# An Analysis Framework based on Random Subspace Sampling for Speaker Verification

*Weiwu Jiang*[1], *Zhifeng Li*[2] *and Helen Meng*[1]

[1] Department of Systems Engineering & Engineering Management, The Chinese University of Hong Kong, China
{wwjiang, hmmeng}@se.cuhk.edu.hk
[2] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, China
zhifeng.li@siat.ac.cn

## Abstract

Using Joint Factor Analysis (JFA) supervector for subspace analysis has many problems, such as high processing complexity and over-fitting. We propose an analysis framework based on random subspace sampling to address these problems. In this framework, JFA supervectors are first partitioned equally and each partitioned subvector is projected on to a subspace by PCA. All projected subvectors are then concatenated and PCA is applied again to reduce the dimension by projection onto a low-dimensional feature space. Finally, we randomly sample this feature space and build classifiers for the sampled features. The classifiers are fused to produce the final classification output. Experiments on NIST SRE08 prove the effectiveness of the proposed framework.

**Index Terms**: random sampling, supervector, subspace framework, joint factor analysis, discriminative

## 1. Introduction

Joint Factor Analysis (JFA) [1] has been applied to the generative Gaussian Mixture Model (GMM) [2] for speaker verification. It addresses problems of speaker characteristics and channel variability, thus performance improvement has been achieved. Specifically, the approach assumes that the noise variability is located in a low-dimensional subspace and the effects of speaker and channel are additive in the feature space.

Although success has been achieved by JFA, recent works show that neither JFA supervector-based subspace analysis [3] nor the combination of JFA supervector and SVM [4] achieve the significant improvements for speaker verification system. It is believed that this is due to numeric instability and sparsity of the supervectors in the high-dimensional space. These facts may adversely affect the model training process due to overfitting.

Based on JFA, Dehak [5] proposed an i-vector feature-based speaker verification system and achieved great success on the NIST evaluation. The i-vectors are extracted from low-dimensional subspace of the supervectors. Therefore, the i-vector-based system substantially reduces the execution time of the speaker recognition task.

Our previous work [6][7] proposed an enhancement by applying the Fishervoice framework to the JFA supervector space to make use of both the high-dimensional JFA supervector and discriminative information for training. Unlike traditional subspace analysis that is based on the whole supervector space, multiple discriminant projections are derived equally for the partitioned subspaces of the supervector to achieve fast and effective matching. There are two major problems encountered in this approach: 1) the dimensionality of the final projected subspace is much higher than the i-vector approach; 2) the performance of a single classifier with limited training samples is unstable.

In this work, we propose to investigate the combination of the JFA supervector [1] and random subspace sampling [11] to address two problems mentioned above. We randomly sample the feature space into several subspaces. Classifiers are built for each subspace and their results are integrated to obtain the final decision. The proposed approach is inspired by the success in subspace modeling for face recognition [10][13][14][15].

The rest of the paper is organized as follows: Section 2 presents the background of previous work. Section 3 presents details of the random subspace sampling approach for speaker verification. Implementation and experimental results on the NIST SRE08 are presented in sections 4 and 5 respectively. The final section gives the conclusion of our study.

## 2. Background

### 2.1. Joint Factor Analysis

The JFA theory [1] assumes that the speaker and channel noise components, which reside in the speaker-and-channel-dependent supervectors respectively, have Gaussian distributions. Let $M_{ih}$ denote the speaker and session-dependent supervector of mean for the *h-th* utterance from speaker *i*. $M_{ih}$ is further assumed to be made up of four supervectors as shown below:

$$M_{ih} = m + Vy_{ih} + Dz_{ih} + Ux_i \tag{1}$$

where $m$ is the mean supervector of the background model, $U$ is the eigenchannel matrix, $V$ is the Eigenvoice matrix, $D$ is the diagonal residual scaling matrix, $x_i$ is speaker-dependent eigenchannel factor, $y_{ih}$ is the speaker-and-session-dependent eigenvoice factor and $z_{ih}$ is the speaker residuals. We also define $s_{ih}$ as speaker vector by the first three terms in Eq. (1):

