# Rapid Discriminative Acoustic Model Based on Eigenspace Mapping for Fast Speaker Adaptation

Bowen Zhou, *Member, IEEE,* and John H. L. Hansen, *Senior Member, IEEE*

***Abstract*—It is widely believed that strong correlations exist across an utterance as a consequence of time-invariant characteristics of speaker and acoustic environments. It is verified in this paper that the first primary eigendirections of the utterance covariance matrix are speaker dependent. Based on this observation, a novel family of fast speaker adaptation algorithms entitled Eigenspace Mapping (EigMap) is proposed. The proposed algorithms are applied to continuous density Hidden Markov Model (HMM) based speech recognition. The EigMap algorithm rapidly constructs discriminative acoustic models in the test speaker's eigenspace by preserving discriminative information learned from baseline models in the directions of the test speaker's eigenspace. Moreover, the adapted models are compressed by discarding model parameters that are assumed to contain no discrimination information. The core idea of EigMap can be extended in many ways, and a family of algorithms based on EigMap is described in this paper. Unsupervised adaptation experiments show that EigMap is effective in improving baseline models using very limited amounts of adaptation data with superior performance to conventional adaptation techniques such as MLLR and block diagonal MLLR. A relative improvement of 18.4% over a baseline recognizer is achieved using EigMap with only about 4.5 s of adaptation data. Furthermore, it is also demonstrated that EigMap is additive to MLLR by encompassing important speaker dependent discriminative information. A significant relative improvement of 24.6% over baseline is observed using 4.5 s of adaptation data by combining MLLR and EigMap techniques.**

***Index Terms*—Discriminative acoustic model, eigenspace mapping, hidden Markov models, rapid speaker adaptation, speech recognition.**

## I. Introduction

RAPID speaker adaptation for large vocabulary continuous speech recognition (LVCSR) has been an interesting and challenging problem for last decade. The task of how to adapt a set of speaker independent (SI) hidden Markov models (HMMs) to a new speaker with a small amount of adaptation data is very important in many applications, such as speech recognition in changing car environments [14] or data mining in an audio information retrieval system [32]. The main challenge of rapid speaker adaptation is to improve speech recognition performance by adjusting the speaker independent recognition system toward a target speaker, where a range of speaker specific acoustical information must be learned from a very limited amount of adaptation data.

Currently, the most commonly-used speaker adaptation algorithms include transformation-based techniques and Bayesian learning. The typical approach of the former is maximum likelihood linear regression (MLLR) [20], which is achieved with affine transformations using maximum likelihood estimation. The representative approach of the latter is maximum *a posterior* (MAP) [11], which combines adaptation data with some *a priori* knowledge concerning the model parameters that was represented by *a priori* distribution. In addition, there are also several extensions or combinations of these two schemes that have been extensively investigated in recent years that include regression based model prediction (RMP) [1], Structural MAP [25], block-diagonal MLLR [21], MAP linear regression (MAPLR) [5], [6] and structural MAPLR [26], discounted likelihood linear regression (DLLR) [3], and others (refer to the review in [29] for more comparisons). These two families of algorithms are able to obtain direct adaptation for the test speaker by transforming only the SI models, which is obviously one of the desirable properties for speaker adaptation technologies. Both MLLR and MAP adaptation have been successfully applied to many speaker adaptation situations where sufficient amounts of adaptation data are available. For relatively small amounts of adaptation data, transformation-based schemes have demonstrated superior performance over MAP due to its global adaptation via transformation sharing. On the other hand, MAP adaptation is more desirable for its asymptotic convergence to maximum likelihood estimation when the amount of adaptation data continues to increase [11]. However, both MLLR and MAP have not been able to show comparable improvements when only a very limited amount of adaptation data is available (e.g., around 5 s of adaptation data observed in the first utterances from the test speaker for an HMM system with 100 K component Gaussians), which is very important for many real world applications such as telephony or spoken dialog systems which require rapid adaptation.

Recently, a family of cluster based speaker adaptation schemes has received much attention [17], [27]. In this family of approaches, the correlations among different training speakers are explored and adaptation is based on obtaining the appropriate linear combination of acoustic models of some "canonical" speakers. Eigenvoice, which is based on *a priori* knowledge of speaker variation, is a typical example of such cluster based speaker adaptation [17], [18]. In this method, all mean vectors from a single set of acoustic models are combined into a "supervector" and then the speaker space is constructed by spanning a $K$-space via the principal component analysis

Manuscript received December 18, 2001; revised November 17, 2003. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Shrikanth Narayanan. This work was supported by the NSF under Cooperative Agreement IIS-9817485. The opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of NSF.

B. Zhou was with the Robust Speech Processing Group, University of Colorado, Boulder, CO 80302 USA. He is now with IBM, Yorktown Heights, NY 10598 USA.

J. H. L. Hansen was with the Robust Speech Processing Group, University of Colorado, Boulder, CO 80302 USA. He is now with the University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: john.hansen@colorado.edu).

Digital Object Identifier 10.1109/TSA.2005.845808

(PCA) of a reasonable set of supervectors. The eigenvoices with the first $K$ largest eigenvalues are chosen as the basis set of this speaker space. Next, the adapted acoustic models for the test speaker is represented, and hence obtained, as a point in this $K$-dimensional eigenspace through a maximum likelihood eigen-decomposition algorithm. This family of schemes was shown to produce better speaker adaptation performance than MLLR or MAP when only a small amount of adaptation data was available.

However, speaker cluster based schemes require either the entire training corpus to be available on-line for the adaptation process, or a set of well-formed speaker dependent (SD) models, or *a priori* knowledge about speaker class information extracted from a large amount of training speakers. These issues impact the practical application of model adaptation due to either large data storage requirements or insufficient data to obtain reliable *a priori* speaker class information. If this family of model adaptation methods are collectively compared, it becomes apparent that an algorithm that directly adapts acoustic models from a single set of SI models and requires minimal resources is more attractive.

With advances in applying speech technology to different tasks, many speech applications require rapid deployment of speech recognition with minimal resources. In such speech systems, it is desirable that the acoustic model be dynamically improved based only on the baseline model and a very limited amount of adaptation data. In other words, no training data, no speaker dependent models or any *a priori* speaker clustering information is demanded for adaptation (note that this makes the proposed algorithm fundamentally different from speaker class based methods), and the computational and storage overhead of the adaptation process in the desirable algorithm should be inexpensive. On the other hand, the desirable algorithm should be effective in adaptation using a very limited amount of adaptation data.

Our goal in this paper, is therefore to develop a novel algorithm to meet these requirements. To achieve these goals, the proposed algorithm should sufficiently capitalize on information contained in the baseline model, and also be able to discover sufficient speaker knowledge within a very limited amount of adaptation data. One of the motivations of this paper is from the widely believed fact that strong correlations exist across an utterance as a consequence of time-invariant characteristics of speaker and/or acoustic environments. Given a sequence of observation feature frames from an utterance, there are at least two types of correlation that exist over the observations: the temporal correlation between feature frames, and the correlation between feature components. However, state-of-the-art speech recognition technologies ignore such correlations. For example, it is usually assumed that observations are independent in both acoustic model training and decoding. The use of dynamic feature components [9] partly captures some correlation between feature frames, but it is limited to neighboring frames. On the other hand, for many practical considerations, such as storage and computation, acoustic models typically assume diagonal covariance. This assumption ignores the correlations between feature components. It is expected that bringing these correlations into consideration should produce more accurate acoustic modeling. For example, linear discriminant analysis (LDA) [19], [24] and maximum likelihood linear transform (MLLT) [10], [13] have been used to improve acoustic model training.

The focus of this paper is to introduce a method that dynamically incorporates the correlation at the decoding phase for rapid model adaptation. It is noted that directly modeling the correlation is too expensive and not computationally practical. Alternatively, the proposed method constrains model parameters implicitly based on correlation. The question of how to capture speaker information from limited amounts of adaptation data, and how to impose the speaker information appropriately into baseline acoustic models are the key problems investigated in this paper. The existence of strong correlation within an utterance has long been noted by researchers in the literature [2]. The motivation for using long distance correlation for rapid speaker adaptation is that the correlation should be speaker dependent. Intuitively, the manner by which speech frames affect each other is highly related to the vocal tract movement and speaking styles, which are largely dependent on the speaker [22]. In Section II, a set of experiments are designed to verify this claim. As one might expect, it is observed from our experiments that the first primary eigendirections of the utterance covariance matrix encode significant speaker information.

