Acoustic Factor Analysis for Robust Speaker Verification

Taufiq Hasan, *Student Member, IEEE*, and John H. L. Hansen, *Fellow, IEEE*

***Abstract*—Factor analysis based channel mismatch compensation methods for speaker recognition are based on the assumption that speaker/utterance dependent Gaussian Mixture Model (GMM) mean super-vectors can be constrained to reside in a lower dimensional subspace. This approach does not consider the fact that conventional acoustic feature vectors also reside in a lower dimensional manifold of the feature space, when feature covariance matrices contain close to zero eigenvalues. In this study, based on observations of the covariance structure of acoustic features, we propose a *factor analysis* modeling scheme in the acoustic feature space instead of the super-vector space and derive a mixture dependent feature transformation. We demonstrate how this single linear transformation performs feature dimensionality reduction, de-correlation, normalization and enhancement, at once. The proposed transformation is shown to be closely related to signal subspace based speech enhancement schemes. In contrast to traditional front-end mixture dependent feature transformations, where feature alignment is performed using the highest scoring mixture, the proposed transformation is integrated within the speaker recognition system using a probabilistic feature alignment technique, which nullifies the need for regenerating the features/retraining the Universal Background Model (UBM). Incorporating the proposed method with a state-of-the-art i-vector and Gaussian Probabilistic Linear Discriminant Analysis (PLDA) framework, we perform evaluations on National Institute of Science and Technology (NIST) Speaker Recognition Evaluation (SRE) 2010 core telephone and microphone tasks. The experimental results demonstrate the superiority of the proposed scheme compared to both full-covariance and diagonal covariance UBM based systems. Simple equal-weight fusion of baseline and proposed systems also yield significant performance gains.**

***Index Terms*—Acoustic feature enhancement, factor analysis, probabilistic principal component analysis, speaker verification.**

I. Introduction

MISMATCH between training and test conditions represent one of the most challenging problems facing speaker recognition researchers today. There can be considerable sources of mismatch present including: transmission channel differences [1], [2], handset variability [3], background noise [4], session variability due to physical stress [5], vocal effort such as whisper [11], [12], Lombard effect [13], non-stationarity environment [10], spontaneity of speech, but to name a few. Various compensation strategies have been proposed in the past to reduce unwanted variability between training and test utterances, while retaining the speaker identity information. The current trend in state-of-the-art speaker recognition systems is to model the acoustic features with a GMM-UBM, use utterance dependent adapted GMM [7] mean super-vectors [14] as the features representing the speech segments, and model the super-vectors using various *latent factor analysis* techniques [1], [6], [15]. In [16], the aim was to identify the lower dimensional speaker and channel dependent subspaces, termed Eigenvoice [15], [17] and Eigenchannel [1], in the super-vector domain. In [1], an alternative was considered where speaker and channel variabilities were jointly modeled. The recently proposed i-vector [6] scheme utilizes a factor analysis framework [15], [18] to perform dimensionality reduction on the super-vectors while retaining important speaker discriminant information. This lower dimensional i-vector representation enables the development of full Bayesian techniques [19], [20], using a single model to represent the speaker and channel variability.

One limitation of the conventional GMM super-vector domain representation and subsequent factor analysis modeling is that, it does not take into account the fact that the original acoustic features contain redundancy. In general, the speech short-time spectrum is known to be representable in a lower dimensional subspace, which motivates a separate class of speech enhancement methods known as *signal subspace* approaches [21], [22]. Linear correlation among the speech spectral components are quite high, which justifies the success of these methods. This phenomenon is also valid for popular acoustic features, such as Mel-frequency Cepstral Coefficients (MFCC) [23], [24], even though these features are processed through Discrete Cosine Transform (DCT) for de-correlation before use in training or test.

A. Motivation

To motivate the proposed work, we first demonstrate that the conventional acoustic features can be constrained to reside in a lower dimensional subspace. For this purpose, we train a 1024 mixture full covariance GMM UBM using 60 dimensional MFCC features on a large background speech data set.[1] For a

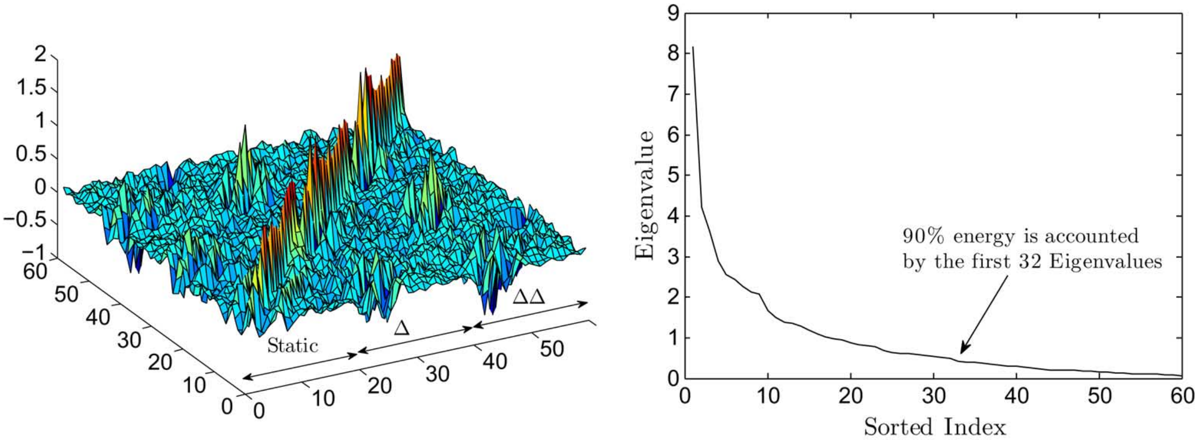
Manuscript received May 05, 2012; revised August 19, 2012; accepted October 09, 2012. Date of publication October 23, 2012; date of current version January 23, 2013. This work was supported by AFRL under Contract FA8750-12-1-0188 (approved for public release, distribution unlimited), and in part by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. H. L. Hansen. The authors note that preliminary investigations on Acoustic Factor Analysis (AFA) was presented in [8] and [9]. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Rodrigo Guido.

The authors are with the Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas, Richardson, TX 75252 USA (e-mail: taufiq.hasan@utdallas.edu; john.hansen@utdallas.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2012.2226161

[1] More details on feature extraction and development data are given in Sections V-A and V-B, respectively



(a)

(b)

Fig. 1. Analysis of full covariance matrices of a UBM trained using 60-dimensional MFCC feature ($20\ \mathrm{static} + \Delta + \Delta\Delta$). (a) A 3-D surface plot of the covariance matrix showing high values in the diagonal and significant off-diagonal values indicating correlation among different feature coefficients. (b) Sorted eigenvalues of the same covariance matrix demonstrating that most of the energy is accounted for by in the first few dimensions.

typical mixture of this UBM, the covariance matrix and distribution of its eigenvalues is shown in Fig. 1. From Fig. 1(a) it is clear that the full covariance matrix, which shows strong diagonal terms, has significant non-zero off-diagonal elements, indicating that the feature coefficients are not fully uncorrelated. Fig. 1(b) shows the sorted eigenvalues of the same covariance matrix revealing that most of it's energy is accounted for by the first few dimensions only. This shows that the acoustic feature space is actually lower dimensional and features can thus be further compacted or enhanced by using a *factor analysis* model. Also, it is known that the first few directions obtained by the Eigen-decomposition of acoustic feature covariance matrices are mostly speaker dependent (e.g. see Zhou and Hansen [25] for a quantitative analysis), while other directions are more phoneme dependent. In this study, considering these noted observations on the acoustic features, we aim at investigating a factor analysis scheme on acoustic features for speaker recognition. We would like to name this method *acoustic factor analysis*.

B. Limitations of Conventional Factor Analysis

Before proceeding with the formulation of the factor analysis scheme in the front-end features, we first defend the argument that the traditional factor analysis schemes do not take full advantage of the acoustic feature covariances. In a standard i-vector system, the GMM super-vectors are dimensionality reduced by a total factor analysis model, which is based on the idea that utterance super-vectors lie in a lower dimensional subspace. Let $\mathbf{m}_s$ denote a GMM super-vector extracted from an utterance s, and $\mathbf{x}_n$ would denote the acoustic features. For a randomly chosen utterance s, it is generally assumed that $\mathbf{m}_s$ is normally distributed with mean $\mathbf{m}_0$ and covariance matrix $\mathbf{B}$ [15]. Here, $\mathbf{m}_0$ denotes the speaker independent mean vector obtained by concatenating the UBM mean vectors $\mathbf{m}_{0[g]}$. Let the UBM covariance matrices be $\mathbf{\Sigma}_g$, where g denotes the mixture number. The main motivation of both Eigenvoice and total variability modeling, is that the super-covariance matrix $\mathbf{B}$ contains zero eigenvalues and thus some dimensions of $\mathbf{m}_s$ can be disregarded. For the g-th Gaussian mixture, the utterance dependent mean vector $\mathbf{m}_{s[g]}$ is estimated from the posterior mean of the acoustic features that belong to s, that is $\mathbf{x}_n \in s$. This is a deterministic parameter. However, for a randomly selected utterance s, the sub-vectors $\mathbf{m}_{s[g]}$ are normally distributed random vectors having covariance matrix $\mathbf{B}_{[g]}$, which is the g-th sub-matrix of the super-covariance matrix $\mathbf{B}$. Clearly, the matrices $\mathbf{B}_{[g]}$ are not related to the feature covariance matrices $\mathbf{\Sigma}_g$, since the former represents the covariance of the mean sub-vectors $\mathbf{m}_{s[g]}$ obtained from different utterances, while the latter represents the covariance of the acoustic features $\mathbf{x}_n$ which is independent of the utterance.[2] Thus, assuming that the matrix $\mathbf{B}$ contains zero eigenvalues is not equivalent to assuming the same for the $\mathbf{\Sigma}_g$ matrices. Though this reasoning is based on full covariance UBM models, similar arguments can be made for a diagonal covariance based system.

