SPEAKER VERIFICATION WITH THE MIXTURE OF GAUSSIAN FACTOR ANALYSIS BASED REPRESENTATION

Ming Li

SYSU-CMU Joint Institute of Engineering, Sun Yat-Sen University, Guangzhou, China
SYSU-CMU Shunde International Joint Research Institute, Guangdong, China
liming46@mail.sysu.edu.cn, mingli1@cmu.edu

ABSTRACT

This paper presents a generalized i-vector representation framework using the mixture of Gaussian (MoG) factor analysis for speaker verification. Conventionally, a single standard factor analysis is adopted to generate a low rank total variability subspace where the mean supervector is assumed to be Gaussian distributed. The energy that can't be represented by the low rank space is modeled by a single multivariate Gaussian. However, due to the sparsity of the frame level posterior probability and the short duration characteristics, some dimensions of the first-order statistics may not be Gaussian distributed. Therefore, we replace the single Gaussian with a mixture of Gaussians to better represent the residual energy. Experimental results on the NIST SRE 2010 condition 5 female task and the RSR 2015 part 1 female task show that the MoG i-vector outperforms the i-vector baseline by more than 10% relatively for both text independent and text dependent speaker verification tasks, respectively.

Index Terms— Speaker verification, factor analysis, mixture of Gaussian, i-vector

1. INTRODUCTION

Total variability i-vector modeling has gained significant attention in both speaker verification (SV) and language identification (LID) domains due to its excellent performance, compact representation and small model size [1, 2, 3]. In this modeling, first, zero-order and first-order Baum-Welch statistics are calculated by projecting the MFCC features on those Gaussian Mixture Model (GMM) components using the occupancy posterior probability. Second, in order to reduce the dimensionality of the concatenated statistics vectors, a single factor analysis is adopted to generate a low dimensional total variability space which jointly models language, speaker and channel variabilities all together [1]. Third, within this i-vector space, variability compensation methods, such as Within-Class Covariance Normalization (WCCN) [4], Linear Discriminative Analysis (LDA) and Nuisance Attribute Projection (NAP) [5], are performed to reduce the variability for the subsequent modeling methods (e.g., Support Vector Machine [6], Sparse Representation [7], Probabilistic Linear Discriminant Analysis (PLDA) [8, 9, 10], etc.).

Conventionally, in the i-vector framework, the tokens for calculating the zero-order and first-order Baum-Welch statistics are the MFCC features trained GMM components. Such choice of token units may not be the optimal solution. Recently, the generalized i-vector framework [11, 12, 13, 14, 15] has been proposed. In this framework, the tokens for calculating the zero-order statistics have been extended to tied triphone states, monophone states, tandem features trained GMM components, bottleneck features trained GMM components, etc. The features for calculating the first-order statistics have also been extended from MFCC to feature level acoustic and phonetic fused features [13]. The phonetically-aware tokens trained by supervised learning can provide better token separation and discrimination. This enables the system to compare different speakers' voices token by token with more accurate token alignment, which leads to significant performance improvement on the text independent speaker verification task [11, 12, 13, 14, 15].

In both the traditional and the generalized i-vector frameworks, after statistics calculation, a single standard factor analysis is adopted to generate a low rank total variability subspace for dimension reduction. In this generative model, the first-order statistics vector is assumed to be Gaussian distributed and the residual that cannot be represented by the low rank total variability space is modeled by a single multivariate gaussian distribution. However, due to the sparsity of the frame level posterior probability vectors on those tokens, the uneven distribution of zero-order statistics on different tokens, and the short duration characteristics, some dimensions of the first-order statistics may not be Gaussian distributed. Therefore, we propose a generalized factor analysis framework by replacing the single Gaussian with a mixture of Gaussians to better model the residual noises. This idea was originally proposed in [16] to fit the complex residual energy in the robust principal component analysis (PCA) framework. In this work, we extend the MoG residual noise fitting method [16] from PCA to factor analysis. The MoG factor analysis model parameters and the MoG i-vectors are trained and extracted by Expectation-Maximization (EM) algorithm and point estimate, respectively.

Furthermore, in the short duration speaker verification scenario, the zero-order statistics on certain tokens may be too small to robustly make the first-order statistics Gaussian distributed. Therefore, the proposed MoG i-vector framework could also be adopted in short duration speaker verification tasks.