If every component Gaussian distribution in the acoustic model is viewed as a class, then a well-trained baseline model can be assumed to maintain a fair discrimination power between different Gaussians, in the sense of providing a reasonable between-class covariance $\mathbf{B_x}$. $\mathbf{B_x}$ can be decomposed into the sum of variances along its different eigendirections. Among them, the variances that belong to the first primary eigendirection reflect the dominant power for discrimination. This paper proposes an algorithm to construct the discriminative acoustic models for the test speaker, by preserving the dominant discriminating power from the baseline model along the test speaker's first primary eigendirection of the specific speaker's between-class covariance $\mathbf{B_y}$. Other constraints are also imposed on the adapted means to minimize the shift from the baseline model due to insufficient observations of adaptation data within the context of rapid adaptation. The adaptation process is performed through a linear transformation in the model space using a method entitled Eigenspace Mapping (EigMap). Based on the core idea of EigMap, a number of algorithms can be extended using different objective functions. Some typical examples include one algorithm entitled Structural maximum likelihood Eigenspace mapping (SMLEM) [30], [31], which will also be developed in this paper. Experimental results show that EigMap is effective in improving the baseline model using very limited amounts of adaptation data with superior performance to MLLR. Moreover, EigMap is highly additive to MLLR by bringing additional discrimination information into the adapted acoustic model that maximizes the adaptation data likelihood.

The remainder of this paper is organized as follows: Section II investigates the speaker information in utterances, and shows from experiments that the first primary eigendirections of the observation covariance matrix encode significant speaker information; Section III develops the eigenspace mapping algorithm, and points out the relationship between EigMap and LDA; Section IV introduces some extensions of EigMap algorithm such as SMLEM; Section V evaluates the proposed algorithm with multiple experiments in standard applications using both na-

tive and nonnative speakers from the Wall Street Journal (WSJ) corpus; Section VI is a discussion of algorithm issues and Section VII summarizes the paper contributions.

## II. Speaker Information in Utterances

In speech recognition, raw speech from a speaker is typically first parameterized into the Mel-cepstrum. One interesting observation is that the covariance matrix of the feature vectors from a specific speaker encapsulates a range of speaker dependent features. An example can be found in [4], [33] where the statistics based on covariances are used successfully to detect speaker turns in audio streams. Previous work by other researchers have also shown that the statistics based on the covariance matrix can be applied successfully in the task of speaker identification and tracking [16].

The utterance covariance matrix $\mathbf{B_u}$ of feature vectors $\{\mathbf{O}_i \mid i = 1, 2, \ldots, t\}$ represents the variance among feature dimensions, and the first few eigenvectors[1] in the ordered set of of eigenvectors $\mathbf{e}_{ui}$ of $\mathbf{B_u}$ will therefore indicate the directions in the feature space, in decreasing order, that contribute most to the variances between feature vectors [15]. A reasonable assumption from the above observations is that the directions of the first few eigenvectors encapsulate a range of those speaker specific traits. In other words, what directions contribute most to the variances between feature vectors reflects the speaker's primary acoustic characteristics.

Typically, dependence in feature observations exist between more than two feature components, and Principal Component Analysis (PCA) can help extract the most important dimensions of variations. In our study, we claim that the first primary eigendirections encode more significant speaker information than phonemic information.

We design the following experiments to verify our claim.

1. First, we select a set of speakers, $\mathbf{S} = \{s_1, s_2, \ldots, s_L\}$, and randomly select an identical set of utterances $\mathbf{U} = \{u_1, u_2, \ldots, u_L\}$ produced by each speaker in $S$. A well-trained speaker independent acoustic model $\mathbf{\Lambda}$ with 100 K component Gaussians is used to represent the acoustic space.
2. For each utterance $u$ in $\mathbf{S} \times \mathbf{U}$, we estimate the covariance matrix $\mathbf{B_u}$ of the observation frames in the standard MFCC feature domain. The covariances are estimated independently for the static cepstrum (12 MFCC plus energy), delta, and double delta feature streams.
3. Next, the first $p$ eigendirections $[\mathbf{e}_{u1}, \mathbf{e}_{u2}, \ldots, \mathbf{e}_{up}]$ of $\mathbf{B_u}$ are derived using PCA.
4. To measure the relative position of an eigenvector $\mathbf{e}_{uk}$ in this space, each Gaussian mean $\mathbf{x}_i$ in $\Lambda$ is projected onto the eigendirection to obtain an inner product $d_{iuk} = \mathbf{x}_i \cdot \mathbf{e}_{uk}$.
5. Next, $v_{si}$, the variance of $d_{iuk}$ across the speaker set $\mathbf{S}$, and $v_{ui}$, the variance across the utterance set $\mathbf{U}$, are estimated respectively to determine which dimension possesses speaker dependent versus utterance dependent information.

The goal is to compare $V_s$ and $V_u$, the averaged variances of $v_{si}$ and $v_{ui}$ across all component Gaussian projections. If the claim is correct, and the eigendirections of the utterance covariance matrix are more likely to be speaker dependent, one might expect to observe that the former should be higher than the latter.

Part (a) and (b) in Fig. 1 compare the averaged projection variances onto the first and second eigendirections respectively. Clearly, the averaged variance across different speakers with the same utterance, $V_s$, is higher than the averaged variance across different utterances from the same speakers, $V_u$. This observation strongly supports the claim that the first primary eigendirections are more likely to be speaker dependent and are less affected by the phoneme context in utterances. In addition, it is interesting to note that the ratio $V_s/V_u$ are in different ranges for each feature stream (i.e., the ratio is more than five for the static features, above two for the delta stream, while only slightly above one for double delta), as indicated in the lower part of Fig. 1. This again verifies the common knowledge that the static feature stream carries the most significant speaker traits. The experimental results in Fig. 1 also suggest that the feature streams should be treated separately in such eigenspace processing, to assure that we are extracting appropriate speaker information from each stream.

## III. Eigenspace Mapping (EigMap)

For the task of model adaptation, the improved model is achieved by adjusting the baseline model parameters based on adaptation data. From the previous section, it is assumed that the speaker dependent information can be learned from the first primary eigendirections. On the other hand, a well-trained baseline model $\Lambda$ is assumed to maintain a fair model discrimination between component Gaussian means $\{\mathbf{x}_i \mid i = 1, 2, \ldots, N\}$, in the sense of providing a reasonable between-class covariance $\mathbf{B_x}$

$$\mathbf{B_x} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^T - \bar{\mathbf{x}}\bar{\mathbf{x}}^T \tag{1}$$

where every component Gaussian $\mathbf{x}_i$ is treated as a single class, and $\bar{\mathbf{x}} = (1/N)\sum_{i=1}^{N} \mathbf{x}_i$. $\mathbf{B_x}$ can be decomposed as the sum of variations along its eigendirections $\{\mathbf{e}_{x1}, \mathbf{e}_{x2}, \ldots, \mathbf{e}_{xn}\}$

$$\log(\det(\mathbf{B_x})) = \sum_{i=1}^{n} \log \lambda_i \simeq \sum_{i=1}^{p} \log \lambda_i \tag{2}$$

where $n$ is the Gaussian dimension, and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$ are the rank ordered eigenvalues of the symmetric semi-positive definite matrix $\mathbf{B_x}$. The variance of the $i$th principal component is $\lambda_i$, and in a loose sense, this component "accounts for" a proportion $\lambda_i / \sum_{j=1}^{n} \lambda_j$ of the total variances. It is assumed

[1]All the eigenvectors $\mathbf{e}_i$ mentioned in this paper are normalized, i.e., $|\mathbf{e}_i| = 1$. To emphasize the directions pointed by the eigenvectors, the terms "eigenvector" and "eigendirection" will be used interchangeably in this paper.

