PLDA Modeling in the Fishervoice Subspace for Speaker Verification

Jinghua Zhong[1], *Weiwu Jiang*[1], *Wei Rao*[2], *Man-Wai Mak*[2], *Helen Meng*[1]

[1]Department of Systems Engineering & Engineering Management,
The Chinese University of Hong Kong, Hong Kong SAR of China,
[2]Department of Electronic and Information Engineering,
The Hong Kong Polytechnic University, Hong Kong SAR of China

[1]{jhzhong, wwjiang, hmmeng}@se.cuhk.edu.hk
[2]ellen.wei-rao@connect.polyu.hk, enmwmak@polyu.edu.hk

Abstract

We have previously developed a Fishervoice framework that maps the JFA-mean supervectors into a compressed discriminant subspace using nonparametric Fishers discriminant analysis. It was shown that performing cosine distance scoring (CDS) on these Fishervoice projected vectors (denoted as f-vectors) can outperform the classical joint factor analysis. Unlike the i-vector approach in which the channel variability is suppressed in the classification stage, in the Fishervoice framework, channel variability is suppressed when the f-vectors are constructed. In this paper, we investigate whether channel variability can be further suppressed by performing Gaussian probabilistic discriminant analysis (PLDA) in the classification stage. We also use random subspace sampling to enrich the speaker discriminative information in the f-vectors. Experiments on NIST SRE10 show that PLDA can boost the performance of Fishervoice in speaker verification significantly by a relative decrease of 14.4% in minDCF (from 0.526 to 0.450).

Index Terms: supervector, joint factor analysis, random sampling, Fishervoice, Probabilistic Linear Discriminant Analysis

1. Introduction

Gaussian Mixture Model (GMM) [1] based Joint Factor Analysis (JFA) [2] has laid down the foundation for many state-of-the-art speaker recognition systems. The goal of JFA is to find an optimal linear subspace that represents speaker information with minimal influence of channel noise and inter-session effects. To deal with the problems of high processing complexity and over-fitting, we have proposed a novel speaker recognition framework named Fishervoice in [3][4]. Based on nonparametric Fisher's discriminant analysis, the framework maps JFA-mean supervectors[1] into multiple discriminant subspaces. Such algorithm can reduce dimensionality through reducing unfavorable intra-speaker variability; it can also exploit the discriminative information such as classification boundaries in the multiple discriminative subspaces.

Besides, since the dimension of the subspaces is relatively high when compared to the number of training samples, the constructed subspace classifier is often biased and unstable. We proposed to use random subspace sampling to address this problem [5] (denoted hereafter by random Fishervoice). We randomly sampled the feature space into a number of subspaces. For every discriminant random subspace, we used Fishervoice to model the intrinsic vocal characteristics. The complex speaker characteristics were modeled through multiple f-vectors. Then multiple subspace dependent classifiers constructed for these f-vectors were fused to produce a more powerful classifier that covers most of the feature space.

Based on JFA, Dehak et al. [6] proposed an i-vector speaker verification system that compressed both channel and speaker information into a low-dimensional space called total variability space, and accordingly projected the GMM-supervector to a total-factor feature vector called the i-vector. Then Linear Discriminant Analysis (LDA) [6] and Probabilistic LDA (PLDA) [7] were applied to the i-vectors for inter-session compensation. To better model the data distribution, heavy-tailed PLDA [8] was proposed by assuming that the priors on the latent variables in the PLDA model follow a Student's t distribution. Later, it was found that Gaussian based PLDA with length normalization [9] achieves similar performance as heavy-tailed PLDA with less computation resource. However, as demonstrated in [10][11], there are some limitations in the i-vector representation of speech segments, such as sensitivity to segment durations.

PLDA has shown to be a good inter-session compensation method for the i-vector framework. It is also a subspace modeling method for dimension reduction. In [12], PLDA was used in the supervector space. But the results show that PLDA may not work well for high-dimensional supervectors. In our previous work, the JFA based GMM supervectors were projected by Fishervoice transformation to a low dimensional vector – the f-vector. This inspires us to explore the use of PLDA on f-vectors for further inter-session compensation in this study. We also investigate using random subspace sampling to enrich the speaker discriminative information in the f-vectors.

The rest of the paper is organized as follows. In Section 2, we describe the background of Fishervoice and Gaussian PLDA. Then we describe the details of the proposed framework in Section 3. Implementation and experimental results on the NIST SRE10 male core task (common conditions 5 and 6) are presented in Sections 4 and 5. Finally, the conclusions are presented in Section 6.