The remainder of the paper is organized as follows. The baseline and the proposed algorithms are explained in Section 2. Experimental results and discussions are presented in Section 3 while conclusions are future works are provided in Section 4.

2. METHODS

First, we will introduce the i-vector baseline and the statistics calculation for the generalized i-vector approach. Second, the details of the proposed MoG i-vector framework are provided. Finally, the scoring backend is described.

This research is supported in part by the National Natural Science Foundation of China No. 61401524, Natural Science Foundation of Guangdong Province, China, SYSU-CMU Shunde International Joint Research Institute, CMU-SYSU Collaborative Innovation Research Center.

2.1. The i-vector baseline

In the total variability space, there is no distinction between the speaker effects and the channel effects. Rather than separately using the eigenvoice matrix $\mathbf{V}$ and the eigenchannel matrix $\mathbf{U}$ [17], the total variability space simultaneously captures the speaker and channel variabilities [2]. Given a C component GMM UBM model λ with $\lambda_c = \{p_c, \mu_\mathbf{c}, \mathbf{\Sigma_c}\}, c = 1, \cdots, C$ and an utterance with a L frame feature sequence $\{\mathbf{y_1}, \mathbf{y_2}, \cdots, \mathbf{y_L}\}$, the 0^{th} and centered 1^{st} order Baum-Welch statistics on the UBM are calculated as follows:

$$N_c = \sum_{t=1}^{L} P(c|\mathbf{y_t}, \lambda) \tag{1}$$

$$\mathbf{F_c} = \sum_{t=1}^{L} P(c|\mathbf{y_t}, \lambda)(\mathbf{y_t} - \mu_\mathbf{c}) \tag{2}$$

where $c = 1, \cdots, C$ is the GMM component index and $P(c|\mathbf{y_t}, \lambda)$ is the occupancy probability for $\mathbf{y_t}$ on λ_c. The corresponding centered mean supervector $\tilde{\mathbf{F}}$ is generated by concatenating all the $\tilde{\mathbf{F}_\mathbf{c}}$ together:

$$\tilde{\mathbf{F}_\mathbf{c}} = \frac{\sum_{t=1}^{L} P(c|\mathbf{y_t}, \lambda)(\mathbf{y_t} - \mu_\mathbf{c})}{\sum_{t=1}^{L} P(c|\mathbf{y_t}, \lambda)}. \tag{3}$$

The centered GMM mean supervector $\tilde{\mathbf{F}}$ can be projected as follows:

$$\tilde{\mathbf{F}} \rightarrow \mathbf{Tx}, \tag{4}$$

where $\mathbf{T}$ is a rectangular total variability matrix of low rank and $\mathbf{x}$ is the so-called i-vector [2]. Considering a C-component GMM and D dimensional acoustic features, the total variability matrix $\mathbf{T}$ is a $CD \times K$ matrix which can be estimated the same way as learning the eigenvoice matrix $\mathbf{V}$ in [18] except that here we consider that every utterance is produced by a new speaker [2].

Given the centered mean supervector $\tilde{\mathbf{F}}$ and total variability matrix $\mathbf{T}$, the i-vector is computed as follows [2]:

$$\mathbf{x} = (\mathbf{I} + \mathbf{T^t \Sigma^{-1} N T})^{-1} \mathbf{T^t \Sigma^{-1} N \tilde{F}} \tag{5}$$

where $\mathbf{N}$ is a diagonal matrix of dimension $CD \times CD$ whose diagonal blocks are $N_c\mathbf{I}, c = 1, \cdots, C$ and $\mathbf{\Sigma}$ is a diagonal covariance matrix of dimension $CD \times CD$ estimated in the factor analysis training step. It models the residual variability not captured by the total variability matrix $\mathbf{T}$ [2]. Covariance $\mathbf{\Sigma}$ is also updated iteratively.