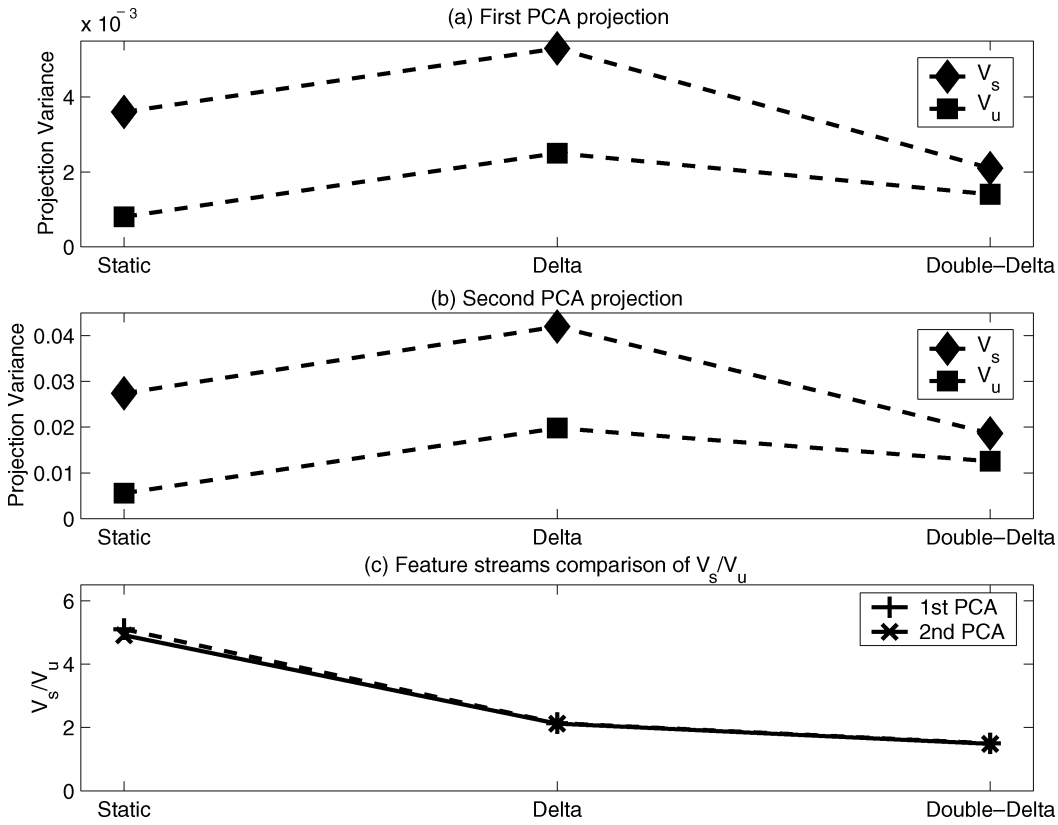



Fig. 1. Comparison of averaged Gaussian mean projection variances across the speaker set $S$ ($V_s$) and across the utterance set $U$ ($V_u$): (a) Variances of projection onto first PCA of speaker's eigenspace, (b) variances of projection onto second PCA of speaker's eigenspace, and (c) the ratio of $V_s/V_u$ for different feature streams.

that $p < n$ is the number of primary eigenvalues that contribute dominant variations, and hence the variations along the corresponding eigendirections $\{\mathbf{e}_{x1}, \mathbf{e}_{x2}, \ldots, \mathbf{e}_{xp}\}$ provide the most significant discrimination power, among any $p$ eigendirections, in the sense of maximizing the Fisher ratio

$$F = \frac{\det(\mathbf{B})}{\det(\mathbf{W})} \tag{3}$$

where $\mathbf{W}$ is the averaged within-class covariance matrix.

### A. EigMap

The basic idea of EigMap is to maintain the between-class variances (i.e, the discrimination power) of the baseline Gaussian means unchanged along the first primary eigendirections in the test speaker's eigenspace. Given the primary eigendirections $\{\mathbf{e}_{y1}, \mathbf{e}_{y2}, \ldots, \mathbf{e}_{yp}\}$ of the test speaker's observation covariance matrix $\mathbf{B}_\mathbf{y}$, the adapted Gaussian means $\{\mathbf{y}_i \mid i = 1, 2, \ldots, N\}$ are expected to satisfy

$$\sum_{j=1}^{n} y_{ij} e_{ymj} = \sum_{j=1}^{n} x_{ij} e_{xmj}, m = 1, \ldots, p. \tag{4}$$

For every component Gaussian $\mathbf{x}_i$ in the model $\Lambda$, all possible adapted means $\mathbf{y}_i$ that satisfies (4) form a $(n-p)$-dimensional subplane $\Omega(\mathbf{x}_i)$ in the acoustic space that is given by

$$\Omega(\mathbf{x}_i) = \left\{ \mathbf{y}_i \mid \sum_{j=1}^{n} y_{ij} e_{ymj} = \sum_{j=1}^{n} x_{ij} e_{xmj}, m = 1, \ldots, p \right\}. \tag{5}$$

In the task of rapid model adaptation where observation data is sparse, aggressive assumptions based on insufficient adaptation data often tend to be unreliable. Alternatively, a more conservative approach is to minimize the shift from the well-trained baseline model parameters, given the constraint of no loss of discrimination power along the first dominant eigendirections in the test speaker eigenspace

$$\mathbf{y}_i = \underset{\mathbf{y}_i \in \Omega(\mathbf{x}_i)}{\operatorname{argmin}} (\mathbf{x}_i - \mathbf{y}_i)^T (\mathbf{x}_i - \mathbf{y}_i). \tag{6}$$

By substituting (5) into (6) and minimizing the objective function using the Lagrange Multiplier method, the adapted mean $\mathbf{y}_i$ can be obtained from $\mathbf{x}_i$ using a linear transformation: $f : \Re^n \rightarrow \Re^n$, $\mathbf{y} = f(\mathbf{x}) = \mathbf{\Theta}\mathbf{x}$, with $\mathbf{\Theta}$ an $n \times n$ nonsingular matrix given by

$$\mathbf{\Theta} = \mathbf{I}_n - \sum_{i=1}^{p} (-1)^{(i-1)} \mathbf{e}_{yi}^T (\mathbf{e}_{yi} - \mathbf{e}_{xi}) \tag{7}$$

where $\mathbf{I}_n$ is an $n \times n$ identity matrix. Considering the orthogonality between eigenvectors (i.e., $\mathbf{e}_{yi} \cdot \mathbf{e}_{yi} = 1$; $\mathbf{e}_{yi} \cdot \mathbf{e}_{yj} = 0$, $\forall i \neq j$), one can show that $\mathbf{\Theta} = \mathbf{E}_\mathbf{y}^{-1} \mathbf{E}_\mathbf{m}$, where

$$\mathbf{E}_\mathbf{y}^{-1} = [\mathbf{e}_{y1}^T, \mathbf{e}_{y2}^T, \ldots, \mathbf{e}_{yn}^T] \tag{8}$$

and

$$\mathbf{E}_\mathbf{m} = [\mathbf{e}_{x1}, \ldots, \mathbf{e}_{xp}, \mathbf{e}_{y(p+1)}, \ldots, \mathbf{e}_{yn}]^T. \tag{9}$$

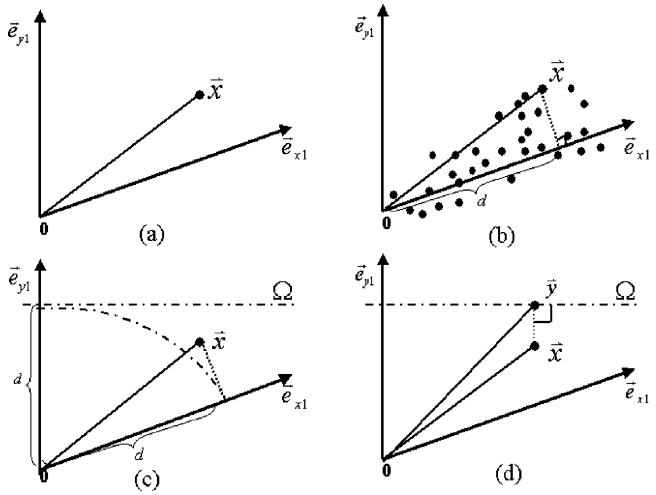



Fig. 2. Illustration of adapting the baseline model Gaussian mean $\mathbf{x}$ to the test speaker specified Gaussian mean $\mathbf{y}$ using the Eigenspace Mapping (EigMap). (a) The first eigenvector of baseline model $\mathbf{e}_{x1}$ and of the test speaker $\mathbf{e}_{y1}$. (b) The projection of $\mathbf{x}$ onto $\mathbf{e}_{x1}$, where $d$ is the most significant discriminative power to separate $\mathbf{x}$ from other baseline Gaussian means. (c) Preserve the first principal component $d$ in the test speaker's eigenspace and obtain a subplane $\Omega$ by rotating $d$ onto the $\mathbf{e}_{y1}$. (d) To be conservative, project $\mathbf{x}$ onto the $\Omega$ to obtain adapted mean $\mathbf{y}$.

