Odyssey 2014:
The Speaker and Language Recognition Workshop
16-19 June 2014, Joensuu, Finland

An Integration of Random Subspace Sampling and Fishervoice for Speaker Verification

Jinghua Zhong[1], *Weiwu Jiang*[1], *Helen Meng*[1,2], *Na Li*[2,3] *and Zhifeng Li*[2]

[1]Department of Systems Engineering & Engineering Management,
The Chinese University of Hong Kong, Hong Kong SAR of China,
[2]Shenzhen Institutes of Advanced Technology, CAS, China,
[3]Northwestern Polytechnical University, Xian, China

{jhzhong,wwjiang,hmmeng}se.cuhk.edu.hk, {na.li, zhifeng.li}siat.ac.cn

Abstract

In this paper, we propose an integration of random subspace sampling and Fishervoice for speaker verification. In the previous random sampling framework [1], we randomly sample the JFA feature space into a set of low-dimensional subspaces. For every random subspace, we use Fishervoice to model the intrinsic vocal characteristics in a discriminant subspace. The complex speaker characteristics are modeled through multiple subspaces. Through a fusion rule, we form a more powerful and stable classifier that can preserve most of the discriminative information. But in many cases, random subspace sampling may discard too much useful discriminative information for high-dimensional feature space. Instead of increasing the number of random subspace or using more complex fusion rules which increase system complexity, we attempt to increase the performance of each individual weak classifier. Hence, we propose to investigate the integration of random subspace sampling with the Fishervoice approach. The proposed new framework is shown to provide better performance in both NIST SRE08 and NIST SRE10 evaluation corpora. Besides, we also apply Probabilistic Linear Discriminant Analysis (PLDA) on the supervector space for comparision. Our proposed framework can improve PLDA performance by a relative decrease of 12.47% in EER and reduced the minDCF from 0.0216 to 0.0210.

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

1. Introduction

In the field of the speaker verification, Gaussian Mixture Model (GMM) [2] based Joint Factor Analysis (JFA) [3] has become a popular approach for many systems. It achieves performance improvements by addressing issues relating to speaker characteristics and channel variability. However, due to numeric instability and sparsity of the supervectors in a high-dimensional space, the approach also has over-fitting problems during the model training process, which limits performance improvement.

Based on JFA, Dehak et al. [4] proposes an i-vector feature-based speaker verification system which not only reduces the computational time but also achieves high performance on the NIST evaluation. The system compressed both channel and speaker information into a low-dimensional space called total variability space, and accordingly projectes each supervector to a total factor feature vector called the i-vector. Then Linear Discriminant Analysis (LDA) [4] and Probabilistic LDA (PLDA) [5] are applied on the i-vectors for further dimension reduction. To better model the data distribution, heavy-tailed PLDA [6] 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 [7] achieves similar performance as heavy-tailed PLDA with less computation resource.

Due to the complexity of the speaker recognition problem, it is difficult to pursue a single optimal classifier to meet all requirements. Therefore, instead of developing a single optimal classifier, we proposed an ensemble learning framework based on random subspace sampling [8][9] on JFA feature space [1] (denoted hereafter by JFA + subspace). For every random subspace, we use Fishervoice [10][11] to model the intrinsic vocal characteristics in a discriminant subspace. The complex speaker characteristics are modeled through multiple subspaces. Random subspace [8] is a popular random sampling technique to strengthen weak classifiers. Random subspace sampling can alleviate the overfitting problem since it samples a set of low-dimensional subspaces to reduce the discrepancy between the training set size and the feature vector length. Then multiple stable Fishervoice classifiers constructed in each random subspaces are fused to produce a more powerful classifier that covers most of the feature space. Such an algorithm is inspired by the success of subspace modeling in face recognition [12][13][14].

In many cases, random subspace sampling may discard too much useful discriminative information for a high-dimensional feature space. Although increasing the number of random subspace and using more complex fusion rules may address this issue to maintain performances, the method will also increase system complexity and computational burden. A better approach to improve the combined classifiers is to increase the performance of each individual weak classifier. Hence, we propose to investigate the integration of random subspace sampling with the Fishervoice approach (denoted hereafter by Fishervoice + subspace). Such combination may lessen unnecessary loss of information and maximize class separability to build classifiers for the sampled features. Finally, we apply linear fusion for the classifiers to produce the final classification output.

