# Maximum Likelihood Acoustic Factor Analysis Models for Robust Speaker Verification in Noise

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

***Abstract*—Recent speaker recognition/verification systems generally utilize an utterance dependent fixed dimensional vector as features to Bayesian classifiers. These vectors, known as i-Vectors, are lower dimensional representations of Gaussian Mixture Model (GMM) mean super-vectors adapted from a Universal Background Model (UBM) using speech utterance features, and extracted utilizing a Factor Analysis (FA) framework. This method is based on the assumption that the speaker dependent information resides in a lower dimensional sub-space. In this study, we utilize a mixture of Acoustic Factor Analyzers (AFA) to model the acoustic features instead of a GMM-UBM. Following our previously proposed AFA technique ("Acoustic factor analysis for robust speaker verification," by Hasan and Hansen, *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 4, April 2013), this model is based on the assumption that the speaker relevant information lies in a lower dimensional subspace in the multi-dimensional feature space localized by the mixture components. Unlike our previous method, here we train the AFA-UBM model directly from the data using an Expectation-Maximization (EM) algorithm. This method shows improved robustness to noise as the nuisance dimensions are removed in each EM iteration. Two variants of the AFA model are considered utilizing an isotropic and diagonal covariance residual term. The method is integrated within a standard i-Vector system where the hidden variables of the model, termed as *acoustic factors*, are utilized as the input for *total variability* modeling. Experimental results obtained on the 2012 National Institute of Standards and Technology (NIST) Speaker Recognition Evaluation (SRE) core-extended trials indicate the effectiveness of the proposed strategy in both clean and noisy conditions.**

***Index Terms*—Acoustic factor analysis, mixture of factor analyzers, speaker verification.**

## I. Introduction

ONE of the most challenging problems facing the speaker recognition research is environmental noise. Among other extrinsic degradations affecting speech system performance, transmission channel mismatch [1], [2], handset variability [3], reverberation [8], and non-stationarity environments [9] reflect core sources of distortion/mismatch. There are also intrinsic sources of mismatch including, physical task stress induced variability [4], low vocal effort/whispered speech [10], [11], Lombard effect [12], spontaneity of speech, etc. Speaker recognition research efforts have been largely driven by the SRE efforts organized by NIST [13]. During the more recent evaluations, transmission channel and microphone variability have been given emphasis, leading to the most recent evaluation task in 2012 where additive noise and mixed duration test utterances were introduced [13].

The issue of channel variability has been thoroughly studied over the last few years [2], [5] leading to several breakthroughs in this area. Various compensation strategies have been proposed in the past to reduce unwanted variability between training and test utterances, while retaining the speaker identity information. To address issues related to noisy and channel degraded conditions, most effective techniques operate on the utterance models, including GMM super-vectors [14] and various factor analysis schemes built in this domain [2], [15], as well as i-Vectors with Probabilistic Linear Discriminant Analysis (PLDA) based classifiers along with various pre-processing techniques [16]–[18]. Robust feature development [19]–[21], enhancement [22]–[24], effective front-end compensation methods [25]–[27] and score domain techniques have also been considered [3], [28] for mismatch compensation. Many techniques have evolved and are being replaced by new variants over the last decade, but for short-term spectrum based systems, a GMM has almost always been used as the background model.

Since the advent of i-Vectors, the most effective and convenient way of dealing with mismatched conditions has been to include degraded data similar to test utterances in the PLDA training. Such utterances can also be included during the UBM and i-Vector extractor training. We have observed this during the recent NIST SRE 2012, where additive noise and mixed duration utterances were introduced in test conditions. One straightforward solution is to add noisy and mixed duration data into the PLDA training phase [29], [30]. Even though PLDA is a linear model, it seems to be quite effective for additive noise, convoluational channel and duration variability. This work, however, is motivated by the presumption that improved solutions to noise robustness can lie in earlier stages of the system, especially where the degraded features are being modeled for the first time.

In our recent studies [7], [31], we proposed a factor analysis scheme for front-end features that operates on different mixtures of the UBM, termed Acoustic Factor Analysis (AFA). The principal motivation of the approach was the assumption that

Manuscript received June 17, 2013; revised September 11, 2013; accepted November 05, 2013. Date of publication November 22, 2013; date of current version December 31, 2013. This work was supported in part by the Air Force Research Laboratory under Contract FA8750-12-1-0188 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 an initial study on ML-AFA was presented in [6]. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Thomas Fang Zheng.

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/TASLP.2013.2292356

acoustic features reside in a lower dimensional subspace, similar to the assumption made on the GMM super-vectors. The technique operates on the first order Baum-Welch statistics in each mixture with a transformation matrix, effectively reducing the feature dimension within the model. Integrated within an i-Vector system, this method led way towards a two-stage factor analysis scheme for speaker recognition. We also showed the similarity of the AFA technique between signal sub-space based speech enhancement schemes [7].

In this study, we take the AFA concept further by completely replacing the traditional UBM with a Mixture of Factor Analyzers (MFA) model and propose an i-Vector extraction strategy that utilizes the first order statistics of the hidden variables (i.e., the *acoustic factors*), instead of the acoustic features. In our past studies [7], [32], we derived an AFA model from a full-covariance UBM which utilized an isotropic residual term, making it equivalent to a Probabilistic Principal Component Analyzer (PPCA) [33] model. The method was interpreted as a transformation of acoustic features in different mixture components, which in effect would also transform the UBM.

In our experiments during the NIST SRE 2012 evaluation, we observed that extracting the AFA model parameters from the full-covariance UBM would degrade system performance in noisy conditions. Our UBM dataset was clean, which led us to believe that the sub-spaces learned from the eigen-decomposition of the full-covariances is not as useful in the separation of the signal from the noisy sub-space. Later, we added noisy data into the UBM, but that by itself was not sufficient for the method to be effective. Next, we hypothesize that this could be due to the full-covariance model training which considered the full feature space in each iteration, leading to a mixture model which is already affected by the noisy directions. This implies that the noisy directions in each mixture should have been removed in each iteration.

Motivated by these observations made during the NIST SRE 2012 preparation, in this study we propose to utilize a mixture of AFA model in place of a UBM to develop an i-Vector system. We consider the scenarios where the model residual is isotropic and uncorrelated (diagonal covariance), leading to a mixture of PPCA model [33] and MFA model [34], respectively. These models are iteratively trained using an EM algorithm. The advantage of using these models when training a UBM with noisy data is that they only consider the dominant directions of the feature space in each mixture, providing more robustness to the noisy test data. However, as will be demonstrated shortly, significant improvement can be obtained through the method if only the posterior statistics of the hidden variables are utilized for the i-Vector extraction. This confirms the original motivation of the earlier AFA method presented in [7], that speaker dependent information resides within the first few dominant directions in the feature space. It maybe noted that this observation was previously revealed for model adaptation for speech recognition [35].

The organization of this paper is as follows. In Section II, we describe the Acoustic Factor Analysis method in generic way, leading to its variations due to assumptions made on the residual term. We also describe here how the proposed method is integrated within an i-Vector system and provide some insightful interpretation of the model. Section III describes the various components of the baseline speaker verification system and corpora utilization. In Section IV. experimental results are presented and finally, Section V concludes the study.

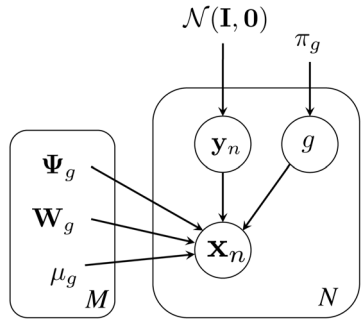


Fig. 1. Probabilistic graphical model of a Mixture of Factor Analyzer (MFA) model for acoustic features. The box on the right denotes a 'plate' representing a dataset of $N$ independent observations of acoustic features $\mathbf{x}_n$. Here, $\mathbf{y}_n$ are the hidden variables, or *acoustic factors*, and $g$ indicate the responsible mixture component in the model. The box on the left represent the parameters of the $g$-th model component out of a total of $M$ mixtures.

## II. Acoustic Factor Analysis

In this section, we describe the proposed model of acoustic features, discuss its formulation and EM-training steps and application in an i-Vector based speaker verification system.

### A. Formulation

Let $\mathbf{x} \in \mathbb{R}^d$ represent the acoustic feature vectors and $\mathcal{X} = \{\mathbf{x}_n | n = 1 \cdots N\}$ denote the collection of development data. Using a standard factor analysis model [36], [37], the feature vector $\mathbf{x}$ can be represented by,

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

Here, $\mathbf{W}$ is a $d \times q$ factor loading matrix that represents $q < d$ bases spanning the sub-space corresponding to the important variability in the feature space, and $\mu$ is the $d \times 1$ mean vector. Following our terminology in [7], [31], [32], 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 $q \times 1$. The remaining variability in the data is modeled by the noise component $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{\Psi})$. In this model, the feature vectors are normally distributed such that, $\mathbf{x} \sim \mathcal{N}(\mu, \mathbf{\Psi} + \mathbf{W}\mathbf{W}^T)$.