$$s_{ih} = m + Vy_{ih} + Dz_{ih} \tag{2}$$

### 2.2. Intersession Compensation in Subspace Analysis

The traditional Linear Discriminant Analysis (LDA) seeks to determine an optimal projection $W$, which maximizes the ratio of the determinant of the between-class scatter matrix $S_b$ to that of the within-class scatter matrix $S_w$. Given the supervectors from all training speakers, let $C$ denote the total number of speakers, $x_{i,h}$ be $h$-th supervector from speaker $i$, $H_i$ be the number of samples (or sessions) in the speaker $i$, $\xi_i$ be the sample mean of the class $i$ and $\xi$ be the sample mean of all development data. The optimal projection $W_{lda}$ for LDA is calculated as follows:

$$W_{lda} = \arg\max_W \left(\frac{\left\|W^T S_b W\right\|}{\left\|W^T S_w W\right\|}\right)$$

$$S_w = \sum_{i=1}^{C}\sum_{h=1}^{H_i}(x_{i,h} - \xi_i)(x_{i,h} - \xi_i)^T,\ S_b = \sum_{i=1}^{C} H_i(\mu_i - \xi)(\mu_i - \xi)^T \tag{3}$$

Within-Class Covariance Normalization (WCCN) [8] aims to minimize the expectation of speaker verification errors in both false acceptance and false rejection. This approach has been successfully applied to the i-vector-based speaker verification system for channel noise compensation in [5]. To perform WCCN, we first apply LDA to derive the optimal projection. The within-class covariance projection matrix $B$ is then estimated by using development data through

$$\begin{aligned} W_{wccn} &= \frac{1}{C}\sum_{i=1}^{C}\frac{1}{H_i}\sum_{h=1}^{H_i}(W_{lda}^T x_{i,h} - \overline{\xi_i})(W_{lda}^T x_{i,h} - \overline{\xi_i})^T \\ BB^T &= W_{wccn}^{-1} \end{aligned} \tag{4}$$

where $W_{lda}$ is the LDA projection matrix and $\overline{\xi_i}$ is the LDA projected sample mean of the class $i$. When LDA and WCCN are both applied, the final projection matrix $W_P$ is given by:

$$W_P = W_{lda}B \tag{5}$$

### 2.3. Fishervoice Discriminant Analysis

Fishervoice [9] aims to enhance performance by extracting discriminant information from the scatter matrices $S_w$ and $S_b$ effectively. The overall projection matrix of Fishervoice can be considered as three components:

1) Subspace projection matrix $W_1$ for dimension reduction using PCA — the subspace projection $f_1$ result is obtained by:

$$f_1 = W_1^T x, \text{ where } W_1 = \arg\max_W \left\| W^T \Psi W \right\| \tag{6}$$

where $x$ is an any supervectors and $\Psi$ is the covariance matrix of all supervectors in the development set.

2) Whitening matrix $W_2$ for reducing intra-speaker variations — from the above projected subspace, $f_1$ is whitened as $f_2$ according to the equation:

$$f_2 = W_2^T f_1, \text{ where } W_2^T S_w W_2 = I,\ W_2 = \Phi\Lambda^{-1/2} \tag{7}$$

where $S_w$ is the standard within-class scatter matrix, $\Phi$ is the normalized eigenvector matrix of $S_w$, and $\Lambda$ is the eigenvalue matrix of $S_w$.

3) Subspace projection matrix $W_3$ for discriminative speaker class boundaries — this is obtained by using the nonparametric between-class scatter matrix $S_b^{'}$ according to Eq. (8-9) [9] from the whitened subspace above as:

$$f_3 = W_3^T f_2, \text{ where } W_3 = \arg\max_W \left\| W^T S_b^{'} W \right\| \tag{8}$$

Finally, the overall transformation matrix $W_{NF}$ for nonparametric Fisher discriminant analysis is given by:

$$W_{NF} = W_1 W_2 W_3 \tag{9}$$

In the WCCN method, the optimal projection matrix $B$ is obtained by decomposing the expected with-class covariance matrix $W_{WCCN}$ via Cholesky decomposition algorithm. This implies that the covariance matrix $W_{WCCN}$ can be normalized to the identity matrix by $B^T W_{WCCN} B = I$. In the second step of Fishervoice, the whitening projection matrix $W_2$ also tries to normalize the within-class covariance scatter matrix $S_w$ by $W_2^T S_w W_2 = I$. Comparing the equations for these two projection matrices, we can see that projection matrices $B$ and $W_2$ produce the same function (the latent factor for compensation) but with different solutions. Theoretically, the whole Fishervoice framework can improve the classification accuracy more than the LDA+WCCN method, since the projection matrix $W_3$ emphasizes structural information of speaker boundaries through the nonparametric between-class scatter matrix $S_b^{'}$. This is verified in the experiments section.

