An Enhanced Fishervoice Subspace Framework for Text-independent Speaker Verification

Weiwu JIANG, Helen MENG
Dept. of Systems Engineering & Engineering Management
Chinese University of Hong Kong
Hong Kong SAR of China
{wwjiang, hmmeng }@se.cuhk.edu.hk

Zhifeng LI
Dept. of Computer Science and Engineering
Michigan State University
Michigan, USA
zfli@msu.edu

Abstract—**In this paper, we propose an enhancement to Fishervoice approach [6] for speaker verification. In this framework, we first represent each utterance with a fix-length super-vector using Joint Factor Analysis (JFA) Gaussian Mixture Models (GMM). Multiple discriminant projections are then used on partitioned vectors of the super-vector for fast and effective matching. Experiments are presented on the core test of NIST SRE 2008 male corpora. We achieve 10.3% relative improvement on EER compared to the state-of-art JFA. Finally, by integrating the enhanced Fishervoice model and the JFA, the matching performance can be further improved. This demonstrates the effectiveness of the proposed framework.**

Keywords---speaker verification; subspace model; joint factor analysis; fishervoice; discriminant analysis

I. INTRODUCTION

Recently, speaker recognition has attracted great attention due to increasing demands for real-world applications. There are two main kinds of speaker recognition tasks: speaker identification and speaker verification. In this paper, we focus on the speaker verification task.

In the field of the speaker verification, Gaussian mixture model (GMM) [1] based Joint Factor Analysis (JFA) [2] [3] and discriminative Support Vector Machine (SVM) became popular methods for many systems. However, in the last two years, the combination of JFA and GMM supervector based SVM was not very successful [4]. A possible reason is that supervectors in high-dimensional space present a challenge for model training with SVM approach.

Recently, feature-based speaker verification system with i-vectors has been proposed for training stage in low-dimensional space [5]. It tries to represent both speaker and channel variability by extracting a low-dimensional subspace from the GMM supervector space and thus reduce the execution time of the recognition task substantially.

In order to make use of both the high-dimensional JFA supervector and discriminative training information, we proposed the Fishervoice [6] approach for speaker verification. It maps a supervector into a compressed subspace by nonparametric Fisher's discriminant analysis [7], which is performed in an attempt to suppress intra-speaker variations and to emphasize the discriminative information for speaker recognition. Besides, the Fishervoice framework can be applied directly in the testing stage to compute the distance between an input test sample and the reference vector of each known speaker.

The rest of the paper is organized as follows: In section 2 we describe the general setup for standard speaker verification systems. In section 3 and 4, we introduce and discuss the Fishervoice approach for speaker verification. Implementation details and experiments on the NIST 2008 male core test (cc=6) are then presented in section 5 and 6, respectively. Finally, the conclusion and future directions are presented in last section.

II. THEORETICAL BACKGROUND

A. Joint factor analysis

In the JFA theory [2], the basic assumption is that speaker- and channel- dependent GMM supervectors are Gaussian distributed with the speaker and the nuisance components (usually called *channel* or *session variability*). Suppose the arbitrary utterance h from speaker i who contains multiple H_i sessions (utterances), we consider M_{ih} as speaker and session-dependent supervector of GMM mean. Therefore, M_{ih} can be decomposed into a sum of four supervectors as follows:

$$M_{ih} = m + Vy_{ih} + Dz_{ih} + Ux_i \quad (1)$$

where m is the UBM supervector mean, U is Eigenchannel matrix, V is Eigenvoice matrix, D is diagonal residual scaling matrix, x_i is speaker dependent Eigenchannel factor, y_{ih} is the session and speaker dependent Eigenvoice factor and z_{ih} is the session and speaker dependent speaker-residuals. The term U, V, D are estimated from a sufficiently large data set while the latent variables x_i, y_{ih}, z_{ih} are estimated for each utterance.