Naturally, acoustic features extracted from speech data containing many different channel/noise variations are better modeled using clusters in the feature space. Thus, we utilize a mixture of AFA models [7] similar to a traditional GMM-UBM. In this case, the probability density function of $\mathbf{x}_n$ is given by,

$$p(\mathbf{x}_n) = \sum_{g=1}^{M} \pi_g p(\mathbf{x}_n | g), \tag{2}$$

where $\pi_g$ is the weight corresponding to the $g$-th mixture component, $M$ is the total number of mixtures, and $p(\mathbf{x}_n|g) \sim \mathcal{N}(\mu_g, \mathbf{C}_g)$. Here, the model covariance matrix for each component is given by,

$$\mathbf{C}_g = \mathbf{\Psi}_g + \mathbf{W}_g \mathbf{W}_g^T. \tag{3}$$

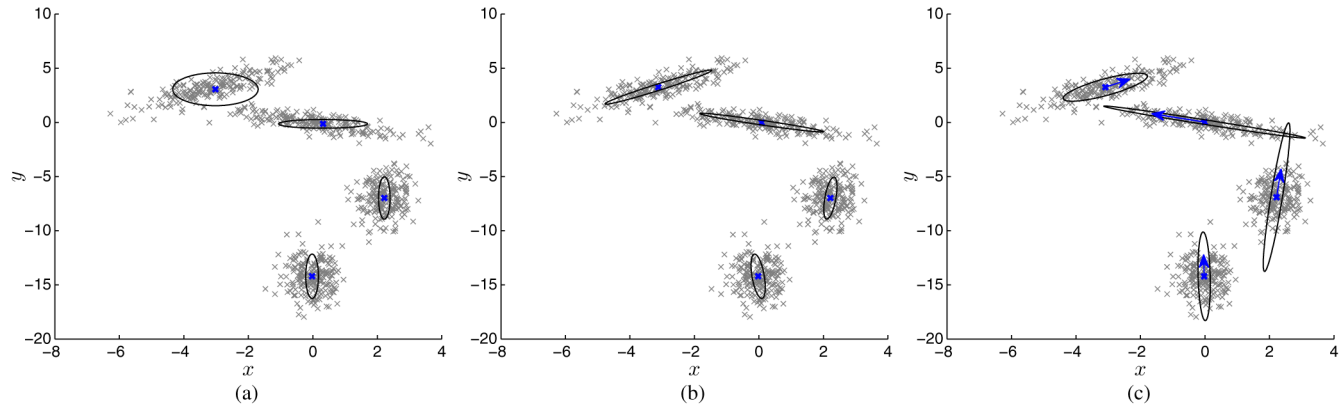


Fig. 2. Scatter plot of synthetic 2D Gaussian data with four clusters and trained mixture models. Means are shown as blue points while ellipses depict the covariance matrices. (a) Diagonal covariance GMM, (b) full covariance GMM, (c) Mixture of PPCA model showing the principal directions in each mixture.

Fig. 1 shows a probabilistic graphical model of this mixture model. In our previous studies [7], [31], [32], we assumed $\epsilon$ to be isotropic, that is $\mathbf{\Psi}_g = \sigma_g^2\mathbf{I}$ where $\sigma_g^2$ denotes the average noise power, and the AFA model parameters were derived from a full-covariance GMM-UBM. In this study, we obtain Maximum-Likelihood (ML) formulations of the mixture of AFA model assuming $\mathbf{\Psi}_g$ to be isotropic and diagonal. This model, trained similar to a GMM, essentially replaces the UBM model of the speaker verification system and leads to a new method for extracting i-Vectors.

The learning behavior of these models is illustrated in Fig. 2. Synthetic 2-dimensional Gaussian data points distributed in four clusters are used for this example. The diagonal-covariance GMM, full-covariance GMM and a mixture of PPCA model are used with four mixtures, and the mean and covariances of the models are shown in Figs. 2(a-c) as points and ellipsoids, respectively. As expected, the diagonal model is insensitive to the dominant direction of the data, while the full covariance model is able to take this into account. The PPCA model with $q = 1$ finds the dominant direction of the data and considers the other direction as noise. The covariance shown here in Fig. 2(c) corresponds to the model covariance $\mathbf{C}_g$. In the proposed approach, we only consider data variation in the dominant direction as detected by the model.

### B. Isotropic Residual Noise

In this scenario, we assume that the noise covariance matrix in each mixture $\mathbf{\Psi}_g = \sigma_g^2\mathbf{I}$ is isotropic. This leads to the standard PPCA model as derived in [33]. The EM algorithm procedure for a mixture of PPCA model is as follows. In the first step, the following parameters are computed given the initial or old parameter estimates $\Lambda = \{\pi_g, \mu_g, \mathbf{C}_g\}$:

$$\gamma_n(g) = p(g|\mathbf{x}_n, \Lambda) = \frac{p(\mathbf{x}_n|g,\Lambda)\pi_g}{p(\mathbf{x}_n|\Lambda)}, \tag{4}$$

$$\tilde{\pi}_g = \frac{1}{N}\sum_{n=1}^{N}\gamma_n(g), \tag{5}$$

$$\tilde{\mu}_g = \frac{\sum_{n=1}^{N}\gamma_n(g)\mathbf{x}_n}{\sum_{n=1}^{N}\gamma_n(g)}, \quad \text{and} \tag{6}$$

$$\mathbf{S}_g = \frac{1}{N\tilde{\pi}_g}\sum_{n=1}^{N}\gamma_n(g)(\mathbf{x}_n - \tilde{\mu}_g)(\mathbf{x}_n - \tilde{\mu}_g)^T. \tag{7}$$

Here, $\tilde{\pi}_g$ and $\tilde{\mu}_g$ are the new estimates for the weights and mean vectors, respectively. Next, the new values, $\tilde{\mathbf{W}}_g$ and $\tilde{\sigma}_g^2$ can be obtained by:

$$\tilde{\mathbf{W}}_g = \mathbf{S}_g\mathbf{W}_g(\sigma_g^2\mathbf{I} + \mathbf{M}_g^{-1}\mathbf{W}_g^T\mathbf{S}_g\mathbf{W}_g)^{-1}, \quad \text{and} \tag{8}$$

$$\tilde{\sigma}_g^2 = \frac{1}{d}\mathrm{tr}(\mathbf{S}_g - \mathbf{S}_g\mathbf{W}_g\mathbf{M}_g^{-1}\tilde{\mathbf{W}}_g^T), \tag{9}$$

where $\mathbf{M}_g = \sigma_g^2\mathbf{I} + \mathbf{W}_g^T\mathbf{W}_g$. The posterior covariance matrix of the distribution $p(\mathbf{y}_n|\mathbf{x}_n, g)$ is given by $\sigma_g^2\mathbf{M}_g^{-1}$. The posterior distribution of the acoustic factors for the $g$-th mixture is given by:

$$p(\mathbf{y}_n|\mathbf{x}_n, g) = \mathcal{N}\left(\mathbf{y}_n|\mathbf{M}_g^{-1}\mathbf{W}_g^T(\mathbf{x}_n - \mu_g), \sigma_g^2\mathbf{M}_g^{-1}\right). \tag{10}$$

The updated model covariance is obtained using (3). Using the updated parameters, (4)–(7) are utilized in the next iteration. We denote this model by: $\text{ML-AFA}_{\text{iso}}$.