# 3. Proposed Framework

In this work, we propose a random subspace sampling approach in a multi-subsystems fusion framework for the speaker verification task as an extension to [6][9]. We employ various classifiers in the randomly sampled subspace of the high-dimensional JFA mean vector and perform classifier fusion in the projected low-dimensional discriminant space (see Figure 1).

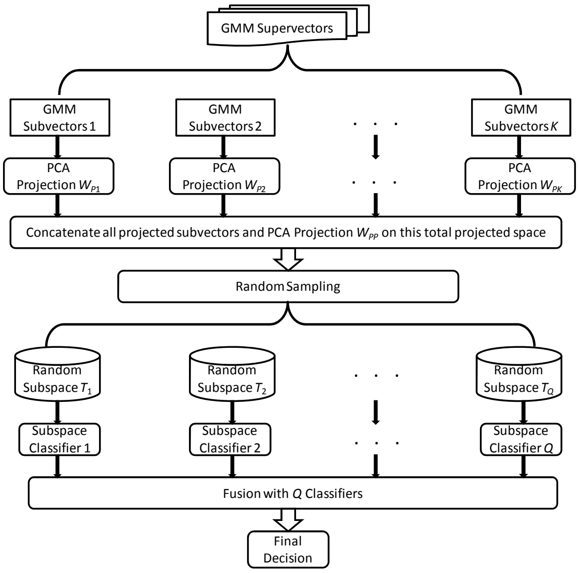

Figure 1: *Overall organization of the proposed multi-classifier framework based on random sampling*

### 3.1. Supervector Extraction

We believe that the structure of JFA speaker supervectors can capture the probabilistic distribution of acoustic feature classes in the whole acoustic space. We also assume that the whole acoustic space can be characterized by a set of acoustic classes with Gaussian models that correspond to some broad phonetic events. Based on the result of [6], we represent the utterance $h$ from the speaker $i$ by a $GF$ dimensional speaker supervector $s_{i,h}$ in Eq. (2).

### 3.2. Training Stage

Figure 1 illustrates the overall organization of the proposed framework. It is a multi-classifier fusion framework based on random subspace sampling. The details of this framework are presented as follows:

1) For each utterance from the development set, obtain the corresponding input feature vector according to the supervector extraction process.
2) Divide the whole supervector into $K$ slices (subvectors) equally and sequentially, and then perform PCA on each subvectors subspace to reduce the dimensionality to $L$ for each of the $K$ subspaces, $W_{Pk}$ ($k$=1, 2, ... , $K$).
3) In order to further reduce the number of vector dimensions, all projected subvectors are first concatenated sequentially to a $K$×$L$ dimension vector. Then a second level PCA projection ($W_{PP}$) is performed onto a $J$-dimensional subspace. $O_{i,h}$=[ $o_1$, $o_2$, … , $o_J$] denotes the projected supervector (candidate vectors) from the $h$-th utterance of speaker $i$ for constructing the random subspaces.
4) Construct $Q$ random subspaces $T_q$ ($q$=1, 2, … , $Q$) with each spanned by a total ($E_1$+$E_2$) dimensions. The primary $E_1$ dimensions are fixed so as to keep the first $E_1$ largest

eigenvalues in $W_{PP}$, which can preserve the main intra-personal variations. The remaining $E_2$ dimensions are *randomly* selected from the remaining ($J$-$E_1$) dimension space.

