (6)$$

Define

$$\hat{\boldsymbol{L}} = \boldsymbol{L} \bigotimes \boldsymbol{I} = \begin{bmatrix} l_{11}\boldsymbol{I} & l_{12}\boldsymbol{I} & \dots & l_{1N}\boldsymbol{I} & \boldsymbol{I} \\ l_{21}\boldsymbol{I} & l_{22}\boldsymbol{I} & \dots & l_{2N}\boldsymbol{I} & \boldsymbol{I} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ l_{M1}\boldsymbol{I} & l_{M2}\boldsymbol{I} & \dots & l_{MN}\boldsymbol{I} & \boldsymbol{I} \end{bmatrix}, \quad (7)$$

and

$$\boldsymbol{v}(s) = \mathrm{rvec}\left(\boldsymbol{V}(s)\right), \quad (8)$$

which is the concatenation of the SD eigenphones $\{\boldsymbol{v}(n,s)\}_{n=1}^{N}$ and the subspace origin $\overline{\boldsymbol{v}}(0,s)$, then (6) can be rewritten as

$$\boldsymbol{y}(s) = \boldsymbol{\mu} + \hat{\boldsymbol{L}}\boldsymbol{v}(s). \quad (9)$$

$\boldsymbol{v}(s)$ is called *speaker dependent eigenphone supervector* in this paper. Formulation in this way will make the adaptation process similar to that of the eigenvoice method and simplify the derivation of the adaptation formula. A comparison of the eigenvoice decomposition method and the proposed eigenphone decomposition method is shown in Fig.1 and Fig.2.

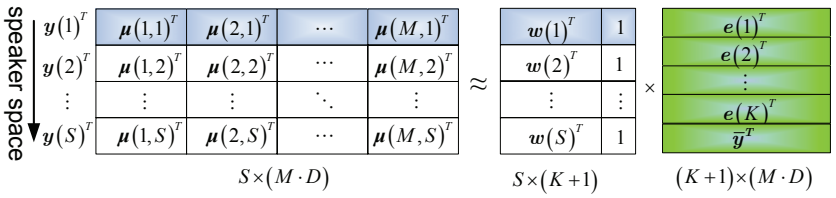

Fig. 1. Eigenvoice decomposition of the training speaker supervectors. $\boldsymbol{e}(k)$ denotes the kth eigenvoice, $\overline{\boldsymbol{y}}$ is the mean of the training speaker supervectors and $\boldsymbol{w}(s)$ is the speaker factor of speaker s. The rightmost matrix (the green part) shows speaker independent eigenvoices and the first rows of the left and middle matrices (the blue part) indicate the decomposition of the first training speaker.

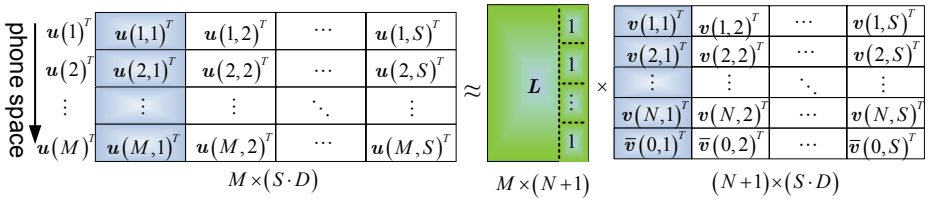

Fig. 2. Eigenphone decomposition of the training speaker phone supervectors. The middle matrix (the green part) is the speaker independent phone coordinate matrix and the first columns of the left and right matrices (the blue part) indicate the decomposition of the first training speaker.

B. Relations to MLLR and 2DPCA-based method

The conventional MLLR method can be viewed as a special case of eigenphone adaptation if we change our viewpoint of its formulation. Let's consider the case in which there is a global maximum likelihood transformation matrix. For a particular speaker s, let $\boldsymbol{A}(s)$ denote the global transformation matrix and $\boldsymbol{b}(s)$ denote the transformation bias vector. The phone variation $\boldsymbol{u}(m,s)$ of this speaker is given by

$$\boldsymbol{u}(m,s) = \boldsymbol{\mu}(m,s) - \boldsymbol{\mu}_m = (\boldsymbol{A}(s) - \boldsymbol{I})\,\boldsymbol{\mu}_m + \boldsymbol{b}(s). \quad (10)$$

From (10), it can be observed that if we view $\boldsymbol{b}(s)$ as the origin of the speaker dependent 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 just same as the estimation of $(D+1)\cdot D$ dimensional eigenphone supervector .

Our method also has some relationships with the recently 2DPCA-based method in [5]. In fact, the phone coordinate matrix $\boldsymbol{L}$ plays the the same role as the fixed matrix obtained by 2DPCA (the $\boldsymbol{\Phi}$ matrix of [5]). But our method is easier to interpret and implement.

