Odyssey 2012
The Speaker and Language Recognition Workshop
25-28 June 2012, Singapore

Factor Analysis of Acoustic Features using a Mixture of Probabilistic Principal Component Analyzers for robust Speaker Verification

*Taufiq Hasan and John H. L. Hansen**

Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering and Computer Science
University of Texas at Dallas, Richardson, TX, 75252 USA.
{Taufiq.Hasan, John.Hansen}@utdallas.edu

Abstract

Robustness due to mismatched train/test conditions is one of the biggest challenges facing speaker recognition today, with transmission channel/handset and additive noise distortion being the most prominent factors. One limitation of the recent speaker recognition systems is that they are based on a latent factor analysis modeling of the GMM mean super-vectors alone. Motivated by the covariance structure of cepstral features, in this study, we develop a factor analysis model in the acoustic feature space instead of the super-vector domain. The proposed technique computes a mixture dependent feature dimensionality reduction transform and is directly applied to the first order Baum-Welch statistics for effective integration with a conventional i-vector-PLDA system. Experimental results on the telephone trials of the NIST SRE 2010 demonstrate the superiority of the proposed scheme.

1. Introduction

Mismatch between training and test conditions represent one of the most challenging problems faced by speaker recognition researchers today. There can be different sources of mismatch present including: transmission channel differences [1], handset variability, background noise, session variability due to physical stress [2], non-stationarity environment [3], different levels of vocal effort or spontaneity of speech, 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 of the state-of-the-art speaker recognition system is to model acoustic features with Gaussian Mixture Models (GMM) [4], use utterance dependent adapted GMM [4] mean super-vectors [5] as features representing speech segments, and model the super-vectors using various latent factor analysis techniques [1, 6, 7]. In [8], the aim was to identify the speaker and channel dependent subspaces, termed Eigenvoice [6] and Eigenchannel [1], in the super-vector domain. In [1], speaker and channel variabilities were jointly modeled. With the introduction of i-vectors [7], research trend shifted towards directly applying compensation techniques on these lower dimensional features representing a speech segment. In simple terms, the i-vector scheme utilizes a factor analysis framework [6, 9] to perform dimensionality reduction on the super-vectors while retaining important speaker discriminant information. This lower dimensional i-vector representation enabled the development of fully Bayesian techniques [10, 11] using a single model to represent the speaker and channel variability.

From the success of Bayesian modeling of i-vectors [10, 11], it is clear that including more speaker discriminant information in the i-vector is the key to achieving better performance. Often higher dimensional i-vectors are extracted [12] to achieve this goal. Another way of improving speaker discriminant ability of an i-vector can be by suppressing speaker irrelevant/nuisance components from it. In this study, diverting our attention from the super-vectors, we attempt to model the acoustic features using a factor analysis framework, and thereby aim to reduce nuisance components in this domain.

Acoustic factor analysis has been previously explored in the areas of speech recognition, speech production modeling, etc. [13, 14]. In the speaker recognition community, this avenue has been somewhat unexplored from the popular belief that cepstral feature coefficients can be modeled sufficiently well by a diagonal covariance GMM model. However, full covariance GMM models have shown to provide advantage in speaker recognition system performance in recent studies [11], confirming the fact that there are speaker discriminatory information in the covariance structure of the acoustic features. Also, in [15], it was shown that the first few directions obtained by the Eigen-decomposition of the feature covariance matrix is mostly speaker dependent, while other directions are phoneme dependent, indicating that speaker relevant information lies in a lower dimensional subspace of the acoustic features. These facts demand a closer investigation of latent factor analysis modeling in the acoustic feature space and develop strategies to possibly represent speaker dependent information more compactly in the current speaker recognition frame-work.