After transforming the baseline model mean $\mathbf{x}$ into $\mathbf{y}$ using (7), the discrimination information is assumed to be mostly encapsulated in the first $q$ dimensions, where $p < q < n$, hence the last $n-q$ dimensions of $\mathbf{y}$ can be discarded. In the model space, this is equivalent to setting the last $n-q$ rows of $\mathbf{\Theta}$ to zeros

$$\bar{\mathbf{\Theta}} = [\mathbf{\Theta}_{q\times n}, \mathbf{0}_{(n-q)\times n}]^T \tag{10}$$

and the adapted Gaussian mean $\mathbf{y}$ is achieved through following transformation:

$$\mathbf{y}_{q\times 1} = \bar{\mathbf{\Theta}}_{q\times n}\mathbf{x}_{n\times 1}. \tag{11}$$

The values of $p$ and $q$ are determined based on the distributions of adaptation data and baseline model parameters. More specifically, the selection is affected by the distribution of eigenvalues of $\mathbf{B_x}$ and $\mathbf{B_y}$. The $p$ is selected to ensure that the first $p$ eigenvalues are sufficient to represent overall variations in $\mathbf{B_x}$ and $\mathbf{B_y}$ while avoiding the use of subsequent eigendirections that are unreliably estimated. Similarly, $q$ is determined to achieve the balance of removing the noise in the model parameters after the linear transformation and maintaining sufficient model discrimination.

It is important to conclude from the above equation that the baseline model is not only adapted through the transformation, but also compressed with reduced Gaussian dimensions of the model mean, which further suggests that faster recognition speed can also be achieved using the adapted model due to reduced Gaussian computations.

To illustrate the underlying principles of the EigMap, a graphic example of mapping a baseline Gaussian mean onto the test speaker's eigenspace is further explained in Fig. 2, where a simple 2-dimensional eigenspace is used, and only the first PCA is considered [i.e., let $n = 2$ and $p = 1$ in (5)]. The first eigendirections for the baseline and test speaker, $\mathbf{e}_{x1}$ and $\mathbf{e}_{y1}$ respectively, are first estimated. Given any baseline Gaussian mean $\mathbf{x}$, its first PCA $d$ in the baseline's eigenspace is the most significant discriminative factor to separate this Gaussian from any others in the baseline model. To maintain this discrimination power, this PCA is preserved for the test speaker in his eigenspace, which produces a $n-1$ dimensional subplane $\Omega$ in the new eigenspace. On the other hand, to be conservative, the baseline model mean $\mathbf{x}$ is projected onto the subplane $\Omega$ to minimize the shift from the original model parameters. Thus, the adapted mean $\mathbf{y}$ is obtained as the projection of $\mathbf{x}$ onto $\Omega$.

### B. Multistream Processing Approach

An important issue in the procedure of eigenspace mapping is multi-stream processing. The feature vectors used in many state-of-the-art continuous density HMM based speech recognition systems are composed of three base feature vectors: 12 static MFCCs plus energy, followed by their first-derivatives and second-derivatives. These three base feature vectors are referred to as static, delta, and double delta streams respectively in this paper. Following the observations from Section II, that different feature streams encode different amounts of speaker information, it is suggested that EigMap processing should be performed independently for different streams.

There are a number of facts to explain the need of multi-stream processing. First, the delta and double delta streams possess different spectral-temporal information. Secondly, the static and dynamic streams are of different numerical range, and principal components are scale dependent. Typically, the variations of static feature components is much higher than those of dynamic feature components. Mixing different feature streams together will make the first principal component dominated by the static stream, and the speaker information in the dynamic streams will therefore be overlooked.

To address this issue, each feature stream is treated separately in EigMap, and the nonsingular transformation $\bar{\mathbf{\Theta}}$ is estimated and applied independently for each stream of Gaussian means in the adaptation.

### C. Between-Class Variances Estimation

One of the key points in the EigMap scheme is how to estimate the between-class variances $\mathbf{B_y}$ for the test speaker, and accordingly, the $\mathbf{B_x}$ for the baseline model given the adaptation data.

One approach is based on Viterbi forced alignment. In this approach, the best state sequence of the adaptation data is first found through Viterbi alignment

$$Q(q_1,\ldots,q_t) = \underset{q_1,\ldots,q_t}{\operatorname{argmax}}\ P(q_1,\ldots,q_t;\mathbf{o}_1,\ldots,\mathbf{o}_t \mid \Lambda). \tag{12}$$

$\mathbf{B_y}$ is directly computed from the observed adaptation speech frames $\mathbf{o}_i$ as follows:

$$\mathbf{B_y} = \mathbf{B_o} = \frac{1}{t}\sum_{i=1}^{t}\mathbf{o}_i\mathbf{o}_i^T - \bar{\mathbf{o}}\bar{\mathbf{o}}^T \tag{13}$$

where $t$ is the number of observed *speech* frames, and $\bar{\mathbf{o}} = (1/t)\sum_{i=1}^{t}\mathbf{o}_i$. At the baseline model side, a "simulated" observation from the perspective of baseline models is, given the best state $q_i$ at each time $i$

$$\hat{\mathbf{o}}_i = \sum_{q_i:m} w_{q_i m}\mathbf{x}_{q_i m}, i = 1, 2, \ldots, t \tag{14}$$

where $w_{q_i m}$ is the $m$-th mixture weight of component Gaussian $\mathbf{x}_{q_i m}$ at state $q_i$ with the constraint: $\sum_m w_{q_i m} = 1$. Next, $\mathbf{B}_{\mathbf{x}}$ is estimated from these "simulated" observations as follows:

$$\mathbf{B}_{\mathbf{x}} = \frac{1}{t}\sum_{i=1}^{t}\hat{\mathbf{o}}_i\hat{\mathbf{o}}_i{}^T - \bar{\hat{\mathbf{o}}}\bar{\hat{\mathbf{o}}}^T \tag{15}$$

where $\bar{\hat{\mathbf{o}}} = (1/t)\sum_{i=1}^{t}\hat{\mathbf{o}}_i$.

### D. EigMap and LDA

LDA [8], or more recently, Heteroscedastic Discriminant Analysis (HDA) [19], is used by researchers to improve acoustic model discrimination before Maximum Likelihood (ML) based model training. The goal of LDA is to find the linear transformation $\theta$ in the *feature space* $f : \Re^n \rightarrow \Re^p$, $y = f(x) = \theta x$, where $\theta$ is a $p \times n$ nonsingular matrix with $p < n$. The $\theta$ is obtained to maximize the following objective function:

$$J(\boldsymbol{\theta}) = \frac{\det(\boldsymbol{\theta}B\boldsymbol{\theta}^T)}{det(\boldsymbol{\theta}W\boldsymbol{\theta}^T)}. \tag{16}$$

However, EigMap seeks a linear transformation $\boldsymbol{\Theta}$ in the *model space* for rapid model adaptation, and is applied in the decoding phase. Therefore, no training data is required in EigMap. If the Gaussian variance is not adapted, then the within-class covariance $W$ is unchanged after EigMap transformation. In this sense, the EigMap transformation $\boldsymbol{\Theta}$ can be viewed as a solution that maximizes the same objective function in (16) with the constraint that the baseline model discrimination is preserved along the speaker's first primary eigendirections.

## IV. Extensions to EigMap

Based on the core idea of EigMap, a number of extensions can be derived. This section will introduce some of them.

