Acoustic Factor Analysis based Universal Background Model for Robust Speaker Verification in Noise

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

Center for Robust Speech Systems (CRSS)
Erik Jonsson School of Engineering & Computer Science
The University of Texas at Dallas (UTD), Richardson, TX 75080-3021, USA
{Taufiq.Hasan,John.Hansen}@utdallas.edu

Abstract

The Universal Background Model (UBM) is known as a speaker independent Gaussian Mixture Model (GMM) trained on a large speech corpus containing many speakers' recordings in various conditions. When noisy test data is involved, UBM trained on clean data is generally not optimal. Using noisy data for UBM training, however, creates a bias towards the specific development noise samples resulting in degraded speaker recognition performance in unseen noise types. In this study, we utilize an Acoustic Factor Analysis (AFA) based UBM that iteratively learns the dominant feature sub-spaces in each mixture component, resulting in a more robust model. We explore two variants of the model: one with an isotropic and the other with a diagonal residual noise. The Maximum-Likelihood (ML) training formulations of the models are provided. The latent variables of the model, termed *acoustic factors*, are used as features to train the second stage of factor analysis parameters, i.e., the traditional i-vector extractor. Experiments performed on the 2012 National Institute of Standards and Technology (NIST) Speaker Recognition Evaluation (SRE) indicate the effectiveness of the proposed strategy in both clean and noisy conditions.

Index Terms: speaker verification, NIST SRE 2012, noisy data, acoustic factor analysis

1. Introduction

Gaussian Mixture Models (GMM) have become the standard technique for modeling acoustic features for speaker recognition over the last decade. From the classical GMM-UBM system [1] to the recent i-vector system [2], almost all the approaches depend on the GMM based back-ground model that is expected to cover the entire acoustic space. To deal with noisy and channel degraded conditions, most effective techniques operate on the utterance models, including GMM supervectors [3] and various factor analysis schemes built in this domain [4, 5], and i-vectors with Probabilistic Linear Discriminant Analysis (PLDA) based classifiers along with various pre-processing techniques [6–8]. Robust feature development [9–11], enhancement [12–15], effective front-end compensation methods [16–18] and score domain techniques have also been considered [19, 20] for mismatch compensation. Many techniques evolved and have been 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.

*This project was funded by AFRL under contract FA8750-12-1-0188 and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. H. L. Hansen.

The UBM is defined as a speaker independent model trained on a large data-set representing many acoustic conditions [1]. Previously in [21], we pointed out two important aspects of UBM training for robust speaker recognition: data variability and balancing. The more diverse the UBM dataset is, the more likely it becomes for an unseen test utterance features to find appropriate mixture components. Data balancing is important so that the UBM is not dominated by specific type of recordings. If clean utterances are used for training the UBM, the test features frames may not align well with the different mixture components (i.e., the posterior probability of a mixture component given an acoustic feature frame may be too low). This in turn will cause the zero and first order Baum-Welch statistics collected using this UBM to be unreliable for i-vector extraction [2]. This problem may be alleviated by using some noisy data in the UBM training itself. However, this can cause the system to be biased towards those specific noise samples involved in UBM training. Thus, a reasonable solution should be to train the UBM using a modeling scheme that learns the behavior of noisy speech data from the development set, but does not over-train towards the development noise samples.

In our recent studies [22–24], we proposed the Acoustic Factor Analysis (AFA) scheme that operates on different mixtures of the UBM as a feature transformation. The principal motivation of this approach was the assumption that traditional acoustic features reside in a lower dimensional subspace, and therefore, the GMM mean super-vector representation of an utterance contain redundancies. The technique operated on the first order Baum-Welch statistics in each mixture with a transformation matrix, effectively reducing the feature dimension within the model. Integrated with an i-vector system, this method led way towards a two-stage factor analysis scheme for speaker recognition.

In this paper, we proceed further with the AFA concept by completely replacing the traditional GMM-UBM with a Mixture of Factor Analyzers (MFA) [25, 26] model and propose an i-vector extraction strategy that utilizes the statistics of the model's hidden variables, termed the *acoustic factors*. This model is somewhat similar in nature to sub-space GMMs proposed for speech recognition [27]. The proposed AFA-UBM model is trained using an Expectation-Maximization (EM) algorithm, which iteratively removes the relatively less important sub-spaces in each mixture component, in contrast to our previous approach [22] where the AFA parameters were extracted from a full-covariance GMM-UBM.