2.2. Statistics calculation in the generalized i-vector framework

In the generalized i-vector framework [13], the zero-order statistics and centered mean supervector for the j^{th} utterance are calculated as follows:

$$N_c = \sum_{t=1}^{L} P(c|\mathbf{z_t^j}, \hat{\lambda}) \tag{6}$$

$$\tilde{\mathbf{F}_\mathbf{c}} = \frac{\sum_{t=1}^{L} P(c|\mathbf{z_t^j}, \hat{\lambda})(\mathbf{y_t^j} - \hat{\mu}_\mathbf{c})}{\sum_{t=1}^{L} P(c|\mathbf{z_t^j}, \hat{\lambda})} \tag{7}$$

$$\hat{\mu}_\mathbf{c} = \frac{\sum_{j=1}^{J}\sum_{t=1}^{L} P(c|\mathbf{z_t^j}, \lambda)\mathbf{y_t}}{\sum_{j=1}^{J}\sum_{t=1}^{L} P(c|\mathbf{z_t^j}, \lambda)}. \tag{8}$$

where $c = 1, \cdots, C$ is the new token index and $P(c|\mathbf{z_t^j}, \hat{\lambda})$ is the posterior probability for the j^{th} utterance's feature vector at time t on the c^{th} token. Note that the feature $(\mathbf{z_t})$ used to calculate the posterior probability $P(c|\mathbf{z_t}, \hat{\lambda})$ and the feature $(\mathbf{y_t})$ for cumulating the first-order statistics $\mathbf{F_c}$ are not necessarily the same. The global mean $\hat{\mu}_\mathbf{c}$ is computed using all the training data in the same way as the mean parameter estimation in GMM.

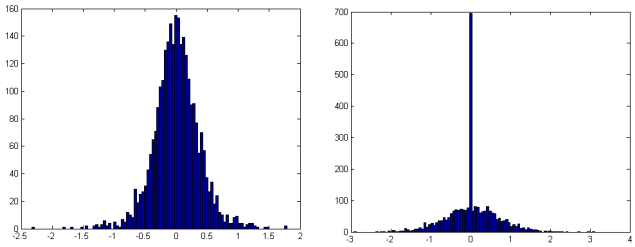

Fig. 1. The histogram of the 1^{st} and 100^{th} dimension of the centered mean supervector $\tilde{\mathbf{F}}$ in the NIST SRE 2005 database.

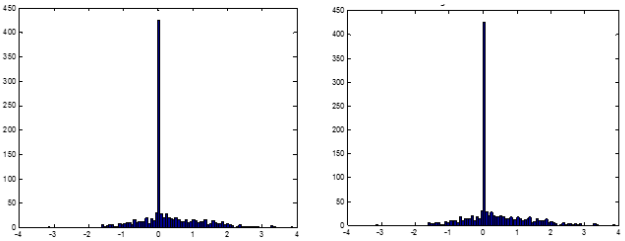

Fig. 2. The histogram of the 1^{st} and 100^{th} dimension of the centered mean supervector $\tilde{\mathbf{F}}$ in the short duration RSR 2015 database.

2.3. The proposed Mixture of Gaussian factor analysis

From Fig.1 and Fig.2, we can see that in real data the centered mean supervector $\tilde{\mathbf{F}}$ may not be Gaussian distributed, especially in the short duration scenario. There are certain number of utterances that have very small zero order statistics on a particular dimension which generate the high peak in the histogram. Therefore, we use the mixture of Gaussians to better fit the residual noises.

In standard factor analysis, the i^{th} dimension of the j^{th} utterance's mean supervector $\tilde{F_{ij}}$ can be considered as the reconstruction using the i^{th} row of $\mathbf{T}$ and the j^{th} utterance's i-vector $\mathbf{x_j}$. The residual that cannot be represented by $\mathbf{T}$ is described as a single gaussian variable ϵ_{ij}.

$$\tilde{F_{ij}} = \mathbf{T_i x_j} + \epsilon_{ij} \tag{9}$$

The corresponding generative model is defined the same way as in [19], where $N_{i,j}$ denotes the corresponding zero-order statistics:

$$P(\mathbf{x_j}) = \mathcal{N}(\mathbf{0}, I), \qquad P(\tilde{F_{ij}}|\mathbf{x_j}) = \mathcal{N}(\mathbf{T_i x_j}, \frac{{\sigma_i}^2}{N_{ij}}) \tag{10}$$

In the Mixture of Gaussian factor analysis, we apply a mixture of Gaussians with K components to describe the residual noises.

$$P(\mathbf{x_j}) = \mathcal{N}(\mathbf{0}, I), \qquad P(\tilde{F_{ij}}|\mathbf{x_j}) = \sum_{k=1}^{K} \pi_{ik}\mathcal{N}(\mathbf{T_i x_j}, \frac{{\sigma_{ik}}^2}{N_{ij}}) \tag{11}$$