### C. Diagonal Covariance Residual Noise

Here, we assume that $\mathbf{\Psi}_g$ is diagonal. In this case, the $q$ dominant directions represented by the factor loading matrix $\mathbf{W}_g$ are no longer the principal components. This, essentially becomes a standard Mixture of Factor Analyzers (MFA) [34], [37] model applied to acoustic features. Similar to the PPCA case, the update equations for the diagonal covariance AFA model can be obtained through maximization of the complete data likelihood function. A two step EM strategy similar to [33] is followed in order to obtain compact update equations as in (8) and (9). Details of this derivation are provided in Appendix A. The new values of $\tilde{\pi}_g$ and $\tilde{\mu}_g$ are obtained through equations (4)-(7) as before. Update equations for $\tilde{\mathbf{W}}_g$ and $\tilde{\mathbf{\Psi}}_g$ are as follows:

$$\tilde{\mathbf{W}}_g = \mathbf{S}_g\mathbf{\Psi}_g^{-1}\mathbf{W}_g\left[\mathbf{I} + \mathbf{M}_g^{-1}\mathbf{W}_g^T\mathbf{\Psi}_g^{-1}\mathbf{S}_g\mathbf{\Psi}_g^{-1}\mathbf{W}_g\right]^{-1} \text{and} \tag{11}$$

$$\tilde{\mathbf{\Psi}}_g = \mathrm{diag}\left(\mathbf{S}_g - \mathbf{S}_g\mathbf{\Psi}_g^{-1}\mathbf{W}_g\mathbf{M}_g^{-1}\tilde{\mathbf{W}}_g^T\right), \tag{12}$$

where

$$\mathbf{M}_g = \mathbf{I}_q + \mathbf{W}_g^T\mathbf{\Psi}_g^{-1}\mathbf{W}_g. \tag{13}$$

The diag$(\cdot)$ operation in (12) retains only the diagonal elements of the matrix. In this case, the posterior distribution of the acoustic factors for the $g$-th mixture is given by:

$$p(\mathbf{y}_n|\mathbf{x}_n, g) = \mathcal{N}\left(\mathbf{y}_n|\mathbf{M}_g^{-1}\mathbf{W}_g^T\mathbf{\Psi}_g^{-1}(\mathbf{x}_n - \mu_g), \mathbf{M}_g^{-1}\right). \tag{14}$$

As before, the updated model covariance is obtained using (3) and (4)–(7) are utilized in the next iteration with new parameters. We denote this variant of the model as: $\text{ML-AFA}_{\text{diag}}$.

### D. i-Vector Extraction

Conventionally, the i-Vectors are extracted using the zero and first order statistics calculated from the features with respect to the UBM model. Next, as we replace the UBM model with the AFA model (isotropic/diagonal), it is still possible to proceed as before by computing the statistics in the traditional way considering the model as a GMM with parameters $\Lambda = \{\pi_g, \mu_g, \mathbf{C}_g\}$. In this case, the model covariance matrices $\mathbf{C}_g$ are restricted depending on the type of model used (isotropic/diagonal). As an alternative, we propose to model the acoustic factors for each of the mixtures as input to the next stage of the factor analyzer (i.e., the i-Vector extractor). This is motivated by the assumption that the variation in the acoustic factors contain the most important speaker dependent information. In this approach, we essentially develop a two stage factor analysis scheme for speaker verification, where the second stage (i-Vector extractor) utilizes the posterior mean and covariance matrices of the hidden variables (acoustic factors) of the first modeling stage. Later in Section IV-A, we experimentally show that modeling the acoustic factors in this way provides superior performance.

The proposed strategy is somewhat similar to the Deep Mixture of Factor Analyzers (DMFA) approach [36], where the later stage of factor analyzer uses the posterior mean of the latent factors obtained from the earlier stage as features. However, in the current scenario, the second stage of the factor analyzer is trained at the *utterance* level, whereas the first stage is trained at the *frame* level.

Proceeding with the above method, for an utterance $s$, the zero order statistics are extracted as:

$$N_s(g) = \sum_{n \in s} \gamma_n(g), \tag{15}$$

which follows the conventional approach [5]. Here, $\gamma_n(g)$ is extracted as in (4) utilizing model parameters $\Lambda$. Conventionally, the first order statistics are extracted as:

$$\mathbf{F}_s(g) = \sum_{n \in s} \gamma_n(g)\mathbf{x}_n. \tag{16}$$

Using the proposed model, the first order statistics are extracted as:

$$\begin{aligned}\hat{\mathbf{F}}_s(g) &= \sum_{n \in s} \gamma_n(g)\mathbf{A}_g^T(\mathbf{x}_n - \mu_g) \\ &= \mathbf{A}_g^T\left[\mathbf{F}_s(g) - N_s(g)\mu_g\right] = \mathbf{A}_g^T\bar{\mathbf{F}}_s(g),\end{aligned} \tag{17}$$

where $\mathbf{A}_g^T = \mathbf{M}_g^{-1}\mathbf{W}_g^T$ for the isotropic model and $\mathbf{M}_g^{-1}\mathbf{W}_g^T\mathbf{\Psi}_g^{-1}$ for the diagonal model (using appropriate definitions of $\mathbf{M}_g$ in each case). Also, $\bar{\mathbf{F}}_s(g)$ represents the centralized first order statistics computed using the model parameters $\Lambda$.

The remaining procedure for the i-Vector extractor/total variability matrix training follows the exact same principles as outlined in [7]. However, when the acoustic factors $\mathbf{y}_n$ are used as features for the i-Vector extractor, the mean and covariance for the UBM parameter is set to $(\mathbf{0}, \mathbf{I})$, following the original definition of the term.

### E. Model Interpretation

In order to gain further insight towards understanding the mechanisms of the proposed method, we aim to compare the super-covariance matrices (covariance matrices obtained from GMM mean super-vectors) in an i-Vector system using the conventional approach and the proposed AFA integrated approach. This will illustrate the effect of using the acoustic factors as features in the total variability model.

Using the total variability model, for a randomly chosen utterance $s$, the GMM super-vector $\mathbf{m}_s$ can be represented by,

$$\mathbf{m}_s = \mathbf{m}_0 + \mathbf{T}\mathbf{w}_s, \tag{18}$$

where $\mathbf{m}_0 \in \mathbb{R}^{Md}$ is the speaker independent mean super-vector (i.e., concatenated UBM mean vectors $\mu_g$), $\mathbf{T}$ is an $Md \times R$ rectangular matrix ($R < Md$) of low rank whose columns span the so called *total variability* space [5], and $\mathbf{w}_s \in \mathbb{R}^R$ is a standard normal random vector, known as the *total factors*. The posterior mean vector of $\mathbf{w}_s$ given an utterance data is considered as an i-Vector. In this model, the covariance matrix of $\mathbf{m}_s$ is given by, $\mathbf{B} = \mathbf{T}\mathbf{T}^T$. Since, it is known that i-Vectors are effective lower dimensional representations of the GMM super-vectors, we are interested in observing this approximate super-covariance matrix for specific mixture components. The training data and algorithms used for the UBM and $\mathbf{T}$ matrix is provided in Section III-D and III-E, respectively. A full-covariance GMM-UBM is used in this analysis.

In Fig. 3(a) the estimated super-covariance matrix for the first mixture is shown. In other words, this is the first sub-matrix of $\mathbf{B}$ including $d \times d$ components from the upper left corner. From Fig. 3(a), we observe the following: (i) the covariance matrix in this part and the covariance of the UBM in Fig. 3(c), are not the same, and thus a factor analysis model in these two domains are not equivalent [7]; (ii) a strong peak is observed near the component (20,20) of the matrix (this pattern is observed in other mixture blocks of the matrix $\mathbf{B}$ as well). This indicates that strong correlations are present in specific feature components of the GMM super-vectors, $\mathbf{m}_s$, collected over a large number of utterances.

When an AFA model (isotropic noise with $q = 42$) is utilized and the acoustic factors are the inputs to the total variability model, the partial super-covariance matrix $\mathbf{B}$ is shown in Fig. 3(b). Interestingly, this matrix does not contain any dominant peaks as observed in Fig. 3(a). This further justifies the inclusion of the first stage factor analyzer which takes into account the correlation among feature coefficients (independent of

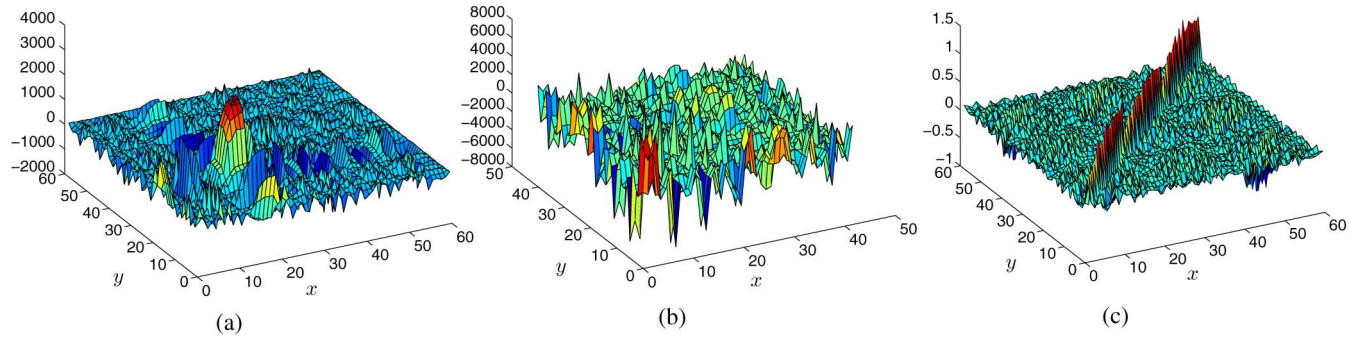


Fig. 3. Partial super-covariance matrices and a UBM covariance matrix obtained from a GMM and AFA model. The super-covariance is estimated using the *total variability* matrix $\mathbf{T}$. (a) Partial super-covariance matrix of mixture-1 for a full covariance GMM-UBM. (b) Partial super-covariance matrix of mixture-1 for an $\mathrm{AFA_{iso}}$ UBM model ($q = 42$). (c) The full-covariance matrix of the GMM-UBM obtained from mixture-1.

the utterance) and provides a de-correlated input (acoustic factors) to the second stage. These two stages are thus complimentary in nature, and can be expected to provide superior results.