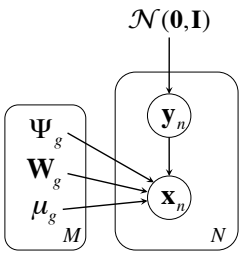

Figure 1: Probabilistic graphical model of a Mixture of Factor Analyzer (MFA) model for acoustic features. The box on the right denotes a 'plate' representing a dataset of N independent observations of acoustic features $\mathbf{x}_n$. Here, $\mathbf{y}_n$ are the hidden variables, or *acoustic factors*.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

2. Acoustic Factor Analysis

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

2.1. 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 [25,26], 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 $q < d$ dominant directions in the feature space, and μ is the mean vector. Following our terminology in [22–24], 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 [22] similar to a GMM-UBM. The probability density function of $\mathbf{x}_n$ is given by:

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

where π_g is the weight of the g-th mixture component, M is the total number of mixtures, and the model covariance matrix in each mixture is given by:

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

A graphical representation of this model is shown in Fig. 1. In our previous studies [22], we assumed ϵ to be isotropic, that is $\mathbf{\Psi}_g = \sigma_g^2\mathbf{I}$, where σ_g^2 denotes the average noise power [28]. Furthermore, the AFA model parameters were derived from a pre-trained full-covariance UBM instead of direct training from the available data. In this paper, we obtain the 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 way of extracting the i-vectors.

2.2. 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 [28]. The EM algorithm procedure for mixture of PPCA model is as follows. In the first step, the following parameters are computed given the initial/old parameter estimates:

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

$$\tilde{\pi}_g = \frac{1}{N}\sum_{n=1}^{N}\gamma_{ng}, \tag{5}$$

$$\tilde{\mu}_g = \frac{\sum_{n=1}^{N}\gamma_{ng}\mathbf{x}_n}{\sum_{n=1}^{N}\gamma_{ng}}, \text{ and} \tag{6}$$

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

Here, $\tilde{\pi}_g$ and $\tilde{\mu}_g$ are the new estimate for the weights and mean vectors, respectively. The new values of the factor loading matrix and noise variance, $\tilde{\mathbf{W}}_g$ and $\tilde{\sigma}_g^2$, can be obtained by:

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

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

where $\mathbf{M}_g = \sigma_g^2\mathbf{I} + \mathbf{W}_g^T\mathbf{W}_g$. The posterior distribution of the acoustic factors for the g-th mixture is given by:

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

We denote this model by ML-AFA$_{\text{iso}}$.

2.3. Diagonal Residual Noise

In this case, we assume that the noise covariance $\mathbf{\Psi}_g$ is diagonal. Here, the q dominant directions represented by the factor loading matrix $\mathbf{W}_g$ no longer contains the principal components. The update equations for this AFA model can be obtained through maximization of the complete data likelihood function in a similar way as in the PPCA case [28]. The new values, $\tilde{\pi}_g$ and $\tilde{\mu}_g$, are obtained through equations (4)-(7) as described in the previous section. Update equations for $\tilde{\mathbf{W}}_g$ and $\mathbf{\Psi}_g$ can be shown to take the following form:

$$\tilde{\mathbf{W}}_g = \mathbf{S}_g\mathbf{\Psi}_g^{-1}\mathbf{W}_g\left(\mathbf{I} + \mathbf{M}_g^{-1}\mathbf{W}_g^T\mathbf{\Psi}_g^{-1}\mathbf{S}_g\mathbf{\Psi}_g^{-1}\mathbf{W}_g\right)^{-1} \tag{11}$$

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

where

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

In this case, the posterior distribution of the acoustic factors for the g-th mixture is given by:

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

We denote this variant of the model as ML-AFA$_{\text{diag}}$.

2.4. I-vector extraction

Conventionally, the i-vectors are extracted from the zero and first order statistics calculated from the features with respect

to the GMM-UBM model. Now, as we replace the GMM-UBM model with the AFA-UBM model (isotropic/diagonal), we could still proceed as before by computing the statistics in the usual way assuming 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 in some way depending on the type model used (isotropic/diagonal). In our initial experiments, we did not observe much difference in performance using this approach compared to a full-covariance GMM-UBM. Here, we propose to model the acoustic factors $\mathbf{y}_n$ in each mixtures as features for the next stage of 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 variability. As an added benefit, the dimension of the statistics are of a lower dimension, leading to a reduction in computational resources in the i-vector extraction process.