For the purpose of fast calculation, we implement JFA only with speaker factors and channel factors, without the diagonal matrix D. We define the first two parts of Eq. (1) on the right hand side as speaker vector s_{ih}:

$$s_{ih} = m + Vy_{ih} \quad (2)$$

The rest Furthermore, we use the log-likelihood ratio (LLR) for scoring in the verification stage, which is similar to [8].The implementation of this approach is to subtract the estimated noise in the feature level, which means that feature frames from

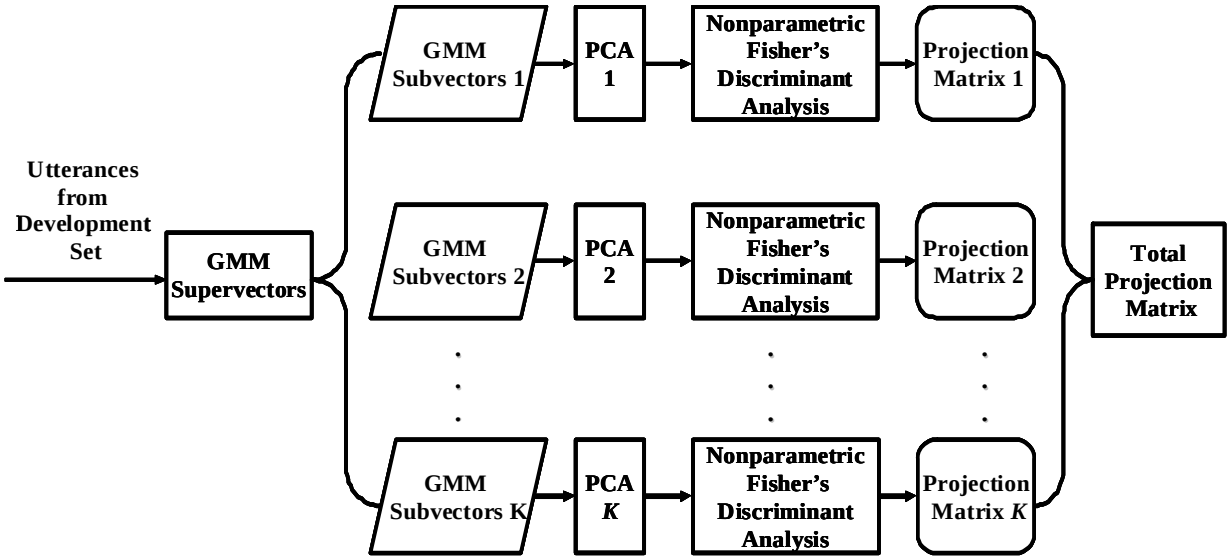

Figure 1. Overall organization of the proposed Fishervoice framework

the test utterance are extractedg out by the estimated channel noise via following formula:

$$t' = t - \sum_{g=1}^{G} \gamma_g(t)\{U \cdot x_{ih}\}_{[g]} \quad (3)$$

where t is original feature frame, t' is feature frame with channel noise subtracted, $\gamma_g(t)$ is the posteriori probability of the g-th Gaussian for the feature frame t and G is the total number of Gaussian mixtures. Theoretically, the performance of applying JFA in model level (Equation (2)) and feature level (Equation (3)) for testing should be the same.

B. SVM-JFA in GMM supervector space

The SVM is a binary classifier which tries to find a separator. The basic idea of SVM is to project input vectors into a feature space in which a hyperplane can separate each classes linearly. This projection is carried out by using a mapping function. In practice, SVMs use kernel functions to perform the scalar product computation in the feature space.

In order to apply SVM with JFA using speaker GMM supervector as input, Campbell *et al.* [9] proposed a classical linear Kullback-Leibler (KL) divergence based kernel between two GMM supervector u_a and u_b:

$$K(u_a, u_b) = \sum_{g=1}^{G} (\sqrt{w_g}\,\Sigma_g^{-\frac{1}{2}} u_g^a)(\sqrt{w_g}\,\Sigma_g^{-\frac{1}{2}} u_g^b)^t \quad (4)$$

where u_g^a (or u_g^b) is the mean vector of the g-th Gaussian for speaker GMM, w_g and Σ_g corresponds to the g-th UBM mixture weights and diagonal covariance matrix.

During the experiment, we also test supervectors with conventional cosine kernel [4] given as follows:

$$K(u_a, u_b) = \frac{\langle u_a, u_b \rangle}{\|u_a\|\|u_b\|} \quad (5)$$

III. FISHERVOICE IN SPEAKER VERIFICATION

Following [6], we propose an enhanced Fishervoice framework for speaker verification task by projecting the high dimensional JFA-based supervector into a low-dimensional discriminant subspace to model speaker characteristics. More specifically, we explore different types of JFA GMM mean vector feature representations with nonparametric Fisher's discriminant analysis.