2. Background

2.1. Fishervoice

Fishervoice [4] aims to enhance performance by extracting discriminant information from the within-speaker scatter matrices S_ω and between-speaker scatter matrix S_b effectively. The overall projection matrix of Fishervoice comprises three transfor-

[1]The JFA-mean supervector of an utterance is a GMM supervector obtained from the JFA model.

mations:

1. Perform PCA for dimension reduction with the subspace projection W_1, producing f_1:

$$f_1 = W_1^T x, \text{where } W_1 = \underset{W:\|\boldsymbol{w}_i\|=1}{\arg\max} \left\| W^T \Psi W \right\| \quad (1)$$

where x is an arbitrary GMM-supervector and Ψ is the covariance matrix of all of the supervectors in the training set.

2. Apply whitening to reduce intra-speaker variations with the matrix W_2, producing f_2:

$$f_2 = W_2^T f_1, \text{where } W_2^T S_\omega W_2 = I, W_2 = \Phi \Lambda^{-\frac{1}{2}} \quad (2)$$

where S_ω is the standard within-class scatter matrix, Φ is the normalized eigenvector matrix of S_ω, and Λ is the eigenvalue matrix of S_ω.

3. Extract discriminative speaker class boundaries information by subspace projection matrix W_3 — from the above whitened subspace, f_3 is obtained using the nonparametric between-class scatter matrix S_b' according to Eqs. 8 – 9 in [3]:

$$f_3 = W_3^T f_2, \text{where } W_3 = \underset{W:\|\boldsymbol{w}_i\|=1}{\arg\max} \left\| W^T S_b' W \right\| \quad (3)$$

Finally, the overall subspace projection matrix W_{NF} is given by:

$$W_{NF} = W_1 W_2 W_3 \quad (4)$$

2.2. Gaussian PLDA

The traditional Linear Discriminant Analysis (LDA) aims to find a linear transformation that maximizes the between-class separation and minimizes the within-class variation. Recently, Ioffe [13] and Prince et al. [7] proposed a probabilistic approach called probabilistic LDA (PLDA) which applies generative factor analysis modeling to solve the subspace recognition problem. Kenny et al. [8] introduced heavy-tailed PLDA which uses Student's t distributions instead of Gaussian distribution to model the i-vectors. Significant performance improvement was demonstrated, but the system was complicated and computationally demanding. Later, a simple length normalization scheme [9] was proposed to deal with the non-Gaussian behavior of i-vectors, which allows the use of probabilistic models with Gaussian assumptions. This non-linear transformation simplifies the second step of Radial Gaussianization proposed in [14] by scaling the length of each whitened i-vector to unit length. In this way, PLDA with Gaussian assumptions can achieve a performance comparable to that of heavy-tailed PLDA. In this paper, we focus on PLDA with Gaussian assumptions, named Gaussian PLDA.

Suppose each speaker i has H_i utterances. The Gaussian PLDA model assumes that each length-normalized speaker vector η_{ih} can be decomposed as

$$\eta_{ih} = m + \Phi\beta_i + \Gamma\alpha_{ih} + \epsilon_{ih} \quad (5)$$

where m is a global offset, the columns of Φ provides a basis for the speaker-specific subspace (i.e. eigenvoices), Γ provides a basis for the channel subspace (i.e. eigenchannels), β_i and α_{ih} are the corresponding latent vectors and ϵ_{ih} is a residual term. Besides, β_i and α_{ih} are both assumed to have standard normal distributions, and ϵ_{ih} follows a Gaussian distribution with zero mean and diagonal covariance matrix Σ. If Σ is assumed to be a full covariance matrix, then the eigenchannels can be absorbed into Σ and the modified model becomes:

$$\eta_{ih} = m + \Phi\beta_i + \epsilon_{ih} \quad (6)$$

3. PLDA for Fishervoice Subspace

In this section, we describe the three fundamental components of our speaker recognition system. Namely, supervector extraction, f-vector transformation, PLDA modeling and verification score computation. Figures 1 and 2 illustrate the overall organization of the proposed framework.