The rest of the paper is organized as follows: In Section 2 we describe the background of JFA, PLDA and our previous JFA + subspace method. Then we describe the details of the proposed framework. Implementation and experimental results on the NIST SRE08 male short2-short3 task of tel-tel condition

and NIST SRE10 core-core task of common condition 6 are presented respectively in sections 4 and 5. Finally, the conclusions are presented in section 6.

2. Background

2.1. Joint Factor Analysis (JFA)

According to the JFA theory [3], the speaker and channel noise components, which reside in the speaker- and channel-dependent supervectors respectively, have Gaussian distributions. Given the Gaussian mean of the utterance h from the speaker i who has data from H_i utterances, we get a $G \times F$ dimension supervector $M_{i,h}$ by concatenating the GMM-UBM speaker vectors.

$$M_{i,h} = [s_{i,h,1}\ s_{i,h,2}\ ...\ s_{i,h,g}\ ...\ s_{i,h,G}]^T \quad (1)$$

where $s_{i,h,g}$ is the F-dimensional GMM-UBM speaker vector for the g-th Gaussian mixture. Then $M_{i,h}$ is further decomposed into four supervectors:

$$M_{ih} = m + Vy_i + Dz_{ih} + Ux_{ih} \quad (2)$$

where m is the UBM supervector mean, U is the eigenchannel matrix, V is the eigenvoice matrix, D is the diagonal residual scaling matrix, x_{ih} is the channel- and session-dependent eigenchannel factor, y_i is the speaker-dependent eigenvoice factor and z_{ih} is the speaker residuals. Based on the result of [11], the GMM speaker supervector is substituted by the first three parts of Eq.(1) as follows:

$$s_{ih} = m + Vy_i + Dz_{ih} \quad (3)$$

2.2. Probabilistic Linear Discriminant Analysis (PLDA)

Recently, Ioffe [15] and Prince et al. [5] proposed a novel approach called probabilistic LDA (PLDA) which applied generative factor analysis modeling to solve the subspace recognition problem. Suppose each speaker i has multiple H_i utterances. The PLDA theory assumes that each speaker vector η_{ih} can be decomposed as

$$\eta_{ih} = m + \Phi\beta_i + \Gamma\alpha_{ih} + \epsilon_{ih} \quad (4)$$

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 corresponding latent identification vectors and ϵ_{ih} is a residual term. In [4] , β_i and α_{ih} are both assumed to have standard normal distributions, ϵ_{ih} is Gaussian with zero mean and diagonal covariance matrix Σ. Under Gaussian assumptions, α_{ih} can always be eliminated and the modified model becomes:

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

Kenny et al. [6] introduced heavy-tailed PLDA which used student's t distributions to model β_i, α_{ih} and ϵ_{ih}. Then a simple length normalization [7] was proposed to deal with non-Gaussian behavior of the speaker vector, which allowed the use of probabilistic models with Gaussian assumptions. This non-linear transformation simplifies the second step of Radial Gaussianization proposed in [16] by scaling the length of each whitening transformed speaker vector η_{wht} to unit length. Most frameworks based on PLDA work with i-vector representations. Both the PLDA model and i-vector extractor involve dimension reduction through similar methods of subspace modeling. In [17], PLDA was proposed for use in the original supervector space.

2.3. Fishervoice

Since the dimension of the supervector is relatively high when compared with the number of limited training samples, the constructed subspace classifier through Fishervoice [11] is often biased and unstable. The projection vectors may be greatly changed by the slight disturbance of noise in the training set. Besides, the dimension of the final projected subspace is much higher than the i-vector approach [4]. So the original JFA supervector is first processed by PCA dimension reduction to form the principal feature space. Then we randomly sample the PCA projected feature space into a set of low-dimensional subspaces. Classifiers are built for each Fishervoice projected discriminant subspace and their results are integrated to obtain the final decision [1]. The Fishervoice projection can be described as three components:

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

$$f_1 = W_1^T x, \text{where } W_1 = \underset{W:\|\boldsymbol{w}_i\|=1}{\arg\max} \left\| W^T \Psi W \right\| \quad (6)$$

   where x is an arbitrary 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:

$$f_2 = W_2^T f_1, \text{where } W_2^T S_\omega W_2 = I, W_2 = \Phi\Lambda^{-\frac{1}{2}} \quad (7)$$

   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 [10]:

$$f_3 = W_3^T f_2, \text{where } W_3 = \underset{W:\|\boldsymbol{w}_i\|=1}{\arg\max} \left\| W^T S_b' W \right\| \quad (8)$$

Finally, classifiers are constructed in each random subspace. The overall subspace projection matrix W_q for q-th random subspace is given by:

$$W_q = W_{1q} W_{2q} W_{3q} \quad (9)$$

These relative weak classifiers are combined into a powerful and stable classifier afterwards. Such algorithm can therefore preserve nearly all the discriminative information.