The weight for the i^{th} dimension and the k^{th} component is denoted as π_{ik}. Considering the joint likelihood of $\tilde{\mathbf{F}}$ and $\mathbf{x}$ for all those J utterances, we can derive the objective function as follows:

$$\psi = -\sum_{j=1}^{J} \frac{\mathbf{x_j^t x_j}}{2} + \sum_{ij} log(\sum_{k=1}^{K} \pi_{ik}\mathcal{N}(\mathbf{T_i x_j}, \frac{{\sigma_{ik}}^2}{N_{ij}})) \tag{12}$$

Let's denote a hidden variable γ_{ijk} as the posterior probability of the i^{th} dimension of the j^{th} utterance on the k^{th} MoG component:

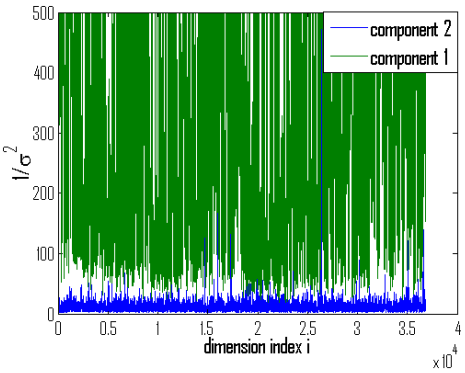

Fig. 3. The inverse covariance $1/\sigma_{ik}^2$ of all the dimensions (36*1024) for two components trained by the RSR 2015 dataset.

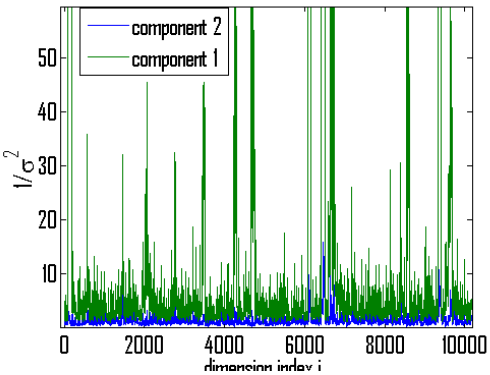

Fig. 4. The inverse covariance $1/\sigma_{ik}^2$ of the first 10000 dimensions for two components trained by the NIST SRE dataset.

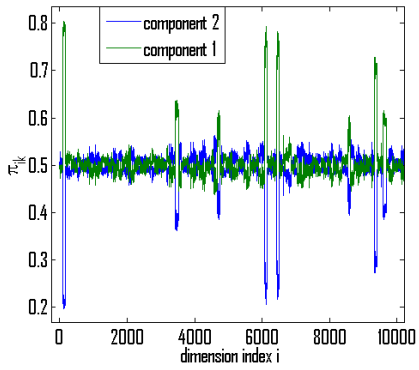

Fig. 5. The MoG component weight π_{ik} of the first 10000 dimensions for two components trained by the NIST SRE dataset.

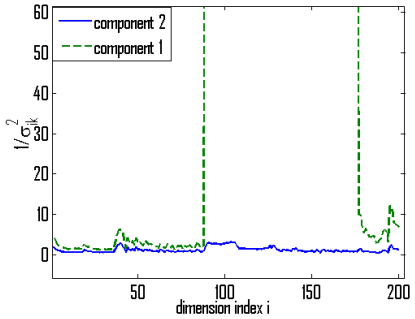

Fig. 6. The inverse covariance $1/\sigma_{ik}^2$ of the first 200 dimensions for two components trained by the NIST SRE dataset.

$$\gamma_{ijk} = \frac{\pi_{ik}\mathcal{N}(\mathbf{T_i x_j}, \frac{{\sigma_{ik}}^2}{N_{ij}})}{\sum_{k=1}^{K}\pi_{ik}\mathcal{N}(\mathbf{T_i x_j}, \frac{{\sigma_{ik}}^2}{N_{ij}})} \tag{13}$$

Using the EM algorithm to solve the maximization problem of equation (12), we can get the following solution.