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

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

Here, $\gamma_g(n)$ is extracted as in (4) utilizing the model parameters Λ. Conventionally, the first order statistics are extracted as:

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

For the proposed models, the first order statistics are obtained from the posterior mean of the acoustic factors $\langle \mathbf{y}_n | \mathbf{x}_n, g \rangle$ obtained from the distributions in (10) and (14):

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

where $\mathbf{A}_g^T = \mathbf{M}_g^{-1} \mathbf{W}_g^T$ for the isotropic model (ML-AFA$_{\text{iso}}$) and $\mathbf{M}_g^{-1} \mathbf{W}_g^T \mathbf{\Psi}_g^{-1}$ for the diagonal model (ML-AFA$_{\text{diag}}$). Also, $\bar{\mathbf{F}}_s(g)$ represents the centralized first order statistics computed using the model parameters Λ. The rest of the procedure for the i-vector extractor training follows the same principles as outlined in [22]. However, when the acoustic factors $\mathbf{y}_n$ are used as features for the i-vector extractor, the mean vector and covariance matrix of the UBM is set to zero and identity, respectively, following the original definition of $\mathbf{y}_n$ in (1).

3. System Description

An i-vector system [2] with a Gaussian Probabilistic Linear Discriminant Analysis (PLDA) [8] classifier, similar to our NIST SRE 2012 submission [29], is used as the baseline system. Specific blocks of the system implementation and details of the proposed scheme are described below.

3.1. Voice activity detection (VAD)

The central VAD algorithm closely follows the method in [30], available through the open-source Voicebox toolkit [31]. For two channel recordings, VAD is performed on both channels, and audio segments are removed from the target speaker's channel if speech detected in the interviewee/other speaker's channel. For the target speaker's channel, 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 a spectral subtraction technique [12] before VAD. Also, the noise power is estimated using methods outlined in [32]. Note that the enhanced utterances are only used for VAD, not for feature extraction.

3.2. Feature Extraction

We extract 60 dimensional Mel-Frequency Cepstral Coefficients (MFCC) using the CTUcopy toolkit [33]. At first, digital zeros are replaced by a uniformly distributed noise floor having a mean zero and amplitude 1.75^{-5}. For segmentation, a 25 ms window with 10 ms frame shift is used. A 24-channel Mel-spaced filterbank is used and 19 static coefficients are retained. The 60 dimensional features are obtained by including log energy, delta and acceleration parameters. Finally, the features are processed through Cepstral Mean and Variance Normalization (CMVN) utilizing a 3-sec sliding window.

3.3. UBM and AFA Model Training

A gender dependent 1024-mixture GMM-UBM with diagonal covariance[1] matrices is used for the baseline system. The proposed AFA-UBM models are trained using the isotropic and diagonal assumptions following the procedures described in Sec. 2.2 and 2.3. Mixture covariance matrix values are floored to 10^{-5} for the baseline GMM-UBM model.

The UBM models are trained on telephone, microphone and interview type utterances selected from SRE'04-08 enrollment data, Switchboard-II Phase 2 and 3, and Switchboard Cellular Part 1 and 2. Artificially generated noisy files containing Heating, Ventilation and Air Conditioning (HVAC) and crowd noise types, and SRE-12 enrollment speaker data are also included in the UBM. The noisy file generation process is discussed in [29]. The UBM utterances are approximately balanced across: (i) clean vs. noisy, (ii) telephone vs. interview/microphone, and (iii) known vs. unknown speakers.

For the EM training, initial four iterations per mixture are gradually increased to 15 for higher order mixtures. We employed data sub-sampling for fast UBM training [21, 34] to perform the experiments. For each 30 frames that are skipped, 3 consecutive frames are selected, resulting in 10% of the original dataset. In this way, the correlation among successive frames are retained. For the proposed AFA models, we use the acoustic factor dimensions $q = 42, 48$ and 54 for comparison.

3.4. I-vector Extractor Training

For training the i-vector extractor, the UBM training dataset with additional data are used, again with both clean and noisy versions. This corresponds to what we used in our SRE-12 system [35]. Here, a 600-dimensional i-vector extractor is trained using 5 EM iterations. The i-vectors are first centralized and then length normalized using radial Gaussianization [8]. Linear Discriminant Analysis (LDA) projection is performed to reduce the i-vector dimension to 150 before the PLDA scoring. LDA is trained on the same data as the i-vector extractor.

3.5. PLDA classifier

In this study, we use a Gaussian PLDA model with a full-covariance residual noise for session variability compensation and scoring [8]. According to this model, an R dimensional

[1] In our initial experiments, full-covariance GMMs performed worse than the diagonal models for noisy data. Thus we use the diagonal covariance model as the baseline.