3. Random subspace sampling in Fishervoice

Following JFA + subspace framework [1], we propose an integration of random subspace sampling and Fishervoice for speaker verification. The first step of Fishervoice can perform dimension reduction without losing as much discriminative information compared to random subspace sampling, while the second and third steps are most essential in reducing within-class variations and emphasizing discriminative class boundary information. So performing random sampling between the first two steps is less likely to sample unfavorable information after removing some useless information through PCA in the first step. Besides, random subspace sampling would supplement some discriminative information which will be dropped by

PCA. Such combination may lessen unnecessary loss of information and maximize class separability for last two subspace projection of Fishervoice to build more accurate individual classifier for the sampled features.

Figure 1 illustrates the overall organization of the proposed framework. The first three parts above the first dotted line are preprocessing procedure similar to the JFA + subspace method [1]. The middle parts between the two dotted lines are integration of random subspace sampling and Fishervoice. The details are illustrated as follows:

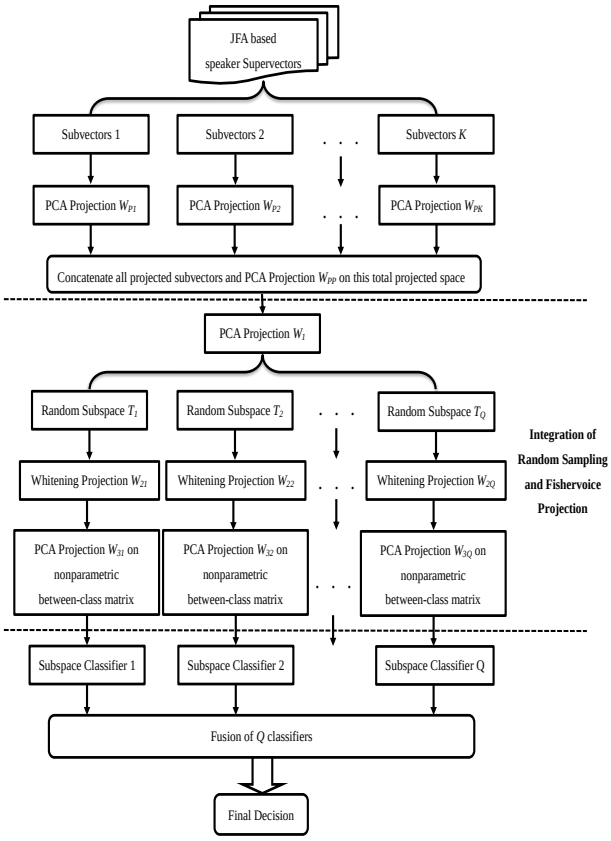

Figure 1: *The overall organization of the proposed framework.*

3.1. Training stage

The training procedure is described as follows:

1. We believe that the structure of JFA speaker supervectors can capture the probabilistic distribution of acoustic feature classes in the overall acoustic space. So according to the JFA-based supervector extraction process in Section 2.1, the speakers' feature supervectors (i.e., the inputs in Figure 1) are extracted from each utterance in the training set.
2. Since the dimensionality of supervector is to high for direct subspace analysis, the high-dimensional supervectors are first divided into K subvectors equally for PCA dimension reduction. The dimensionality of each subvectors is reduced to L by $W_{Pk}(k = 1, 2, ..., K)$. Then all projected subvectors are concatenated to conduct a second level PCA dimension reduction with W_{pp}. The dimensionality of the projected supervector $O_{i,h}$ is further reduced from $K \times L$ to J.
3. We then perform integration of random subspace sampling with Fishervoice. A set of random subspace $\{f_{11}, f_{12}, ..., f_{1Q}\}$ is obtained from the projected subspace f_1 in section 2.2, each of them span $(E_1 + E_2)$ dimensions. The primary E_1 dimensions are fixed so as to keep the first E_1 largest eigenvalues in W_1, which can preserve the main intrapersonal variations. The remaining E_2 dimensions are randomly selected from the remaining $(P - E_1)$ dimensional space, where P is the dimensionality of the supervector after the first step of Fishervoice. Then in each subspace $f_{1q}(q = 1, 2, ..., Q)$, we perform the corresponding projection matrices W_2W_3. Thus we generate Q classifiers in total.
4. During target speaker enrollment, we perform the integration of random subspace sampling and Fishervoice on each projected target speaker supervector from step 2 and form the final Q training reference vector $\theta_{train}(q)$.