## III. System Description

The experiments performed in this study are based on the male trials of the NIST SRE 2012 evaluation. A standard i-Vector system [5] with a Gaussian PLDA [18] similar to our NIST SRE 2012 submission [29] is used as a baseline system. Specific blocks of the baseline system implementation and details of the proposed scheme are described below.

### A. Voice Activity Detection

The VAD algorithm follows the method in [38], available through the open-source Voicebox toolkit [39]. For interview recordings, VAD is performed on both interviewee (A) and interviewer (B) channels, and speech segments detected in channel B are removed from channel A. Since channel B is usually corrupted by a noise floor to mask the interviewee speech, spectral subtraction [22] is always performed before VAD on channel B. For channel A, first the Signal to Noise Ratio (SNR) is estimated using a 2-mixture GMM trained on segment energy. If the SNR is less than 18 dB, the audio channel is enhanced using spectral subtraction [22] before application of VAD. Here, the noise power was estimated using the method outlined in [40].

### B. Feature Extraction

We use 60 dimensional Mel-Frequency Cepstral Coefficients (MFCC) features. At first, digital zeros are replaced by a uniformly distributed noise floor having a mean zero and amplitude $1.75^{-5}$. A 24 channel Mel-spaced filterbank is used and 19 components are retained. The 60 dimensional features are obtained by including log-energy, delta and acceleration coefficients using a 25 ms analysis window with 10 ms frame shift. Finally, the features are processed through Cepstral Mean and Variance Normalization (CMVN) utilizing a 3-sec sliding window.

### C. Noisy File Generation

Since our experiments are performed on the NIST SRE 2012 tasks, we artificially noised our development dataset. We collected 10 HVAC noise files from [41] and generated 10 crowd noise files by summing 500–800 NIST SRE utterances from both male and female speakers. The noise partitioning is described in [29]. We employ our in-house tools to generate the noisy files with a psophometric weighting (ITU-T Recommendation O.41) method as suggested by NIST. The active speech level is measured according to the ITU-T Recommendation P.56. These noisy files are used for speaker enrollment, hyper-parameter estimation, and PLDA training.

TABLE I
Common Conditions in NIST SRE 2012

| No. | Train | Test |
|---|---|---|
| 1 | Multiple speech segments | Clean interview speech |
| 2 | | Clean phone call speech |
| 3 | | Noisy interview speech |
| 4 | | Noisy phone call speech |
| 5 | | Phone call speech collected in noise |

TABLE II
UBM Training List Description for NIST SRE 2012. Number of Files Used in Different Categories are Presented for both Genders

| | Male | Female |
|---|---|---|
| known/unknown | 6000/4323 | 6000/4998 |
| Tel/Int/Mic | 6000/2967/1356 | 6000/2940/2058 |
| Clean/HVAC/crowd | 4941/2687/2695 | 5166/2910/2922 |
| Total | 10323 | 10998 |

### D. UBM and AFA Model Training

Gender dependent 1024-mixture UBMs with full-covariance and the proposed AFA models are trained on telephone utterances selected from the Switchboard-II Phase 2 and 3, Switchboard Cellular Part 1 and 2, and the NIST SRE 2004–06 enrollment data. Noisy files containing HVAC and crowd noise, and SRE 2012 enrollment speaker data are also included in the UBM. The UBM utterances are approximately balanced across: (i) clean vs. noisy, (ii) telephone vs. interview/microphone, and (iii) known vs. unknown speakers. The number of utterances used in UBM training from various data types is summarized in Table II. We employed data sub-sampling for fast UBM training [42], [43] to perform the experiments. For each 30 frames that are skipped, 3 consecutive frames are selected, resulting in use of only 10% of the original dataset. In this way, the correlation among the successive frames are retained. For EM training, the

TABLE III
COMPARISON OF SYSTEM PERFORMANCE WHEN THE PROPOSED MODELS ARE USED AS GMMS VS. AFAS FOR THE I-VECTOR SYSTEM. RESULTS ARE SHOWN FOR FIVE NIST SRE 2012 COMMON CONDITIONS OF THE EXTENDED TRIALS (MALE)

| Model | Method | %EER | | | | | $\min C_{\text{primary}}$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | cc-1 | cc-2 | cc-3 | cc-4 | cc-5 | cc-1 | cc-2 | cc-3 | cc-4 | cc-5 |
| GMM-diag | GMM | 3.243 | 2.819 | 3.127 | 3.113 | 3.228 | 0.264 | 0.312 | 0.130 | 0.271 | 0.307 |
| GMM-full | GMM | 3.302 | 3.714 | 3.328 | 3.770 | 4.142 | 0.273 | 0.378 | 0.137 | 0.318 | 0.354 |
| ML-AFA$_{\text{iso}}(q=42)$ | GMM | 3.375 | 3.924 | 3.261 | 3.948 | 4.348 | 0.270 | 0.392 | 0.130 | 0.327 | 0.381 |
| ML-AFA$_{\text{diag}}(q=42)$ | GMM | 3.522 | 3.717 | 3.213 | 3.946 | 4.143 | 0.271 | 0.390 | 0.125 | 0.317 | 0.368 |
| ML-AFA$_{\text{iso}}(q=42)$ | AFA | 3.298 | 2.642 | 3.118 | 3.007 | 3.080 | 0.245 | 0.304 | 0.123 | 0.260 | 0.294 |
| ML-AFA$_{\text{diag}}(q=42)$ | AFA | 2.993 | 2.655 | 3.242 | 2.928 | 3.027 | 0.221 | 0.291 | 0.107 | 0.257 | 0.282 |

initial four iterations per mixture are gradually increased to 15 for higher order mixtures.

### E. i-Vector Extractor Training

For training the i-Vector extractor, the UBM training dataset and additional SRE 2012 target speakers' data are used (both clean and noisy versions). This corresponds to our NIST SRE 2012 system [29]. Here, 600-dimensional i-Vectors are extracted using 5 EM iterations. The i-Vectors are first mean normalized and then length normalized using radial Gaussianization [18]. Linear Discriminant Analysis (LDA) projection is performed to further reduce the i-Vector dimension to 150 before PLDA scoring.

### F. PLDA Classifier

*1) Model Training:* In this work, we use a Gaussian PLDA model with a full-covariance residual noise [18]. According to this model, an $R$ dimensional i-Vector $\mathbf{w}_s$ extracted from utterance $s$ is expressed as:

$$\mathbf{w}_s = \mathbf{w}_0 + \mathbf{\Phi}\beta + \mathbf{n}. \tag{19}$$

Here, $\mathbf{w}_0 \in \mathbb{R}^R$ is the speaker independent mean vector, $\mathbf{\Phi}$ is the $R \times N_{\text{ev}}$ low rank matrix representing the speaker dependent basis functions or eigenvoices, $\beta \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is an $N_{\text{ev}} \times 1$ hidden variable, and $\mathbf{n} \in \mathbb{R}^R$ is a random vector representing the full covariance residual noise. The model parameter $N_{\text{ev}}$ was set to 150. The data used for i-Vector extractor training are utilized to train this PLDA model. No short duration utterances are included in PLDA training as was the case in [29].

*2) Scoring:* The i-Vectors obtained from each enrollment speaker are first averaged so that one i-Vector per speaker is obtained. The scoring is then performed as described in [44]. To determine if i-Vectors $\mathbf{w}_i$ and $\mathbf{w}_j$ are obtained from the same speaker or not, we evaluate the following likelihood ratio:

$$\mathcal{L}_{i,j} = \frac{P(\mathbf{w}_i, \mathbf{w}_j | H_1)}{P(\mathbf{w}_i | H_0) P(\mathbf{w}_j | H_0)}. \tag{20}$$

The obtained scores are transformed using a compound log-likelihood ratio (LLR) transformation as described in [45]. For this purpose, we set the target prior $P_{\text{known}} = 0.5$ and assume that all speakers are equally likely. We note that, the compound LLR is used only when individual system performances are reported. For fusion of multiple systems, the compound LLR transformation is applied on the final fused scores.