A. Supervector extraction

Inspired by the rationale that the whole acoustic space can be characterized by a set of acoustic classes with a Gaussian model representing some broad phonetic events [10], it is desirable to use these Gaussian mixtures to compress input features. Thus, in the Fishervoice approach, we concatenate the GMM speaker vectors as the input supervector, instead of using the structured score vector (SSV) in [6]. These supervectors are able to leverage the acoustic class structure captured in the GMM speaker vector space in order to extract the "key" information in an input utterance. The structure captures the probabilistic distribution of acoustic feature classes in the overall acoustic space. Therefore, we represent the utterance h from the speaker i in terms of an $G \times F$ dimension vector $x_{i,h}$:

$$x_{i,h} = [s_{i,h,1} \;\; s_{i,h,2} \;\cdots\; s_{i,h,g} \;\cdots\; s_{i,h,G}]^T \quad (6)$$

where $s_{i,h,g}$ is the F-dimensional GMM speaker vector for the g-th Gaussian mixture. For comparison, we also test two other types (weighted supervector and kernel based supervector) which are similar to $s_{i,h,g}$.

$$x_{i,h}' = [s_{i,h,1}' \;\; s_{i,h,2}' \cdots s_{i,h,g}' \cdots s_{i,h,G}']^T \quad s_{i,h,g}' = \sqrt{w_i}\, s_{i,h,g} \quad (7)$$

$$x_{i,h}'' = [s_{i,h,1}'' \;\; s_{i,h,2}'' \cdots s_{i,h,g}'' \cdots s_{i,h,G}'']^T \quad s_{i,h,g}'' = \sqrt{w_i}\,\Sigma_i^{-\frac{1}{2}} s_{i,h,g} \quad (8)$$

B. Traning stage

Figure 1 illustrates the overall organization of the proposed framework. By incorporating all these strategies, a multi-classifier framework is developed. The steps of this algorithm are summarized as follows:

B.1 *For each training sample, obtain the corresponding input feature vector using the supervector extraction technique.*

B.2 *Divide the whole supuervector into K slices equally and then project each subvector via PCA model W_{P_k} respectively for dimension reduction . Construct Fishervoice-based classifiers based on each PCA projected slice. (In our experiments, K = 16, so there are 16 classifiers)*

B.3 *Apply nonparametric Fisher's discriminant analysis on each slice in parallell as introduced in [7]:*

B.3a) Subspace projection for dimension reduction — Compute the PCA projection matrix W_{k1} from the *k*-th slice of the entire development set and use it to project all corresponding subvectors into the PCA subspace. The subspace projection f_{k1} is obtained by:

$$f_{k1} = W_{k1}^T x, \text{ where } W_{k1} = \arg\max_W \left\| W^T \Psi W \right\| \qquad (9)$$

where x is an arbitrary subvectors from Eq. (5), (6) or (7) and Ψ is the covariance matrix of the corresponding *k*-th slice of subvectors in the development set.

B.3b) Subspace projection to reduce intra-speaker variations — In the PCA subspace above, compute the whitened subspace projection f_{k2} and adjust the dimension of the whitened subspace to reduce intra-speaker variability:

$$f_{k2} = W_{k2}^T f_{k1}, \text{ where } W_{k2}^T S_w W_{k2} = I, \; W_{k2} = \Phi \Lambda^{-1/2} \qquad (10)$$

where W_{k2} is the whitening transformation matrix applied to the standard within-class (intra-class) scatter matrix S_w [7], Φ is the normalized eigenvector matrix of S_w, Λ is the eigenvalue matrix of S_w.

B.3c) Subspace projection to extract discriminant speaker class boundary information — Compute the nonparametric between-class scatter matrix S_b' according to Eq. (8-9) in [6]. Perform PCA on S_b' and choose dominant eigenvectors (usually choose number of rank) to form projection matrix W_{k3}. The subspace projection f_3 is obtained by:

$$f_3 = W_{k3}^T f_2, \text{ where } W_{k3} = \arg\max_W \left\| W^T S_b' W \right\| \qquad (11)$$

B.3d) Subspace transformation matrix W_k for the *k*-th slice is denoted as:

$$W_k = W_{k1} W_{k2} W_{k3} \qquad (12)$$

B.4 Finally, we concatenate all projection matrices into a total projection matrix W_{Total} as follows:

$$W_{Total} = \left[W_{P\_1} \cdots W_{P\_k} \cdots W_{P\_K} \right] \left[W_1 \cdots W_k \cdots W_K \right] \qquad (13)$$

B.5 During target speaker enrollment, each speaker's speaker supervector is projected into a low-dimentional ***training reference vector O_{train}*** by the total projection matrix W_{Total}.