3.2. Testing stage

In the testing stage, we obtain the corresponding input feature vector with similar method as the training stage. Then each testing supervector is projected into the testing reference vector $\theta_{test}(r)$ via the q^{th} random subspace in the same manner as the enrollment process. After that, the distance score is calculated between $\theta_{train}(r)$ and $\theta_{test}(r)$ in terms of the normalized correlation (COR) as follows:

$$D(\theta_{train}(r), \theta_{test}(r)) = \frac{\left\|\theta_{train}(r)^T \theta_{test}(r)\right\|}{\sqrt{\theta_{train}(r)^T \theta_{train}(r) \theta_{test}(r)^T \theta_{test}(r)}} \quad (10)$$

Finally, the outputs are weighted and combined. The weights are obtained by grid search based on the training set with values giving the lowest EER.

4. Experimental setup

4.1. Testing protocol

All experiments are performed on the NIST SRE08 male short2-short3 task of tel-tel condition and NIST SRE10 core-core task of common condition 6. For NIST SRE08, each training and testing conversation is telephone apeech with 1,285 true target trials and 11,637 imposter trials. For NIST SRE10, they are also telephone speech utterances with 178 true target trials and 12,825 imposter trials. There is no cross-gender trials. The performance for NIST SRE08 is evaluated using the Equal Error Rate (EER) the minimum decision cost function (DCF), calculated using $C_{miss} = 10$, $C_{FA} = 1$ and $P_{target} = 0.01$. For NIST SRE10, the performance is evaluated using the EER and the new minimum decision cost function (DCF_{new}), calculated using $C_{miss} = 1$, $C_{FA} = 1$ and $P_{target} = 0.001$.

4.2. Feature extraction

First, ETSI Adaptive Multi-Rate (AMR) GSM VAD [18] is applied to prune out the silence region of the speech file. Then the speech signals are segmented into frames by a 25ms Hamming window with 10ms frame shift. The first 16 Mel frequency cepstral coefficients and log energy are calculated; together with their first and second derivatives. A 51-dimensional feature vector is obtained for each frame (the frequency window is restricted to 300-3400 Hz). Finally, feature warping [19] is applied to the MFCC features.

4.3. The baseline system

The baseline system employs 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 train the gender-dependent eigenvoice matrix V using Switchboard II Phase 2 and 3, Switchboard Cellular Part 2, NIST 2004-2006 SRE, including 893 male speakers with 11204 utterances. The rank of the speaker space is set to 300. The eigenchannel matrix U is also trained in a gender-dependent manner from 436 male speakers with 5,410 speech utterances from NIST 2004-2006 SRE. The rank of the channel space is set to 100. We use the expectation maximization (EM) algorithm with 20 iterations for all of the above training. The diagonal residual scaling matrix D is extracted from the UBM covariance without EM estimation.

4.4. Subspace training

The gender-dependent Fishervoice projection matrices are constructed from telephone speeches in NIST 2004-2006, Switchboard II Phase 2, Phase 3 and Switchboard Cellular Part 2. Here, we create two training data sets: 1) 563 male speakers each with 8 different utterances, same as in JFA + subspace (denoted hereafter by standard set). 2) all useful utterances of the above 563 speakers (some speaker with more than 8 utterances) from the above corpus (denoted hereafter by full set). The Fishervoice projection matrices, W_1, W_2 and W_3, have dimensions $[J, P]$, $[(E_1 + E_2), 799]$, $[799, 550]$, respectively. These correspond to the upper limit of their matrix ranks. For both Fishervoice + subsapce and JFA + subspace framework, the parameter (R in Eq.8 of [10]) 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 slices K is set to 16. The parameters L and J for the PCA dimension reduction before Fishervoice are both set to 4,000. The number of random subspaces Q is set to 5.