## IV. EXPERIMENTAL RESULTS

The experiments performed in this study are based on the male portion of the NIST SRE 2012 core-extended trials. We use the SRE 2012 detection cost functions (DCF): $C_{\text{primary}}$, $\min C_{\text{primary}}(P_{\text{known}} = 0.5)$ [13], and the % Equal Error Rate (EER) metric for evaluating the systems [1]. We report the results on the five common conditions of the NIST SRE 2012 extended trials [13]. Definitions of these common conditions are provided in Table I. Note that the clean conditions also contain transmission channel/microphone variability.

### A. Effect of the Modeling Method

We are interested in analyzing system performance of an AFA-UBM model when it is used as a GMM with the parameters $\Lambda = \{\pi_g, \mu_g, \mathbf{C}_g\}$. This means that the effect of the AFA modeling will only be observed in the way the covariance matrix is restricted in (3). The results of these experiments are summarized in Table III. We note here that the full-covariance GMM-UBM based system does not perform as well as the diagonal GMM-UBM. The AFA-UBM models utilized as GMMs (row 4-5 in Table III) are seen to perform close to the full-covariance GMM-UBM. However, when the acoustic factors are utilized for the i-Vector modeling (noted as the AFA method in Table III) we observe a significant improvement in system performance. This confirms our original motivation for using the acoustic factors as inputs to the i-Vector extractor.

### B. Variation of Acoustic Factor Dimension

In this experiment, we intend to observe the effect of changing the acoustic factor dimension on overall system performance. For both model types (ML-AFA$_{\text{iso}}$ and ML-AFA$_{\text{diag}}$), we consider acoustic factor dimensions of: $q = 42$, 48 and 54. These parameter values correspond to 70,80 and 90% of the original 60 dimensional features. The results are provided in Tables IV–VIII obtained from both baseline and proposed systems in the five NIST SRE 2012 common conditions.

The results in Tables IV–VIII clearly demonstrate that, the proposed technique utilizing an AFA-UBM instead of the conventional GMM-UBM provides more robust speaker recognition performance across conditions including clean and noisy test utterances. Except for condition-3 (i.e., the noisy interview

[1]It has been argued that the EER metric is not valid when known non-target speakers are involved during test [45]. However, we still report this performance metric as it is a widely known and understood measure in speaker verification.

TABLE IV
PERFORMANCE COMPARISON BETWEEN BASELINE AND PROPOSED SYSTEMS IN NIST SRE 2012 EXTENDED TRIALS CONDITION-1

| **UBM model** | | %EER | $\min C_{\text{primary}}$ | $C_{\text{primary}}$ |
|---|---|---|---|---|
| GMM-diag | | 3.2428 | 0.2642 | 0.3385 |
| GMM-full | | 3.3020 | 0.2729 | 0.3553 |
| *Method* | $q$ | *Absolute/%relative performance* | | |
| ML-AFA$_{\text{iso}}$ | 42 | 3.298/-1.7 | 0.245/7.3 | 0.336/0.7 |
| | 48 | **2.779/14.3** | 0.241/8.6 | 0.334/1.4 |
| | 54 | 2.874/11.4 | 0.236/10.8 | 0.326/3.8 |
| ML-AFA$_{\text{diag}}$ | 42 | 2.993/7.7 | **0.221/16.5** | **0.316/6.6** |
| | 48 | 3.008/7.2 | 0.242/8.3 | 0.339/-0.2 |
| | 54 | 2.951/9.0 | 0.237/10.4 | 0.331/2.2 |

TABLE V
PERFORMANCE COMPARISON BETWEEN BASELINE AND PROPOSED SYSTEMS IN NIST SRE 2012 EXTENDED TRIALS CONDITION-2

| **UBM model** | | %EER | $\min C_{\text{primary}}$ | $C_{\text{primary}}$ |
|---|---|---|---|---|
| GMM-diag | | 2.8190 | 0.3122 | 0.5482 |
| GMM-full | | 3.7135 | 0.3776 | 0.5863 |
| *Method* | $q$ | *Absolute/%relative performance* | | |
| ML-AFA$_{\text{iso}}$ | 42 | 2.642/6.3 | 0.304/2.5 | 0.541/1.4 |
| | 48 | **2.469/12.4** | 0.286/8.3 | **0.529/3.5** |
| | 54 | 2.596/7.9 | 0.285/8.7 | 0.529/3.5 |
| ML-AFA$_{\text{diag}}$ | 42 | 2.655/5.8 | 0.291/6.7 | 0.536/2.2 |
| | 48 | 2.632/6.6 | 0.289/7.3 | 0.530/3.4 |
| | 54 | 2.553/9.4 | **0.278/10.9** | 0.532/3.0 |

TABLE VI
PERFORMANCE COMPARISON BETWEEN BASELINE AND PROPOSED SYSTEMS IN NIST SRE 2012 EXTENDED TRIALS CONDITION-3

| **UBM model** | | %EER | $\min C_{\text{primary}}$ | $C_{\text{primary}}$ |
|---|---|---|---|---|
| GMM-diag | | 3.1273 | 0.1299 | 0.1421 |
| GMM-full | | 3.3280 | 0.1367 | 0.1460 |
| *Method* | $q$ | *Absolute/%relative performance* | | |
| ML-AFA$_{\text{iso}}$ | 42 | **3.118/0.3** | 0.123/5.5 | 0.134/5.7 |
| | 48 | 3.173/-1.5 | 0.113/13.4 | **0.123/13.4** |
| | 54 | 3.178/-1.6 | 0.114/12.4 | 0.124/13.0 |
| ML-AFA$_{\text{diag}}$ | 42 | 3.242/-3.7 | **0.107/17.9** | 0.126/11.3 |
| | 48 | 3.252/-4.0 | 0.128/1.7 | 0.145/-2.3 |
| | 54 | 3.122/0.2 | 0.112/13.6 | 0.125/11.8 |

TABLE VII
PERFORMANCE COMPARISON BETWEEN BASELINE AND PROPOSED SYSTEMS IN NIST SRE 2012 EXTENDED TRIALS CONDITION-4

| **UBM model** | | %EER | $\min C_{\text{primary}}$ | $C_{\text{primary}}$ |
|---|---|---|---|---|
| GMM-diag | | 3.1130 | 0.2705 | 0.4488 |
| GMM-full | | 3.7704 | 0.3175 | 0.4841 |
| *Method* | $q$ | *Absolute/%relative performance* | | |
| ML-AFA$_{\text{iso}}$ | 42 | 3.007/3.4 | 0.260/4.0 | 0.445/0.9 |
| | 48 | 2.952/5.2 | 0.265/2.2 | 0.443/1.4 |
| | 54 | 3.007/3.4 | 0.266/1.8 | 0.452/-0.6 |
| ML-AFA$_{\text{diag}}$ | 42 | 2.928/6.0 | 0.257/5.1 | 0.450/-0.2 |
| | 48 | 3.119/-0.2 | 0.256/5.4 | 0.439/2.2 |
| | 54 | **2.757/11.4** | **0.247/8.6** | **0.437/2.6** |

TABLE VIII
PERFORMANCE COMPARISON BETWEEN BASELINE AND PROPOSED SYSTEMS IN NIST SRE 2012 EXTENDED TRIALS CONDITION-5

| **UBM model** | | %EER | $\min C_{\text{primary}}$ | $C_{\text{primary}}$ |
|---|---|---|---|---|
| GMM-diag | | 3.2276 | 0.3072 | 0.5941 |
| GMM-full | | 4.1415 | 0.3537 | 0.6243 |
| *Method* | $q$ | *Absolute/%relative performance* | | |
| ML-AFA$_{\text{iso}}$ | 42 | 3.080/4.6 | 0.294/4.4 | 0.582/2.0 |
| | 48 | **2.848/11.8** | 0.275/10.6 | **0.571/3.9** |
| | 54 | 3.105/3.8 | **0.263/14.3** | 0.575/3.3 |
| ML-AFA$_{\text{diag}}$ | 42 | 3.027/6.2 | 0.282/8.1 | 0.584/1.7 |
| | 48 | 3.039/5.8 | 0.281/8.6 | 0.574/3.4 |
| | 54 | 2.850/11.7 | 0.269/12.4 | 0.578/2.7 |

case), the proposed methods provide significantly superior performance compared to the baseline system in all three performance metrics. In general, relative improvements on the order of 5–10% is obtained using the proposed methods. This improved robustness in both clean and noisy conditions justify our motivation for utilizing the ML-AFA models in place of conventional GMM-UBMs, especially since the proposed models attempt to remove the noise in an earlier stage of the system (i.e., within acoustic feature models rather than utterance models).