Table 1: Performance comparison between baseline and proposed systems in NIST SRE 2012 extended trials condition-1

System		%EER	$\min C_{\text{primary}}$	C_{primary}
Baseline		3.3109	0.2684	0.3454
Method	q	*Absolute/%relative performance*		
ML-AFA$_{\text{iso}}$	42	3.370/-1.8	0.248/7.5	0.345/0.1
	48	**2.835/14.4**	0.244/9.1	0.344/0.3
	54	2.931/11.5	0.240/10.6	0.333/3.6
ML-AFA$_{\text{diag}}$	42	3.031/8.5	**0.224/16.7**	**0.328/5.0**
	48	3.078/7.0	0.245/8.6	0.348/-0.7
	54	3.019/8.8	0.239/10.9	0.335/3.0

Table 2: Performance comparison between baseline and proposed systems in NIST SRE 2012 extended trials condition-2

System		%EER	$\min C_{\text{primary}}$	C_{primary}
Baseline		3.0771	0.3272	0.5580
Method	q	*Absolute/%relative performance*		
ML-AFA$_{\text{iso}}$	42	2.903/5.7	0.316/3.4	0.552/1.0
	48	**2.725/11.4**	0.304/7.2	0.540/3.2
	54	2.849/7.4	0.298/9.0	0.541/3.0
ML-AFA$_{\text{diag}}$	42	2.931/4.8	0.306/6.4	0.546/2.1
	48	2.905/5.6	0.301/7.9	**0.540/3.3**
	54	2.737/11.0	**0.292/10.8**	0.543/2.8

Table 3: Performance comparison between baseline and proposed systems in NIST SRE 2012 extended trials condition-3

System		%EER	$\min C_{\text{primary}}$	C_{primary}
Baseline		**3.1564**	0.1317	0.1435
Method	q	*Absolute/%relative performance*		
ML-AFA$_{\text{iso}}$	42	3.190/-1.1	0.123/6.4	0.137/4.8
	48	3.258/-3.2	0.115/13.1	0.129/10.2
	54	3.212/-1.8	0.116/12.1	**0.127/11.4**
ML-AFA$_{\text{diag}}$	42	3.279/-3.9	**0.108/17.8**	0.129/9.8
	48	3.302/-4.6	0.130/1.2	0.149/-4.1
	54	3.173/-0.5	0.114/13.4	0.130/9.5

i-vector $\mathbf{w}_s$ extracted from an utterance s is expressed as:

$$\mathbf{w}_s = \mathbf{w}_0 + \mathbf{\Phi}\beta + \mathbf{n}. \quad (16)$$

Here, $\mathbf{w}_0$ is an $R \times 1$ speaker independent mean vector, $\mathbf{\Phi}$ is the $R \times N_{EV}$ low rank matrix representing the speaker dependent basis functions or eigenvoices, $\beta \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is an $N_{EV} \times 1$ hidden variable, and $\mathbf{n}$ is the $R \times 1$ random vector representing the full covariance residual noise. 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 [35]. 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 [36].

4. Results

The experiments performed in this study are based on the male portion of the NIST SRE 2012 extended trials (containing 27932667). We use the SRE-12 detection cost functions (DCF), C_{primary}, $\min C_{\text{primary}}$(using $P_{\text{known}} = 0.5$) [37] and % Equal Error Rate (EER) for evaluating the systems. Tables 1–5 summarize the results obtained from the baseline and proposed systems in five SRE-12 common test conditions defined as [37]:

Table 4: Performance comparison between baseline and proposed systems in NIST SRE 2012 extended trials condition-4

System		%EER	$\min C_{\text{primary}}$	C_{primary}
Baseline		3.6459	0.3013	0.4556
Method	q	*Absolute/%relative performance*		
ML-AFA$_{\text{iso}}$	42	3.512/3.7	0.289/3.9	0.455/0.2
	48	3.508/3.8	0.298/1.0	0.452/0.7
	54	3.557/2.4	0.302/-0.3	0.461/-1.1
ML-AFA$_{\text{diag}}$	42	3.560/2.4	0.289/4.0	0.460/-1.0
	48	3.575/2.0	0.282/6.4	0.450/1.1
	54	**3.295/9.6**	**0.280/7.2**	**0.449/1.6**

Table 5: Performance comparison between baseline and proposed systems in NIST SRE 2012 extended trials condition-5