III. Speaker adaptation based on speaker dependent eigenphone estimation

The eigenphones proposed in Section II are different from those of [6], which represent the bases of the *phone space* instead of the *phone variation space*. Those phone space bases are fixed in [6] and the phone coordinates are adapted to obtain SD models for speakers in a closed population. Because the bases are inherently speaker dependent, speaker adaptation to a previously unseen speaker cannot be pursued in this manner.

In this section, we propose another speaker adaptation method based on the SD eigenphone decomposition (11). Different from [6], the new method constrains the phone coordinate matrix $\boldsymbol{L}$ to be fixed and estimates a speaker dependent eigenphone supervector $\boldsymbol{v}(s')$ for each new speaker s', which is equivalent to performing an affine transformation of the phone space while keeping the relative position of each phone unchanged. Two estimation methods are proposed, one is based on the maximum likelihood criteria, and the other is based on the maximum posteriori estimation where a Gaussian prior over the eigenphones is assumed.

A. Maximum likelihood estimation scheme

Let $\boldsymbol{O} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, \cdots, \boldsymbol{o}_T\}$ denotes the sequence of feature vectors of the adaptation data, $\boldsymbol{M} = \{m_1, m_2, \cdots, m_T\}$ represents the corresponding mixture sequences. Using expectation-maximization (EM) algorithm, the auxiliary function to be optimized under the maximum likelihood criterion is given as follows

$$\begin{aligned} R(\boldsymbol{y}^{(n)}, \boldsymbol{y}^{(n-1)}) &= E\left[\log p(\boldsymbol{O}, M) | \boldsymbol{y}^{(n-1)}\right] \\ &= \sum_t \sum_m \gamma_m(t) \log p(\boldsymbol{o}(t) | \boldsymbol{y}_m(s')^{(n)}), \end{aligned} \quad (11)$$

where $\boldsymbol{y}_m(s')^{(n)}$ denotes the SD mean vector of mixture m, $\gamma_m(t)$ is the posterior probability of being in mixture m at time t given the observation sequence $\boldsymbol{O}$ and current estimation of SD model $\boldsymbol{y}^{(n-1)}$.

Let $\hat{\boldsymbol{L}}_m$ denotes the part of (7) corresponding to the mth mixture, that is,

$$\hat{\boldsymbol{L}}_m = \begin{bmatrix} l_{m1}\boldsymbol{I} & l_{m2}\boldsymbol{I} & \ldots & l_{mN}\boldsymbol{I} & \boldsymbol{I} \end{bmatrix}, \quad (12)$$

then

$$p\left(\boldsymbol{o}(t) \mid \boldsymbol{y}_m(s')\right) = N\left(\boldsymbol{o}(t) \mid \hat{\boldsymbol{L}}_m \boldsymbol{v}(s') + \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\right). \quad (13)$$

Substituting (13) to (11) yields

$$R(\boldsymbol{v}(s')^{(n)}) = -\frac{1}{2}\sum_t\sum_m \gamma_m(t)\left[\boldsymbol{o}(t) - \hat{\boldsymbol{L}}_m\boldsymbol{v}(s') - \boldsymbol{\mu}_m\right]^T \boldsymbol{\Sigma}_m^{-1}\left[\boldsymbol{o}(t) - \hat{\boldsymbol{L}}_m\boldsymbol{v}(s') - \boldsymbol{\mu}_m\right]. \quad (14)$$

Setting the derivative of (14) with respect to $\boldsymbol{v}(s')$ to zero, we can get

$$\boldsymbol{v}(s') = \left[\sum_m\left(\sum_t \gamma_m(t)\right)\hat{\boldsymbol{L}}_m^T\boldsymbol{\Sigma}_m^{-1}\hat{\boldsymbol{L}}_m\right]^{-1} \left[\sum_m \hat{\boldsymbol{L}}_m^T\boldsymbol{\Sigma}_m^{-1}\sum_t\left(\gamma_m(t)\left(\boldsymbol{o}(t) - \boldsymbol{\mu}_m\right)\right)\right]. \quad (15)$$