C. Feature Dimensionality Reduction

Given that the conventional acoustic features reside in a lower dimensional subspace, it is important now to ask the question how we can use this knowledge to effectively extract utterance level features. Since speaker dependent information is contained in the leading eigen-directions of the acoustic features [25], using all the feature coefficients for modeling channel degraded data will result in retaining some nuisance components along with speaker dependent information in the GMM super-vectors and i-vectors. Therefore, we propose a dimensionality reduction transformation of the acoustic features for each GMM mixture that emphasizes the speaker dependent information in the leading eigenvectors of the corresponding mixture covariance matrix, while suppressing some unwanted channel components. In this manner, the GMM super-vectors will be "enhanced" in the sense that they will be more speaker discriminative, while the subsequently extracted i-vectors will also inherit this quality.

[2] Utterance dependent covariance matrices can also be extracted through MAP adaptation. However, we assume that each utterance GMM shares the common UBM covariance and weights.

Dimensionality reduction of the acoustic features for de-correlation/enhancement is not a new concept. There are many techniques found in the literature that perform this task, including DCT, Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Heteroscedastic LDA (HLDA), to name but a few [16], [26], [27]. The main goal for this process has been to be able to model the features using diagonal covariance matrix GMM/HMMs for speech/speaker recognition. These techniques can be classified in mainly two groups by their mode of operation including: 1) the signal processing domain, and 2) the model domain. In the first scenario, some transformation (supervised/unsupervised) is used at the signal/acoustic feature level in order to achieve improved energy compaction. The most common technique is the application of the DCT for the log-filterbank energies [23] popularized by the MFCC representation. PCA can also be used [26] by learning the principal directions from the Eigen-decomposition of the covariance matrix trained on the utterance data itself. In general, this class of processing only depends on the speech data under consideration and does not use any outside knowledge. In the second scenario, raw acoustic features (e.g., filter-bank energies) are initially used to train a large model, which is then used to derive the feature transformations. One such technique used in speaker recognition is HLDA [16], where first a GMM-UBM is trained on the raw acoustic features. Each mixture is then assumed to represent a separate class, and HLDA transformation is trained so that discrimination between these classes is maximized. In a similar fashion, PCA projections can also be used in each GMM mixture as a transformation [28]. In these methods, after the initial training phase, the acoustic features are aligned to the mixture component providing the highest posterior probability and the corresponding transformation is used for dimensionality reduction.

Both the signal processing domain and model domain feature dimensionality reduction techniques previously used in essence have one common property: they re-generate the acoustic features after a dimensionality reduction. This means, the sub-sequent procedures for the speaker recognition system require that we begin training from these newly extracted features. Model domain dimensionality reduction has an extra inconvenience of mixture-alignment. Speech features are known to be highly intertwined and overlapped in the vector space for different acoustic conditions and generally do not form meaningful clusters [29]. Thus, using the top posterior probability for aligning a feature vector to a single mixture may not be appropriate. To demonstrate this, we select MFCC feature vectors $\mathbf{x}_n$ from 10 development utterances that were used in the UBM training, and for each feature vector, we find the highest posterior probability among the 1024 mixtures of the UBM, $\max_g p(g|\mathbf{x}_n)$. A histogram of these top mixture probabilities is shown in Fig. 2, which clearly demonstrate that only a few frames are unquestionably aligned to a specific Gaussian mixture (indicated by the high peak near $\max_g p(g|\mathbf{x}_n) = 1$). In actuality, a majority of the feature vectors are aligned with more than one mixture, resulting in a top mixture probability in the region of 0.3 $\sim$ 0.8. Thus, using the top scoring mixture for hard alignment of feature vectors to a specific mixture can introduce inaccuracies and should be avoided if possible.

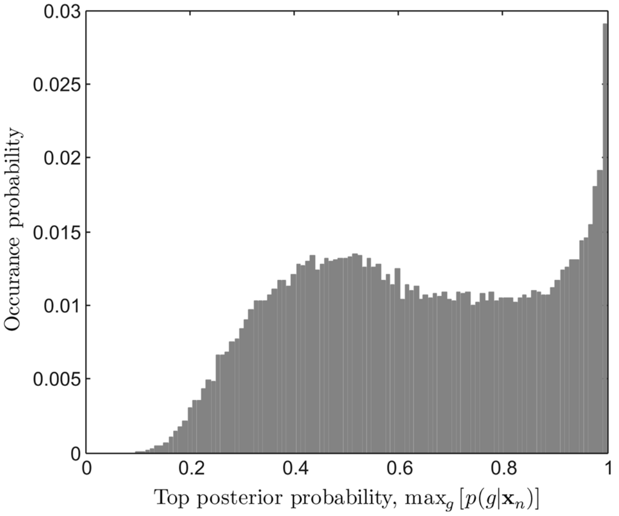


Fig. 2. Distribution of top posterior probabilities $p(g|\mathbf{x}_n)$ obtained from a subset of development data.

D. Further Implications of the Proposed Method

Historically, feature extraction, dimensionality reduction, enhancement and normalization has always been thought of as a separate process from acoustic modeling. In this study, we propose a new modeling scheme of the acoustic features that bridges the gap between these two processing domains through integrated feature dimensionality reduction and modeling. We demonstrate that the proposed method not only performs dimensionality reduction, it also removes the need for hard feature clustering to a specific mixture, and does not require retraining of the UBM from the new features, thereby incorporating a built-in feature normalization and enhancement scheme. All this is achieved using a single linear transformation derived from a pre-trained full covariance matrix UBM and applying this in a probabilistic fashion to the mixture dependent Baum-Welch statistics.

E. Outline

This paper is organized as follows. In Section II, we formulate the proposed Acoustic Factor Analysis (AFA) scheme and derive the mixture-dependent transformation matrices. Section III describes the various properties of the AFA transformation, including normalization and enhancement. In Section IV, we describe how the proposed scheme can be integrated within an i-vector system followed by our system description in Section V. Experimental results are presented in Section VI, and finally, Section VII concludes the study.

II. Acoustic Factor Analysis

In this section, we describe the proposed factor analysis model of acoustic features, discuss its formulation and mixture-wise application for dimensionality reduction.

A. Formulation

Let $\mathcal{X} = \{\mathbf{x}_n | n = 1 \cdots N\}$ be the collection of all acoustic feature vectors from the development set obtained from a large corpus of many speakers' recordings in diverse environment/ channel conditions. Using a factor analysis model, the $d \times 1$ dimensional feature vector $\mathbf{x}$ can be represented by,

$$\mathbf{x} = \mathbf{W}\mathbf{y} + \mu + \epsilon. \tag{1}$$

Here, $\mathbf{W}$ is a $d \times q$ low rank factor loading matrix that represents $q < d$ bases spanning the subspace with important variability in the feature space, and $\boldsymbol{\mu}$ is the $d \times 1$ mean vector of $\mathbf{x}$. We denote the latent variable vector or latent factors $\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, as *acoustic factors*, which is of dimension $q \times 1$. We assume that the remaining noise component $\epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma}^2\mathbf{I})$ is isotropic, and therefore the model is equivalent to PPCA [18]. In this model, the feature vectors are also normally distributed such that, $\mathbf{x} \sim \mathcal{N}(\mu, \sigma^2\mathbf{I} + \mathbf{W}\mathbf{W}^T)$.

The advantage of this model is that the *acoustic factors* $\mathbf{y}$, defining the weights of the factor loadings, explains the correlation between the feature coefficients $\mathbf{x}$, which we believe are more speaker dependent [25], while the noise component ϵ incorporates the residual variance of the data. It should be emphasized that even though we denote the term ϵ as "noise", when used with cepstral features this term actually represents convolutional channel distortion [30]. A mixture of these models [18] can be used to incorporate the variations caused by different phonemes uttered by multiple speakers in distinct noisy/channel degraded conditions, given by,

$$p(\mathbf{x}) = \sum_g w_g p(\mathbf{x}|g) \tag{2}$$

where for the g-th mixture,

$$p(\mathbf{x}|g) = \mathcal{N}\left(\boldsymbol{\mu}_g, \sigma_g^2\mathbf{I} + \mathbf{W}_g\mathbf{W}_g^T\right). \tag{3}$$

Here, μ_g, w_g, $\mathbf{W}_g$ and σ_g^2 represent the mean vector, mixture weight, factor loading matrix, and noise variance for the g-th AFA model, respectively.

B. Mixture Dependent Transformation

One advantage of using the mixture of PPCA for acoustic factor analysis is that, its parameters can be conveniently extracted from a GMM trained using the Expectation-Maximization (EM) algorithm [18]. Thus, we utilize a full covariance UBM to derive the AFA model parameters. The proposed feature transformation and dimensionality reduction procedure is presented below:

1) Universal Background Model: A full covariance UBM model Λ_0, is trained on the development dataset $\mathcal{X} = \{\mathbf{x}_n | n = 1 \cdots N\}$, given by,

$$p(\mathbf{x}|\Lambda_0) = \sum_{g=1}^{M} w_g \mathcal{N}(\mu_g, \boldsymbol{\Sigma}_g) \tag{4}$$

where w_g represents the mixture weights, M is the total number of mixtures, μ_g are the mean vectors and $\boldsymbol{\Sigma}_g$ are the full covariance matrices. The mean and weight parameters of the UBM will be identical to the mixture model of (2).

2) Noise Subspace Selection: We require to set the value of q, which defines the number of principal axes we would like to select. In other words, we assume the lower $d - q$ dimensions of the features will actually represent the noise subspace [21]. Using this value of q, we find the noise variance for the g-th mixture as,

$$\sigma_g^2 = \frac{1}{d-q} \sum_{i=q+1}^{d} \lambda_{g,i} \tag{5}$$

where $\lambda_{g,q+1} \cdots \lambda_{g,d}$ are the smallest eigenvalues of the covariance matrix $\boldsymbol{\Sigma}_g$. Thus, σ_g^2 is essentially the average variance lost per discarded dimension. It may be noted that the model allows the use of different values of q for each mixture. This has been investigated in [9] and also, we elaborate this issue in greater detail in Section IV.

