FISHERVIOCE: A DISCRIMINANT SUBSPACE FRAMEWORK FOR SPEAKER RECOGNITION

Zhifeng Li, Weiwu Jiang and Helen Meng
The Chinese University of Hong Kong, Hong Kong SAR of China

ABSTRACT

We propose a new framework for speaker recognition, referred as Fishervoice. It includes the design of a feature representation known as the structured score vector (SSV), which relates acoustic structures with "key" frames in an input utterance in capturing relevant speaker characteristics. The framework also applies nonparametric Fisher's discriminant analysis to map the SSVs into a compressed discriminant subspace, where matching is performed between a test sample and reference speaker samples to achieve speaker recognition. The objective is to reduce intra-speaker variability and emphasize discriminative class boundary information to facilitate speaker recognition. Experiments based on the XM2VTSDB corpus shows that the Fishervoice framework gave superior performance, compared with other commonly used approaches, e.g. GMM-UBM and Eigenvoice.

Index Terms— speaker recognition, GMM, subspace model, discriminant analysis, Fishervoice

1. INTRODUCTION

Approaches to the speaker recognition have often gravitated towards the use of Gaussian Mixture Models (GMM) [1] and a variety of related techniques, e.g. GMM-UBM [2]. These approaches are faced with several challenging issues that are inherent in the speaker recognition task: (1) the need for an efficient representation of speaker characteristics that reflects structure in the acoustic speech signal that corresponds to different vocal tract configurations; (2) the need for robustness against intra-speaker variabilities due to speaking style differences, noise and other interferences; and (3) the need for discriminant information among speakers that is conducive to improving recognition performance.

Regarding the first issue – the speech signal is laden with a variety of information that are intricately integrated, including the characteristics of the speaker, language, speaking style, etc. Feature representations that are commonly used, e.g. MFCC, PLP, etc. [15] may not readily reflect the structures[1] specific to speaker characterization.

To address the second issue, subspace analysis [3] techniques have been applied. They include Principal Component Analysis (PCA) [9] and Fisher's Discriminant Analysis (FDA) [10]. These techniques work efficiently by adapting each speaker model from the high-dimensional feature space into a reduced dimension subspace. The mean vectors from each speaker's GMM are reconstructed by a linear weighted combination of basis eigenvectors called eigenvoices. In addition, factor analysis (FA) [4] has been used to separate speaker and channel variability and has been shown to be effective in the NIST test set [4]. The approaches proposed in [2-5] have achieved good performance through model adaptation in a GMM-based subspace adaptation. It seems desirable to augment the approach with the use of discriminant information.

To address the third issue, Campbell et al. [6] applied SVMs in speaker recognition that involves a nonlinear mapping from the input space to an SVM expansion space. A similar approach called Nuisance Attribute Projection (NAP) [7] has been applied to suppress channel effects. In addition, Stolcke et al. [8] applied rank-normalization to create nonparametric features and improve SVM performance. The performance of SVMs is critically dependent on the selection of efficient kernel functions.

In this work, we propose a novel speaker recognition framework that applies nonparametric Fisher's discriminant analysis (refered as Fishervoice). The approach is currently introduced for speaker recognition, inspired by work in face recognition [12]. In the Fishervoice framework, we first design a feature representation referred as structured score vector (SSV). The main idea is to consider the unified GMM that models the entire space of speakers' utterances, and use each Gaussian to select a small set of key frames (top-scoring frames) from a speaker's input utterance. The key frames are considered representative of the acoustic class structure represented by the Gaussian. The mean score of the key frames for Gaussian are grouped together to form the SSV. Second, we apply nonparametric Fisher's discriminant analysis that maps SSV into a compressed (reduced dimension) subspace. The analysis is performed in an attempt to suppress intra-speaker variations and to emphasize the discriminative information for speaker recognition. Third, the method of subspace analysis is applied directly to speaker recognition by computing the distance between an input testing sample with the reference vector of each known speaker.

2. THE FISHERVOICE FRAMEWORK

We propose a framework (which we name "Fishervoice") to explore the use of subspace analysis to model speaker

[1] An analogous problem is present in person identification through facial images, where the useful "structures" to extract from the facial image are the eyes, nose and mouth.

characteristics in a low-dimensional discriminant subspace. More specifically, the framework integrates a novel feature representation (which we name "Structured Score Vector"), with nonparametric Fisher's discriminant analysis. Figure 1 illustrates the overall organization of the proposed framework. We will describe the respective components of the framework in the following subsections.