$$E(\psi) \approx -\sum_{j=1}^{J}\frac{\mathbf{x_j^t x_j}}{2} + \sum_{ij}\sum_{k=1}^{K}\gamma_{ijk}(log(\pi_{ik}) \tag{14}$$
$$-\frac{1}{2}log(\frac{2\pi\sigma_{ik}^2}{N_{ij}}) - \frac{N_{ij}(\tilde{F_{ij}} - \mathbf{T_i x_j})^2}{2\sigma_{ik}^2})$$

$$n_{ik} = \sum_{j=1}^{J}\gamma_{ijk} \tag{15}$$

$$\pi_{ik} = \frac{n_{ik}}{\sum_k n_{ik}} \tag{16}$$

$$\sigma_{ik}^2 = \frac{1}{n_{ik}}\sum_{j=1}^{J}\gamma_{ijk}N_{ij}(\tilde{F_{ij}} - \mathbf{T_i x_j})^2 \tag{17}$$

$$\Sigma_{i,j}^{-1} = \sum_{k=1}^{K}\frac{\gamma_{ijk}}{\sigma_{ik}^2} \tag{18}$$

$$\mathbf{x_j} = (\mathbf{I} + \mathbf{T^t\Sigma_j^{-1}N_jT})^{-1}\mathbf{T^t\Sigma_j^{-1}N_j\tilde{F}_j} \tag{19}$$

After several iterations, the MoG factor analysis model is trained. If the mixture number K is 1, then γ_{ijk} is always 1 and the proposed MoG i-vector is exactly the standard factor analysis. Therefore, we can see that the standard factor analysis is a special case in our MoG factor analysis. By increasing the MoG size K, more complex residual noise patterns can be described.

We propose two methods for the MoG i-vector extraction. Method 1: Since γ_{ijk} is hidden for testing utterances, method 1 adopts equation (13)-(19) for 3 iterations and output the MoG i-vector based on the estimated γ_{ijk}.

Method 2: In our speaker verification tasks, if the zero-order statistics are close to 0, then the corresponding first-order statistics can be considered as invalid data. If we use two MoG components ($K = 2$) to fit the residual noises (as shown in Fig. 3 and Fig. 4, the first one associated with the invalid data would have very small variance σ_{ik}^2 and the second component would have normal range of variance to fit the valid data. Therefore, method 2 assumes all testing data come from the second component and we only use the second component to construct the i-vector. Since no iteration is required here, the computational cost is the same as the conversional i-vector extraction.

By comparing Fig. 3 and Fig. 4, we can find out that more dimensions in the RSR 2015 experiment have high inverse covariances which might be due to the short speech duration and small zero-order statistics. Fig. 4 and Fig. 5 show that in the NIST SRE experiment, the weights for most dimensions are around 0.5 except a few dimensions where their inverse covariances are high. Therefore, as shown in Fig. 6 and Fig. 1, if the zero-order statistics of a particular token is sufficient and the corresponding dimensions of the centered mean supervectors are Gaussian distributed, the weight of two components are very similar; and vice versa.

Table 1. Performance of the proposed methods for the 2010 NIST SRE task female part condition 5

System ID	GMM size	MoG K	Extraction method	EER%	norm minDCF 08	norm minDCF 10
1	256	1	baseline	1.69	0.105	0.308
2	256	2	1	1.42	0.105	0.280
3	256	4	1	1.64	0.102	0.348
4	1024	1	baseline	1.98	0.087	**0.210**
5	1024	2	1	1.70	0.089	0.227
6	1024	2	2	1.64	0.078	**0.193**
7	fuse id 4 and id 6			1.69	0.079	**0.182**

2.4. The backend

After MoG i-vectors are extracted, length normalization and simplified PLDA [10] are adopted as the backend. The PLDA training and scoring is exactly the same as the conversional i-vector baseline.

3. EXPERIMENTAL RESULTS

3.1. Text independent speaker verification on NIST SRE 2010

We first conducted experiments on the NIST 2010 speaker recognition evaluation (SRE) corpus [20] for the text independent speaker verification task. Our focus is the female part of the common condition 5 (a subset of tel-tel) in the core task. We used equal error rate (EER) and the 2008 and 2010 normalized minimum decision cost value (norm minDCF) as the metrics for evaluation [20]. We adopt the hybrid-GMM-hybrid feature level fusion strategy in [13]. For cepstral feature extraction, a 25ms Hamming window with 10ms shifts was adopted. Each utterance was converted into a sequence of 36-dimensional feature vectors, each consisting of 18 MFCC coefficients and their first derivatives. For phonetic feature extraction, we employed an English phoneme recognizer [21] to perform the voice activity detection (VAD) and output the frame level monophone states posterior probability. After log, PCA and MVN, the resulted 52 dimensional tandem features are fused with MFCC at the feature level to get the 88 dimensional hybrid feature [13]. Feature warping is applied to mitigate variabilities.