C. Testing stage

C.1 For each test sample, obtain the corsponding input feature vector using similar method as the training stage. Then each supervector is projected into a ***test reference vector O_{test}*** by the total projection matrix W_{Total}.

C.2 Calculate the distance score between projected ***training refrence vector O_{train}*** and ***test refrence vector O_{test}*** in terms of the normalized correlation (COR) shown in Eq. (14):

$$D(O_{train}, O_{test}) = \left\| O_{train}^T O_{test} \right\| \Big/ \sqrt{O_{train}^T O_{train} O_{test}^T O_{test}} \qquad (14)$$

D. Discussion

Essentially, the proposed framework is an extension and improvement of the original Fishervoice approach in [6]. Compared to the approach in [6], the proposed framework is able to generate more discriminant projections for enhanced matching. The idea of separating the long feature into multiple slices with smaller dimension allows us to work on data with more manageable sizes, with consideration in the number of training samples. This will help improve the discriminative ability. The experimental results we will show the advantage of enhanced Fishervoice framework over the original Fishervoice framework, as well as the other state-of-the-art algorithms.

IV. EXPERIMENTAL SET UP

A. 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 and there is no cross-gender trials. Results are given in terms of equal error rate (EER) and minDCF.

B. Feature extraction

First, ETSI Adaptive Multi-Rate (AMR) GSM VAD [11] 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 first 16 Mel frequency cepstral coefficients together with log energy are calculated with their first and second derivatives to form a 51-dimension feature vector (the frequency window is restricted to 300-3400 Hz). Finally, the Gaussianization process is applied to all the MFCCs.

C. Baseline system

The baseline system employs gender-dependent 1024 Gaussian UBMs, which was trained from SRE04 1side-1side and SRE05 lcon4w-1con4w data. The gender-dependent eigenvoice matrix V is trained using LDC releases of 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 also trained gender-dependently from 436 male speakers with 5410 utterance in the SRE04 SRE05 and SRE06. The rank of the channel space is set to 100. Both U and V are trained using the expectation

maximization (EM) algorithm (15 iterations) of the factor analysis and the posteriors of x, y are computed using a single iteration to train a speaker model.

In the SVM system, we use libsvm [12] for implementation. 800 impostors for background training and 300 t-norm models for each gender are taken from the same dataset as in the eigenvoice training.

D. Fishervoice training

For the Fishervoice projection matrix training, the gender-dependent Fishers discriminative projection matrix was constructed on NIST SRE 04-06 telephone data, including 400 male speakers in which each speaker contains 8 different utterances. The Fisher's discriminant subspace projection matrices, W_1, W_2 and W_3, have the dimensions of 299, 298 and 295 respectively, corresponding to the upper limit of their matrix ranks. This means the dimension of original supervector is reduced from 52224 (51*1024) to 4720 (295*16) after Fishervoice projection. The parameter R introduced in [6] is set to 4 according to median number of sessions for each speaker.

E. Score normalization

We apply gender-dependent score normalization (T-norm or TZ-norm) for different speaker verification systems. We adopt the SRE04 and SRE05 corpus as t-norm corpus and the SRE06, Switchboard II Phase 2 and Phase 3 corpus as z-norm corpus. The number of speakers in the corpus for t-norm is 300 and for z-norm is 800.

V. RESULTS

In this section, we present individual and combined results on the NIST SRE 08 male core test (cc=6) from the previously described systems.

A. Different Types of Input Supervectors in Fishervoice

The first experiment investigates the sensitivity of speaker verification performance for the proposed method with regards to the types of input supervectors and score normalization method used. As mentioned before, we apply the Fishervoice framework along with the normalized correlation for distance metric. Table 1 gives the results obtained without score normalization, with T-norm and TZ-norm score normalization on the three types of input supervector for the proposed system. First, we observe that the performance remains stable across input vector types. Second, score normalization may greatly improve system performance. Third, T-norm maybe more suitable for minDCF measurement (12.4% relative improvement in Standard Mean as input) while TZ-norm may achive better results for EER meansurement (9.2% relative improvement in Weighted Mean as input). However, MinDCF of TZ-norm is not improved compare to T-norm. Besides, the compution speed in the test stage is very fast since no likelihood calculation is needed.