From the performance evaluations of Tables IV–VIII, it is apparent that a single AFA model parameter (acoustic factor dimension $q$) or model type (isotropic or diagonal) does not always provide the best result in all conditions. This indicates that an optimal selection of the parameter $q$ in each mixture can provide further benefits [46], [47]. In our previous work, we attempted to derive an automatic selection of the parameter $q$ in [31] using the AFA frame-work proposed in [7].

### C. System Fusion and Calibration

In order to test if the proposed systems can provide complementary information, we perform fusion of several systems using a linear logistic regression method obtained from the Bosaris toolkit [48]. An independent development test set is utilized for training the calibration and fusion parameters. The data-set referred to as the *Eval-Test* in [49] is used here as the development test set. This data-set contains utterances from the enrollment speakers so that target and known non-target trials are present. Also, held out speaker data is included to provide unknown non-target trials. Clean and noisy versions of telephone, interview and microphone recordings are included in this data-set (noise types are HVAC and crowd, following SRE 2012 test data). For training fusion and calibration, we used 15 iterations and an effective prior of 0.001. We select the following systems for fusion: (1) Baseline with diagonal covariance UBM, (2) Baseline with full-covariance UBM, (3) ML-AFA$_{\text{iso}}^{q=48}$ and (4) ML-AFA$_{\text{diag}}^{q=48}$. Three combinations of these systems are used for fusion and the results are summarized in Table IX.

From these results, we observe the complementary nature of the proposed and baseline systems, yielding significant relative improvements of about 20–25% over the baseline system performance with respect to the primary cost metric $C_{\text{primary}}$. For the metric $\min C_{\text{primary}}$, very similar improvements are

TABLE IX
FUSION PERFORMANCE OF BASELINE AND PROPOSED SYSTEMS. ABSOLUTE AND % RELATIVE PERFORMANCE IS SHOWN FOR FUSION SYSTEMS

| ID | System | $\min C_{\text{primary}}$ | | | | | $C_{\text{primary}}$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | cc-1 | cc-2 | cc-3 | cc-4 | cc-5 | cc-1 | cc-2 | cc-3 | cc-4 | cc-5 |
| 1 | GMM-diag | 0.269 | 0.327 | 0.132 | 0.301 | 0.333 | 0.345 | 0.558 | 0.143 | 0.456 | 0.603 |
| 2 | GMM-full | 0.280 | 0.389 | 0.139 | 0.354 | 0.378 | 0.363 | 0.600 | 0.148 | 0.496 | 0.640 |
| 3 | ML-AFA$^{q=48}_{\text{iso}}$ | 0.244 | 0.304 | 0.115 | 0.298 | 0.300 | 0.344 | 0.540 | 0.129 | 0.452 | 0.582 |
| 4 | ML-AFA$^{q=48}_{\text{diag}}$ | 0.245 | 0.301 | 0.130 | 0.282 | 0.305 | 0.348 | 0.540 | 0.149 | 0.450 | 0.585 |
| $\text{Fusion}_{1-3}$ | LR *(Abs.)* | 0.240 | 0.286 | 0.115 | 0.285 | 0.284 | 0.301 | 0.483 | 0.120 | 0.403 | 0.529 |
| | CLR* *(Abs.)* | 0.231 | 0.261 | 0.110 | 0.244 | 0.247 | 0.262 | 0.425 | 0.115 | 0.352 | 0.469 |
| | CLR *(% Rel.)* | 16.0 | 25.5 | 19.8 | 23.3 | 34.6 | 32.1 | 31.2 | 24.6 | 29.6 | 28.7 |
| $\text{Fusion}_{1,2,4}$ | LR *(Abs.)* | 0.241 | 0.285 | 0.126 | 0.275 | 0.289 | 0.301 | 0.479 | 0.128 | 0.393 | 0.524 |
| | CLR* *(Abs.)* | 0.231 | 0.267 | 0.119 | 0.238 | 0.246 | 0.263 | 0.424 | 0.121 | 0.349 | 0.468 |
| | CLR *(% Rel.)* | 16.0 | 22.5 | 10.6 | 26.7 | 35.1 | 31.2 | 31.7 | 18.2 | 30.5 | 28.7 |
| $\text{Fusion}_{1-4}$ | LR *(Abs.)* | 0.238 | 0.276 | 0.117 | 0.274 | 0.276 | 0.298 | 0.471 | 0.121 | 0.393 | 0.516 |
| | CLR* *(Abs.)* | 0.231 | 0.257 | 0.109 | 0.236 | 0.240 | 0.258 | 0.416 | 0.116 | 0.346 | 0.459 |
| | CLR *(% Rel.)* | 14.1 | 21.4 | 17.4 | 21.6 | 27.9 | 25.2 | 25.4 | 18.9 | 24.1 | 23.9 |

*CLR indicates compound likelihood transformed scores.

TABLE X
COMPARISON OF SYSTEM PERFORMANCE BETWEEN ORIGINAL AFA AND ML-AFA BASED I-VECTOR SYSTEMS. RESULTS ARE SHOWN FOR FIVE NIST SRE 2012 COMMON CONDITIONS OF THE EXTENDED TRIALS (MALE)

| Method | %EER | | | | | $\min C_{\text{primary}}$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | cc-1 | cc-2 | cc-3 | cc-4 | cc-5 | cc-1 | cc-2 | cc-3 | cc-4 | cc-5 |
| $\text{AFA}_{\text{iso}}(q=48)$ [7] | 3.332 | 4.249 | 3.299 | 4.893 | 4.790 | 0.317 | 0.408 | 0.168 | 0.376 | 0.412 |
| $\text{AFA}_{\text{iso}}(q=54)$ [7] | 3.842 | 3.780 | 3.440 | 4.460 | 4.600 | 0.278 | 0.386 | 0.151 | 0.354 | 0.394 |
| $\text{ML-AFA}_{\text{iso}}(q=48)$ | 2.779 | 2.469 | 3.173 | 2.952 | 2.848 | 0.241 | 0.286 | 0.113 | 0.265 | 0.275 |
| $\text{ML-AFA}_{\text{iso}}(q=54)$ | 2.874 | 2.596 | 3.178 | 3.007 | 3.105 | 0.236 | 0.285 | 0.114 | 0.266 | 0.263 |

observed for all five conditions. Since the $C_{\text{primary}}$ cost function is the most important metric for the NIST SRE 2012, the results obtained are quite encouraging. Fusion of system 1–3 and 1,2,4, in general provided better results compared to fusing all 4 systems. This indicates that the systems $\text{ML-AFA}^{q=48}_{\text{iso}}$ and $\text{ML-AFA}^{q=48}_{\text{diag}}$ may not fuse well for conditions but can provide significant improvement when individually fused with the baseline systems. However, improvements obtained by fusing all four systems are more uniform. For example, in cc-3, relative improvements in $\min C_{\text{primary}}$ from $\text{Fusion}_{1-3}$ and $\text{Fusion}_{1,2,4}$ are 19.8% and 10.6%, respectively, while $\text{Fusion}_{1-4}$ yields 17.4%. Fusion of all four systems, always improve the performance by at least 14%.

### D. Comparison Between Original AFA and ML-AFA

As the final experiment, in this section, we compare the original AFA approach presented in [7] and the proposed ML-AFA strategy within the current i-Vector system framework. In [7], only an isotropic residual noise was considered in the model formulation. Thus, we use the $\text{ML-AFA}_{\text{iso}}$ model as the UBM in this comparison. The original AFA method is implemented as described in [7], deriving its parameters from a full-covariance UBM. The acoustic factor dimensions $q = 42$ and 48 are considered, and %EER and $\min C_{\text{primary}}$ performance metrics are used for this experiment. The results are summarized Table X.

From these results, we observe that the ML-AFA strategy clearly outperforms the original AFA approach in all conditions. The performance difference is quite significant in conditions 2, 4 and 5. However, it must be noted that the AFA model in [7] was trained using only channel degraded data, whereas in this experiment, both channel degraded and noisy data is used. As discussed in the introduction, when the AFA parameters are extracted from a full-covariance UBM model trained on such data, the transforms estimated as in [7] do not provide improved results. On the other hand, when the iterative EM strategy is used to learn the AFA parameters, as in ML-AFA, the resulting systems provide improved robustness in both channel degraded and noisy test conditions.

## V. CONCLUSIONS

In this work, we have developed the acoustic factor analysis frame-work towards a generative mixture model as an alternative to a conventional GMM based UBM for speaker verification. The proposed modeling scheme was designed to iteratively learn a limited number of dominant feature sub-spaces in different mixture components using clean and noisy training data. Two variations of the proposed model were investigated, one with an isotropic and the other with a diagonal covariance residual noise assumption. The method was integrated within an i-Vector system frame-work where the hidden variables of the proposed model (i.e., *acoustic factors*), were used as input for *total variability* modeling. The interpretation and implication of the proposed method was discussed and analyzed. Extensive experiments were performed on both clean and noisy test conditions from the NIST SRE 2012 extended trials. The proposed methods were found to be superior in multiple noisy conditions in SRE 2012, providing a significant gain in performance when fusion of multiple systems were considered.