The training data for NIST 2010 task include Switchboard II part1 to part3, NIST SRE 2004, 2005, 2006 and 2008 corpora on the telephone channel. We trained two gender-dependent GMM UBM models with 256 and 1024 mixture components, respectively. The sizes of i-vectors and the dimension of speaker-specific subspace in PLDA are 600 and 150, respectively. Simple weighted linear summation is adopted here as the score level fusion.

From Table 1 system id 1-3, we can observe that using a mixture of two Gaussians to fit the residual noises outperforms the single Gaussian standard factor analysis by 10% relatively in terms of EER and 2010 norm minDCF. While further expanding the MoG model (K=4) does not help which might be due to the two mode characteristics of the centered mean supervectors as shown in Fig. 1. Furthermore, by comparing the results of system id 4-6, we can find out that MoG i-vector extraction method 2 performs better than extraction method 1 and the single Gaussian factor analysis baseline. This might be because extraction method 2 only replies on the second component which is trained by the valid data with sufficient zero-order statistics. Finally, by fusing the MoG i-vector system (id6) with the i-vector baseline (id4), the overall system's 2010 norm minDCF is further reduced to 0.18.

3.2. Text dependent speaker verification on RSR 2015 part 1

For the text dependent speaker verification task, we used the Part I female portion of the RSR2015 database as our evaluation dataset

Table 2. Performance on the **development** set of Part I for different definitions of target and non-target trials in terms of EER, old cost and new cost (EER/08 norm min DCF/10 norm min DCF)

Speaker Text	Target T	Target F	Imposter T	Imposter F	MoG K=2 extract method 2	MoG K=1 baseline
Trials	tar	non	-	-	**0.76%/0.05/0.3**	1.07%/0.067/0.3
	tar	-	non	-	**6.08%/0.311/0.8**	6.96%/0.357/0.8
	tar	-	-	non	**0.12%/0/0**	0.19%/0.01/0.1

Table 3. Performance on the **evaluation** set of Part I for different definitions of target and non-target trials in terms of EER, old cost and new cost (EER/08 norm min DCF/10 norm min DCF)

Speaker Text	Target T	Target F	Imposter T	Imposter F	MoG K=2 extract method 2	MoG K=1 baseline
Trials	tar	non	-	-	**0.24%/0.01/0.1**	0.41%/0.02/0.1
	tar	-	non	-	**3.97%/0.195/0.6**	4.88%/0.239/0.7
	tar	-	-	non	**0.05%/0/0**	0.11%/0/0

[22]. RSR 2015 is indeed a short duration database which is suitable to test the performance in the short duration scenario. In the RSR2015 database, the number of speakers in the background, development and evaluation sets are 47, 47 and 49, respectively. We only adopt the 36 dimensional MFCC as our features here. We used the same UBM and PLDA configuration as for our text independent experiments but the UBM, i-vector as well as the PLDA models that we tested were trained on the Part I background data. This consists of parallel recordings of 30 TIMIT phrases uttered by 47 female speakers, each of whom participated in 9 recording sessions on 3 different recording devices. We used the same development and evaluation data in [22] to demonstrate the system performance and we did not use the development data for training.

The number of trials for each of the four text dependent speaker verification conditions on the part I of the RSR 2015 database is shown [22]. We can see that only the target speaker uttering the true lexicon content is considered as the target trial, the other cases are all non-target trials. In order to show the results for all three types of non-target trials, we evaluate the system performance separately for each type of trials the same way as in [22]. The gender-dependent GMM UBM size, the i-vector dimension and the PLDA speaker-specific subspace rank are 1024, 400 and 150, respectively.

Performance on the development and evaluation sets of Part I for different definitions of target and non-target trials are shown in Table 2 and 3. We can see that for all three sets of trials, the proposed MoG i-vector method outperforms the i-vector baseline by more than 13% relatively. Especially on the evaluation set, we observe a larger improvement (close to 20% relatively) compared to the NIST SRE 2010 task. This might be due to the nature of data sparsity in the short duration speaker verification task and more dimensions are affected by the insufficient zero-order statistics.