### A. SMLEM: Structural Maximum Likelihood Eigenspace Mapping

This section outlines the formulation of the proposed SMLEM algorithm, which essentially extends the core EigMap algorithm by imposing a further shift in the model space to maximize the adaptation data likelihood, in addition to the discriminative power obtained by EigMap.

*1) Maximum Likelihood Estimation of Eigenspace Bias:* The objective of the proposed EigMap algorithm is to rapidly impose discriminative information learned from adaptation data onto a baseline model. It is important to note that no adaptation data likelihood information is considered in the EigMap procedure. It is therefore expected that the model can be further adapted after EigMap using an appropriate method to bring the adaptation data likelihood into consideration.

To account for the adaptation data likelihood, the EigMap formulation can be extended by adding a linear bias $\mathbf{b}$ in the test speaker's eigenspace

$$\mathbf{y} = \bar{\boldsymbol{\Theta}}\mathbf{x} + \mathbf{E}_{\mathbf{y}}^{-1}\mathbf{b} \tag{17}$$

where $\mathbf{b}$ is derived in a manner that maximizes the adaptation data likelihood $P(\mathbf{O} \mid \Lambda)$ given the model $\Lambda$. According to the EM algorithm [7], we define an auxiliary function $Q(\Lambda, \hat{\Lambda})$

$$Q(\Lambda, \hat{\Lambda}) = -\frac{1}{2}P(\mathbf{O} \mid \Lambda)\sum_t\sum_s\sum_m \gamma_m^{(s)}(t)\log f(\mathbf{O}(t) \mid \hat{\Lambda}) \tag{18}$$

where $\gamma_m^{(s)}(t)$ is the Gaussian-frame occupancy probability at time $t$ for $m$-th mixture Gaussian component at state $s$, $\Lambda$ is the current model and $\hat{\Lambda}$ is the estimated model. Since we only adapt the Gaussian means, we ignore other model parameters in the auxiliary function, and $\log f(\mathbf{O}(t) \mid \hat{\Lambda})$ from (18) becomes

$$\log f(\mathbf{O}(t) \mid \hat{\Lambda}) = n\log 2\pi + \log|\boldsymbol{\Sigma}_m^{(s)}| + (\mathbf{O}(t) - \mathbf{y}_m^{(s)})^T{\boldsymbol{\Sigma}_m^{(s)}}^{-1}(\mathbf{O}(t) - \mathbf{y}_m^{(s)}) \tag{19}$$

where $n$ is the dimension of the Gaussian model stream means. Substituting $\mathbf{y}_m^{(s)}$ in (19), and using (17), we can rewrite (18) as shown in the equation at the bottom of the page. To maximize $Q(\Lambda, \hat{\Lambda})$ with respect to $\mathbf{b}$, we set

$$\frac{\partial Q(\Lambda, \hat{\Lambda})}{\partial \mathbf{b}} = 0. \tag{20}$$

Therefore, the following accumulation equation can be derived:

$$\sum_t\sum_s\sum_m \gamma_m^{(s)}(t){\boldsymbol{\Sigma}_m^{(s)}}^{-1}(\mathbf{O}(t) - \bar{\boldsymbol{\Theta}}\mathbf{x}_m^{(s)}) = \sum_t\sum_s\sum_m \gamma_m^{(s)}(t){\boldsymbol{\Sigma}_m^{(s)}}^{-1}\mathbf{E}_{\mathbf{y}}^{-1}\hat{\mathbf{b}} \tag{21}$$

where the bias $\hat{\mathbf{b}}$ is tied across sets of states and Gaussian mixtures for accumulation. From (21), it is important to note that only a $(n - q)$-dimensional vector needs to be solved, and only

$$Q(\Lambda, \hat{\Lambda}) = -\frac{1}{2}P(\mathbf{O} \mid \Lambda) \times \sum_t\sum_s\sum_m \left\{\gamma_m^{(s)}(t)\left[n\log 2\pi + \log|\boldsymbol{\Sigma}_m^{(s)}| + (\mathbf{O}(t) - \bar{\boldsymbol{\Theta}}\mathbf{x}_m^{(s)} - \mathbf{E}_{\mathbf{y}}^{-1}\hat{\mathbf{b}})^T{\boldsymbol{\Sigma}_m^{(s)}}^{-1}(\mathbf{O}(t) - \bar{\boldsymbol{\Theta}}\mathbf{x}_m^{(s)} - \mathbf{E}_{\mathbf{y}}^{-1}\hat{\mathbf{b}})\right]\right\}$$

one equation needs to be accumulated. Therefore, the accumulation overhead is limited during online adaptation, and more importantly, robust estimation of the bias can be achieved even with very limited amounts of adaptation data due to the small number of free parameters in the estimation.

*2) Structural Estimation of Maximum Likelihood Bias:* On the other hand, the small number of estimating parameters in (21) suggests that even when only limited amounts of adaptation data are available, a significant number of free parameters can still be reliably estimated. Therefore, the accumulation of both sides of (21) can be tied across a smaller group of Gaussians to achieve more specific bias estimation, rather than across global Gaussians. To automatically determine the degree of tying of the bias estimation, a structural method is employed to hierarchically cluster component Gaussians into a binary tree.

Different methods can be used to generate the clustering tree. In a bottom-up clustering scheme, all of the component Gaussian mean vectors of the well-trained baseline system are clustered into $N$ base classes based on the $K$-means algorithm according to their acoustic similarity using the Euclidian distance measure (see Fig. 3). The average mean of a class with $R$ component Gaussians $\{\mathbf{x}_i \mid i = 1, 2, \ldots, R\}$ is

$$\boldsymbol{\mu} = \frac{1}{\sum_{i=1}^{R} w_i} \sum_{i=1}^{R} w_i \mathbf{x}_i \tag{22}$$

where $w_i$ is the mixture weight associated with component Gaussian $\mathbf{x}_i$. After obtaining base classes, the average of each class is used to represent that class. Next, a binary tree based on those classes is obtained through a greedy search: two closest classes are identified and merged, and then the center is updated for the merged class. This process is repeated until all the classes are clustered into the final root node.

The accumulation of both sides of (21) is first conducted for each base class $\{n \mid n = 1, 2, \ldots, N\}$ by summing all Gaussians that reside in that base class: $\{s, m\} \in n$. Next, according to the hierarchical structure, the base classes that belong to the same space are summed for higher level accumulations until the root of the tree is reached. To determine the adaptation level, we perform a bottom-up traversing of the binary tree and stop at the lowest nodes where $\gamma_n = \sum_{t=1}^{T} \sum_{C^{(n)}} \gamma_m^{(s)}(t)$ is larger than some established threshold $\gamma_s$.

In multi-stream SMLEM, the tree-structured hierarchical spaces are independently constructed for each stream. While processing each stream, only the corresponding feature components in that stream contribute to the distance computation and $K$-means relabeling. Fig. 3 shows how different binary trees are generated using associated feature streams. It is clear that these trees have both different base classes and different tree structures. The dark nodes in the tree are those that meet the stopping criteria while white nodes do not, and the arrows show how lower level nodes map to higher levels to accumulate sufficient adaptation data.

### B. Constrained ML EigMap

For completeness, we consider a constrained ML EigMap scheme, where the eigenspace bias is only allowed to reside on the subplane $\Omega$ (i.e., there is no eigenspace bias allowed along the first primary eigendirections we want to preserve) as follows:

$$\mathbf{y}_i = \underset{\mathbf{y}_i \in \Omega(\mathbf{x}_i)}{\operatorname{argmax}} \; L(\mathbf{O}(t); N(\mathbf{y}_i, \boldsymbol{\Sigma})). \tag{23}$$

Although interesting, experiments with this approach resulted in WER performance that was less successful than SMLEM discussed previously. We therefore, consider it to be a special case of SMLEM.