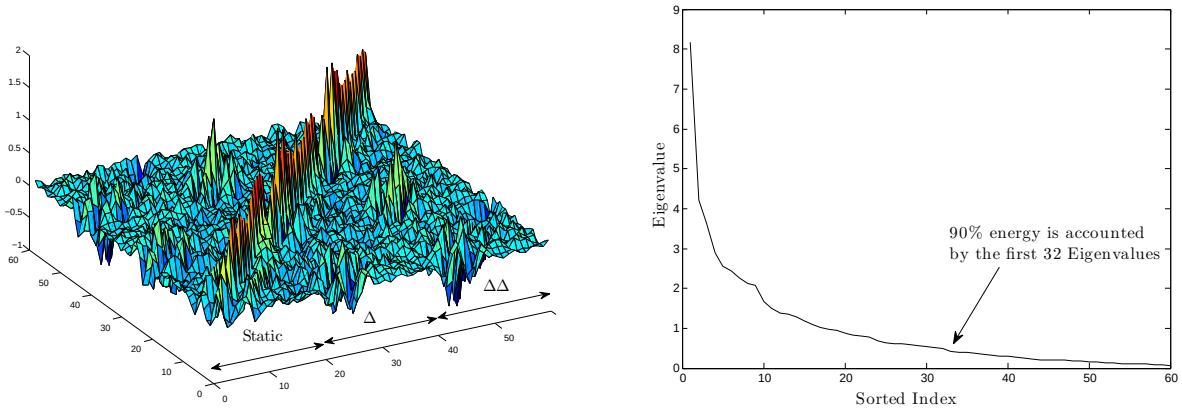
2. Proposed method

In this section we propose a factor analysis model of acoustic features using a mixture of probabilistic principal component analyzers (PPCA) [9] and discuss its integration within an i-vector system framework.

2.1. Motivation

Our intuition is that the acoustic features currently used in speaker recognition systems can be represented by a lower di-

*This project was funded by AFRL through a subcontract to RADC Inc. under FA8750-09-C-0067, and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. Hansen. Approved for Public Release; Distribution Unlimited: 88ABW-2012-0701, 13-Feb-2012



(a) (b)

Figure 1: Analysis of the full covariance matrix of a UBM mixture trained using 60-dimensional MFCC feature (20 static+Δ+ΔΔ). (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.

mensional subspace retaining necessary speaker discriminant information [15]. To test this hypothesis, we train a 1024 mixture full covariance Universal Background Model (UBM) using 60 dimensional Mel Frequency Cepstral Coefficient (MFCC) features on a large background speech data set (details on features and data are given in Sec. 3.1 and 3.2, respectively). From an arbitrary component 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, while shows strong diagonal terms, has significant non-zero off-diagonal elements. This indicates that the feature coefficients are not fully uncorrelated. Furthermore, Fig. 1(b) shows the sorted Eigenvalues of the same covariance matrix indicating that most of it's energy are accounted for by the first few dimensions only. This shows that the acoustic space is indeed lower dimensional and features can thus be further compacted by removing some nuisance dimensions, while retaining speaker dependent information.

2.2. Acoustic Factor Analysis (AFA)

Let $\mathcal{X} = \{\mathbf{x}_n | n = 1 \cdots N\}$ be the collection of all feature vectors from the development set. Using a PPCA based factor analysis model, $\mathbf{x}$ can be represented by,

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

Here, $\mathbf{x}$ represents the $d \times 1$ dimensional feature vector obtained from $\mathcal{X}$, $\mathbf{W}$ 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 μ is the $d \times 1$ mean vector of $\mathbf{x}$. The latent variable vector $\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, termed *acoustic factors*, is of dimension $q \times 1$ and $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ is an isotropic noise vector modeling the residual variance. In this PPCA model, the feature vectors are also normally distributed such that, $\mathbf{x} \sim \mathcal{N}(\mu, \sigma^2\mathbf{I} + \mathbf{W}\mathbf{W}^T)$.

This model of the acoustic features, however, is not sophisticated enough to capture the variations caused by different phonemes uttered by multiple speakers in distinct noisy or channel degraded conditions and thus a mixture of PPCA (MPPCA) models is required. In the MPPCA framework, a combination of PPCA models are used such that,

$$p(\mathbf{x}) = \sum_{i=1}^{M} w_i p(\mathbf{x}|i), \text{ and} \tag{2}$$

$$p(\mathbf{x}|i) = \mathcal{N}(\mu_i, \sigma_i^2\mathbf{I} + \mathbf{W}_i\mathbf{W}_i^T) \tag{3}$$

where μ_i, $\mathbf{W}_i$ and σ_i represent the mean vector, factor loading matrix, and noise variance for the i-th PPCA model, respectively. Our aim is to formulate a dimensionality reduction of acoustic features using this mixture model.