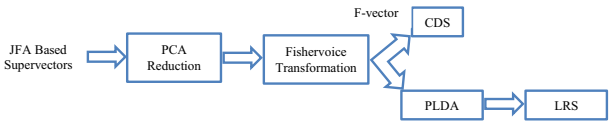

Figure 1: *Framework of f-vectors using cosine distance scoring and f-vectors using PLDA modeling.*

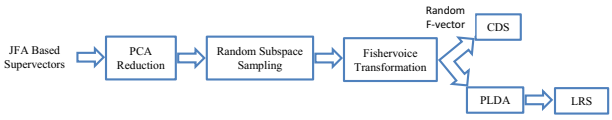

Figure 2: *Framework of random f-vectors using cosine distance scoring and random f-vectors using PLDA modeling.*

3.1. Supervector extraction

The structure of GMM-UBM supervectors captures the probabilistic distribution of acoustic feature classes in the overall acoustic space. Therefore, we concatenate the mean vector of GMM-UBM for JFA modeling. In the JFA theory [2], the speaker- and channel-dependent mean-supervector M_{ih} of the h-th utterance from speaker i is further decomposed into four supervectors:

$$M_{ih} = m + Vy_i + Dz_{ih} + Ux_{ih} \quad (7)$$

where m is the UBM mean supervector, U is an eigenchannel matrix, V is an eigenvoice matrix, D is a diagonal residual scaling matrix, x_{ih} is a channel- and session-dependent eigenchannel factor, y_i is a speaker-dependent eigenvoice factor and z_{ih} is the speaker residuals. The speaker supervector is defined by the first three parts of Eq. 7 as follows:

$$s_{ih} = m + Vy_i + Dz_{ih}, \quad (8)$$

which forms the input to the PCA block in Figures 1 and 2.

3.2. f-vector transformation and PLDA modeling

Since the dimensionality of supervector is too high for direct subspace analysis, the high dimensional feature vector is first divided into K subvectors equally for efficient subspace analysis. Because the dimension of the supervector is relatively high when compared with the number of training samples, the constructed subspace classifier through Fishervoice [4] is often biased and unstable. Therefore we performed PCA on each subvector space to reduce the sub-vectors to L-dimension vectors.

Then all of the projected subvectors are concatenated, followed by a second level PCA dimension reduction to reduce the overall dimension to J. This produces the output of the PCA block in Figures 1 and 2.

To further reduce dimensionality and reduce unfavorable intra-speaker variability and to extract discriminative information such as classification boundaries, we performed Fishervoice on the PCA-projected supervectors to obtain a highly compressed classification subspace, namely f-vector space (see Figure 1). Besides, we randomly sampled the PCA-projected feature space to form a set of low-dimensional subspaces, each spans by primary E_1 dimensions plus randomly selected E_2 dimensions. For every random subsapce, we used Fishervoice to model the intrinsic vocal characteristics. In this way, we obtained a number of random f-vectors (see Figure 2) for each utterances.

To facilitate comparison of f-vectors in a verification trial, we performed Gaussian PLDA to further suppress channel variability in the classification stage.

3.3. Verification score

3.3.1. Cosine Distance Scoring (CDS)

For our previous Fishervoice and random Fishervoice framework, the distance score is calculated between the training and testing f-vectors $(\theta_{train}, \theta_{test})$ in terms of the normalized correlation (COR) as follows:

$$S_{CDS} = \frac{\left\|\theta_{train}^T \theta_{test}\right\|}{\sqrt{\theta_{train}^T \theta_{train} \theta_{test}^T \theta_{test}}}. \quad (9)$$

For the random Fishervoice framework, the outputs are weighted and combined. The weights are obtained by grid search until the lowest minDCF on the training set is found.

3.3.2. Log-likelihood Ratio Scoring (LRS)

The PLDA verification score is calculated as the log-likelihood ratio between two hypotheses $\{H_s, H_d\}$ [9]:

$$S_{LRS} = \log[\frac{p(\eta_1, \eta_2|H_s)}{p(\eta_1, \eta_2|H_d)}] \quad (10)$$

where H_s hypothesizes that $\{\eta_1, \eta_2\}$ belong to the same speaker and H_d hypothesizes that they belong to different speakers. The solution for Eq. 10 is given by:

$$S_{LRS} = \eta_1^T Q \eta_1 + \eta_2^T Q \eta_2 + 2\eta_1^T P \eta_2 + const, \quad (11)$$

with

$$\begin{aligned} Q =& \Sigma_{tot}^{-1} - (\Sigma_{tot} - \Sigma_{ac}\Sigma_{tot}^{-1}\Sigma_{ac})^{-1}, \\ P =& \Sigma_{tot}^{-1}\Sigma_{ac}(\Sigma_{tot} - \Sigma_{ac}\Sigma_{tot}^{-1}\Sigma_{ac})^{-1}, \end{aligned} \quad (12)$$

where $\Sigma_{ac} = \Phi\Phi^T$ and $\Sigma_{tot} = \Phi\Phi^T + \Sigma$.