TABLE I. RESULTS OBTAINED WITH DIFFERENT NORMALIZATION SCHEMES ON THE THREE TYPES OF INPUT SUPERVECTOR. EER(%), MINDCF(X100)

Feature Input	No Norm		T-norm		TZ-norm	
	EER	***minDCF***	***EER***	***minDCF***	***EER***	***minDCF***
Standard Mean	5.95	3.31	5.72	**2.90**	5.47	2.99
Weighted Mean	5.95	3.31	5.68	2.94	**5.40**	2.97
Kernel Mean	5.84	3.30	5.68	2.93	5.46	2.99

B. Comparison with the Other Systems

The second experiment compares the Fishervoice framework with three other standard approaches, namely, JFA [2], JFA-SVM with linear kernel [9] [13] and JFA-SVM with cosine kernel [4]. In the Fishervoice approach, we select the weighted mean as input supervector. Figure 2 shows the results obtained by the above mentioned systems. They suggest that the integration of JFA supervector with nonparametric Fisher's discriminant analysis in the Fishervoice framework leads to superior performance compared to other systems. Compared to a single JFA classifier, the Fishevoice framework improves minDCF results by decreasing the minDCF from 0.0305 to 0.0297 and improves performance of EER by a relative 10.3%. Besides, JFA works better than JFA-SVM where both methods use supervector as input to train discriminative models. The advantage of applying Fishervoice framework is that each high-dimensional input supervector is cut into small slices while multiple subspace analysis works well on the low-dimensional vector without any loss of useful information. These observations motivated us to devise a third experiment that fuse the Fishervoice with above standard systems.

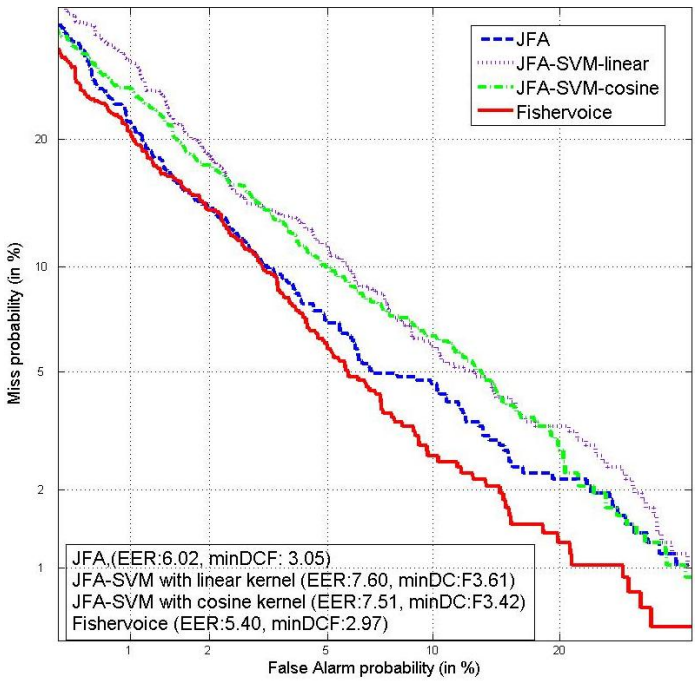

Figure 2. Comparison of Fishervoice and other standard systems on NIST SRE 08 male core task (cc=6,100x minDCF)

C. *Fusion with the Other Systems*

In the third experiment, we fuse the Fishervoice with several standard systems (in Figure 3). We select JFA-SVM with cosine kernel to represent SVM based system. First, it is worth noting that the JFA + JFA-SVM fused systems only achieve comparable results compare to single Fishervoice system. Second, according to EER metric, Fishervoice fused with JFA and JFA-SVM offer best performance compare to single JFA system. It improves results by decreasing the EER from 6.02% to 4.67% by a relative 22.4% and decreasing the minDCF from 0.0305 to 0.0269. Third, we also find that the performance between the fusions (JFA+Fishervoice) and (JFA+Fishervoice+JFA-SVM-cosine) look similar. The possible reason may be that the functions of Fishervoice and SVM are the same since both models in two systems are trained in the discriminative way.