2.3. Feature dimensionality reduction

An MPPCA model can be conveniently extracted from a GMM trained using the Expectation-Maximization (EM) algorithm [9]. Thus we utilize a full covariance UBM to derive the MPPCA model. To obtain the proposed feature dimensionality reduction transform, we proceed as follows:

2.3.1. Universal Background Model

First, a full covariance UBM model, λ_0, is trained on the development dataset $\mathcal{X} = \{\mathbf{x}_n | n = 1 \cdots N\}$, given by,

$$p(\mathbf{x}|\lambda_0) = \sum_{i=1}^{M} w_i \mathcal{N}(\mu_i, \mathbf{\Sigma}_i) \tag{4}$$

where w_i represents the mixture weights, M is the total number of mixtures, μ_i are the mean vectors and $\mathbf{\Sigma}_i$ are the full covariance matrices.

2.3.2. Noise estimation

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 are not important. Using this value of q, we find the noise variance for the i-th mixture as,

$$\sigma_i^2 = \frac{1}{d-q} \sum_{j=q+1}^{d} \lambda_j \tag{5}$$

where $\lambda_{q+1} \cdots \lambda_d$ are the smallest eigenvalues of $\mathbf{\Sigma}_i$. Thus, σ_i^2 is the average variance lost per discarded dimension.

2.3.3. Compute the factor loading matrix

The maximum likelihood estimation of the factor loading matrix $\mathbf{W}_i$ of the i-th mixture of the PPCA model in (2) is given by,

$$\mathbf{W}_i = \mathbf{U}_q^{(i)}(\mathbf{\Lambda}_q^{(i)} - \sigma_i^2\mathbf{I})^{1/2}\mathbf{R} \tag{6}$$

where $\mathbf{U}_q^{(i)}$ is a $d \times q$ matrix whose columns are the q leading eigenvectors of $\mathbf{\Sigma}_i$, $\mathbf{\Lambda}_q^{(i)}$ is a diagonal matrix that contains the corresponding q eigenvalues, and $\mathbf{R}$ is a $q \times q$ arbitrary orthogonal rotation matrix (In this work, we set $\mathbf{R} = \mathbf{I}$).

2.3.4. Compute latent factors

For the i-th PPCA model, the dimensionality reduced version of $\mathbf{x}_n$ can be obtained from the posterior mean of $\mathbf{y}$ as:

$$E\{\mathbf{y}|\mathbf{x}_n, i\} = \langle \mathbf{y}_n^{(i)} \rangle = \mathbf{M}_i^{-1}\mathbf{W}_i^T(\mathbf{x}_n - \mu_i) \tag{7}$$

where

$$\mathbf{M}_i = (\sigma_i^2\mathbf{I} + \mathbf{W}_i^T\mathbf{W}_i). \tag{8}$$

Thus, we are essentially using $\langle \mathbf{y}_n^{(i)} \rangle$ as a mixture dependent transformed acoustic feature, instead of the original vectors $\mathbf{x}_n$.

2.4. Integration within the i-vector system

After feature extraction and UBM training, the first step of training a total variability matrix/i-vector extraction is estimating the zero and first order Baum-Welch statistics. These statistics are computed from acoustic features with respect to the UBM model. Instead using the dimensionality reduction as a front-end processing and then retraining the UBM, as in case of PCA or Heteroscedastic Linear Discriminant Analysis (HLDA) [8], we apply our feature transformation directly on the first order statistics. This also eliminates the premature alignment of $\mathbf{x}_n$ to a specific mixture and thus utilizes the full potential of the probabilistic model. For an utterance S, the zero order statistics is extracted as,

$$N_S(i) = \sum_{n \in S} \gamma_i(n), \text{where } \gamma_i(n) = p(i|\mathbf{x}_n). \tag{9}$$

Using the proposed scheme, the first order statistics $\mathbf{F}_S(i)$, is extracted using the dimensionality reduced feature vectors in the corresponding mixtures instead of the acoustic features.

$$\begin{aligned} \mathbf{F}_S(i) &= \sum_{n \in S} \gamma_i(n)\langle \mathbf{y}_n^{(i)} \rangle = \sum_{n \in S} \gamma_i(n)\mathbf{M}_i^{-1}\mathbf{W}_i^T(\mathbf{x}_n - \mu_i) \\ &= \mathbf{M}_i^{-1}\mathbf{W}_i^T \sum_{n \in S} \gamma_i(n)(\mathbf{x}_n - \mu_i) \end{aligned} \tag{10}$$

As expected, this is simply a transformed version of the centralized first order statistics [12]. From Eq. (7), it can be shown that for the Gaussian component i, the transformed features $\langle \mathbf{y}_n^{(i)} \rangle$ follow a normal distribution with zero mean and diagonal covariance matrix of $\mathbf{I} - \sigma_i^2 {\mathbf{\Lambda}_q^{(i)}}^{-1}$. The derivation is given below.