3) Compute the Factor Loading Matrix: The maximum likelihood estimation of the factor loading matrix $\mathbf{W}_g$ of the g-th mixture of the AFA model in (2) is given by,

$$\mathbf{W}_g = \mathbf{U}_{\mathbf{q}_g}\left(\boldsymbol{\Lambda}_{\mathbf{q}_g} - \sigma_g^2\mathbf{I}\right)^{\frac{1}{2}}\mathbf{R}_g \tag{6}$$

where $\mathbf{U}_{\mathbf{q}_g}$ is a $d \times q$ matrix whose columns are the q leading eigenvectors of $\boldsymbol{\Sigma}_g$, $\boldsymbol{\Lambda}_{\mathbf{q}_g}$ is a diagonal matrix containing the corresponding q eigenvalues, and $\mathbf{R}_g$ is a $q \times q$ arbitrary orthogonal rotation matrix. In this work, we set $\mathbf{R}_g = \mathbf{I}$.

4) Feature Transformation: The posterior mean of the *acoustic factors* $\mathbf{y}_n$ can be used as the transformed and dimensionality reduced version of $\mathbf{x}_n$ for the g-th component of the AFA model. This can be shown to be

$$E\{\mathbf{y}_n|\mathbf{x}_n, g\} = \langle\mathbf{y}_n|\mathbf{x}_n, g\rangle = \mathbf{A}_g^T(\mathbf{x}_n - \boldsymbol{\mu}_g) \triangleq \mathbf{z}_{n,g} \tag{7}$$

where

$$\mathbf{A}_g = \mathbf{W}_g\mathbf{M}_g^{-T} \text{ and} \tag{8}$$

$$\mathbf{M}_g = \sigma_g^2\mathbf{I} + \mathbf{W}_g^T\mathbf{W}_g. \tag{9}$$

We term the matrix $\mathbf{A}_g$ as the g-th *AFA transform*. In this operation, we are essentially replacing the original feature vectors $\mathbf{x}_n$ by the mixture dependent transformed acoustic feature $\mathbf{z}_{n,g}$. Each feature vector $\mathbf{x}_n$ can be transformed by $\mathbf{A}_g$, corresponding to the mixture component it is aligned with and a new set of features can then be obtained. However, as noted earlier, we will not regenerate the acoustic features and instead use a probabilistic soft-alignment in our system. This is described in Section V where we discuss the integration of AFA within an i-vector system.

III. Properties of the AFA Transform

In this section, we discuss the general properties and advantages of the proposed acoustic feature model, the resulting transformation and the transformed features.

A. Probability Distribution of the Transformed Features

Here, we derive the probability distribution of the transformed acoustic features and show how AFA performs feature de-correlation. Let $\mathbf{z}_{n,g} = \langle \mathbf{y}_n | \mathbf{x}_n, g\rangle$ indicate the AFA transformed feature vector for the g-th mixture. We have the following mean vector of $\mathbf{z}_{n,g}$,

$$\begin{aligned}\mu_{\mathbf{z}_g} &= E\left\{\langle \mathbf{y}_n | \mathbf{x}_n, g\rangle\right\} \\ &= E\left\{\mathbf{A}_g^T(\mathbf{x}_n - \mu_g)\right\} = \mathbf{0}\end{aligned} \tag{10}$$

and its corresponding covariance matrix,

$$\begin{aligned}\boldsymbol{\Sigma}_{\mathbf{z_g}} &= E\left\{\mathbf{z}_{n,g}\mathbf{z}_{n,g}^T\right\} - \mu_{\mathbf{z}_g}\mu_{\mathbf{z}_g}^T \\ &= \mathbf{A}_g^T E\left\{(\mathbf{x}_n - \mu_g)(\mathbf{x}_n - \mu_g)^T\right\}\mathbf{A}_g \\ &= \mathbf{A}_g^T \boldsymbol{\Sigma}_g \mathbf{A}_g.\end{aligned} \tag{11}$$

For further simplification, we first substitute the value of $\mathbf{W}_g$ from (6) into (9) and use $\mathbf{R}_g = \mathbf{I}$ to obtain,

$$\begin{aligned}\mathbf{M}_g &= \sigma_g^2\mathbf{I} + \mathbf{W}_g^T\mathbf{W}_g \\ &= \sigma_g^2\mathbf{I} + \left(\boldsymbol{\Lambda}_{\mathbf{q}_g} - \sigma_g^2\mathbf{I}\right)^{\frac{T}{2}} \mathbf{U}_{\mathbf{q}_g}^T \mathbf{U}_{\mathbf{q}_g} \left(\boldsymbol{\Lambda}_{\mathbf{q}_g} - \sigma_g^2\mathbf{I}\right)^{\frac{1}{2}} \\ &= \boldsymbol{\Lambda}_{\mathbf{q}_g}.\end{aligned} \tag{12}$$

Next, substituting the values of $\mathbf{W}_g$ and $\mathbf{M}_g$ from (6) and (12) into (8) we have,

$$\mathbf{A}_g^T = \boldsymbol{\Lambda}_{\mathbf{q}_g}^{-1}\left(\boldsymbol{\Lambda}_{\mathbf{q}_g} - \sigma_g^2\mathbf{I}\right)^{\frac{T}{2}}\mathbf{U}_{\mathbf{q}_g}^T. \tag{13}$$

Using this expression of $\mathbf{A}_g^T$ in (11) we obtain,

$$\begin{aligned}\boldsymbol{\Sigma}_{\mathbf{z_g}} &= \boldsymbol{\Lambda}_{\mathbf{q}_g}^{-1}\left(\boldsymbol{\Lambda}_{\mathbf{q}_g} - \sigma_g^2\mathbf{I}\right)^{\frac{T}{2}}\boldsymbol{\Lambda}_{\mathbf{q}_g}\left(\boldsymbol{\Lambda}_{\mathbf{q}_g} - \sigma_g^2\mathbf{I}\right)^{\frac{1}{2}}\boldsymbol{\Lambda}_{\mathbf{q}_g}^{-T} \\ &= \left(\boldsymbol{\Lambda}_{\mathbf{q}_g} - \sigma_g^2\mathbf{I}\right)\boldsymbol{\Lambda}_{\mathbf{q}_g}^{-T} \\ &= \mathbf{I} - \sigma_g^2\boldsymbol{\Lambda}_{\mathbf{q}_g}^{-1}.\end{aligned} \tag{14}$$

Here, we utilize the expression $\mathbf{U}_{\mathbf{q}_g}^T\boldsymbol{\Sigma}_g\mathbf{U}_{\mathbf{q}_g} = \boldsymbol{\Lambda}_{\mathbf{q}_g}$ and take advantage of the diagonal system. Thus, we show that for a given mixture alignment g, the posterior mean of the *acoustic factors*, or the transformed feature vectors $\mathbf{z}_{n,g}$ follow a Gaussian distribution with zero mean and a diagonal covariance matrix given by $\mathbf{I} - \sigma_g^2\boldsymbol{\Lambda}_{\mathbf{q}_g}^{-1}$. Thus, the AFA transformation de-correlates the mean normalized acoustic features in each mixture.

B. Acoustic Feature Enhancement

In the g-th mixture, the AFA transformation matrix $\mathbf{A}_g^T$ expression given in (13) can be expressed as:

$$\begin{aligned}\mathbf{A}_g^T &= \boldsymbol{\Lambda}_{\mathbf{q}_g}^{-1}\left(\boldsymbol{\Lambda}_{\mathbf{q}_g} - \sigma_g^2\mathbf{I}\right)^{\frac{T}{2}}\mathbf{U}_{\mathbf{q}_g}^T \\ &= \boldsymbol{\Lambda}_{\mathbf{q}_g}^{-\frac{1}{2}}\mathbf{G}_g\mathbf{U}_{\mathbf{q}_g}^T\end{aligned} \tag{15}$$

where we introduced a diagonal gain matrix given by:

$$\mathbf{G}_g = \boldsymbol{\Lambda}_{\mathbf{q}_g}^{-\frac{1}{2}}\left(\boldsymbol{\Lambda}_{\mathbf{q}_g} - \sigma_g^2\mathbf{I}\right)^{\frac{T}{2}}. \tag{16}$$

The i-th diagonal entry of $\mathbf{G}_g$ is given by,

$$G_g(i) = \sqrt{\frac{\left(\lambda_{g,i} - \sigma_g^2\right)}{\lambda_{g,i}}}. \tag{17}$$

Keeping aside the term $\boldsymbol{\Lambda}_{\mathbf{q}_g}^{-(1/2)}$ in (15), we observe that the transformation operation performed by $\mathbf{A}_g^T$ in (7) first computes the inner product of the mean normalized acoustic feature with the q principal eigenvectors of $\boldsymbol{\Sigma}_\mathbf{g}$, then for each i-th eigenvector direction applies the gain function defined by $G_g(i)$. The second term in (17) can be identified as a square-root Wiener gain function [31]. This becomes clearer if we define the classic speech enhancement terminology *a priori* SNR ξ as [21], [32],

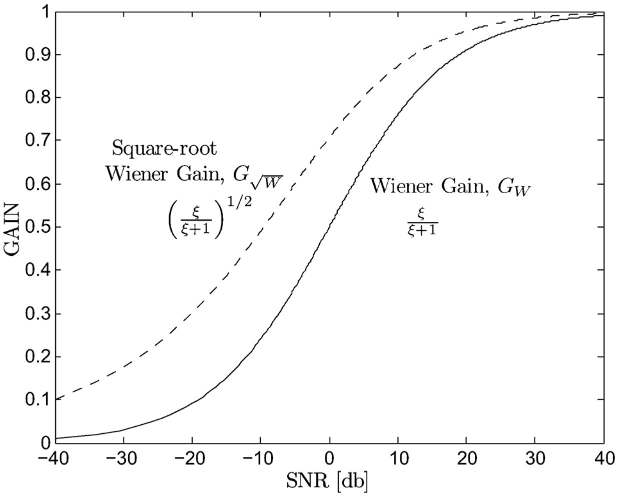


Fig. 3. Input SNR [dB] (ξ) vs. Wiener gains. Wiener gain and square-root Wiener gain are shown with a solid ($-$) and dashed ($-\ -$) line, respectively.