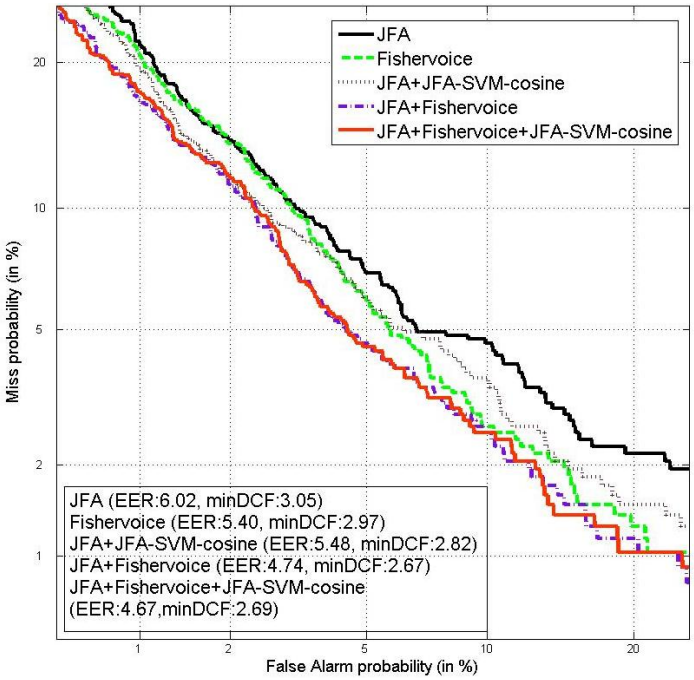

Figure 3. Fusion results with other systems on NIST SRE 2008 male core task (cc=6,100x minDCF)

VI. CONCLUSIONS

This paper enhances our previous work in the Fishervoice approach [6] for speaker verification. The approach includes the application of nonparametric Fisher's discriminant analysis to map the supervector into a discriminant subspace for fast and effective matching. The objective is to reduce intra-speaker variability that is unfavorable for the speaker recognition task, as well as extract discriminant speaker class boundary information that is conducive to the task. The enhancement presented represents each utterance with a fix-length super-vector using JFA Gaussian GMM. Multiple discriminant projections are then used on partitioned vectors of the super-vector for fast and effective matching. Extensive experiments on the NIST08 male core test show the advantage of the proposed framework over the state-of-the-art algorithms.

ACKNOWLEDGMENT

This work is affiliated with the CUHK MoE-Microsoft Key Laboratory of Human-centric Computing and Interface Technologies. The authors are grateful to Dr. Lo Wai-Kit for helpful and informative discussions on this research topic.

REFERENCES

[1] D. Reynolds, T. Quatieri and R. Dunn, "Speaker verification using adapted Gaussian mixture models," Digita Signal Processing, vol. 10, no. 1-3, pp. 1941, 2000.

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

[3] N. Dehak, P. Kenny, R. Dehak, O. Glembek, P. Dumouchel, L. Burget, H. Valiantsina, and F. Castaldo, "Support vector machines and joint factor analysis for speaker verification," Proceedingsof ICASSP 2009,

[4] Johns Hopkins University, SummerWorkshop, "Robust Speaker Recognition Over Varying Channels",2008

[5] 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, November 2009.

[6] Z. Li, W. Jiang and H.Meng "Fishervoice: a discriminant speaker recognition," ICASSP2010

[7] Z. Li, D. Lin, X. Tang, "Nonparametric discriminant analysis for face recognition," *IEEE Trans. on PAMI*, vol. 31, no. 4, pp. 755-761, 2009

[8] D. Matrouf, N. Scheffer, B. Fauve, J.-F. Bonastre "A Straightforward and Efficient Implementation of the Factor Analysis Model for Speaker Verification", in Proc. Interspeech, 2007

[9] W. Campbell, D. Sturim, D. Reynolds, A. Solomonoff, "SVM Based Speaker Verification using a GMM Supervector Kernel and NAP Variability Compensation," IEEE International Conference on Acoustics, Speech and Signal Processing 1, 97–100 (2006)

[10] R. Kuhn, P. Nguyen, J.-C. Junqua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field, and M. Contolini, "Eigenvoices for speaker adaptation," Proceedings of International Conference on Spoken Language Processing, vol. 5, pp. 1771–1774, 1998.

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

[12] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

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