4. CONCLUSIONS

This paper presents a mixture of Gaussian factor analysis based representation framework for speaker verification. Due to the sparsity of the frame level posterior probability and the short duration characteristics, some dimensions of the mean supervector may not be Gaussian distributed. Therefore, we extend the standard factor analysis by replace the single Gaussian with a mixture of Gaussians to better represent the residual noises. If the MoG size equals to 1, then the MoG factor analysis is exactly the same as the conversional factor analysis. If the MoG size is greater than 1, then the proposed MoG factor analysis model has the capability to represent more complex residual noises and therefore achieves better performances on the speaker verification tasks.

5. REFERENCES

[1] N. Dehak, P.A. Torres-Carrasquillo, D. Reynolds, and R. Dehak, "Language recognition via i-vectors and dimensionality reduction," in *Proc. INTERSPEECH*, 2011, pp. 857–860.

[2] N. Dehak, P.J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[3] David Martinez, Oldrich Plchot, Lukas Burget, Ondrej Glembek, and Pavel Matejka, "Language recognition in ivectors space," in *Proc. INTERSPEECH*, 2011, pp. 861–864.

[4] A.O. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *Proc. INTERSPEECH*, 2006, vol. 4, pp. 1471–1474.

[5] W.M. Campbell, DE Sturim, and DA Reynolds, "Support vector machines using gmm supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, 2006.

[6] WM Campbell, DE Sturim, DA Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Proc. ICASSP*, 2006, vol. 1, pp. 97–100.

[7] M. Li, X. Zhang, Y. Yan, and S. Narayanan, "Speaker verification using sparse representations on total variability i-vectors," in *Proc. INTERSPEECH*, 2011, pp. 4548–4551.

[8] S.J.D. Prince and J.H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. ICCV*, 2007, pp. 1–8.

[9] P. Matejka, O. Glembek, F. Castaldo, MJ 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*, 2011, pp. 4828–4831.

[10] Daniel Garcia-Romero and Carol Y Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems.," in *Proc. INTERSPEECH*, 2011, pp. 249–252.

[11] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Proc. ICASSP*, 2014.

[12] P Kenny, V Gupta, T Stafylakis, P Ouellet, and J Alam, "Deep neural networks for extracting baum-welch statistics for speaker recognition," in *Proc. Odyssey*, 2014.

[13] Ming Li and Wenbo Liu, "Speaker verification and spoken language identification using a generalized i-vector framework with phonetic tokenizations and tandem features," in *Proc. INTERSPEECH*, 2014.

[14] Luis Fernando D'Haro, Ricardo Cordoba, Christian Salamea, and Julin David Echeverry, "Extended phone log-likelihood ratio features and acoustic-based i-vectors for language recognition," in *Proc. ICASSP*. IEEE, 2014, pp. 5379–5383.

[15] Haipeng Wang, Cheung-Chi Leung, Tan Lee, Bin Ma, and Haizhou Li, "Shifted-delta mlp features for spoken language recognition," *IEEE Signal Processing Letters*, vol. 20, no. 1, pp. 15–18, 2013.

[16] Qian Zhao, Deyu Meng, Zongben Xu, Wangmeng Zuo, and Lei Zhang, "Robust principal component analysis with complex noise," in *Proc. ICML*, 2014.

[17] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[18] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 345–354, 2005.

[19] Ming Li and Shrikanth Narayanan, "Simplified supervised i-vector modeling with application to robust and efficient language identification and speaker verification," *Computer speech and language*, vol. 28, pp. 940–958, 2014.

[20] NIST, "The NIST 2010 Speaker Recognition Evaluation Plan," *www.itl.nist.gov/iad/mig/tests/spk/2010/index.html*, 2010.

[21] P. Schwarz, P. Matejka, and J. Cernocky, "Hierarchical structures of neural networks for phoneme," in *Proc. ICASSP*, 2006, pp. 325–328, Software available at http://speech.fit.vutbr.cz/software/phoneme-recognizer-based-long-temporal-context.

[22] Anthony Larcher, Kong Aik Lee, Bin Ma, and Haizhou Li, "Text-dependent speaker verification: Classifiers, databases and rsr2015," *Speech Communication*, vol. 60, pp. 56–77, 2014.