$$\xi = \frac{\lambda_{g,i} - \sigma_g^2}{\sigma_g^2} \tag{18}$$

and use this to express the gain equations. The Wiener gain G_w and the square-root Wiener gain $G_{\sqrt{\mathrm{w}}}$ are given by:

$$G_\mathrm{W} = \frac{\xi}{\xi + 1} \text{ and } G_{\sqrt{\mathrm{W}}} = \left(\frac{\xi}{\xi + 1}\right)^{\frac{1}{2}}. \tag{19}$$

Wiener and square-root Wiener gain functions are plotted against ξ in Fig. 3. As discussed in [31] page 179, Sec 6.6.3), in case of additive noise, the square-root Wiener filter is applied, when instead of the magnitude spectrum, the power spectrum of the filtered signal and the clean signal are desired to be equal. The operation performed by the AFA transformation in (15) can be interpreted as a gain function operating on a transformed space defined by the i-th eigenvector to obtain a clean eigenvalue $\lambda_{g,i} - \sigma_g^2$ from the noisy eigenvalue $\lambda_{g,i}$ [9]. Since the eigenvalues can be interpreted as a power spectrum obtained from the principal components [33], it is understandable why $G_{\sqrt{\mathrm{w}}}$ arises in this scenario instead of G_w. Due to this square-root operation on the gain function, the square-root Wiener obviously shows lower attenuation characteristics compared to the standard Wiener filter, as depicted in Fig. 3. It may be noted that conventional factor analysis techniques in the super-vector space can also be interpreted using similar Wiener like gain functions as discussed in [34].

In the signal subspace speech enhancement method [21], a similar gain function is obtained by starting from the same model in (1), except for the standard normal assumption on the latent factors $\mathbf{y}$. In that work, the term $\mathbf{W}\mathbf{y} + \mu \triangleq \mathbf{a}$ in (1) was interpreted as the "clean signal", $\mathbf{x}$ as the noisy signal and ϵ as the additive noise. The goal was to find an estimate of the clean signal $\hat{\mathbf{a}}$ by finding the posterior mean of $\mathbf{a}$ given the noisy signal $\mathbf{x}$ and noise variance. However, in the AFA scheme, the goal is to estimate the posterior mean of the latent factors $\mathbf{y}$ for an "enhanced" and more compact version of the "noisy" (channel degraded) acoustic features $\mathbf{x}$ [18]. This difference between the two approaches yield two different optimization criteria and their resulting gain functions.

Another contrast between the speech enhancement schemes and AFA transformation is the interpretation of noise. In conventional speech enhancement methods the noise statistics are estimated from silence regions between speech segments [35], and thus for the signal subspace based method, noise variance σ_g^2 is assumed to be known in the model (1). In our case, the noise we are attempting to remove or compensate for is actually an additive distortion in the cepstral domain, which will not exist in the silence regions. In addition, even if the silence segments were modeled in the UBM, it is very unlikely that the mixture components modeling the silences would be useful in determining the noise level in other components. Thus, even though the AFA dimension q is related to the noise variance, we resort to set the value of q arbitrarily and compute the corresponding noise variance for each mixture using (5).

C. Acoustic Feature Variance Normalization

Going back to (15), the term $\mathbf{\Lambda}_{\mathbf{q}_g}^{-(1/2)}$ normalizes the variance of the acoustic feature stream in the i-th eigen-direction, since $\lambda_{g,i}$ is the expected feature variance along this direction [36]. This means, the AFA transformation assumes that the features that are closely aligned with the g-th mixture, originates from the same random process, and performs this normalization in addition to the enhancement mentioned in the previous section. This process is interestingly similar to the cepstral variance normalization frequently performed in the front-end. However, feature domain processing considers the temporal movement of the features in performing these normalizations assuming that the feature streams are independent, while AFA groups the features together in a mixture irrespective of their time location and performs the normalization in an orthogonal axis derived from the corresponding mixture covariance matrix. It would be interesting to see how AFA systems perform if the feature domain normalizations are removed from the front-end. Recent studies [37] show that in the full-covariance UBM based i-vector scheme, a very basic scale normalization technique outperforms Cepstral Mean and Variance Normalization (CMVN) and feature Gaussianization [38]. This may be due to the uncorrelated assumption among feature coefficients inherently assumed while applying these normalization schemes. We have yet to perform experiments on comparative feature normalization schemes using AFA and suggest this as a future work.

IV. AFA Integrated i-Vector System

In this section, we describe how the proposed method can be incorporated into a conventional i-vector system [6].

A. UBM and AFA Model Training

First, a full covariance UBM model, Λ_0 given by (4), is trained on the development data vectors. Next, the AFA dimension q is set, which defines the number of principal axes to retain from each mixture component. Using the value of q, we find the noise variance for the g-th mixture using (5). The factor loading matrix $\mathbf{W}_g$ and transformation matrix $\mathbf{A}_g$ are then calculated using (6) and (8), respectively. After applying the transformation as in (7), the posterior means of the *acoustic factors* $\mathbf{z}_{n,g} = \langle \mathbf{y}_n | \mathbf{x}_n, g \rangle$ are used as mixture dependent transformed acoustic features.

B. UBM Transformation

Following the discussion from Section III-A, and using (10) and (14), the AFA transformation would require a new transformed UBM $\hat{\Lambda}_0$ that models $\mathbf{z}_{n,g}$ instead of $\mathbf{x}_n$, such that,

$$p(\mathbf{z}|\hat{\Lambda}_0) = \sum_{i=1}^{M} w_g \mathcal{N}(\mathbf{0}, \hat{\mathbf{\Sigma}}_g) \tag{20}$$

where $\hat{\mathbf{\Sigma}}_g = \mathbf{I} - \sigma_g^2 \mathbf{\Lambda}_{\mathbf{q}_g}^{-1} = \mathbf{\Sigma}_{\mathbf{z_g}}$. This UBM is not an actual acoustic model used to calculate the posterior probabilities or other statistics. Eq. (20) simply indicates how the UBM parameters should be modified/replaced compared to the original UBM Λ_0 given in (4). This transformation only affects the hyper-parameter estimation.

C. Baum-Welch Statistics Estimation

In this step, the zero and first order Baum-Welch statistics are extracted from each feature vector with respect to the UBM. Using the AFA transformed features, extraction of the statistics can be accomplished as follows. The probabilistic alignment of feature $\mathbf{x}_n$ with the g-th mixture is given by:

$$\gamma_g(n) = p(g|\mathbf{x}_n) = \frac{p(\mathbf{x}_n|g) w_g}{p(\mathbf{x}_n)}. \tag{21}$$

For an utterance s, the zero order statistics is extracted as:

$$N_s(g) = \sum_{n \in s} \gamma_g(n), \tag{22}$$

which follows the standard procedure [6], [15]. Conventionally, the first order statistics are extracted as:

$$\mathbf{F}_s(g) = \sum_{n \in s} \gamma_g(n) \mathbf{x}_n.$$

However, with the present AFA transform, the first order statistics $\hat{\mathbf{F}}_s(g)$ is extracted using the transformed features in the corresponding mixtures instead of the original features.

$$\begin{aligned} \hat{\mathbf{F}}_s(g) &= \sum_{n \in s} \gamma_g(n) \mathbf{z}_{n,g} = \sum_{n \in s} \gamma_g(n) \mathbf{A}_g^T (\mathbf{x}_n - \mu_g) \\ &= \mathbf{A}_g^T \left[\mathbf{F}_s(g) - N_s(g) \mu_g \right] = \mathbf{A}_g^T \bar{\mathbf{F}}_s(g) \end{aligned}$$

where $\bar{\mathbf{F}}_s(g)$ is the centralized first order statistics [20]. This transformation of statistics is somewhat similar to the approach in [39], where it was done to normalize the UBM parameters to zero means and identity covariance matrices. However, in [39] the goal was to simplify the i-vector system algorithm, theoretically preserving the procedure results with added computational benefits; whereas in this work, we are performing feature transformation and dimensionality reduction for possible improvement of the i-vector system performance.

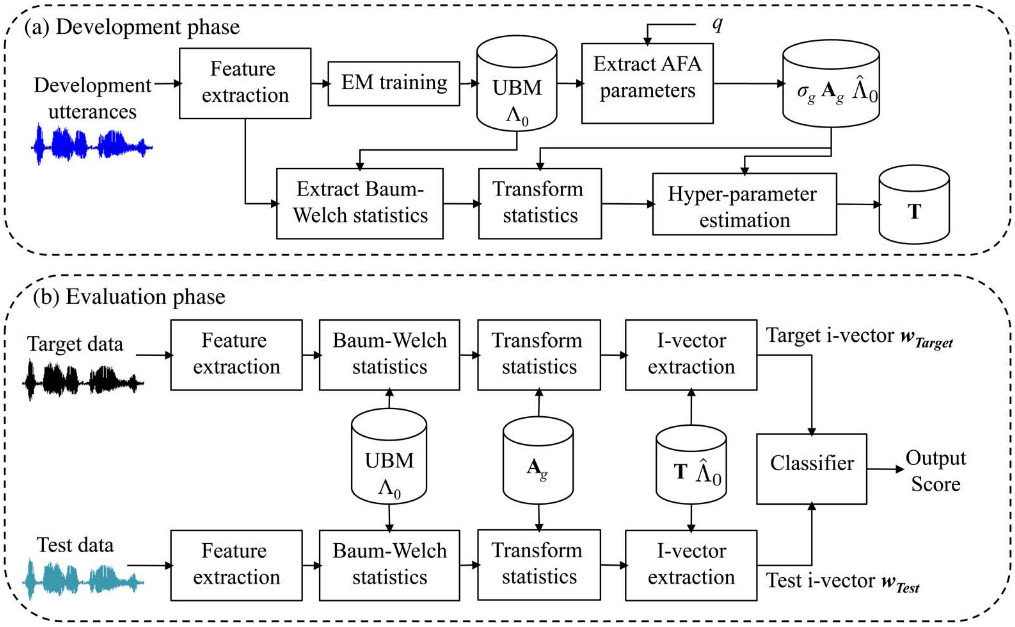


Fig. 4. A block diagram of the proposed AFA integrated i-vector system. The system is shown in two phases: (a) development and (b) evaluation. In the evaluation phase, only i-vector extraction procedure is depicted assuming an arbitrary classifier. For details on the PLDA classifier used, refer to Section V-D.