Let $\mathbf{z_n} = \langle \mathbf{y}_n^{(i)} \rangle$ indicate the transformed feature vector conditioned on the i-th mixture component. Dropping the subscripts i, we have the mean vector of $\mathbf{z_n}$,

$$\begin{aligned} \mu_{\mathbf{z_n}} &= E\{\langle \mathbf{y}_n \rangle\} \\ &= E\{\mathbf{M}^{-1}\mathbf{W}^T(\mathbf{x}_n - \mu)\} = \mathbf{0} \end{aligned}$$

and its covariance matrix,

$$\begin{aligned} \Sigma_{\mathbf{z_n}} &= E\{\mathbf{z_n}{\mathbf{z_n}}^T\} - \mu_{\mathbf{z_n}}{\mu_{\mathbf{z_n}}}^T \\ &= \mathbf{M}^{-1}\mathbf{W}^T E\{(\mathbf{x}_n - \mu)(\mathbf{x}_n - \mu)^T\}\mathbf{W}\mathbf{M}^{-T} \\ &= \mathbf{M}^{-1}\mathbf{W}^T\mathbf{\Sigma}\mathbf{W}\mathbf{M}^{-T}. \end{aligned} \tag{11}$$

From (8), we substitute the value of $\mathbf{W}$ from (6) and obtain,

$$\begin{aligned} \mathbf{M} &= \sigma^2\mathbf{I} + \mathbf{W}^T\mathbf{W} \\ &= \sigma^2\mathbf{I} + (\mathbf{\Lambda}_q - \sigma^2\mathbf{I})^{T/2}{\mathbf{U}_q}^T\mathbf{U}_q(\mathbf{\Lambda}_q - \sigma^2\mathbf{I})^{1/2} \\ &= \sigma^2\mathbf{I} + (\mathbf{\Lambda}_q - \sigma^2\mathbf{I}) = \mathbf{\Lambda}_q. \end{aligned} \tag{12}$$

Here, we use $\mathbf{R} = \mathbf{I}$ and utilize the fact that all the matrices are diagonal and thus symmetric. Substituting the value of $\mathbf{M}$ from (12), $\mathbf{W}$ from (6), and using the relation ${\mathbf{U}_q}^T\mathbf{\Sigma}\mathbf{U}_q = \mathbf{\Lambda}_q$, in (11) we have,

$$\begin{aligned} \Sigma_{\mathbf{z_n}} &= {\mathbf{\Lambda}_q}^{-1}(\mathbf{\Lambda}_q - \sigma^2\mathbf{I})^{T/2}\mathbf{\Lambda}_q(\mathbf{\Lambda}_q - \sigma^2\mathbf{I})^{1/2}{\mathbf{\Lambda}_q}^{-T} \\ &= (\mathbf{\Lambda}_q - \sigma^2\mathbf{I}){\mathbf{\Lambda}_q}^{-T} \\ &= \mathbf{I} - \sigma^2{\mathbf{\Lambda}_q}^{-1}. \end{aligned} \tag{13}$$

Since the posterior means of the *acoustic factors* $\mathbf{y}_n$ is used as mixture dependent features, the the UBM λ_0 is replaced by a transformed UBM model λ_0^{AFA}, that follows the distribution of $\mathbf{z}_n$. Even though it is not strictly a correct expression, for convenience, we write the new UBM equation as,

$$p(\mathbf{z}|\lambda_0^{\text{AFA}}) = \sum_{i=1}^{M} w_i\mathcal{N}(\mathbf{0}, \mathbf{I} - \sigma_i^2{\mathbf{\Lambda}_q^{(i)}}^{-1}). \tag{14}$$

This is similar to the approach in [12], when UBM is normalized to zero means and identity covariance matrices. However, in [12] the goal was to simplify the i-vector system procedure while in this work, we are performing feature transformation and dimensionality reduction. The proposed normalization of the UBM should not be interpreted such that (14) refers to a GMM model for which all the mean vectors are zero and thus the mixtures are on top of each other. Eq. (14) simply indicates how the UBM parameters should be modified. The posterior probabilities of the mixture i are calculated using the original feature vectors $\mathbf{x}_n$ and UBM λ_0, not $\mathbf{z}_n$ and UBM λ_0^{AFA}. The rest of the i-vector system procedure exactly follow the conventional approach, except: (i) feature dimension is now q instead of d and (ii) UBM model is λ_0^{AFA} instead of λ_0.