B. Maximum a posteriori estimation scheme

When the amount of the adaptation data is limited, maximum likelihood criteria cannot generate reliable estimations. In this case, prior information has to be applied to constrain the variation of the estimation parameters. For simplicity, we assume the prior distribution of the SD eigenphones is Gaussian with zero mean and diagonal variance $\sigma^2\boldsymbol{I}$. Then the auxiliary function of the EM algorithm is

$$R(\boldsymbol{v}(s')^{(n)}) = -\frac{1}{2}\sum_t\sum_m \gamma_m(t)\left[\boldsymbol{o}(t) - \hat{\boldsymbol{L}}_m\boldsymbol{v}(s') - \boldsymbol{\mu}_m\right]^T \boldsymbol{\Sigma}_m^{-1}\left[\boldsymbol{o}(t) - \hat{\boldsymbol{L}}_m\boldsymbol{v}(s') - \boldsymbol{\mu}_m\right] - \frac{1}{2\sigma^2}\boldsymbol{v}(s')^T\boldsymbol{v}(s'). \quad (16)$$

Setting the derivative of (16) with respect to $\boldsymbol{v}(s')$ to zero yields

$$\boldsymbol{v}(s') = \left[\sum_m\left(\sum_t \gamma_m(t)\right)\hat{\boldsymbol{L}}_m^T\boldsymbol{\Sigma}_m^{-1}\hat{\boldsymbol{L}}_m + \sigma^2\boldsymbol{I}\right]^{-1} \left[\sum_m \hat{\boldsymbol{L}}_m^T\boldsymbol{\Sigma}_m^{-1}\sum_t\left(\gamma_m(t)\left(\boldsymbol{o}(t) - \boldsymbol{\mu}_m\right)\right)\right]. \quad (17)$$

IV. EXPERIMENTS

Performance of the proposed method was evaluated with Mandarin Chinese continuous speech recognition experiments on the Microsoft speech database [7]. Utterances from 100 male speakers comprised the training data set and those from the other 25 male speakers were used for evaluation. Each training speaker contributed 200 sentences for training (about 33 hours in total) and each test speaker has 20 sentences, from which 10 sentences were drawn for adaptation and the other 10 were reserved for testing (the average duration of each sentence is about 5 seconds). The frame length and frame step were set as 25ms and 10ms, 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 3-state left-to-right HMM without skips. The SI model was trained using the HTK (v 3.4.1) [8] tool set. After state clustering, there were 2392 different Gaussian mixtures in the system. Then each mixture is incrementally split to have 8 Gaussian components. A standard regression class tree based MLLR method was used to obtain the 100 SD HMM models for the training speakers. In the recognition experiments, HVite was used as the decoder without any language models. We drew 2, 4, 6, 8 or 10 sentences from each testing speaker for adaptation, tonal syllable recognition rate is averaged among the remaining 10 sentences.

TABLE I
AVERAGE TONAL SYLLABLE RECOGNITION RATE (%) AFTER SPEAKER ADAPTATION USING CONVENTIONAL METHODS

Methods	Settings	Number of adaptation sentences				
		2	4	6	8	10
MAP	$\tau = 10$	52.33	52.48	52.37	53.17	53.15
	$\tau = 20$	52.27	52.60	**52.62**	**53.27**	**53.50**
	$\tau = 40$	**52.37**	**52.60**	52.39	53.11	53.15
MLLR-diagonal	$RC = 16$	53.52	54.45	54.34	54.41	54.55
	$RC = 32$	53.52	54.45	54.34	54.41	54.62
	$RC = 64$	53.52	54.45	54.34	54.41	54.62
MLLR-block	$RC = 16$	54.68	57.41	57.81	58.81	58.92
	$RC = 32$	**54.68**	**57.41**	**57.93**	**58.83**	**58.98**
	$RC = 64$	54.68	57.41	57.81	58.83	58.92
MLLR-full	$RC = 16$	53.04	56.57	57.93	58.31	58.75
	$RC = 32$	53.04	56.57	57.93	58.35	58.79
	$RC = 64$	53.04	56.57	57.93	58.35	58.79
Eigenvoice	$K = 20$	56.38	56.61	56.90	57.11	57.05
	$K = 40$	56.59	57.03	57.26	57.62	57.45
	$K = 60$	57.01	57.15	57.36	57.87	57.95
	$K = 80$	56.97	**57.39**	57.45	58.14	58.18
	$K = 100$	**57.11**	57.24	**57.53**	**57.91**	**58.39**

TABLE II
AVERAGE TONAL SYLLABLE RECOGNITION RATE (%) AFTER SPEAKER ADAPTATION BASED ON ML EIGENPHONE ESTIMATION

Eigenphone Num.	Number of adaptation sentences				
	2	4	6	8	10
10	**56.71**	56.95	57.41	57.87	58.12
25	55.73	57.99	**59.36**	59.34	59.57
50	51.38	**58.16**	59.00	**59.84**	**60.62**
100	41.46	54.30	57.91	59.44	60.13