D. *Hyper-Parameter Estimation*

Training of the Total Variability (TV) matrix $\mathbf{T}$ for the i-vector system follows a very similar procedure as discussed in [6]. In this system, an utterance dependent super-vector s is expressed as,

$$\mathbf{m}_s = \mathbf{m}_0 + \mathbf{T}\mathbf{w}_s \tag{23}$$

where the Md dimensional vector $\mathbf{m}_0$ denotes the speaker independent mean super-vector (i.e., concatenation of the UBM means $\mu_g = \mathbf{m}_{0[g]}$), $\mathbf{T}$ is an $Md \times R$ low rank matrix $(R < Md)$ whose columns span the total variability space, and $\mathbf{w}_s$ is a normal distributed random vector of size R, known as the *total factors*. The posterior mean vector of $\mathbf{w}_s$ given an utterance data is know as an i-vector.

1) Initialization: Depending on the AFA parameter q, the size of the matrix $\mathbf{T}$ needs to be defined. In the AFA based i-vector system, the super-vector dimension becomes $K = Mq$ instead of Md. Thus, the $\mathbf{T}$ matrix size needs to be set to $K \times R$, and randomly initialized. We define a parameter, super-vector compression (SVC) ratio $\alpha = K/Md = q/d$, measuring compaction obtained through AFA transformation.

2) EM Iterations: For each utterance $s \in \mathcal{S}$, $R \times R$ precision matrix $\mathbf{L}_s$ and $R \times 1$ vector $\mathbf{B}_s$ are estimated as [40]:

$$\mathbf{L}_s = \mathbf{I} + \sum_{g=1}^{M} N_s(g)\mathbf{T}_{[g]}^T \hat{\boldsymbol{\Sigma}}_g^{-1} \mathbf{T}_{[g]} \text{ and} \tag{24}$$

$$\mathbf{B}_s = \sum_{g=1}^{M} N_s(g)\mathbf{T}_{[g]}^T \hat{\boldsymbol{\Sigma}}_g^{-1} \hat{\mathbf{F}}_s(g) \tag{25}$$

respectively, where $\mathbf{T}_{[g]}$ is the g-th sub-matrix of $\mathbf{T}$ of dimension $q \times R$, $\hat{\boldsymbol{\Sigma}}_g$ is the $q \times q$ AFA transformed UBM covariance matrix. The total factors for the utterance s are estimated as:

$$\mathbf{w}_s = \mathbf{L}_s^{-1}\mathbf{B}_s. \tag{26}$$

In each iteration, the g-th block of the $\mathbf{T}$ matrix is updated using the following equation:

$$\mathbf{T}_{[g]} = \sum_{s \in \mathcal{S}} \hat{\mathbf{F}}_s(g)\mathbf{w}_s^T \left[\sum_{s \in \mathcal{S}} \left(\mathbf{L}_s^{-1} + \mathbf{w}_s\mathbf{w}_s^T \right) N_s(g) \right]^{-1} \tag{27}$$

which follows the same procedure as a conventional i-vector system [6], [40].

V. System Description

We perform our experiments on the male trials of the NIST SRE 2010 telephone and microphone conditions (core conditions 1–5, extended trials). A standard i-vector system [6] with a Gaussian Probabilistic Linear Discriminant Analysis (PLDA) [41] back-end is used for evaluation. Specific blocks of the baseline system implementation and details of the proposed scheme are described below. An overall block diagram of the proposed system is included in Fig. 4.

A. *Feature Extraction*

In order to remove the silence frames, an independent Hungarian phoneme recognizer [42] combined with an energy based voice activity detection (VAD) scheme is used. A 60-dimensional feature vector $(19 \text{ MFCC} + \text{Energy} + \Delta + \Delta\Delta)$ is extracted using a 25 ms analysis window with subsequent 10 ms shifts, and then Gaussianized utilizing a 3-s sliding window [38].

B. *UBM Training*

Gender dependent UBMs having full and diagonal-covariance matrices with 1024 mixtures are trained on telephone utterances selected from the Switchboard II Phase 2 and 3, Switchboard Cellular Part 1 and 2, and the NIST 2004, 2005, 2006 SRE enrollment data. We use the HTK toolkit for training with 15 iterations per mixture split. The UBM full covariance values were floored to 10^{-5} using the $-v$ option in HTK HERest toolkit [43].

C. Total Variability Modeling

For the TV matrix training, the UBM training dataset is utilized. Five iterations are used for the EM training. We use 400 total factors (i.e., our i-vector size was 400). All i-vectors are first whitened and then length normalized using radial Gaussianization [41].

D. Session Variability Compensation and Scoring

A Gaussian probabilistic linear discriminant analysis (PLDA) model with a full-covariance noise process is used for session variability compensation and scoring [41]. In this generative model, an R dimensional i-vector $\mathbf{w}_s$ extracted from a speech utterance s is expressed as:

$$\mathbf{w}_s = \mathbf{w}_0 + \mathbf{\Phi}\boldsymbol{\beta} + \mathbf{n} \tag{28}$$

where $\mathbf{w}_0$ is an $R \times 1$ speaker independent mean vector, $\mathbf{\Phi}$ is the $R \times N_{EV}$ rectangular matrix representing a basis for the speaker-specific subspace/eigenvoices, β is an $N_{EV} \times 1$ latent vector having a standard normal distribution, and $\mathbf{n}$ is the $R \times 1$ random vector representing the full covariance residual noise. The only model parameter here is the number of eigenvoices N_{EV}, that is the number of columns in the matrix $\mathbf{\Phi}$. I-vectors extracted from the UBM training dataset and additional microphone data selected from SRE 2004 and 2005, are utilized to train this PLDA model.

VI. Evaluation Results

A. Performance Evaluation of AFA Systems

In this experiment, in four different runs we retain q =36, 42 and 48 coefficients from the $d = 60$ dimensional features using the proposed AFA method. We vary the number of eigenvoices N_{EV} in the PLDA model from 50 to 400 in 50 step increments. The performance metrics used are %Equal Error Rate (EER) and minimum Detection Cost Functions (DCF) defined in NIST SRE 2008 [44] (DCF_{old}) and NIST SRE 2010 [45] (DCF_{new}). The results are summarized in the plot shown in Fig. 5 and a subset of these results, organized by performance metrics, is also shown in Table I. The proposed systems are compared against our baseline full-covariance and diagonal covariance UBM based i-vector systems, referred to as "Baseline full-cov" and "Baseline diag-cov", respectively.

From Fig. 5(a)–(c), we observe that for $q = 42$ and for almost all values of N_{EV}, the proposed AFA system performs better than both baseline systems with respect to all three performance metrics. For $q = 48$, the AFA system is superior to the baselines in DCF_{new}, but very close with respect to the other performance measures. For $q = 42$ and $N_{EV} = 200$, we achieve the best EER performance of 1.73% which is 11.28% lower relative to the corresponding Baseline full-cov system EER. The results in Fig. 5 and Table I indicate that the proposed AFA transformation of the acoustic features are successfully able to reduce nuisance directions in the feature space, producing i-vectors with better speaker discriminating ability. We also note that our full-covariance baseline system and AFA based systems perform significantly better than the diagonal-covariance system.

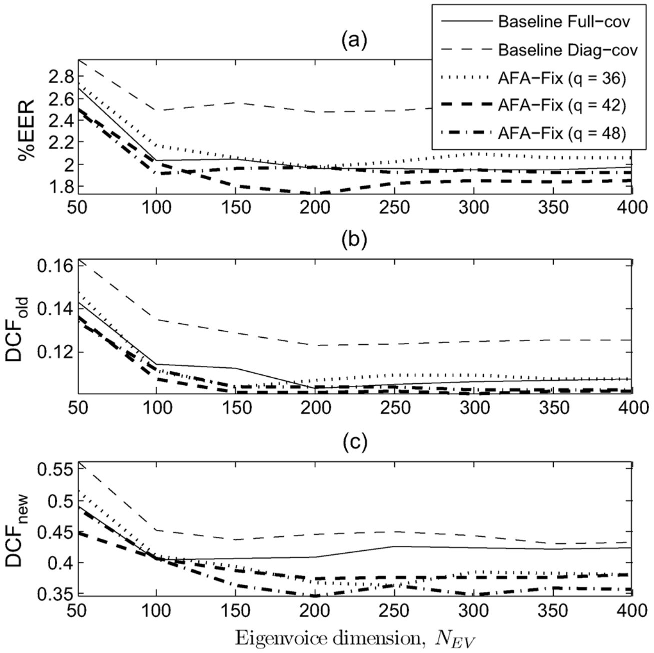


Fig. 5. Performance comparison between proposed AFA and baseline i-vector system with respect to (a) %EER, (b) DCF_{old} and (c) DCF_{new} for different eigenvoice size N_{EV} of the PLDA model. Evaluation is performed on NIST SRE 2010 core condition-5 using the extended trials.