4.5. Score normalization

We use gender-dependent score normalization (TZ-norm) for different speaker verification systems. The SRE04, SRE05 and SRE06 corpora are adopted as the T-norm corpus and Switchboard II Phase 2 and 3 corpora as the Z-norm corpus. The number of speakers is 400 for T-norm and 622 for Z-norm.

5. Results

In this section, we present the individual and combined results based on the NIST SRE08 male short2-short3 task of tel-tel condition and NIST SRE10 core-core task of cc6 for the systems described above.

5.1. Integration of random subspace sampling and Fishervoice

We first explore the sensitivity of the Fishervoice + subsapce approach under different dimensions of P, E_1 and E_2 on NIST SRE08. The parameter of P is set at 1300, 1400 and 1500. $(E_1 + E_2)$ is constrained to a constant value of 800 for dimension reduction and complementary discriminative information. As mentioned before, we apply integration of random subspace sampling and Fishervoice for final subspace analysis along with the normalized correlation for distance metric. Table 1 summarizes the results (EER and minDCF $\times$ 100) obtained with the best and worst individuals and fused systems on the seven combinations of (E_1, E_2) input for the proposed framework.

From the table, we observe that: First, making full use of data generates better performance both in EER and minDCF. This is because when there are more samples for each speaker, matrix S_ω can be better estimated. Second, $P = 1400$ shows slightly better performance in most cases. Third, the best and worst individual and fused systems all obtain better results, as compared with random sampling on JFA feature subspace. This verifies the prevous hypothesis stated in Section 1. Fourth, fusion results under different combinations of E_1 and E_2 validate the stability of the fused classifiers. Fifth, the best EER achieved across all individual systems is 3.97% in (500,300), while the best minDCF obtained is 0.0201 in (300,500), both for the full set. These results indicate that the best performance is dependent on the selection of primary eigenvectors, which is unknown to us.

5.2. Comparison with other systems

We also compared the Fishervoice + subspace method with three existing standard approaches, namely, enhanced Fishervoice[11], JFA + subspace[1] and PLDA [7]. The input of these four frameworks are all the second level PCA projected supervectors $O_{i,h}$ in section 3.1. In an attempt to make the fusion process balanced and to avoid the risk of worst system case, we also select all the best/worst individual systems form each combination of (E_1, E_2) separately and fuse all of them together to form total best/worst fusion. For enhanced Fishervoice, we select dimension combination of (900,899,550) which achieved the best performance. The best performance of PLDA is obtained when the dimensionality of LDA is set to 900, the number of eigenvoices in PLDA is set to 550 and when the log-likelihood ratio is used as the score metrics. The parameters under the best performances are similar for the enhanced Fishervoice and PLDA. We think that it may be due to the similarity in their processing. First, PLDA performs LDA for dimension reduction initially, which is similar to the first step of Fishervoice. Second, whitening is applied in both approaches.

Table 2: *Comparison of the integration method with other standard systems on NIST SRE08 male core task (tel-tel). The performance is reputed in EER(%), 100 $\times$ minDCF.*

System Type		standard set	full set
Enhanced Fishervoice [11](900,899,550)		4.36, 2.09	4.04, 2.03
JFA + subspace [1]	Total worst fusion	4.28, 2.11	4.03, 2.08
	Total best fusion	4.12, 2.16	3.94, 2.07
Fishervoice + subspace ($P = 1300$)	Total worst fusion	4.18, 2.12	4.09, **2.05**
	Total best fusion	4.09, 2.09	3.93, 2.06
Fishervoice + subspace ($P = 1400$)	Total worst fusion	4.17, 2.12	4.10, 2.09
	Total best fusion	**4.00**, 2.10	**3.87**, 2.07
Fishervoice + subspace ($P = 1500$)	Total worst fusion	4.19, 2.15	4.02, 2.08
	Total best fusion	4.09, **2.09**	3.95, 2.07
PLDA (LDA+whiten+length)		4.75, 2.33	4.28, 2.20
PLDA (LDA+whiten+length+tznorm)		4.57, 2.16	4.43, 2.07
standard PLDA (whiten+length+tznorm)[17]		4.59, 2.37	

Table 2 shows that the Fishervoice + subspace method generates the best performance. The best performing systems by both EER and minDCF are highlighted in each column. Compared with the JFA + subspace method, the Fishervoice +

Table 1: *Results obtained with the best and worst individuals and fused systems on NIST08 male core task. EER(%), 100 × minDCF*