Speaker adaptation experiments were performed in supervised mode. For the purpose of comparison, we carried out three comparative experiments using conventional MAP, MLLR and eigenvoice adaptation methods respectively. For conventional MAP adaptation, the weighting factor of the SI model (τ) was varied between 10 and 40. For MLLR, three types of transformation matrix (a diagonal matrix, a 3-block-diagonal matrix and a full matrix) with different number of regression classes (RC) are evaluated. For eigenvoice adaptation, K eigenvoices were obtained from the 100 training speaker supervectors using PCA, and the maximum likelihood eigen decomposition (MLED) formula [3] was adopted for adaptation. For our speaker dependent eigenphone adaptation method, the phone subspace dimension of 10, 25, 50, 100 were tested. Adaptation experiment results of the three conventional methods are shown in Table I, and those based on the ML eigenphone adaptation are summarized in Table II. The recognition accuracy of the SI model is 53.04% (the baseline reference result reported in [7] is 51.21%).

TABLE III
AVERAGE TONAL SYLLABLE RECOGNITION RATE (%) AFTER SPEAKER ADAPTATION BASED ON MAP EIGENPHONE ESTIMATION

Eigenphone Num.	σ^2	Number of adaptation sentences				
		2	4	6	8	10
25	0.5	56.32	57.83	**59.30**	**59.30**	**59.61**
	0.1	**56.34**	**57.99**	59.11	59.13	59.32
	0.05	56.19	57.68	58.88	58.88	59.21
50	0.5	52.14	58.13	**59.23**	59.82	**60.60**
	0.1	53.67	**58.43**	59.11	59.78	60.45
	0.05	**53.71**	58.31	59.17	**59.86**	60.34

From Table I, it can be seen that for conventional MAP adaptation methods, recognition results have little improvement over the SI model for the limited adaptation data available. For MLLR method, best results are obtained when 3-block-diagonal transformation matrix is used with 32 regression classes, and the performance is consistently improved when more adaptation data is available. Speaker adaptation using eigenvoice method yields best recognition result when the adaptation data is limited to 2 sentences (about 10 seconds).

From Table II, it can be observed that the ML speaker dependent eigenphone adaptation shows better performance than the other three methods when the amount of the adaptation data is sufficient, with best results obtained at a 50 dimension setting; but when the adaptation data is limited, smaller subspace is preferred as better SD model can be obtained if there are fewer free parameters to be estimated. However in this low dimensional subspace setting (e.g. $N = 10$), performance is not better than the conventional MLLR method when there are 6 or more sentences for adaptation.

One drawback of the ML eigenphone adaptation is that, in the case of limited adaptation data, the recognition result is worse than for the SI model when the number of eigenphones become large. This may be due to the unconstrained estimation manner of the ML criterion. So we perform the MAP adaptation scheme proposed in Section III-B, and the recognition results are presented in Table III with different settings of the prior variance (σ^2). If the amount of the adaptation data is limited, it can be shown that the MAP adaptation scheme achieves comparable performance to the eigenvoice method when the dimension of the phone subspace is set to 25, and the recognition rate can be prevented from degenerating compared to the baseline SI system when the dimension of the phone subspace is set to 50. In both cases, high recognition rates are maintained when sufficient adaptation data is provided.

V. CONCLUSION

We have proposed a new speaker adaptation approach for speech recognition through estimation of a set of speaker dependent eigenphones. These eigenphones represent the main phone variation patterns for a specific speaker. The coordination matrix of the whole phone set is fixed across different speakers, which contains the intra-speaker coordination information. With a Gaussian prior assumption, the speaker dependent eigenphones can be estimated under the MAP criterion. Experimental results show that the new method has better performance than conventional ones. Future work will look at using speaker adaptative training to obtain the phone coordination matrix and incorporating the inter-speaker information to further improve the adaptation performance.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China under Grants No. 60872142 and No. 61005019.

REFERENCES

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

[2] C. J. Leggetter and P. C. Woodland, "Flexible speaker adaptation using maximum likelihood linear regression," in *Proc. ARPA Spoken Language Technology Workshop*. Morgan Kaufmann, 1995, pp. 110–115.

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

[4] B. Mak, T.-C. Lai, and R. Hsiao, "Improving reference speaker weighting adaptation by the use of maximum-likelihood reference speakers," in *Proc. of ICASSP*, vol. 1, May. 2006, pp. I –I.

[5] Y. Jeong and H. S. Kim, "New speaker adaptation method using 2-d pca," *IEEE Signal Process. Lett.*, vol. 17, no. 2, pp. 193 –196, Feb. 2010.

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

[7] E. Chang, Y. Shi, J. Zhou *et al.*, "Speech lab in a box : a Mandarin speech toolbox to jumpstart speech related research," in *Proc. of Eurospeech*, 2001, pp. 2799–2802.

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