5) In each subspace $T_q$, perform Fishervoice or LDA+WCCN based discriminant subspace analysis and project all vectors from the training data onto the random subspace via the projection matrix $W_{Sq}$. Hence, we generate $Q$ classifiers in total.
6) During target enrollment, each projected target speaker supervector (in step 3) resulting from the second level PCA projection is fed to the $Q$ subspace classifiers to form the final $Q$ parallel feature vectors $\theta'_{train}(q)$ ($q$=1, 2, … , $Q$). We define $\theta_{train}(q)$ as the training reference vector projected by $\theta'_{train}(q)$ via projection matrix $W_{Sq}$.

### 3.3. Testing Procedures

In the testing stage, each testing supervector is projected to the testing vector $\theta_{test}(q)$ via the $q$-th random subspace in the same manner as the enrollment process. Then a distance score is calculated between $\theta_{train}(q)$ and $\theta_{test}(q)$ in terms of the normalized correlation (COR) as shown in Eq. (12):

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

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 core data set (cc=6). Each training and testing conversation has an average duration of 5 minutes with 874 true target trials and 11,637 imposter trials. We use the EER and minDCF as metrics for performance evaluation.

### 4.2. Feature extraction

First, ETSI Adaptive Multi-Rate (AMR) GSM VAD [12] is applied to prune out silence. Then the speech is segmented into frames by a 25 ms Hamming window shifting with 10 ms frame rate. The passing frequency band is restricted to 300-3400 Hz. The first 16 Mel frequency cepstral coefficients (MFCC) with log energy are calculated with their first and second derivatives to form a 51-dimension feature. Finally, the Gaussianization process is applied to all the MFCCs.

### 4.3. Baseline system

The baseline system employs the JFA [1] with the enhanced Fishervoice framework [6]. During the training phase:

2048-Gaussian gender-dependent UBMs were created from SRE04 1side-1side and SRE05 lcon4w-1con4w data.

The eigenvoice matrix $V$ is trained using LDC Switchboard II Phase 2, Phase 3, Switchboard Cellular Parts 2, SRE04, SRE05 and SRE06, including 893 male speakers with 11204 utterances. The rank of the speaker space is set to 300.

The eigenchannel matrix $U$ is trained from 436 male speakers with 5410 utterance in the SRE04 SRE05 and SRE06. The rank of the channel space is set to 100.

The diagonal residual scaling matrix $D$ is extracted from the UBM covariance.

### 4.4. Subspace Training

The Fishervoice projection matrices ($W_1$, $W_2$ and $W_3$) are trained on telephone utterances from the NIST SRE04, SRE05, SRE06, LDC releases of Switchboard II Phase 2, Phase 3 and Switchboard Cellular Parts 2. This amounts to 563 male speakers altogether, each with 8 different utterances. The projection matrices, $W_1$, $W_2$ and $W_3$, have dimensions ($E_1$+$E_2$) × 800, 800 × 799 and 799 × 550, respectively. These correspond to the upper limits of their matrix ranks. The parameter $R$ which controls the number of nearest neighbors for constructing $S'_b$ in [9] was set to 4, according to the median number of sessions for each speaker. For training the LDA and WCCN matrices, we use the same dataset as Fishervoice. For both the proposed framework and the enhanced Fishervoice framework, the number of slices $K$ is set to 16. The parameters $L$ and $J$ for PCA dimension reduction are both set to 4000.

### 4.5. Score normalization

The scores of all evaluated speaker verification systems were normalized by gender-dependent TZ-norm. We adopt the SRE04, SRE05 and SRE06 corpora as the t-norm corpus and Switchboard II Phase 2 and Phase 3 corpora as the z-norm corpus. The number of speakers in the corpus is 400 for t-norm and the 622 for z-norm.

## 5. Results

In this section, we present individual and combined results on the NIST SRE08 male core task (cc=6) from the systems described above.

### 5.1. Random Subspace Sampling

The first experiment investigates the sensitivity of speaker verification performance for the proposed method with regards to the different dimensions of $E_1$ and $E_2$. We constrain the dimensionality of ($E_1$+$E_2$) to a constant value of 2500 for the dimension reduction. Besides, full space analysis without random sampling is also considered in the experiment. As mentioned before, we apply Fishervoice or LDA+WCCN for final subspace analysis in the selected random subspace along with the normalized correlation for distance metric. Table 1 summarizes the results (EER and minDCFx100) obtained with the best individual and fused systems on the seven combinations of ($E_1$, $E_2$) input for the proposed framework. For each combination of ($E_1$, $E_2$), we create 3 subspaces randomly for training and evaluation.