TABLE I
PERFORMANCE COMPARISON BETWEEN BASELINE I-VECTOR AND PROPOSED AFA SYSTEMS FOR DIFFERENT VALUES OF N_{EV} AND q. EVALUATION PERFORMED ON NIST SRE 2010 CORE CONDITION-5 EXTENDED TRIALS

PLDA	**Baseline system**		**AFA system**		
N_{EV}	**full-cov**	**diag-cov**	*q = 36*	*q = 42*	*q = 48*
		% Equal Error Rate (EER)			
100	2.0274	2.4896	2.1706	2.0115	1.9117
150	2.0396	2.5548	2.0632	1.7944	1.9554
200	1.9535	2.4750	1.9756	**1.7322**	1.9706
250	1.9551	2.4854	2.0216	1.8233	1.9183
300	1.9467	2.5343	2.0980	1.8497	1.9422
		DCF_{old} (NIST SRE 2008)			
100	0.1145	0.1348	0.1110	0.1073	0.1120
150	0.1124	0.1285	0.1039	**0.1014**	0.1037
200	0.1033	0.1229	0.1071	0.1015	0.1040
250	0.1053	0.1237	0.1090	0.1017	0.1035
300	0.1061	0.1247	0.1092	0.1009	0.1024
		DCF_{new} (NIST SRE 2010)			
100	0.4050	0.4526	0.4103	0.4056	0.4057
150	0.4056	0.4365	0.3928	0.3869	0.3635
200	0.4093	0.4444	0.3678	0.3732	**0.3468**
250	0.4251	0.4501	0.3639	0.3765	0.3620
300	0.4234	0.4428	0.3844	0.3750	0.3473

B. Effect of Different AFA Dimension

In Fig. 6, AFA system performance is compared with the Baseline full-cov system for different values of q, keeping the parameter N_{EV} fixed at 150. Here we use q = 24, 30, 36, 42, 48 and 54, yielding super-vector compression (SVC) ratios of

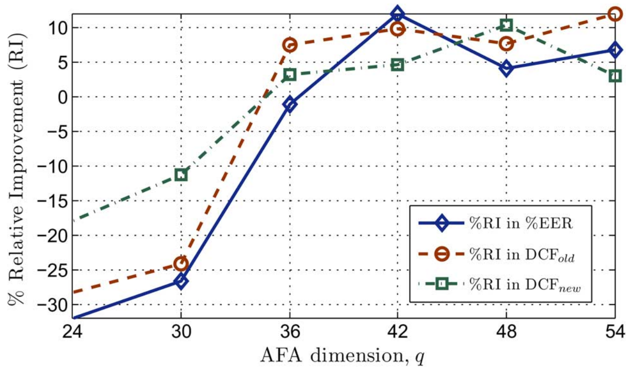


Fig. 6. Performance comparison of AFA system for different values of q with respect to % Relative Improvements (RI) in %EER, DCF_{old} and DCF_{new} compared to the corresponding baseline system performance metric. Evaluation is performed on NIST SRE 2010 core condition-5 using the extended trials. The figure clearly reveals that the system performance drastically degrades as the value of q is reduced.

α =0.4, 0.5, 0.6, 0.7, 0.8 and 0.9, respectively. From this figure, we observe that the system performance is quite sensitive to the q parameter of the proposed AFA method, though performance improvement is achieved compared to the baseline system in almost all cases. If the value of q is too low, some speaker dependent information is removed by the AFA transform and system performance degrades. Values of q close to feature dimension d yields performances similar to the baseline system. We observe consistent improvements in the system performance by setting q close to 42 $\sim$ 48 for the AFA systems. In this region, relative improvement values of all three performance metrics are in the range of 4 $\sim$ 12%. We believe the fluctuation of performance is due to the fact that a different value of q is suitable for each mixture component. Thus, methods of selecting the optimal AFA dimension can be a viable future work, especially since the model allows different values of q for each mixture.

C. Effect of UBM Variance Flooring

It is known that full covariance UBM based speaker recognition systems can be very sensitive to small values in the UBM covariance matrices [20]. In [20], a variance flooring algorithm [46] was used to tackle with this issue. As mentioned in Section V-B, we performed UBM variance flooring by limiting the minimum value of a covariance matrix component to 10^{-5} using HTK. We refer to this flooring method as "vFloor-1". To observe the effect of an alternate variance flooring on the AFA systems, we trained the UBM as described in [20]. In each EM iteration, the full covariance matrices were processed using the flooring function described in Table II [20], [46]. We used the floor matrix $\mathbf{F} = f\bar{\boldsymbol{\Sigma}}$, where

$$\bar{\boldsymbol{\Sigma}} = \frac{1}{M}\sum_{g=1}^{M}\boldsymbol{\Sigma}_g \tag{29}$$

is the average covariance matrix, and $f = 0.1$ is set as in [20]. We refer to this flooring method as "vFloor-2". Baseline and AFA system results using these two different UBM flooring methods are summarized in Table III. In this experiment, PLDA size N_{EV} was set to 150.

TABLE II
UBM COVARIANCE MATRIX FLOORING FUNCTION (VFLOOR-2) [20]

Function: $\tilde{\mathbf{S}} = \mathrm{floor}(\mathbf{S}, \mathbf{F})$	
1.	Cholesky decomposition: $\mathbf{F} = \mathbf{L}\mathbf{L}^T$
2.	Normalize target matrix: $\mathbf{Q} \leftarrow \mathbf{L}^{-1}\mathbf{S}\mathbf{L}^{-T}$
3.	Eigenvalue decomposition: $\mathbf{Q} = \mathbf{U}\mathbf{D}\mathbf{U}^T$
4.	Obtain diagonal matrix $\tilde{\mathbf{D}}$ by flooring $\mathbf{D}$ to 1: $\tilde{d}_{ii} = \max(d_{ii}, 1)$
5.	Return to full matrix: $\tilde{\mathbf{Q}} \leftarrow \mathbf{U}\tilde{\mathbf{D}}\mathbf{U}^T$
6.	De-normalization: $\tilde{\mathbf{S}} \leftarrow \mathbf{L}\tilde{\mathbf{Q}}\mathbf{L}^T$

TABLE III
PERFORMANCE COMPARISON BETWEEN BASELINE I-VECTOR AND DIFFERENT AFA SYSTEMS USING ALTERNATE UBM FLOORING. EVALUATIONS PERFORMED ON NIST SRE 2010 CORE CONDITION-5 EXTENDED TRIALS

System		EER	DCF$_{old}$	DCF$_{new}$
UBM variance flooring using vFloor-1				
Baseline full-cov		2.03961	0.11236	0.40556
AFA	$q = 36$	2.04234	0.10417	**0.35646**
	$q = 42$	**1.79444**	0.10143	0.38688
	$q = 48$	1.95537	0.10375	0.36349
	$q = 54$	1.90054	**0.09884**	0.39321
UBM variance flooring using vFloor-2 [20]				
Baseline full-cov		1.93923	0.10315	0.41917
AFA	$q = 36$	1.90352	0.11016	0.39275
	$q = 42$	2.05354	0.10328	0.38583
	$q = 48$	2.00150	0.10168	0.38034
	$q = 54$	1.92451	0.10138	0.39755

From the results, we observe that the variance flooring vFloor-2 [20] provides slightly improved baseline system performance compared to vFloor-1, with respect to %EER and DCF_{old} but degrades in DCF_{new} measure. The proposed AFA transformation achieves much better performance over the baseline system when using vFloor-1. AFA provides improvement over the baseline system using vFloor-2 only for $q = 54$, whereas performance improvement is observed for $q = 42$, 48 and 54 when vFloor-1 is used. This deterioration of AFA system performance can be expected, since the vFloor-2 algorithm modifies the eigenvalues of the covariance matrices on which the AFA approach directly relies on. Noting that AFA with vFloor-1 provides the best overall performance and vFloor-2 does not provide sufficient advantage over vFloor-1, we use vFloor-1 method in all of our subsequent experiments.

D. Performance in Microphone Conditions

In this section we present evaluation results of the proposed systems on the NIST SRE 2010 core conditions 1–4 using the extended trials. In these experiments, additional microphone data from SRE 2005 and 2006 corpora was included for UBM and TV matrix training. The PLDA model was trained using both telephone and microphone data as before. The results are given in Tables IV–VII. We compare the following systems: Baseline full-cov, and AFA with $q = 36$, 42, 48 and 54. The PLDA parameter N_{EV} was set to 150. We did not evaluate the diagonal UBM system in these conditions.

From the results, again we observe that the proposed AFA systems consistently outperform the baseline system, especially for conditions 1–3. However, it seems a single parameter setting of q does not always provide the best performance across all the

TABLE IV
PERFORMANCE COMPARISON BETWEEN BASELINE I-VECTOR AND DIFFERENT AFA SYSTEMS. EVALUATION PERFORMED IN NIST SRE 2010 CORE CONDITION-1 EXTENDED TRIALS

System		EER	DCF$_{old}$	DCF$_{new}$
Baseline full-cov		2.09767	0.08539	0.31712
AFA	$q = 36$	2.26966	0.08024	0.28560
	$q = 42$	2.07210	0.07921	**0.28063**
	$q = 48$	**1.93024**	**0.07849**	0.28756
	$q = 54$	2.01850	0.07914	0.33058

TABLE V
PERFORMANCE COMPARISON BETWEEN BASELINE I-VECTOR AND DIFFERENT AFA SYSTEMS. EVALUATION PERFORMED IN NIST SRE 2010 CORE CONDITION-2 EXTENDED TRIALS

System		EER	DCF$_{old}$	DCF$_{new}$
Baseline full-cov		3.75464	0.16353	0.53167
AFA	$q = 36$	3.78150	0.16862	0.48979
	$q = 42$	**3.58186**	**0.15783**	**0.48376**
	$q = 48$	3.67975	0.16186	0.50176
	$q = 54$	3.80477	0.16084	0.50948

TABLE VI
PERFORMANCE COMPARISON BETWEEN BASELINE I-VECTOR AND DIFFERENT AFA SYSTEMS. EVALUATION PERFORMED IN NIST SRE 2010 CORE CONDITION-3 EXTENDED TRIALS

System		EER	DCF$_{old}$	DCF$_{new}$
Baseline full-cov		3.17154	0.15207	0.45750
AFA	$q = 36$	3.45395	0.16002	0.48452
	$q = 42$	3.15838	0.14754	0.44873
	$q = 48$	3.10171	**0.14656**	**0.42633**
	$q = 54$	**2.89653**	0.14827	0.43774

TABLE VII
PERFORMANCE COMPARISON BETWEEN BASELINE I-VECTOR AND DIFFERENT AFA SYSTEMS. EVALUATION PERFORMED IN NIST SRE 2010 CORE CONDITION-4 EXTENDED TRIALS

System		EER	DCF$_{old}$	DCF$_{new}$
Baseline full-cov		2.05830	0.09356	**0.26975**
AFA	$q = 36$	2.01237	0.09314	0.30255
	$q = 42$	2.01237	0.09314	0.30255
	$q = 48$	**1.80459**	0.09456	0.28637
	$q = 54$	1.82594	**0.08728**	0.27816

performance metrics. Considering the best %EER values, the proposed systems achieved 8.14%, 6.43%, 8.67% and 12.33% relative improvements in conditions 1, 2, 3 and 4, respectively. These results demonstrate the effectiveness of the proposed scheme in the microphone mismatched conditions as well.