3. Experiments

We perform our experiments on the male trials of NIST SRE 2010 telephone train/test condition (condition 5, normal vocal effort). Different blocks of the baseline system implementation and the details of the proposed scheme is described below.

3.1. Feature Extraction

For voice activity detection (VAD), a phoneme recognizer [16] and energy based scheme is used. A 60-dimension feature vec-

tor (19 MFCC +Energy + $\Delta + \Delta\Delta$) is extracted, using a 25 ms analysis window with 10 ms shift and filtered by feature warping using a 3-s sliding window [17].

3.2. UBM Training

A gender dependent full-covariance UBM with 1024 mixtures were trained on utterances selected from Switchboard II Phase 2 and 3, Switchboard Cellular Part 1 and 2, and the NIST 2004, 2005, 2006 SRE enrollment data. We used the HTK toolkit for training with 15 iterations per mixture split. The main diagonal of the full covariances were floored to 10^{-5} using the *-v* option in HTK HERest toolkit.

3.3. Acoustic Factor Analysis

For three different runs, we set the AFA parameter q to 54, 48 and 42, to reduce feature dimensionality from $d = 60$ to q. For each value of q, the procedure from Sec. 2.3 is followed to compute the transformation matrices for each mixture of the UBM. No transformation was used in the baseline system.

3.4. Total variability modeling

For the total variability (TV) matrix training, the UBM training dataset is utilized. i-vector dimension was set to 400. All i-vectors are first whitened and then length normalized [18]. Five iterations were used for the EM training.

3.5. Channel 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 [18][1]. In this generative model, an R dimensional i-Vector $\mathbf{w}_s$ extracted from a speech utterance s is expressed as:

$$\mathbf{w}_s = \mathbf{w}_0 + \mathbf{\Phi}\beta + \mathbf{n} \tag{15}$$

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 dataset is once again used for PLDA training.

4. Results

Here, we vary the number of Eigenvoices N_{EV}, in the PLDA model from 50 to 300 in 50 step increments. The performance metrics used are %EER, normalized minimum Detection Cost Function (DCF) proposed in NIST SRE 2008 (MinDCF$_{old}$) [19] and NIST SRE 2010 (MinDCF$_{new}$) [20]. The results are summarized in Fig. 2 and Table 1.

From the results we observe that for all values of q and N_{EV}, the proposed system performs significantly better than the baseline system. For $q = 48$ and $N_{EV} = 100$ we achieve the best EER performance of 2.15%. For $N_{EV} = 200$ and $q = 48$, the proposed scheme achieves 20.35%, 17.97% and 9.16% relative improvement in EER, MinDCF$_{old}$ and MinDCF$_{new}$, respectively, compared to baseline. This indicates that the proposed transformation in the acoustic features was successfully able to reduce some nuisance directions in the feature space producing i-vectors with better speaker discriminating ability. We observed that using a smaller value of q starts degrading the system performance. Simple equal-weight linear fusion of baseline and multiple AFA systems results in further performance gain, as shown in Table 2, reaching the best EER of 1.94%. In this fusion, the only calibration performed is mean and variance normalization to (0,1) for the individual sub-systems.

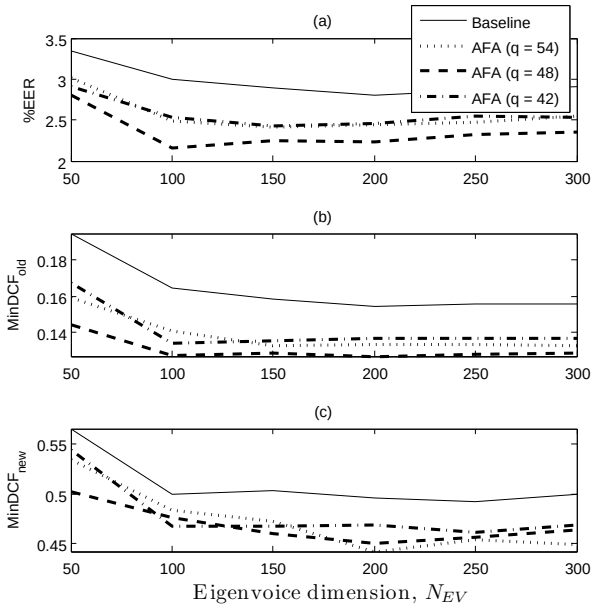


Figure 2: Performance comparison between proposed AFA and baseline i-vector system with respect to (a) %EER, (b) MinDCF$_{old}$ and (c) MinDCF$_{new}$ for different Eigenvoice sizes.