4. Experimental Setup

4.1. Testing protocol

All experiments were performed on the NIST 2010 SRE male core-core data set of Common Conditions 5 and 6. The training and testing speech files comprise telephone conversations with 353 true target trials and 13,707 imposter trials for cc5 and 178 true target trials and 12,825 imposter trials for cc6. There is no cross-gender trials. Performance evaluation is given in terms of Equal Error Rates (EER) and the new Minimum Detection Cost Function (DCF).

4.2. Feature extraction

First, ETSI Adaptive Multi-Rate (AMR) GSM VAD [15] was applied to prune out the silence region of the speech files. Then the speech signals were segmented into frames by a 25ms Hamming window with a 10ms frame shift. The first 16 Mel frequency cepstral coefficients and log energy were calculated; together with their first and second derivatives. A 51-dimensional feature vector was obtained for each frame (the frequency window was restricted to 300-3400 Hz). Finally, feature warping [16] was applied to the MFCC features.

4.3. The baseline system

The baseline system employed gender-dependent 2,048-Gaussian UBMs with JFA. First, we trained the UBMs using NIST 2004-2006 SRE male telephone speech utterances, including 4,222 recordings.

Then, for the JFA part, we trained the gender-dependent eigenvoice matrix V using Switchboard II Phases 2 and 3, Switchboard Cellular Part 2, NIST 2004-2006 SRE, including 893 male speakers with 11,204 utterances. The rank of the speaker space was set to 300. The eigenchannel matrix U was also trained in a gender-dependent manner from 436 male speakers with 5,410 utterances from NIST 2004-2006 SRE. The rank of the channel space was set to 100. The diagonal residual scaling matrix D was extracted from the UBM covariance without EM estimation. We used an expectation maximization (EM) algorithm with 20 iterations for all of the above training.

4.4. f-vector transformation and PLDA modeling

For the Fishervoice and random Fishervoice frameworks, the gender-dependent Fishervoice projection matrices were constructed from telephone speeches in NIST 2004-2006 SRE, Switchboard II Phase 2, Phase 3 and Switchboard Cellular Part 2. This amounts to 563 male speakers altogether, each with not less than 8 different utterances. For random Fishervoice, the Fishervoice projection matrices, W_1, W_2 and W_3, have dimensions $(E_1 + E_2) \times 1200$, 1200×1199, 1199×550, respectively. These correspond to the upper limit of their matrix ranks. The parameter (R in Eq. 8 of [3]) that controls the number of nearest neighbors for constructing nonparametric between-class scatter matrix S_b' was set to 4, according to the median number of sessions for each speaker. The number of slice K is set to 16. Besides, the parameters L and J for the PCA dimension reduction before Fishervoice was set to 4,000.

4.5. Score normalization

We used gender-dependent score normalization (TZ-norm) for cosine distance scoring. The SRE04, SRE05 and SRE06 corpora were adopted for T-norm and Switchboard II Phases 2 and 3 for Z-norm. The number of speakers was 400 for T-norm and 622 for Z-norm. As reported in [8], s-norm is more effective than standard score normalization methods for likelihood ratio scoring. In this work, all speakers for T-norm and Z-norm were used for s-norm.

5. Results

5.1. f-vector/PLDA approach

The first experiment is to investigate the effectiveness of PLDA modeling in the Fishervoice framework. We studied the system performance with regard to the dimension of Fishervoice

projection. We restricted the dimension of the last nonparametric between-class projection matrix to a constant value of 550. This dimensionality keeps $98\% \sim 99\%$ of the variational energy retained in the eigenspace matrices. Our previous experience suggests that this arrangement can achieve the best performance.