Key observations include: First, the performance of multi-classifier fusion remains stable across different randomization of the dimensions ($E_1$, $E_2$). Second, the Fishervoice framework shows slightly better performance than LDA+WCCN most of the time. Third, as discussed above, when the data is of high dimensionality, a single classifier constructed on the limited training samples is unstable. Traditional subspace methods suffer from training data sparsity and fusion results clearly verify that the proposed random subspace sampling framework can further improve the whole system, as compared with the best individual system as well as subspace analysis system within the full dimensional space (4000 dimensions).

### 5.2. Comparison with Other Systems

We also compared the proposed random subspace sampling framework with two existing standard approaches, namely, JFA and the enhanced Fishervoice framework. In the random subspace sampling approach, we fuse all 12 Fishervoice subsystems from (300, 2200) to (600, 1900) – this interval has more stable and better performance. Figure 2 shows the results obtained from these systems. They suggest that random subspace sampling and classifier fusion lead to better performance compared to other systems. Compared with the use of a single JFA classifier, the use of random subspace sampling on Fishervoice improves results by decreasing the minDCF from 0.0284 to 0.0239, as well as a relative decrease in EER by 11.33%. Besides, the random subspace sampling framework works better than the enhanced Fishervoice since multiple classifiers are integrated to overcome instability due to the use of a single classifier.

*Table 1. Results obtained with the best individual and fused systems. EER (%), minDCFx100*

| Types of | Fishervoice | | LDA+WCCN | |
|---|---|---|---|---|
| ($E_1$, $E_2$) | Best | Fused | Best | fused |
| (100,2400) | **5.24**, **2.74** | **4.66**, **2.55** | 5.37, 2.76 | 4.79, 2.60 |
| (200,2300) | **5.00**, **2.71** | **4.68**, 2.65 | 5.02, 2.71 | 4.83, **2.58** |
| (300,2200) | **4.57**, **2.45** | **4.56**, **2.43** | 4.85, 2.47 | 4.67, 2.47 |
| (400,2100) | 4.74, **2.43** | 4.66, **2.41** | **4.68**, 2.48 | **4.56**, 2.43 |
| (500,2000) | **4.57**, **2.39** | **4.56**, **2.39** | 4.68, 2.40 | 4.56, 2.40 |
| (600,1900) | **4.76**, **2.44** | **4.68**, 2.42 | 4.80, 2.44 | 4.77, **2.40** |
| (700,1800) | **4.69**, 2.42 | 4.68, **2.41** | 4.69, 2.44 | **4.57**, 2.43 |
| Full Space | **4.69**, 2.42 | / | 5.03, 2.52 | / |

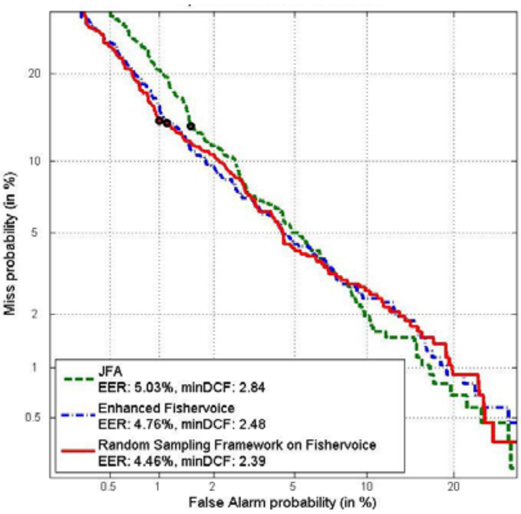

Figure 2: *Comparison of Fishervoice and other standard systems on NIST SRE 08 male core task (cc=6, 100x minDCF)*

### 5.3. Fusion with the Other Systems

In the third experiment, we fuse the random subspace sampling based system in section 5.2 with two other standard systems (see Figure 3). We select JFA and the enhanced Fishervoice framework to represent the baseline systems. It is worth noting that JFA gives comparable performance with the two other systems. Fusion of the enhanced Fishervoice framework with the system in section 5.2 only achieves little improvement since both of the systems use classifiers based on Fishervoice. Third, according to the EER and minDCF metrics, three systems fused totally offer the best performance among all fused systems. It improves JFA results by a relative decrease of 13.72% in EER (from 5.03% to 4.34%) and reduced the minDCF from 0.0284 to 0.0234.