E. Fusion of Multiple Systems

We select three of our systems for fusion: (i) Baseline full-cov, (ii) AFA $(q = 42)$ and (iii) AFA $(q = 48)$. The PLDA N_{EV} parameter was set to 150 for all systems. Simple equal-weight linear fusion was used with mean and variance normalization of individual system scores to (0, 1) for calibration. Results are shown for NIST SRE 2010 core condition 5 and pooled condition (combining all trials from condition 1–5) in Tables VIII and IX, respectively.

TABLE VIII
LINEAR EQUAL-WEIGHT SCORE FUSION PERFORMANCE OF BASELINE I-VECTOR AND PROPOSED SYSTEMS FOR NIST SRE 2010 CORE CONDITION-5

Individual system performances				
	System	%EER	DCF$_{old}$	DCF$_{new}$
(i)	Baseline full-cov	2.03961	0.11236	0.40556
(ii)	AFA $(q = 42)$	1.79444	0.10143	0.38688
(iii)	AFA $(q = 48)$	1.95537	0.10375	0.36349
Fusion system performances				
1	Fusion of (i) & (ii)	1.77162	0.09704	0.36610
2	Fusion of (i) - (iii)	**1.70258**	**0.09610**	**0.34839**

TABLE IX
LINEAR EQUAL-WEIGHT SCORE FUSION PERFORMANCE OF BASELINE I-VECTOR AND PROPOSED SYSTEMS FOR NIST SRE 2010 CORE CONDITIONS 1–5 POOLED

Individual system performances				
	System	%EER	DCF$_{old}$	DCF$_{new}$
(i)	Baseline full-cov	3.02720	0.13995	0.46022
(ii)	AFA $(q = 42)$	2.86091	0.13316	0.43030
(iii)	AFA $(q = 48)$	2.88596	0.13615	0.43086
Fusion system performances				
1	Fusion of (i) & (ii)	2.69742	0.12199	0.41459
2	Fusion of (i) - (iii)	**2.61094**	**0.12035**	**0.40591**

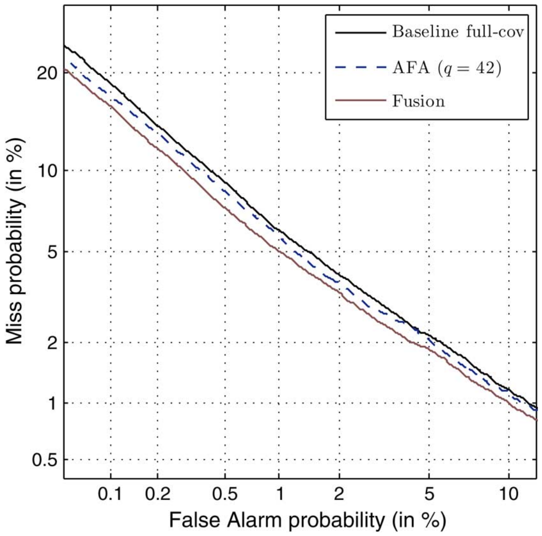


Fig. 7. Performance comparison of baseline, AFA and fusion systems using DET curves. Evaluation is performed by pooling results of the core conditions 1–5 of NIST SRE 2010 extended trials. (i) Baseline i-vector system using Full Covariance UBM (Baseline full-cov), (ii) AFA i-vector system $(q = 42)$, and (iii) Equal-weight linear fusion of systems (i) and (ii).

From the results, fusion performance of systems (i) and (ii) clearly reveal that AFA and baseline system have complementary information, since %EER and the DCF values improve. This is observed for both telephone and pooled condition. The best result is achieved by fusing systems (i)–(iii), to obtain 16.52%, 14.47% and 14.09% relative improvement in %EER, DCF$_{old}$ and DCF$_{new}$, respectively, compared to the baseline system in condition-5. In the pooled condition, this fusion provides 13.75%, 14.0% and 11.80% relative improvement in %EER, DCF$_{old}$ and DCF$_{new}$, respectively. Performance comparison of the systems (i), (ii) and their fusion for the pooled condition is shown in Fig. 7 using Detection Error Trade-off (DET) curves. Here, again we observe the superiority of the proposed AFA system over the baseline system while the fusion of these systems consistently provide further improvement in the full DET range.

F. Computational Advantages

In our experiments, we observe that the TV matrix training process using the AFA transform is computationally less expensive compared to the conventional process. This is expected since the computational complexity of an i-vector system is proportional to the super-vector size Md [39], which is reduced to Mq for an AFA based system. Thus, the computational complexity of the proposed system is theoretically reduced by a factor of $1/\alpha$ $(0 < \alpha < 1)$ compared to the baseline system.

VII. Conclusions

In this study, we have proposed an alternate modeling technique to address and compensate for transmission channel mismatch in speaker recognition. Motivated by the covariance structure of conventional acoustic features, we developed a factor analysis technique which operates within the acoustic feature domain utilizing a well trained UBM with full covariance matrices. We advocated that conventional super-vector domain factor analysis methods fail to take advantage of the observation that speech features reside in a lower dimensional manifold in the acoustic space. The proposed acoustic factor analysis scheme was utilized to develop a mixture-dependent feature transformation that performs dimensionality reduction, de-correlation, normalization and enhancement at the same time. Finally, the transformation was effectively integrated within a standard i-vector-PLDA based speaker recognition system using a probabilistic feature alignment technique. The superiority of the proposed method was demonstrated by experiments performed using the NIST SRE 2010 extended trials of five core conditions. Measurable improvements over two baseline systems were shown in terms of EER, min DCFs and DET curves.

References

[1] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, “Joint factor analysis versus Eigenchannels in speaker recognition,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1435–1447, May 2007.

[2] D. Reynolds, M. Zissman, T. Quatieri, G. O’Leary, and B. Carlson, “The effects of telephone transmission degradations on speaker recognition performance,” in *Proc. ICASSP*, 1995, pp. 329–332.

[3] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, “Score normalization for text-independent speaker verification systems,” *Digital Signal Process.*, vol. 10, no. 1–3, pp. 42–54, 2000.

[4] R. Rose, E. Hofstetter, and D. Reynolds, “Integrated models of signal and background with application to speaker identification in noise,” *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 245–257, Apr. 1994.

[5] J. H. L. Hansen, “Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition,” *Speech Commun.*, vol. 20, no. 1–2, pp. 151–173, 1996.

[6] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 788–798, May 2010.

[7] D. Reynolds, T. Quatieri, and R. Dunn, “Speaker verification using adapted Gaussian mixture models,” *Digital Signal Process.*, vol. 10, no. 1–3, pp. 19–41, 2000.

[8] T. Hasan and J. H. L. Hansen, “Factor analysis of acoustic features using a mixture of probabilistic principal component analyzers for robust speaker verification,” in *Proc. Odyssey*, Singapore, Jun. 2012.

[9] T. Hasan and J. H. L. Hansen, “Integrated feature normalization and enhancement for robust speaker recognition using acoustic factor analysis,” in *Proc. Interspeech*, Portland, OR, Sept. 2012.

[10] T. Hasan and J. H. L. Hansen, “Robust speaker recognition in non-stationary room environments based on empirical mode decomposition,” in *Proc. Interspeech*, Florence, Italy, Oct. 2011, pp. 2733–2736.

[11] X. Fan and J. H. L. Hansen, “Speaker identification within whispered speech audio streams,” *IEEE Trans. Speech Audio Process.*, vol. 19, no. 5, pp. 1408–1421, Jul. 2011.

[12] C. Zhang and J. H. L. Hansen, “Whisper-island detection based on unsupervised segmentation with entropy-based speech feature processing,” *IEEE Trans. Speech Audio Process.*, vol. 19, no. 4, pp. 883–894, May 2011.

[13] J. H. L. Hansen and V. Varadarajan, “Analysis and compensation of lombard speech across noise type and levels with application to in-set/out-of-set speaker recognition,” *IEEE Trans. Speech Audio Process.*, vol. 17, no. 2, pp. 366–378, Feb. 2009.

[14] W. Campbell, D. Sturim, and D. Reynolds, “Support vector machines using GMM supervectors for speaker verification,” *IEEE Signal Process. Lett.*, vol. 13, no. 5, pp. 308–311, May 2006.

[15] P. Kenny, G. Boulianne, and P. Dumouchel, “Eigenvoice modeling with sparse training data,” *IEEE Trans. Speech Audio Process.*, vol. 13, no. 3, pp. 345–354, May 2005.

[16] L. Burget, P. Matejka, P. Schwarz, O. Glembek, and J. Cernocky, “Analysis of feature extraction and channel compensation in a GMM speaker recognition system,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. ASSP-15, no. 7, pp. 1979–1986, Sep. 2007.