Table 1 shows the performance comparison of f-vectors using PLDA modeling (denoted as f-vector/PLDA) with f-vector using cosine distance scoring (denoted as f-vector/CDS) on male core task of common condition 5 in NIST 2010 SRE. The best EER and minDCF are highlighted for both systems. The results suggest that PLDA modeling can significantly improve the performance in minDCF with a small loss in EER performance. Besides, results of f-vector/CDS suggest that keeping the rank of W_1 and W_2 high seems to give better performance. However, it takes longer to train the Fishervoice transformation matrices. On the other hand, for f-vector/PLDA, keeping the rank of W_1 and W_2 small improves performance.

Table 2 shows the results of common condition 6. These results further verify that PLDA modeling can improve the performance in minDCF with only a small loss in EER performance. For the combination of (800,799,550), f-vector/PLDA with snorm performs better than f-vector/CDS in terms of both EER and minDCF.

Table 1: *Comparison of f-vector/PLDA with f-vector/CDS on NIST SRE10 male core task (cc=5). The performance is reported in EER(%) and 1000 × minDCF.*

Fishervoice dim (W_1, W_2, W_3)	f-vector/CDS	f-vector/PLDA	
		no norm	snorm
(800,799,550)	3.63,0.617	4.18,**0.442**	4.20,**0.493**
(900,899,550)	3.39,0.610	4.24,0.450	**3.88**,0.501
(1000,999,550)	3.68,0.571	4.24,0.470	3.96,0.546
(1100,1099,550)	3.67,0.560	**4.15**,0.479	3.96,0.543
(1200,1199,550)	3.68,0.552	4.42,0.464	4.17,0.527
(1300,1299,550)	**3.39**,0.537	4.53,0.479	4.24,0.535
(1400,1399,550)	3.64,**0.526**	4.41,0.450	4.25,0.507

Table 2: *Comparison of f-vector/PLDA with f-vector/CDS on NIST SRE10 male core task (cc=6). The performance is reported in EER(%) and 1000 × minDCF.*

Fishervoice dim (W_1, W_2, W_3)	f-vector/CDS	f-vector/PLDA	
		no norm	snorm
(800,799,550)	4.78,**0.819**	**5.05,0.786**	**4.49,0.814**
(900,899,550)	5.05,0.831	5.44,0.814	4.93,0.825
(1000,999,550)	5.00,0.831	5.60,0.819	5.05,0.831
(1100,1099,550)	4.78,0.847	5.62,0.797	5.05,0.831
(1200,1199,550)	4.75,0.853	5.62,0.825	5.05,0.836
(1300,1299,550)	**4.49**,0.864	5.61,0.825	5.05,0.831
(1400,1399,550)	4.49,0.864	5.50,0.836	5.05,0.831

5.2. Random f-vector/PLDA approach

This set of experiments are to investigate the effectiveness of PLDA modeling in random Fishervoice framework with regards to the different dimensions of E_1 and E_2. We restricted the dimensionality of $(E_1 + E_2)$ to a constant value of 2500 for dimension reduction. Tables 3 and 4 summarize the results obtained with the best/worst individual and fused systems on the five combinations of (E_1, E_2). No score normalization was performed for random f-vector/PLDA. For each combination of $(E_1 + E_2)$, we created 5 different subspaces randomly and applied linear fusion to fuse the scores arising from the subspace PLDA models. In addition, in an attempt to make the fusion process balanced and to assess the risk of using the worst system for fusion in actual deployment, we also selected the best/worst individual systems from each combination of (E_1, E_2) and fused them together to produce the fusion results, namely total best/worst fusion.

From the table, we observe that: (1) PLDA modeling can significantly improve the performance in terms of minDCF with little loss in EER. (2) The lowest minDCF achieved across all fused results of random f-vector/PLDA is 0.425 in total best fusion for cc5 and 0.713 in (300,2200) for cc6. (3) Compared to f-vector/PLDA, random f-vector/PLDA can achieve a further reduction of 3.8% (from 0.442 to 0.425) for cc5 and 9.3% (from 0.786 to 0.713) for cc6 in minDCF.