System		%EER	$\min C_{\text{primary}}$	C_{primary}
Baseline		3.5350	0.3329	0.6029
Method	q	*Absolute/%relative performance*		
ML-AFA$_{\text{iso}}$	42	3.474/1.7	0.322/3.3	0.593/1.7
	48	3.154/10.8	0.300/10.0	**0.582/3.5**
	54	3.441/2.7	**0.292/12.2**	0.585/3.1
ML-AFA$_{\text{diag}}$	42	3.395/4.0	0.309/7.3	0.594/1.5
	48	3.477/1.6	0.305/8.4	0.585/3.0
	54	**3.090/12.6**	0.295/11.4	0.587/2.6

1) clean interview speech, 2) clean phone call speech, 3) artificially noised interview speech, 4) artificially noised phone call speech, and 5) phone call speech collected in a noisy environment. The artificially added noise samples are of three types: i) crowd noise, ii) HVAC noise, and iii) single speaker noise. We use multiple utterances in their clean and noisy versions (using in-house noise samples) for speaker enrollment.

From Tables 1–5, it is clear that the proposed technique of utilizing an AFA-UBM instead of a conventional GMM-UBM provides more robust speaker recognition performance across conditions including clean and noisy test scenarios. Except for condition-3, that is noisy interview case, the proposed methods provide superior performance compared to the baseline system in all three performance metrics. Generally, relative improvements in the order of $5 - 10\%$ is obtained using the proposed methods. However, a single model parameter (acoustic factor dimension q) or model type (isotropic or diagonal) does not always provide the best result in all conditions. Nevertheless, these results are encouraging and point towards the need for further research and development in this direction.

5. Conclusions

In this study, we have proposed an acoustic factor analysis based mixture model as an alternative to a conventional GMM-UBM for speaker verification in noise. The proposed model was shown to be robust when trained with a combination of clean and noisy data, due to learning only a limited number of sub-spaces in different mixture components. Two variants of the proposed model was studied, with an isotropic and diagonal residual noise assumption. The method was integrated with a conventional i-vector system where the zero and first order statistics of the so called *acoustic factors* were used instead of the conventional Baum-Welch statistics. Experimental results obtained from the clean and noisy test conditions of the NIST SRE 2012 extended trials demonstrate the effectiveness of the proposed approach.

6. References

[1] 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.

[2] 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. 99, pp. 788 – 798, May 2010.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] S. O. Sadjadi, T. Hasan, and J. H. L. Hansen, "Mean Hilbert Envelope Coefficients (MHEC) for Robust Speaker Recognition," in *Proc. InterSpeech*, Portland, OR, Sept. 2012.

[11] 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, February 2008.

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

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

[14] T. Hasan and M. K. Hasan, "An MMSE estimator for speech enhancement considering the constructive and destructive interference of noise," *Signal Processing, IET*, vol. 4, no. 1, pp. 1 –11, Feb. 2010.

[15] ——, "Suppression of residual noise from speech signals using empirical mode decomposition," *IEEE Signal Process. Lett.*, vol. 16, no. 1, pp. 2–5, Jan. 2009.

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

[17] 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.*, pp. 1379–1393, Sep. 2010.

[18] ——, "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.

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

[20] 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, Canada, May. 2013.

[21] 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.

[22] ——, "Acoustic factor analysis for robust speaker verification," *IEEE Trans. Audio Speech Lang. Process.*, Oct. 2012.

[23] ——, "Integrated feature normalization and enhancement for robust speaker recognition using acoustic factor analysis," in *Proc. InterSpeech*, Portland, OR, Sept. 2012.

[24] ——, "Factor analysis of acoustic features using a mixture of probabilistic principal component analyzers for robust speaker verification," in *Proc. Odyssey*, Singapore, June 2012.

[25] Y. Tang, R. Salakhutdinov, and G. Hinton, "Deep mixtures of factor analysers," in *Proceedings of the 29th International Conference on Machine Learning, 2012, Edinburgh, Scotland*, 2012.

[26] Z. Ghahramani, G. Hinton *et al.*, "The em algorithm for mixtures of factor analyzers," Technical Report CRG-TR-96-1, University of Toronto, Tech. Rep., 1996.

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

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

[29] T. Hasan, S. O. Sadjadi, G. Liu, N. Shokouhi, H. Boril, and J. H. Hansen, "CRSS Systems for 2012 NIST Speaker Recognition Evaluation," in *Proc. IEEE ICASSP*, Vancouver, Canada, May. 2013.

[30] 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, 1999.

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

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

[33] P. Fousek, "CTUCopy – universal speech enhancer and feature extractor," 2007. [Online]. Available: http://noel.feld.cvut.cz/speechlab/

[34] 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, March 2010, pp. 4494 – 4497.

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

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

[37] "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