## APPENDIX

### A. EM Algorithm for Diagonal Covariance AFA

Given the mixture of AFA model, or more generally, the mixture of factor analyzers model in (1), first we obtain the posterior probability density functions (PDFs) of $\mathbf{x}_n$ given the latent variables $\mathbf{y}_n$. Here, we assume only one mixture component at this stage.

$$p(\mathbf{x}_n|\mathbf{y}_n)=\frac{\exp\left[-\frac{1}{2}(\mathbf{x}_n-\mathbf{W}\mathbf{y}_n-\mu)^T\mathbf{\Psi}^{-1}(\mathbf{x}_n-\mathbf{W}\mathbf{y}_n-\mu)\right]}{(2\pi)^{d/2}|\mathbf{\Psi}|^{\frac{1}{2}}},$$

$$p(\mathbf{y}_n)=(2\pi)^{-q/2}\exp\left(-\frac{1}{2}\mathbf{y}_n\mathbf{y}_n^T\right).$$

The model covariance matrix is given by, $\mathbf{C}=\mathbf{\Psi}+\mathbf{W}\mathbf{W}^T$. Thus, the data model is given by,

$$\begin{aligned}p(\mathbf{x}_n)&=\int p(\mathbf{x}_n|\mathbf{y}_n)p(\mathbf{y}_n)d\mathbf{y}_n\\&=\frac{\exp\left[-\frac{1}{2}(\mathbf{x}_n-\mu)^T\mathbf{C}^{-1}(\mathbf{x}_n-\mu)\right]}{(2\pi)^{d/2}|\mathbf{C}|^{\frac{1}{2}}}.\end{aligned}\tag{21}$$

Now, the posterior probability of the hidden variables $\mathbf{y}_n$ is:

$$\begin{aligned}p(\mathbf{y}_n|\mathbf{x}_n)&=\frac{p(\mathbf{x}_n|\mathbf{y}_n)p(\mathbf{y}_n)}{p(\mathbf{x}_n)}\\&=\frac{(2\pi)^{-q/2}}{|\mathbf{C}|^{-\frac{1}{2}}|\mathbf{\Psi}|^{\frac{1}{2}}}\exp\left[-\frac{1}{2}(\mathbf{y}_n-\mathbf{M}^{-1}\mathbf{W}^T(\mathbf{x}_n-\mu))^T\right.\\&\quad\left.\times\mathbf{M}(\mathbf{y}_n-\mathbf{M}^{-1}\mathbf{W}^T(\mathbf{x}_n-\mu))\right].\end{aligned}\tag{22}$$

where

$$\mathbf{M}=\mathbf{I}_q+\mathbf{W}^T\mathbf{\Psi}^{-1}\mathbf{W}.\tag{23}$$

It can be shown that, $|\mathbf{M}|=|\mathbf{\Psi}^{-1}\mathbf{C}|=|\mathbf{I}_q+\mathbf{\Psi}^{-1}\mathbf{W}\mathbf{W}^T|$. Thus, from (22) we observe that $\mathbf{M}^{-1}$ is the posterior covariance of $\mathbf{y}_n$, providing its Gaussian PDF:

$$p(\mathbf{y}_n|\mathbf{x}_n)=\mathcal{N}(\mathbf{M}^{-1}\mathbf{W}^T(\mathbf{x}_n-\mu),\mathbf{M}^{-1}).\tag{24}$$

Thus, the first and second order moment of $\mathbf{y}_n$ given $\mathbf{x}_n$ is given by:

$$\langle\mathbf{y}_n\rangle=\mathbf{M}^{-1}\mathbf{W}^T\mathbf{\Psi}^{-1}(\mathbf{x}_n-\mu)\tag{25}$$

$$\langle\mathbf{y}_n\mathbf{y}_n^T\rangle=\mathbf{M}^{-1}+\langle\mathbf{y}_n\rangle\langle\mathbf{y}_n\rangle^T\tag{26}$$

For a mixture of AFA model, these moments will have different values in each mixture component. We denote them as $\langle\mathbf{y}_{ng}\rangle$ and $\langle\mathbf{y}_{ng}\mathbf{y}_{ng}^T\rangle$, respectively. For a mixture of factor analyzers, the expected complete-data log likelihood plus the Lagrangian multiplier is given by,

$$\begin{aligned}&\sum_{n=1}^{N}\sum_{g=1}^{M}\gamma_n(g)\left[\ln\pi_g-\frac{1}{2}\ln|\mathbf{\Psi}_g|-\frac{1}{2}\mathrm{tr}\left(\langle\mathbf{y}_{ng}\mathbf{y}_{ng}^T\rangle\right)-\right.\\&\frac{1}{2}\mathrm{tr}\left(\mathbf{\Psi}_g^{-1}(\mathbf{x}_n-\mu_g)(\mathbf{x}_n-\mu_g)^T\right)+\\&\left.\langle\mathbf{y}_{ng}\rangle^T\mathbf{W}_g^T\mathbf{\Psi}_g^{-1}(\mathbf{x}_n-\mu_g)-\frac{1}{2}\mathrm{tr}\left(\mathbf{W}_g^T\mathbf{\Psi}_g^{-1}\mathbf{W}_g\langle\mathbf{y}_{ng}\mathbf{y}_{ng}^T\rangle\right)\right]\\&+\lambda\left(\sum_{g=1}^{M}\pi_g-1\right).\end{aligned}\tag{27}$$

We note that the last term in the equation with the Lagrangian multiplier is required to constrain the sum of the mixture weights $\pi_g$ to be equal to unity. Differentiation of (27) with respect to $\pi_g$ and $\lambda$, setting to zero, and solving the equations provide the new value of the model parameter $\tilde{\pi}_g$ given by,

$$\tilde{\pi}_g=\frac{1}{N}\sum_{n=1}^{N}\gamma_n(g).\tag{28}$$

Next, maximizing (27) with respect to $\mu_g$, we obtain,

$$\tilde{\mu}_g=\frac{\sum_{n=1}^{N}\gamma_n(g)\left(\mathbf{x}_n-\tilde{\mathbf{W}}_g\langle\mathbf{y}_{ng}\rangle\right)}{\sum_{n=1}^{N}\gamma_n(g)}.\tag{29}$$

Maximizing for $\mathbf{W}_g$ we obtain its update equation,

$$\tilde{\mathbf{W}}_g=\left[\sum_{n=1}^{N}\gamma_n(g)(\mathbf{x}_n-\tilde{\mu}_g)\langle\mathbf{y}_{ng}\rangle^T\right]\left(\sum_{n=1}^{N}\gamma_n(g)\langle\mathbf{y}_{ng}\mathbf{y}_{ng}^T\rangle\right)^{-1}.\tag{30}$$

Finally, differentiating (27) with respect to $\mathbf{\Psi}_g^{-1}$ and setting to zero gives the update equation for $\mathbf{\Psi}_g$ as,

$$\begin{aligned}\tilde{\mathbf{\Psi}}_g=\frac{1}{N\tilde{\pi}_g}\sum_{n=1}^{N}[&(\mathbf{x}_n-\mu_g)(\mathbf{x}_n-\mu_g)^T\\&-2(\mathbf{x}_n-\mu_g)\langle\mathbf{y}_{ng}\rangle^T\tilde{\mathbf{W}}_g^T+\tilde{\mathbf{W}}_g\langle\mathbf{y}_{ng}\mathbf{y}_{ng}^T\rangle\tilde{\mathbf{W}}_g^T].\end{aligned}\tag{31}$$

These solutions are very similar to what was obtained in the PPCA case, except for the noise covariance term $\mathbf{\Psi}_g$, which is now considered to be a diagonal matrix for optimization[2].