Table 3: *Comparison of random f-vector/PLDA with random f-vector/CDS on NIST SRE10 male core task (cc5), in terms of EER(%) and 1000 × minDCF.*

(E_1, E_2)	random f-vector/CDS			random f-vector/PLDA		
	best	worst	fused	best	worst	fused
(300,2200)	3.39 0.566	3.65 0.608	3.39 0.563	4.43 **0.442**	4.77 **0.503**	3.96 **0.436**
(400,2100)	3.26 0.583	3.39 0.617	3.39 0.580	4.14 **0.445**	4.25 **0.464**	4.14 **0.425**
(500,2000)	3.38 0.560	3.32 0.600	3.38 0.560	4.17 **0.439**	3.96 **0.467**	4.13 **0.433**
(600,1900)	3.56 0.585	3.62 0.608	3.56 0.585	3.96 **0.439**	3.96 **0.493**	3.96 **0.439**
(700,1800)	3.57 0.602	3.68 0.611	3.39 0.596	4.12 **0.439**	4.25 **0.453**	4.12 **0.439**
Total best fusion	3.37,0.557			3.95,**0.425**		
Total worst fusion	3.29,0.600			3.96,**0.436**		

Table 4: *Comparison of random f-vector/PLDA with random f-vector/CDS on NIST SRE10 male core task (cc6), in terms of EER(%) and 1000 × minDCF.*

(E_1, E_2)	random f-vector/CDS			random f-vector/PLDA		
	best	worst	fused	best	worst	fused
(300,2200)	5.04 0.797	5.05 0.847	4.79 0.797	5.05 **0.718**	5.61 **0.780**	5.04 **0.713**
(400,2100)	5.54 0.803	4.81 0.852	5.33 0.803	5.97 **0.746**	5.62 **0.792**	5.60 **0.746**
(500,2000)	5.39 0.819	5.37 0.853	5.39 0.819	5.60 **0.769**	6.09 **0.830**	5.56 **0.769**
(600,1900)	5.44 0.819	5.05 0.859	5.05 0.819	5.55 **0.774**	5.40 **0.819**	5.55 **0.774**
(700,1800)	5.34 0.825	5.05 0.836	4.99 0.814	5.40 **0.780**	5.37 **0.814**	5.45 **0.774**
Total best fusion	5.04,0.791			5.05,**0.718**		
Total worst fusion	4.86,0.825			5.60,**0.780**		

6. Conclusions

This paper performs Gaussian PLDA modeling on two Fishervoice-based frameworks, Fishervoice and random Fishervoice, for further inter-session compensation. Experiments on the NIST SRE10 male core-core data set showed that f-vector/PLDA led to significant reduction in minDCF, although the performance did not improve consistently over the whole DET curve. Besides, we have achieved a further reduction in minDCF by using random subspace sampling.

7. References

[1] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital signal processing*, vol. 10, no. 1, pp. 19–41, 2000.

[2] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 980–988, 2008.

[3] Z. Li, W. Jiang, and H. Meng, "Fishervioce: A discriminant subspace framework for speaker recognition," in *2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010, pp. 4522–4525.

[4] W. Jiang, H. Meng, and Z. Li, "An enhanced fishervoice subspace framework for text-independent speaker verification," in *2010 7th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2010, pp. 300–304.

[5] W. Jiang, Z. Li, and H. M. Meng, "An analysis framework based on random subspace sampling for speaker verification." in *Interspeech*, 2011, pp. 253–256.

[6] N. Dehak, P. 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.

[7] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *IEEE 11th International Conference on Computer Vision*, 2007, pp. 1–8.

[8] P. Kenny, "Bayesian speaker verification with heavy tailed priors," in *Speaker and Language Recognition Workshop (IEEE Odyssey)*, 2010.

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

[10] P. Kenny, T. Stafylakis, P. Ouellet, M. Alam, P. Dumouchel *et al.*, "PLDA for speaker verification with utterances of arbitrary duration," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7649–7653.

[11] P. Kenny, T. Stafylakis, P. Ouellet, and M. J. Alam, "JFA-based front ends for speaker recogntion," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[12] Y. Jiang, K.-A. Lee, Z. Tang, B. Ma, A. Larcher, and H. Li, "PLDA modeling in i-vector and supervector space for speaker verification." in *Interspeech*, 2012.

[13] S. Ioffe, "Probabilistic linear discriminant analysis," in *ECCV*, 2006, pp. 531–542.

[14] S. Lyu and E. P. Simoncelli, "Nonlinear extraction of independent components of natural images using radial Gaussianization," *Neural Computation*, vol. 21, no. 6, pp. 1485–1519, 2009.

[15] *Digital cellular telecommunication system (Phase 2+); Voice Activity Detect or VAD for AdaptiveMulti Rate (AMR) speech traffic channels; General description*. ETSI, GSM 06.94, 1999.

[16] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. ISCA Workshop on Speaker Recognition: A Speaker Oddyssey*, 2001.