Types of (E_1, E_2)	standard set									full set								
	$P = 1300$			$P = 1400$			$P = 1500$			$P = 1300$			$P = 1400$			$P = 1500$		
	best	worst	fused	best	worst	fused	best	worst	fused	best	worst	fused	best	worst	fused	best	worst	fused
(300,500)	4.20	4.51	**4.03**	4.28	4.43	4.19	4.20	4.50	4.12	4.09	4.36	4.03	4.04	4.28	3.96	4.20	4.36	4.12
	2.13	2.18	2.13	2.23	2.16	2.13	2.08	2.24	**2.09**	2.08	2.04	2.08	2.06	**2.01**	**2.06**	2.05	2.02	**2.03**
(350,450)	4.20	4.43	4.16	4.20	4.50	4.11	4.20	4.33	**4.11**	4.12	4.43	4.04	4.12	4.36	4.03	4.18	4.28	**3.96**
	2.11	2.18	2.10	2.12	2.20	2.14	2.18	2.18	2.14	2.07	2.04	**2.06**	2.16	2.11	2.11	2.10	2.10	2.06
(400,400)	4.25	4.43	4.10	4.12	4.51	**4.03**	4.32	4.43	4.12	4.12	4.28	4.03	4.19	4.43	4.03	4.04	4.28	3.97
	2.11	2.17	2.09	2.09	2.14	2.09	2.14	2.17	2.13	2.08	2.05	2.08	2.10	2.13	2.10	2.08	2.13	2.11
(450,350)	4.20	4.43	4.16	4.27	4.51	4.11	4.27	4.51	4.24	4.04	4.36	3.96	4.12	4.28	4.01	4.04	4.19	3.96
	2.15	2.11	2.14	2.12	2.16	2.09	2.21	2.16	2.10	2.10	2.11	2.09	2.08	2.07	2.09	2.10	2.02	2.09
(500,300)	4.25	4.36	4.19	4.26	4.43	4.18	4.25	4.51	4.18	**3.97**	4.26	3.96	4.11	4.27	**3.95**	4.18	4.25	4.03
	2.11	2.13	2.10	2.18	2.17	2.15	2.14	2.17	2.11	2.08	2.11	2.09	2.14	2.04	2.07	2.13	2.10	2.13
(550,250)	4.28	4.36	4.19	4.25	4.28	4.11	4.34	4.50	4.19	4.11	4.43	4.03	4.04	4.28	3.96	4.12	4.28	4.08
	2.08	2.17	2.14	2.12	2.14	2.12	2.09	2.18	2.12	2.12	2.15	2.11	2.11	2.11	2.11	2.12	2.11	2.11
(600,200)	4.33	4.51	4.27	4.28	4.43	4.24	4.28	4.40	4.19	4.04	4.20	**3.94**	4.03	4.20	4.02	4.11	4.20	4.02
	2.13	2.13	**2.08**	2.07	2.09	**2.06**	2.15	2.12	2.12	2.03	2.14	2.07	2.09	2.11	2.09	2.12	2.09	2.07

subspace framework can decrease EER by 2.91%. On the other hand, the first two rows of PLDA results in the table demonstrate that PLDA gives relative low performance when working with supervector representations. However, it still shows a relative improvement compared to the results in [17], which applies PLDA on the original supervector.

Table 3: *Comparison of the integration method with other standard systems on NIST SRE10 male core-core task (cc6). The performance is reputed in EER(%),* $1000 \times minDCF_{new}$)

System Type		full set
Enhanced Fishervoice [11](900,899,550)		5.05,0.831
JFA + subspace [1]	Total worst fusion	4.93,0.814
	Total best fusion	4.48,0.819
Fishervoice + subspace	Total worst fusion	4.49,0.831
	Total best fusion	**4.48**,0.819
PLDA (LDA+whiten+length)		6.10,0.837
PLDA (LDA+whiten+length+tznorm)		5.48,**0.758**

We also take the NIST 2010 SRE male data for performance comparision (see Table 3). The experiment setup is the same as that performs on NIST 2008 SRE. Except that for Fishervoice + subspace, P is constrained to a constant value of 1400. From the table, we can see that random subspace sampling can improve the performance of Fishervoice method and the Fishervoice + subspace framework is more stable than JFA + subspace framework. Besides, the Fishrvoice + subspace framework gives best performance in terms of EER.