## 6. Conclusions

This paper enhances our previous work in the Fishervoice approach [6] for speaker verification. In order to overcome the problems caused by the high dimensionality of JFA supervector, we developed a classification framework that incorporates: 1) a random sampling method in the principle subspace and 2) discriminant subspace analysis in the randomly sampled subspace. By applying the proposed framework, the dimensionality of the original JFA supervector is reduced from 104448 (51×2048) to 550 after discriminant subspace projection (e.g. Fishervoice), to facilitate fast and effective matching. Finally, parallel classification outputs are combined into a final classification decision output. Extensive experiments on the NIST08 male core test show the advantage of the proposed framework over the state-of-the-art algorithms.

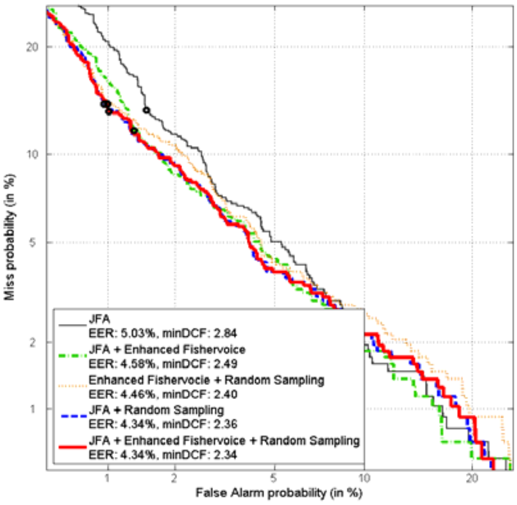

Figure 3: *Fusion results with other systems on NIST SRE 2008 male core task (cc=6, 100x minDCF)*

## 7. References

[1] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A Study of Inter-Speaker Variability in Speaker Verification," *IEEE Trans. on Audio, Speech and Language Processing*, 2008.

[2] D. Reynolds, T. Quatieri and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *DSP*, 2000.

[3] Johns Hopkins University, Summer Workshop, "Robust Speaker Recognition Over Varying Channels," 2008.

[4] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, "SVM Based Speaker Verification using a GMM Supervector Kernel and NAP Variability Compensation," *ICASSP*, 2006.

[5] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Trans. on Audio, Speech and Language Processing*, November, 2009.

[6] W. Jiang, H. Meng and Z. Li, "An Enhanced Fishervoice Subspace Framework for Text Independent Speaker Verification," *ISCSLP*, 2010.

[7] W. Jiang, M. Mak, W. Rao and H. Meng, "The HKCUPU system for the NIST 2010 speaker recognition evaluation," *ICASSP*, 2011.

[8] A. Hatch, S. Kajarekar and A. Stolcke, "Within-class covariance normalization for svm-based speaker recognition," *Proceedings of Interspeech,* 2006.

[9] Z. Li, W. Jiang and H.Meng "Fishervoice: a discriminant speaker recognition," *ICASSP*, 2010.

[10] Z. Li, D. Lin, and X. Tang, "Nonparametric Discriminant Analysis for Face Recognition," *IEEE Trans. on PAMI*, vol. 31, no. 4, pp. 755-761, 2009.

[11] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Trans. on PAMI*, 1998.

[12] GSM 06.94, "Digital cellular telecommunication system (Phase 2+); Voice Activity Detector VAD for AdaptiveMulti Rate (AMR) speech traffic channels; General description," 1999.

[13] X. Wang and X. Tang, "Random sampling LDA for face recognition," *CVPR,* 2004.

[14] X. Wang and X. Tang, "Random sampling for subspace face recognition," IJCV, 2006.

[15] X. Wang and X. Tang, "A unified framework for subspace face recognition," *IEEE Trans. On PAMI* , 2004