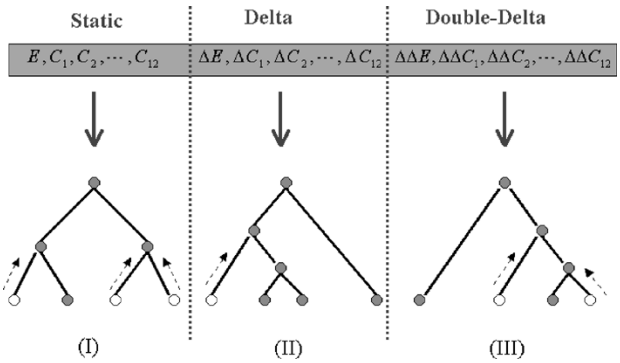



Fig. 3. Portions of the binary tree-structured hierarchical eigenspaces for different streams. The binary trees are constructed independently for each feature stream through centroid bottom-up search from base classes that are generated from K-means clustering. Note that each stream will have different base classes and tree structures.

### C. EigMap With Minimized Mahalanobis Distance

One more extension to EigMap is to employ a minimized Mahalanobis distance. When we bring the component Gaussian covariance $\boldsymbol{\Sigma}_i$ into consideration, it is reasonable to minimize the Mahalanobis distance between $\mathbf{x}_i$ and $\mathbf{y}_i$ as follows:

$$\mathbf{y}_i = \underset{\mathbf{y}_i \in \Omega(\mathbf{x}_i)}{\operatorname{argmin}} \; (\mathbf{y}_i - \mathbf{x}_i)^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{y}_i - \mathbf{x}_i). \tag{24}$$

Similarly, (24) can be minimized using the Lagrange Multiplier approach to obtain the linear transformation[2] $\Psi$. For example, $\Psi$ is given by the following relation, when $p = 1$:

$$\Psi = \mathbf{I_n} - \frac{\mathbf{e}_{y1}^T (\mathbf{e}_{y1} - \mathbf{e}_{x1}) \boldsymbol{\Sigma}_i}{\mathbf{e}_{y1} \boldsymbol{\Sigma}_i \mathbf{e}_{y1}^T}. \tag{25}$$

### D. MLLR+EigMap: EigMap to Extend Maximum Likelihood Models

Another way to use the EigMap method is to combine the algorithm with other conventional techniques where the discrimination information has not been considered. For example, MLLR has been widely used for adaptation by adjusting the baseline model parameters to maximize the adaptation data likelihood. However, adaptation frames are treated independently in MLLR and no discrimination information has been learned from the adaptation data. In this case, applying EigMap to the

[2]Thanks to Juan Yuan, University of Colorado, for helpful discussions.

MLLR adapted models can thus further enhance the discrimination power of the acoustic model. Thus, complementary performance improvement can be expected.

## V. Evaluations

### A. Experimental Setup

The adaptation experiments reported in this paper are all conducted in an *unsupervised* manner, on the WSJ Spoke3 and Spoke4 corpus. The baseline speaker independent acoustic model has 6275 context-dependent tied states, each of which has 16 mixture component Gaussians (i.e., in total 100,400 diagonal mixture component Gaussians exist in the acoustic model). The baseline system uses a feature of 39 dimensions with 13 static cepstral coefficients plus delta and double-delta. The baseline speech recognition system used for experiments is the CMU Sphinx-3.2 recognizer [23]. The language model is a standard 5000-word back-off trigram. In offline multi-stream tree-structured Gaussian clustering, the 100,400 component Gaussians in the SI models are grouped into 300 base classes and hence a binary tree with 599 nodes is used to represent the structural space for each stream.

The Spoke4 corpus is collected from four native speakers of American English with balanced gender. The Spoke3 data consists of nonnative speakers. Each speaker of Spoke3 provides a set of adaptation utterances, and another set of 40 utterances for testing. We select the last six speakers from Spoke3 for our experiments[3] (approximately 3900 words in the test set). For Spoke4, all speakers and all 50 test utterances from group G of each speaker are used in the evaluation (approximately 3300 words in the test set). Since we are primarily interested in rapid adaptation, only a single utterance from each speaker is allowed to be used as adaptation data to improve the baseline model. To account for variability in the small amount of data, and to obtain statistically representable results, three randomly selected adaptation utterances are identically used for each test speaker in adaptation. The adaptation data ranges from 3.7 to 5 s of speech for different utterances and speakers. All experimental results presented are obtained by averaging all open experiments.

The EigMap algorithm was compared with the block diagonal MLLR (BD-MLLR) scheme, since the amount of adaptation data is very limited, and it is shown from experiments that BD-MLLR achieves better performance than conventional MLLR [31] due to the reduced parameters to be estimated. For the same reason, one global regression class is used for BD-MLLR adaptation. For a fair comparison, EigMap also uses *a global eigenspace* for both test speaker and baseline model for the mapping. In our experiments, $n$ is set as 13 for static, delta and double-delta streams. The values of $p$ and $q$ are selected automatically for each stream by comparing the eigenvalues of $\mathbf{B_x}$ and $\mathbf{B_y}$ with some thresholds, which are determined through experiments with some evaluation test speakers. Once they are determined, these thresholds are fixed for all speakers and adaptation data across all the experiments reported in this paper.

[3]The first four speakers demonstrate a relatively high Word Error Rate (WER) that is above 65% for the baseline system. We believe this may be in conflict with our assumption for the EigMap algorithm that the SI models are reasonably well-trained for the test speakers. Therefore, we exclude the first four speakers.

TABLE I
WER (%) of Native Speakers (WSJ Spoke4) With About 4 s of Unsupervised Adaptation Data in Average

| Speaker | 4o6 | 4o7 | 4o8 | 4o9 | Average |
|---|---|---|---|---|---|
| Baseline | 4.4 | 3.8 | 8.0 | 6.2 | 5.6 |
| BD-MLLR | 4.9 | 3.5 | 8.2 | 6.2 | 5.8 |
| EigMap | 4.0 | 3.2 | 7.8 | 6.0 | 5.2 |

### B. Experimental Results

*1) Adaptation for Native Speakers:* Table I shows the performance comparison using Spoke4 corpus with about 4 s of adaptation data. Due to the very limited amount of adaptation data and the close match between the baseline model and test data, BD-MLLR achieves no improvement over baseline on average, while EigMap obtains consistent improvement for all speakers, with an average of 7% relative improvement from baseline. This observation suggests that even if the test data matches the acoustic model well, the discriminative power introduced by EigMap is still able to improve the acoustic model for more accurate classification.

*2) Adaptation for Non-Native Speakers:* The experimental results on Spoke3 corpus are summarized in Table II. In average, about 4.5 s of adaptation data are used. Due to the mismatch between the model and test data, the averaged baseline model WER performance is as high as 20.7%. Table II clearly shows that EigMap consistently improves the recognition for all nonnative speakers. On average, the proposed algorithm effectively enhances the baseline by a relative improvement of 18.4%, while BD-MLLR achieves a 15.9% relative improvement. By applying SMLEM, to maximize adaptation data likelihood after EigMap, the overall relative performance gain is further improved to 21.7%. Moreover, EigMap is highly additive to MLLR by bringing additional discrimination information into the adapted acoustic model that maximizes the adaptation data likelihood. As showm in the last column of Table II, by applying EigMap to the MLLR adapted model, the average WER is reduced to 15.6% and a significant relative improvement of 24.6% is observed in the experiments.

*3) More Results of SMLEM:* The performance of SMLEM is affected by the threshold setting of $\gamma_s$, which needs to be tuned to achieve the balance of reliable estimation and specific adaptation. Fig. 4 shows the adaptation performance of different speakers with threshold $\gamma_s$ varied from 5 to 15. As observed from this figure, different speakers demonstrate varied sensitivities to the value of $\gamma_s$, and on average, the optimal performance is achieved when $\gamma_s = 12$.