5.3. Fusion with other systems

In the third experiment, we fuse the PLDA result on supervector with our Fishervoice + subspace method (see Figure 2). The weights are obtained by grid search with values giving the lowest EER. Here we select the NIST SRE08 results using full set. According to the EER and minDCF metrics, the best performance is achieved when the worst fusion of Fishervoice + subsapce method is fused with PLDA. It improves PLDA performance by a relative decrease of 8.17% in EER (from 4.28% to 3.93%) and reduced the minDCF by a relative decrease of 9.54% (from 0.0220 to 0.0199).

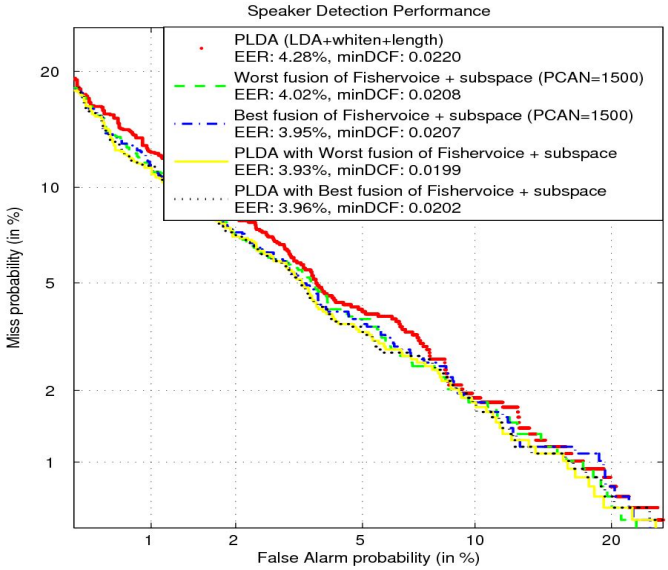

Figure 2: *Fusion results with other systems on NIST SRE08 male tel-tel task (projection matrix trained with full set)*

6. Conclusions

This paper presents an enhancement of our previous work of the JFA + subspace method for speaker verification. The proposed framework is refered as the Fishervoice + subspace method. The approach effectively stabilizes the Fishervoice classifier and makes use of almost all the discriminative information in the high-dimensional space, since multiple classifiers can cover most of the speaker feature space. In this study, we use the simplest linear fusion scheme to combine multiple classifiers and achieve notable improvement. Extensive experiments on the NIST SRE08 and NIST SRE10 male core test show the advantage of the proposed framework over state-of-the-art algorithms. In future work, we will investigate the relationship between PLDA and Fishervoice and seek to combine their respective advantages for further performance improvement.

7. Acknowledgement

This work is partly supported by National Natural Science Foundation of China (61103164).

8. References

[1] Weiwu Jiang, Zhifeng Li, and Helen M Meng, "An analysis framework based on random subspace sampling for speaker verification.," in *Interspeech*, 2011, pp. 253–256.

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

[3] Patrick Kenny, Pierre Ouellet, Najim Dehak, Vishwa Gupta, and Pierre 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.

[4] Najim Dehak, Patrick Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[5] Simon JD Prince and James H Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *IEEE 11th International Conference on Computer Vision (ICCV)*, 2007, pp. 1–8.

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

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

[8] Tin Kam Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832–844, 1998.

[9] Tin Kam Ho, "Nearest neighbors in random subspaces," in *Advances in Pattern Recognition*, pp. 640–648. Springer, 1998.

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

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

[12] Xiaogang Wang and Xiaoou Tang, "Random sampling LDA for face recognition," in *2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2004, vol. 2, pp. II–259.

[13] Xiaogang Wang and Xiaoou Tang, "Random sampling for subspace face recognition," *International Journal of Computer Vision*, vol. 70, no. 1, pp. 91–104, 2006.

[14] Xiaogang Wang and Xiaoou Tang, "A unified framework for subspace face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1222–1228, 2004.

[15] Sergey Ioffe, "Probabilistic linear discriminant analysis," in *Computer Vision–ECCV 2006*, pp. 531–542. Springer, 2006.

[16] Siwei Lyu and Eero P Simoncelli, "Nonlinear extraction of independent components of natural images using radial Gaussianization," *Neural computation*, vol. 21, no. 6, pp. 1485–1519, 2009.

[17] Ye Jiang, Kong-Aik Lee, Zhenmin Tang, Bin Ma, Anthony Larcher, and Haizhou Li, "PLDA modeling in i-vector and supervector space for speaker verification.," in *Interspeech*, 2012.

[18] *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.

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