5. Conclusions

In this study, we have proposed an alternate modeling technique to address and compensate for transmission channel mismatch in speaker recognition. We have developed a dimensionality reduction transform for acoustic features using a factor analysis model derived from a full-covariance UBM. Instead of applying as a front-end processing, the proposed transform has been integrated within an i-vector speaker recognition framework. Experimental results have demonstrated the superiority of the proposed scheme compared to the baseline i-vector system.

6. References

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

[2] J. H. L. Hansen, "Analysis and compensation of speech

[1]We would like to thank Daniel Garcia-Romero from University of Maryland for providing us with the Gaussian PLDA software (https://sites.google.com/site/dgromeroweb/software).

Table 2: Linear score fusion performance of baseline i-vector and AFA systems

Fused systems	**No. of Sys.**	**%EER**	**MinDCF$_{old}$**	**MinDCF$_{new}$**
Baseline$_{N_{EV}=200}$ & AFA$_{q=48,N_{EV}=100}$	2	2.07302	0.11487	0.42622
Baseline$_{N_{EV}=200}$ & AFA$_{q=48,N_{EV}=100}$ & AFA$_{q=54,N_{EV}=100}$	3	1.99411	0.11178	0.43107
Baseline$_{N_{EV}=200}$ & AFA$_{q=48,N_{EV}=100}$ & AFA$_{q=48,N_{EV}=200}$	3	**1.93956**	**0.11010**	**0.42030**

Table 1: Performance comparison between baseline i-vector and AFA systems with respect to %EER, MinDCF$_{old}$ and MinDCF$_{new}$ for different values of N_{EV} and q.

Eigenvoice dimension, N_{EV}	**Baseline system**	**Proposed AFA system**		
		q = 54	*q = 48*	*q = 42*
% Equal Error Rate (EER)				
50	3.3397	3.0105	2.8094	2.9129
100	2.9989	2.4843	**2.1542**	2.5367
150	2.8904	2.4162	2.2403	2.4337
200	2.8060	2.4377	2.2350	2.4625
250	2.8656	2.4687	2.3264	2.5477
300	2.9078	2.5409	2.3578	2.5356
MinDCF$_{old}$				
50	0.1951	0.1593	0.1436	0.1671
100	0.1645	0.1403	0.1270	0.1338
150	0.1586	0.1323	0.1281	0.1352
200	0.1541	0.1328	**0.1264**	0.1360
250	0.1553	0.1332	0.1274	0.1364
300	0.1555	0.1320	0.1278	0.1363
MinDCF$_{new}$				
50	0.5653	0.5341	0.5020	0.5434
100	0.4997	0.4833	0.4759	0.4679
150	0.5027	0.4722	0.4608	0.4682
200	0.4954	**0.4424**	0.4500	0.4694
250	0.4923	0.4536	0.4562	0.4617
300	0.4992	0.4496	0.4636	0.4691

under stress and noise for environmental robustness in speech recognition," *Speech Comm.*, vol. 20, no. 1-2, pp. 151–173, 1996.

[3] 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.

[4] 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.

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

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

[7] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 19, no. 99, pp. 788 – 798, May 2010.

[8] L. Burget *et. al.*, "Analysis of feature extraction and channel compensation in a GMM speaker recognition system," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 15, no. 7, pp. 1979–1986, Sept. 2007.

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

[10] 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.

[11] 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.

[12] 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.

[13] J.-T. Chien and C.-W. Ting, "Acoustic factor analysis for streamed hidden markov modeling," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 17, no. 7, pp. 1279 – 1291, Sep. 2009.

[14] X. Lu and J. Dang, "Vowel production manifold: Intrinsic factor analysis of vowel articulation," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 18, no. 5, pp. 1053 – 1062, July 2010.

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

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

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

[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] "The NIST year 2008 speaker recognition evaluation plan," 2008. [Online]. Available: http://www.nist.gov

[20] "The NIST year 2010 speaker recognition evaluation plan," 2010. [Online]. Available: http://www.nist.gov