To ensure the improvements of SMLEM is not dominantly contributed by structural maximum likelihood bias (SMLB) $\hat{\mathbf{b}}$, the performance of SMLB is compared with EigMap and SMLEM in Table III. Comparing EigMap with SMLEM shows that applying the structural maximum likelihood bias after EigMap provides additional benefits. On the other hand, as

TABLE II
WER (%) OF NON-NATIVE SPEAKERS (WSJ SPOKE3) WITH ABOUT 4.5 s OF UNSUPERVISED ADAPTATION DATA ON AVERAGE

| Spkr | Baseline | BD-MLLR | EigMap | SMLEM | BD-MLLR+EigMap |
|---|---|---|---|---|---|
| 4n5 | 23.5 | 20.2 | 21.4 | 20.8 | 20.2 |
| 4n8 | 16.4 | 13.0 | 13.6 | 12.5 | 13.3 |
| 4n9 | 21.6 | 18.9 | 16.7 | 16.0 | 15.0 |
| 4na | 11.9 | 10.3 | 8.0 | 8.4 | 7.5 |
| 4nb | 32.0 | 28.3 | 25.8 | 26.9 | 25.8 |
| 4nc | 18.7 | 13.6 | 15.9 | 12.4 | 11.6 |
| Avg | 20.7 | 17.4 | 16.9 | 16.2 | 15.6 |
| Rel. Imp | — | 15.9% | 18.4% | 21.7% | 24.6% |

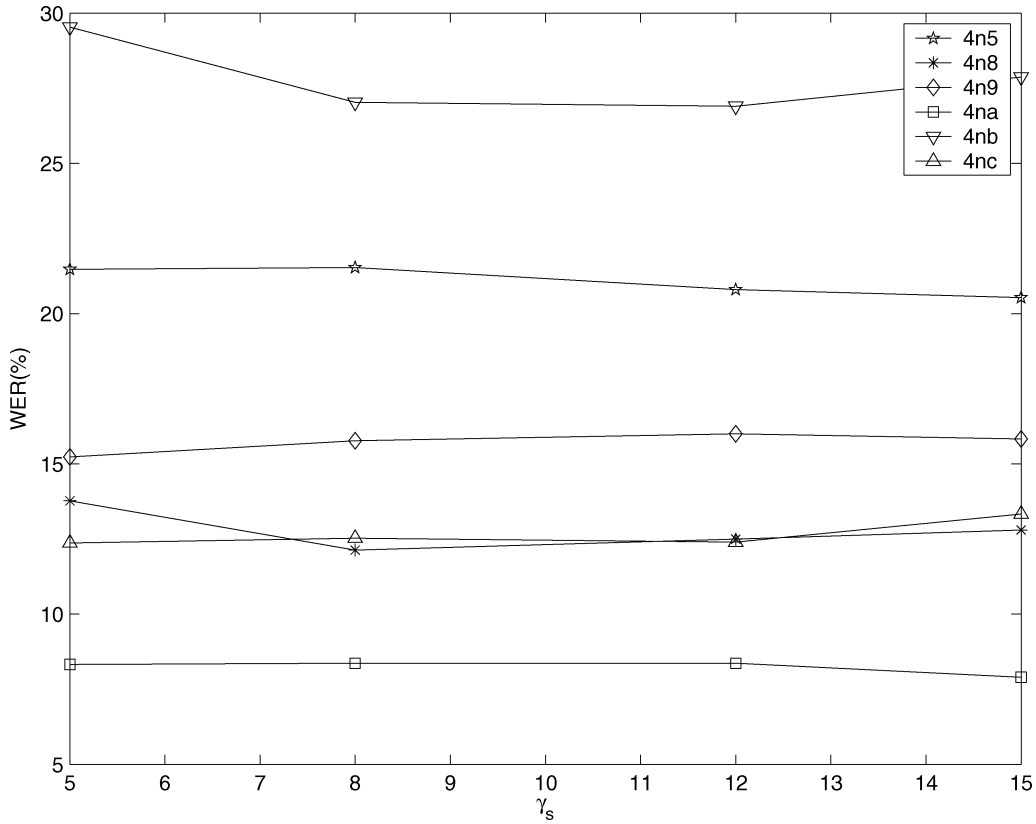



Fig. 4. Performance of SMLEM as a function of $\gamma$ for different speakers.

illustrated in this table, though SMLEM achieves consistent and significant improvement over the baseline recognizer for all speakers, using only the structural maximum likelihood bias obtains a slight performance gain on average, while providing worse performance for some speakers. Therefore, while the bias can help, it is clear that EigMap is the key contributor of the SMLEM algorithm.

TABLE III
A COMPARISON OF WER PERFORMANCE OF SMLEM, EIGMAP AND STRUCTURAL MAXIMUM LIKELIHOOD BIAS (SMLB) (WITH $\gamma_s = 12$)

| Spkr | 4n5 | 4n8 | 4n9 | 4na | 4nb | 4nc | Avg |
|---|---|---|---|---|---|---|---|
| SMLB | 21.9 | 13.7 | 21.4 | 11.8 | 33.4 | 17.6 | 20.0 |
| EigMap | 21.4 | 13.6 | 16.7 | 8.0 | 25.8 | 15.9 | 16.9 |
| SMLEM | 20.8 | 12.5 | 16.0 | 8.4 | 26.9 | 12.4 | 16.2 |

## VI. DISCUSSIONS AND FUTURE WORK

The family of EigMap algorithms rapidly constructs discriminative acoustic models based on the strong correlations existing over adaptation utterances, which, we believe, encode significant speaker information. The algorithm might also be applied to background noise adaptation, in cases where the correlations over adaptation frames carry a range of background acoustic characteristics. Experiments will be conducted for adapting different noise conditions.

It is observed from experiments that the bias vector, estimated using maximum likelihood criterion such as that was derived in SMLEM, is not linearly additive to EigMap. Therefore, it is interesting to explore other methods for bias estimation in the future. Among others, discriminative biases based on minimal

classification error (MCE) and maximum mutual information estimation (MMIE) are of special interest to us.

The issue of how to optimally determine the number of model dimensions that should be discarded after EigMap transformation is not clear yet. In our experiments, the value of $q$ is determined empirically. It is observed that nonnative speakers typically require a larger $q$ than native speakers. This may be explained by fact that EigMap transformations for nonnative speakers are more dramatic than those for native speakers, due to more significant mismatch with the baseline model which existed for nonnative speakers. In the future, a more theoretical method that optimally determines the values of $q$ for different tasks will be expected.

## VII. Conclusions

This paper has introduced a novel family of algorithms based on Eigenspace Mapping (EigMap) for rapid speaker adaptation. EigMap constructs a discriminative acoustic model for the test speaker by preserving the discrimination power of the baseline model in the test speaker's eigenspace with constraints. Unsupervised adaptation experiments show that EigMap can effectively improve the baseline model with very limited amounts of adaptation data. In addition, the EigMap algorithm can be extended in many ways. The algorithm entitled structural maximum likelihood Eigenspace mapping (SMLEM) achieves extra benefits by incorporating a linear bias to maximize adaptation data likelihood. Moreover, EigMap is able to provide additional performance gain to existing methods such as MLLR. By combining MLLR and EigMap, a significant 24.6% relative improvement is achieved using only about 4.5 s of adaptation data.

## Acknowledgment

The authors thank the anonymous reviewers and the associate editor Dr. S. Narayanan for their helpful and constructive comments during the review process.

## References

[1] S. M. Ahadi and P. C. Woodland, "Combined bayesian and predictive techniques for rapid speaker adaptation of continuous density hidden Markov models," *Comput. Speech Lang.*, vol. 11, pp. 187–206, 1997.

[2] M. Blomberg, "Within-utterance correlation for speech recognition," in *Proc. Eurospeech*, Budapest, Hungary, 1999, pp. 2479–2482.

[3] W. Byrne and A. Gunawardana, "Discounted likelihood linear regression for rapid speaker adaptation," in *Proc. Eurospeech*, Budapest, Hungary, 1999, pp. 203–206.

[4] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian information criterion," in *Proc. Broadcast News Transcription Understanding Workshop*, Feb. 1998, pp. 127–132.

[5] C. Chesta, O. Siohan, and C. H. Lee, "Maximum a posterior linear regression for hidden Markov model adaptation," in *Proc. Eurospeech*, Budapest, Hungary, 1999, pp. 203–206.

[6] W. Chou, "Maximum a posterior linear regression with elliptically symmetric matrix priors," in *Proc. Eurospeech*, Budapest, Hungary, 1999, pp. 1–4.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood estimation from incomplete data via the EM algorithm," *J. R. Statist. Soc.*, vol. B39, pp. 1–38, 1977.

[8] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. New York: Academic, 1990.

[9] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 1, pp. 52–59, 1986.

[10] M. J. F. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 7, no. 3, 1999.

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

[12] N. C. Giri, *Multivariate Statistical Analysis*. New York: Marcel Dekker, 1995, ch. 10.