[17] R. Kuhn, J. Junqua, P. Nguyen, and N. Niedzielski, “Rapid speaker adaptation in Eigenvoice space,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 8, no. 6, pp. 695–707, Nov. 2000.

[18] M. Tipping and C. Bishop, “Mixtures of probabilistic principal component analyzers,” *Neural Comput.*, vol. 11, no. 2, pp. 443–482, 1999.

[19] J. Villalba and N. Brummer, “Towards fully Bayesian speaker recognition: Integrating out the between-speaker covariance,” in *Proc. Interspeech*, Florence, Italy, Oct. 2011, pp. 505–508.

[20] P. Matejka, O. Glembek, F. Castaldo, M. Alam, O. Plchot, P. Kenny, L. Burget, and J. Cernocky, “Full-covariance UBM and heavy-tailed PLDA in i-vector speaker verification,” in *Proc. ICASSP*, Florence, Italy, Oct. 2011, pp. 4828–4831.

[21] Y. Ephraim and H. Van Trees, “A signal subspace approach for speech enhancement,” *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 251–266, Jul. 1995.

[22] A. Rezayee and S. Gazor, “An adaptive KLT approach for speech enhancement,” *IEEE Trans. Speech Audio Process.*, vol. 9, no. 2, pp. 87–95, Feb. 2001.

[23] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.

[24] M. Kuhne, D. Pullella, R. Togneri, and S. Nordholm, “Towards the use of full covariance models for missing data speaker recognition,” in *Proc. ICASSP*, Las Vegas, NV, Apr. 2008, pp. 4537–4540.

[25] B. Zhou and J. H. L. Hansen, “Rapid discriminative acoustic model based on Eigenspace mapping for fast speaker adaptation,” *IEEE Trans. Speech Audio Process.*, vol. 13, no. 4, pp. 554–564, Jul. 2005.

[26] E. Batlle, C. Nadeu, and J. Fonollosa, “Feature decorrelation methods in speech recognition. A comparative study,” in *Proc. ICSLP*, Sydney, Australia, 1998, vol. 7, pp. 2907–2910.

[27] T. Eisele, R. Haeb-Umbach, and D. Langmann, “A comparative study of linear feature transformation techniques for automatic speech recognition,” in *Proc. ICSLP*, 1996, vol. 1, pp. 252–255.

[28] K. Y. Lee, “Local fuzzy PCA based GMM with dimension reduction on speaker identification,” *Pattern Recogn. Lett.*, vol. 25, pp. 1811–1817, December 2004.

[29] T. Kinnunen, I. Kärkkäinen, and P. Fränti, “Is speech data clustered?—Statistical analysis of cepstral features,” in *Proc. Interspeech*, Aalborg, Denmark, Sep. 2001, pp. 2627–2630.

[30] A. Oppenheim and R. Schafer, “From frequency to quefrency: A history of the cepstrum,” *IEEE Signal Process. Mag.*, vol. 21, no. 5, pp. 95–106, Sep. 2004.

[31] S. Vaseghi, *Advanced Signal Processing and Digital Noise Reduction*. New York: Wiley, 1996.

[32] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[33] M. Hayes, *Statistical Digital Signal Processing and Modeling*. New York: Wiley, 2009.

[34] A. McCree, D. Sturim, and D. Reynolds, “A new perspective on GMM subspace compensation based on PPCA and Wiener filtering,” in *Proc. Interspeech*, Florence, Italy, Oct. 2011, pp. 145–148.

[35] R. Martin, “Noise power spectral density estimation based on optimal smoothing and minimum statistics,” *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[36] T. Anderson, "Asymptotic theory for principal component analysis," *Ann. Math. Statist.*, vol. 34, no. 1, pp. 122–148, 1963.

[37] M. Alam, P. Ouellet, P. Kenny, and D. O'Shaughnessy, "Comparative evaluation of feature normalization techniques for speaker verification," *Adv. Nonlinear Speech Process.*, pp. 246–253, 2011.

[38] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. Odyssey*, 2001, pp. 213–218.

[39] O. Glembek, L. Burget, P. Matejka, M. Karafiat, and P. Kenny, "Simplification and optimization of i-vector extraction," in *Proc. ICASSP*, Florence, Italy, Oct. 2011, pp. 4516–4519.

[40] D. Matrouf, N. Scheffer, B. Fauve, and J. Bonastre, "A straightforward and efficient implementation of the factor analysis model for speaker verification," in *Proc. Interspeech*, Antwerp, Belgium, Aug. 2007, pp. 1242–1245.

[41] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proc. Interspeech*, Florence, Italy, Oct. 2011, pp. 249–252.

[42] P. Schwarz, P. Matejka, and J. Cernocky, "Hierarchical structures of neural networks for phoneme recognition," in *Proc. ICASSP*, May 2006, vol. 1, pp. 325–328.

[43] S. Young, *HTK Reference Manual*. Cambridge, U.K.: Cambridge Univ. Eng. Dept., 1993.

[44] The NIST Year 2008 Speaker Recognition Evaluation Plan, 2008 [Online]. Available: http://www.nist.gov

[45] The NIST Year 2010 Speaker Recognition Evaluation Plan, 2010 [Online]. Available: http://www.nist.gov

[46] D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafit, A. Rastrow, R. C. Rose, P. Schwarz, and S. Thomas, "The subspace Gaussian mixture modelA structured model for speech recognition," *Comput. Speech Lang.*, vol. 25, no. 2, pp. 404–439, 2011.

Taufiq Hasan received his B.Sc. and M.Sc. degrees in Electrical and Electronic Engineering from Bangladesh University of Engineering and Technology, Dhaka, Bangladesh, in 2006 and 2008, respectively. He was a Lecturer in the Electrical and Electronic Engineering Department at United International University, Dhaka, Bangladesh from December 2006 to June 2008. Currently has been pursuing his Ph.D. degree as a Research Assistant in the Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas (UTD), Richardson, U.S.A. since August 2008. He is a member of the Center for Robust Speech Systems (CRSS) at UTD. His research interests are robust speaker recognition in noise and channel mismatch, speech enhancement and automatic video summarization.

John H. L. Hansen (S'81–M'82–SM'93–F'07) received the Ph.D. and M.S. degrees in Electrical Engineering from Georgia Institute of Technology, Atlanta, Georgia, in 1988 and 1983, and B.S.E.E. degree from Rutgers University, College of Engineering, New Brunswick, N.J. in 1982.

He joined University of Texas at Dallas (UTD), Erik Jonsson School of Engineering and Computer Science in the fall of 2005, where he is presently serving as Jonsson School Associate Dean for Research, as well as Professor of Electrical Engineering and also holds the Distinguished University Chair in Telecommunications Engineering. He previously served as Department Head of Electrical Engineering from 2005–12, overseeing a 5x increase in research expenditures with a 20% increase in enrollment and the addition of 18 T/TT faculty, growing UTDallas to be the 8th largest EE program from ASEE rankings. He also holds a joint appointment He also holds a joint appointment as Professor in the School of Behavioral and Brain Sciences (Speech & Hearing). At UTD, he established the Center for Robust Speech Systems (CRSS) which is part of the Human Language Technology Research Institute. Previously, he served as Department Chairman and Professor in the Dept. of Speech, Language and Hearing Sciences (SLHS), and Professor in the Dept. of Electrical & Computer Engineering, at Univ. of Colorado Boulder (1998–2005), where he co-founded the Center for Spoken Language Research. In 1988, he established the Robust Speech Processing Laboratory (RSPL) and continues to direct research activities in CRSS at UTD. In 2007, he was named IEEE Fellow for contributions in "Robust Speech Recognition in Stress and Noise," and is currently serving as Member of the IEEE Signal Processing Society Speech Technical Committee (2005–08; 2010–13; elected Chair-elect in 2010), and Educational Technical Committee (2005–08; 2008–10). Previously, he has served as Technical Advisor to U.S. Delegate for NATO (IST/TG-01), IEEE Signal Processing Society Distinguished Lecturer (2005/06), Associate Editor for IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (1992–99), Associate Editor for IEEE SIGNAL PROCESSING LETTERS (1998–2000), Editorial Board Member for the IEEE Signal Processing Magazine (2001–03). He has also served as guest editor of the Oct. 1994 special issue on Robust Speech Recognition for IEEE TRANSACTIONS ON SPEECH & AUDIO PROCESSING. He has served on the Speech Communications Technical Committee for the Acoustical Society of America (2000–03), and is serving as a member of the ISCA (Inter. Speech Communications Association) Advisory Council. In 2010, he was recognized as ISCA Fellow, for contributions on "research for speech signals under adverse conditions." His research interests span the areas of digital speech processing, analysis and modeling of speech and speaker traits, speech enhancement, feature estimation in noise, robust speech recognition with emphasis on spoken document retrieval, and in-vehicle interactive systems for hands-free human-computer interaction. He has supervised 59 (27 PhD, 32 MS/MA) thesis candidates, was recipient of The 2005 University of Colorado Teacher Recognition Award as voted on by the student body, author/co-author of 433 journal and conference papers and 8 textbooks in the field of speech processing and language technology, coauthor of the textbook *Discrete-Time Processing of Speech Signals*, (IEEE Press, 2000), co-editor of *DSP for In-Vehicle and Mobile Systems* (Springer, 2004), *Advances for In-Vehicle and Mobile Systems: Challenges for International Standards* (Springer, 2006), *In-Vehicle Corpus and Signal Processing for Driver Behavior* (Springer, 2008), and lead author of the report "The Impact of Speech Under 'Stress' on Military Speech Technology," (NATO RTO-TR-10, 2000). He also organized and served as General Chair for ICSLP/Interspeech-2002: International Conference on Spoken Language Processing, Sept. 16–20, 2002, and served as Co-Organizer and Technical Program Chair for IEEE ICASSP-2010, Dallas, TX.