As the M-step update equations of $\tilde{\mathbf{W}}_g$ and $\tilde{\mathbf{\Psi}}_g$ (i.e. Eqs. (30) and (31)), are coupled, we proceed in the same way as in [33]. First, we ignore the latent variables $\mathbf{y}_n$ and maximize the likelihood function for $\mu_g$ and $\pi_g$. This gives us the update equations (5) and (6) as in the isotropic noise case. Next, to update $\mathbf{W}_g$ and $\mathbf{\Psi}_g$, we only seek to increase the likelihood function instead of maximizing it, which is in principal similar to the Generalized Expectation Maximization (GEM) method. The parameters $\tilde{\mu}_g$ and $\tilde{\pi}_g$ are assumed to be fixed. Also, the statistics $\langle\mathbf{y}_{ng}\rangle$ and $\langle\mathbf{y}_{ng}\mathbf{y}_{ng}^T\rangle$ are obtained from the estimated parameters in the first step using equations (25) and (26) for each mixture. In this case, the parameters $\mathbf{W}$, $\mathbf{\Psi}$ and $\mathbf{M}$ are also considered mixture dependent. Now, when the maximization is carried out assuming these parameters as pre-computed constants, we obtain a new set of simplified update equations for $\tilde{\mathbf{W}}_g$ and $\tilde{\mathbf{\Psi}}_g$, given by:

$$\tilde{\mathbf{W}}_g=\mathbf{S}_g\mathbf{\Psi}_g^{-1}\mathbf{W}_g\left[\mathbf{I}+\mathbf{M}_g^{-1}\mathbf{W}_g^T\mathbf{\Psi}_g^{-1}\mathbf{S}_g\Psi_g^{-1}\mathbf{W}_g\right]^{-1}\text{and}\tag{32}$$

$$\tilde{\mathbf{\Psi}}_g=\mathrm{diag}\left(\mathbf{S}_g-\mathbf{S}_g\mathbf{\Psi}_g^{-1}\mathbf{W}_g\mathbf{M}_g^{-1}\tilde{\mathbf{W}}_g^T\right).\tag{33}$$

Here, the $\mathrm{diag}(\cdot)$ operation retains only the diagonal elements of the matrix that it operates on. The value of $\mathbf{S}_g$ is obtained from (7).

[2]The approaches presented in [37] and [34] could also be followed in this procedure. However, we chose to utilize the methods in [33] for the EM formulation to obtain a set of compact M-step equations.

## ACKNOWLEDGMENT

The authors would like to thank the CRSS members Seyed Omid Sadjadi and Keith W. Godin for maintaining the high performance computation cluster (HPCC) that is used for this research. We also thank Hynek Bořil for his insights on front-end feature extraction. Finally, we thank Seyed Omid Sadjadi for preparing the noisy files, originally developed during the CRSS NIST SRE 2012 submission.

## REFERENCES

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

[2] 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.

[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, Jan. 2000.

[4] 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, Nov. 1996.

[5] 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.

[6] T. Hasan and J. H. L. Hansen, "Acoustic factor analysis based universal background model for robust speaker verification in noise," in *Proc. InterSpeech*, Lyon, France, Aug. 2013.

[7] T. Hasan and J. H. L. Hansen, "Acoustic factor analysis for robust speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 4, pp. 842–853, Apr. 2013.

[8] Q. Jin, T. Schultz, and A. Waibel, "Far-field speaker recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 7, pp. 2023–2032, Sep. 2007.

[9] 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.

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

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

[12] 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. Audio, Speech, Lang. Process.*, vol. 17, no. 2, pp. 366–378, Feb. 2009.

[13] NIST, "The NIST year 2012 speaker recognition evaluation plan," 2012 [Online]. Available: http://www.nist.gov/itl/iad/mig/upload/NIST_SRE12_evalplan-v17-r1.pdf

[14] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Proc. IEEE ICASSP*, May 2006, pp. 97–100.

[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] 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.

[17] 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. IEEE ICASSP*, Prague, Czech Republic, May 2011, pp. 4828–4831.

[18] 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.

[19] E. ETSI, "202 050 v1. 1.3: Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," ETSI standard 2002.

[20] S. O. Sadjadi, T. Hasan, and J. H. L. Hansen, "Mean Hilbert Envelope Coefficients (MHEC) for robust speaker recognition," in *Proc. InterSpeech*, Portland, OR, USA, Sep. 2012, pp. 1696–1699.

[21] U. H. Yapanel and J. H. L. Hansen, "A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition," *Speech Commun.*, vol. 50, pp. 142–152, Feb. 2008.

[22] 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.

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

[24] T. Hasan and M. K. Hasan, "Suppression of residual noise from speech signals using empirical mode decomposition," *IEEE Signal Process. Lett.*, vol. 16, no. 1, pp. 2–5, Jan. 2009.

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

[26] H. Bořil and J. H. L. Hansen, "Unsupervised equalization of Lombard effect for speech recognition in noisy adverse environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 6, pp. 1379–1393, Sep. 2010.

[27] H. Bořil and J. H. L. Hansen, "UT-scope: Towards LVCSR under Lombard effect induced by varying types and levels of noisy background," in *Proc. IEEE ICASSP*, Prague, Czech Republic, May 2011, pp. 4472–4475.

[28] T. Hasan, R. Saeidi, J. H. L. Hansen, and D. A. van Leeuwen, "Duration mismatch compensation for i-Vector based speaker recognition systems," in *Proc. IEEE ICASSP*, Vancouver, BC, Canada, May 2013.

[29] T. Hasan, G. Liu, S. O. Sadjadi, N. Shokouhi, H. Bořil, A. Misra, K. W. Godin, and J. H. Hansen, "UTD-CRSS systems for 2012 NIST speaker recognition evaluation," in *Proc. NIST 2012 Speaker Recognition Evaluation Workshop*, Orlando, FL, USA, Dec. 2012.

[30] Y. Lei, L. Burget, L. Ferrer, M. Graciarena, and N. Scheffer, "Towards noise-robust speaker recognition using probabilistic linear discriminant analysis," in *Proc. IEEE ICASSP*, Mar. 2012, pp. 4253–4256.

[31] 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, USA, Sep. 2012, pp. 1568–1571.

[32] 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.

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

[34] G. J. McLachlan and D. Peel, "Mixtures of factor analyzers," in *Proc. ICML*, Jun. 2000, pp. 599–606.

[35] 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.

[36] Y. Tang, R. Salakhutdinov, and G. Hinton, "Deep mixtures of factor analysers," in *Proc. ICML*, Edinburgh, U.K., Jun. 2012.

[37] Z. Ghahramani and G. Hinton *et al.*, "The EM algorithm for mixtures of factor analyzers," Univ. of Toronto, Toronto, ON, Canada, Tech. Rep. CRG-TR-96-1, 1996.

[38] J. Sohn, N. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[39] M. Brooks, VOICEBOX: Speech Processing Toolbox for MATLAB [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[40] 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.

[41] [Online]. Available: www.freesound.org

[42] T. Hasan, Y. Lei, A. Chandrasekaran, and J. H. L. Hansen, "A novel feature sub-sampling method for efficient universal background model training in speaker verification," in *Proc. IEEE ICASSP*, Dallas, TX, USA, Mar. 2010, pp. 4494–4497.

[43] T. Hasan and J. H. L. Hansen, "A study on universal background model training in speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, pp. 1890–1899, Sep. 2011.

[44] P. Kenny, "Bayesian speaker verification with heavy tailed priors," in *Proc. Odyssey*, Brno, Czech Republic, 2010.

[45] N. Brummer, 2012, SRE'12 - BOSARIS Toolkit [Online]. Available: https://sites.google.com/site/bosaristoolkit/sre12

[46] C. M. Bishop, "Bayesian PCA," *Adv. Neural Inf. Process. Syst.*, pp. 382–388, 1999.

[47] S. Nakajima, M. Sugiyama, and D. Babacan, "On Bayesian PCA: Automatic dimensionality selection and analytic solution," in *Proc. ICML*, Bellevue, WA, USA, 2011, pp. 497–504.
[48] N. Brummer and E. de Villiers, "The BOSARIS toolkit: Theory, algorithms and code for surviving the new DCF," in *Proc. NIST SRE Analysis Workshop*, Atlanta, GA, USA, Dec. 2011.
[49] T. Hasan, S. O. Sadjadi, G. Liu, N. Shokouhi, H. Bořil, and J. H. Hansen, "CRSS systems for 2012 NIST speaker recognition evaluation," in *Proc. IEEE ICASSP*, Vancouver, BC, Canada, May. 2013, pp. 6783–6787.

**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 has been a Lecturer in the Electrical and Electronic Engineering Department at United International University, Dhaka, Bangladesh from December 2006 to June 2008. Currently he is 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 served as the lead student in the NIST SRE 2012 submission from the CRSS group. His research interests focus on robust speaker recognition in noisy and channel mismatched conditions, front-end processing, speech enhancement and multi-modal signal processing.

**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 Professor and Department Head of Electrical Engineering, and holds the Distinguished University Chair in Telecommunications Engineering. 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 and served as TC Chair in 2011-2012, presently serving as Past-TC Chair in 2013), 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 TRANS. SPEECH & 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 TRANS. SPEECH & AUDIO PROC. 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 63 (30 Ph.D., 33 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 494 journal and conference papers and 11 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.