[13] R. Gopinath, "Maximum likelihood modeling with Gaussian distributions for classification," in *Proc. ICASSP*, Seattle, WA, 1998.

[14] J. H. L. Hansen, P. Angkititrakul, J. Plucienkowski, S. Gallant, U. Yapanel, B. Pellom, W. Ward, and R. Cole, "CU-move: Analysis & corpus development for interactive in-vehicle speech systems," in *Proc. Eurospeech*, Aalborg, Denmark, Sep. 2001.

[15] Z. Hu, "Understanding and Adapting to Speaker Variability using Correlation-Based Principal Component Analysis," Ph.D. dissertation, Oregon Graduate Institute of Science and Technology, Portland, 1999.

[16] S. Johnson, "Speaker Tracking," M.S. thesis, Univ. Cambridge, Cambridge, U.K., 1997.

[17] R. Kuhn, P. Nguyen, J. C. Junqua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field, and M. Contolini, "EigenVoices for speaker adaptation," in *Proc. ICSLP*, Sydney, Australia, Nov. 1998.

[18] R. Kuhn, F. Perronnin, P. Nguyen, J.-C. Junqua, and L. Rigazio, "Very fast adaptation with a compact context-dependent eigenvoice model," in *Proc. ICASSP*, Salt Lake City, UT, May 2001.

[19] N. Kumar, "Investigation of Silicon-Auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition," Ph.D. dissertation, Johns Hopkins Univ., Baltimore, MD, 1997.

[20] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech Lang.*, vol. 9, pp. 171–185, 1995.

[21] L. R. Neumeyer, A. Sankar, and V. V. Digalakis, "A comparative study of speaker adaptation techniques," in *Proc. Eurospeech*, 1995, pp. 1127–1130.

[22] *Invariance and Variability in Speech Processes*, J. S. Perkell and D. H. Klatt, Eds., Lawrence Erlbaum, Princeton, NJ, 1986.

[23] P. Placeway, S. Chen, M. Eskenazi, U. Jain, V. Parikh, B. Raj, M. Ravishankar, R. Rosenfeld, K. Seymore, M. Siegler, R. Stern, and E. Thayer, "The 1996 hub-4 sphinx-3 system," in *Proc. DARPA Speech Recognition Workshop*, 1997.

[24] G. Saon, M. Padmanabhan, R. Gopinath, and S. Chen, "Maximum likelihood discriminant feature spaces," in *Proc. ICASSP*, Istanbul, Turkey, Jun. 2000.

[25] K. Shinoda and C. H. Lee, "Structural MAP speaker adaptation using hierarchical priors," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, Santa Barbara, CA, 1997, pp. 381–388.

[26] O. Siohan, T. A. Myrvoll, and 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.

[27] R. J. Westwood, "Speaker Adaptation using Eigenvoices," M.Phil. thesis, Univ. Cambridge, Cambridge, U.K., 1999.

[28] K. Wong and B. Mak, "Rapid speaker adaptation using MLLR and subspace regression classes," in *Proc. Eurospeech*, Aalborg, Denmark, Sep. 2001.

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

[30] B. Zhou and J. H. L. Hansen, "A novel algorithm for rapid speaker adaptation based on structural maximum likelihood eigenspace mapping," in *Proc. Eurospeech*, vol. 2, Aalborg, Denmark, Sep. 2001, pp. 1215–1218.

[31] ——, "Improved structural maximum likelihood eigenspace mapping for speaker adaptation," in *Proc. ICSLP'2002*, Denver, CO, 2002, pp. 1433–14367.

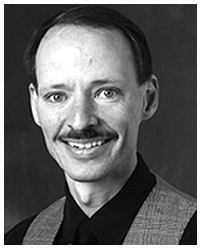
[32] ——, "SpeechFind: An experimental on-line spoken document retrieval system for historical audio archives," in *Proc. ICSLP'2002*, vol. 3, Denver, CO, 2002, pp. 1969–1972.

[33] ——, "Unsupervised audio stream segmentation and clustering via the Bayesian information criterion," in *Proc. ICSLP'2000*, Beijing, China, Oct. 2000, pp. 714–717.

**Bowen Zhou** (M'03) received the B.S. degree from the University of Science and Technology of China in 1996, the M.S. degree from the Chinese Academy of Sciences in 1999, and the Ph.D. degree from the University of Colorado at Boulder in 2003, all in electrical engineering.

He was a Research Assistant with the Robust Speech Processing Group-Center for Spoken Language Research (RSPG-CSLR) during his graduate studies at the University of Colorado. He was an invited speaker for IBM User Interface Technology Student Symposium in November 2002. He joined the Department of Human Language Technologies at IBM Thomas J. Watson Research Center, Yorktown Heights, NY, in March 2003. His current research interest includes automatic speech recognition, natural language understanding, speech-to-speech machine translation, spoken information retrieval, and machine learning.

Dr. Zhou has served as a reviewer for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING.



**John H. L. Hansen** (S'81–M'82–SM'93) was born in Plainfield, NJ. He received the Ph.D. and M.S. degrees in electrical engineering from Georgia Institute of Technology, Atlanta, in 1988 and 1983, respectively, and the B.S.E.E. degree with highest honors from Rutgers University, New Brunswick, NJ, in 1982.

He is Professor with the Departments of Speech, Language, and Hearing Sciences, and Electrical and Computer Engineering, at the University of Colorado at Boulder. He also serves as Department Chairman of Speech, Language and Hearing Sciences. In 1988, he established the Robust Speech Processing Laboratory (RSPL), which is now the Robust Speech Processing Group at the Center for Spoken Language Research (CSLR), which he co-founded and serves as Associate Director. He was a faculty member with the Departments of Electrical and Biomedical Engineering, Duke University, Durham, NC, for 11 years before joining the University of Colorado in 1999. In the fall of 2005, he joined the University of Texas at Dallas, Richardson, as Professor and Department Chair of Electrical Engineering, where he will establish the Center for Robust Speech Systems (CRSS). His research interests span the areas of digital speech processing, analysis and modeling of speech and speaker traits, speech enhancement and feature estimation in noise, robust speech recognition with current emphasis on robust recognition and training methods for spoken document retrieval in noise, accent, stress/emotion, and Lombard effect, and speech feature enhancement in hands-free environments for human-computer interaction. He has served as a technical consultant to industry and the U.S. Government, including ATT Bell Labs, IBM, Sparta, Signalscape, BAE Systems, ASEC, VeriVoice, and DoD in the areas of voice communications, wireless telephony, robust speech recognition, and forensic speech/speaker analysis. He is the author of more than 172 journal and conference papers in the field of speech processing and communications, coauthor of the textbook *Discrete-Time Processing of Speech Signals* (New York: IEEE Press, 2000) and lead author of the report "The Impact of Speech Under 'Stress' on Military Speech Technology," (NATO RTO-TR-10, 2000, ISBN: 92-837-1027-4), and co-editor of *DSP for In-Vehicle and Mobile Systems* (Norwell, MA: Kluwer, 2004). He also organized and served as General Chair for ICSLP-2002: International Conference on Spoken Language Processing, Denver, CO, Oct. 2002.

Dr. Hansen was an invited tutorial speaker for IEEE ICASSP-95 and the 1995 ESCA-NATO Speech Under Stress Research Workshop, Lisbon, Portugal, the 2004 IMI-COE Symposium (Nagoya, Japan). He has served as Technical Advisor to U.S. Delegate for NATO (IST/TG-01: Research Study Group on Speech Processing, 1996–1999), Chairman for the IEEE Comm. and Signal Proc. Society of N.C. (1992–1994), Advisor for the Duke University IEEE Student Branch (1990–1997), Tutorials Chair for IEEE ICASSP-96, Associate Editor for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (1992–1998), Associate Editor for IEEE SIGNAL PROCESSING LETTERS (1998–2000), Member of the Editorial Board for *IEEE Signal Processing Magazine* (2001–2003). He has also served as guest editor of the Oct. 1994 special issue on Robust Speech Recognition for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING. He was the recipient of a Whitaker Foundation Biomedical Research Award, an NSF Research Initiation Award, and has been named a Lilly Foundation Teaching Fellow for "Contributions to the Advancement of Engineering Education."