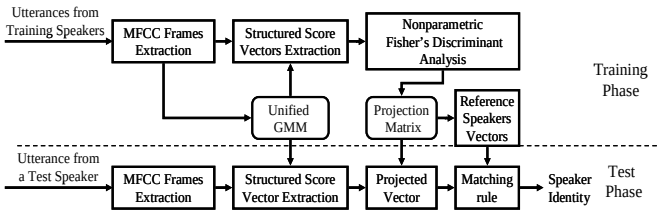

Figure 1. Illustration of the proposed framework for speaker recognition framework, to which we refer as "Fishervoice".

2.1 Structured score vector (SSV) extraction

The SSV is designed to leverage the acoustic class structure captured in the unified GMM (see Figure 1) in order to locate the "key" frames in an input utterance. This unified GMM is trained from all training data (akin to the universal background model commonly used in speaker recognition), which captures the probabilistic distribution of acoustic feature classes (or *structures*[2]) in the overall acoustic space in terms of a mixture of Gaussians. Given an input utterance, we extract specific MFCC frames that have high likelihood scores with respective to each Gaussian distribution, with the assumption that these are the more representative frames (i.e. "key" frames) for the corresponding acoustic class. Mathematically, suppose there are H_i utterances in total for the speaker i across all recording sessions, of which the h-th utterance consists of $N_{i,h}$ MFCC frames $\{v_{i,h}(n)|n = 1,2,\ldots N_{i,h}\}$. We can compute the likelihoods scores of these frames for each Gaussian distribution to form the matrix $B_{i,h}$:

$$B_{i,h} = \begin{bmatrix} b_1(v_{i,h}(1)) & \cdots & b_1(v_{i,h}(n)) & \cdots & b_1(v_{i,h}(N_{i,h})) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ b_j(v_{i,h}(1)) & \cdots & b_j(v_{i,h}(n)) & \cdots & b_j(v_{i,h}(N_{i,h})) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ b_M(v_{i,h}(1)) & \cdots & b_M(v_{i,h}(n)) & \cdots & b_M(v_{i,h}(N_{i,h})) \end{bmatrix} \quad (3)$$

$$b_j(v_{i,h}(n)) = w_j (2\pi)^{-\frac{D}{2}} \left|\sum_j\right|^{-\frac{1}{2}} e^{-\frac{1}{2}(v_{i,h}(n)-\mu_j)^T \sum_j^{-1}(v_{i,h}(n)-\mu_j)}$$

where M is the number of Gaussian mixtures and $b_j(v_{i,h}(n))$ is the weighted likelihood score of the n-th frame with respect to the j-th Gaussian that has mean vector μ_j, covariance matrix $\sum_j$ and weight w_j.[3] If we assume that a higher likelihood score for an MFCC frame is more representative of the acoustic structure characterized by the Gaussian, then we may regard the top-scoring K frames to be the *key frames* in the utterance for the specific acoustic structure. Let $t_{j,e}$ denote the original index of the e-th (e = 1,2,…K) key frame of the j-th Gaussian, we compute the mean likelihood score $s_{i,h,j}$ of K key frames to represent each structure as follows:

$$s_{i,h,j} = \frac{1}{K}\sum_{e=1}^{K} b_j(v_{i,h}(t_{j,e})) \quad (4)$$

Furthermore, we represent the h-th utterance from the i-th speaker in terms of an M dimensional vector $x_{i,h}$:

$$x_{i,h} = [s_{i,h,1} \;\; s_{i,h,2} \;\; \cdots \; s_{i,h,j} \;\; \cdots \; s_{i,h,M}]^T \quad (5)$$

We refer to $x_{i,h}$ as the *structured score vector* (SSV), which represents its acoustic structure with selected *key* frames to capture relevant speaker characteristics. This step differs slightly from [16], which uses Gaussian scores of all frames.

2.2 Nonparametric Fisher's discriminant analysis

The SSV represents each input utterance in terms of a score vector in a high-dimensional space (M dimensions). If we assume that there is a low-dimensional discriminant subspace for effective speaker discrimination, we may attempt to apply nonparametric Fisher's discriminant analysis for mapping into the subspace.

The traditional Fisher's discriminant analysis (FDA) aims to maximize class separability [10]. It 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 SSVs from all training speakers, we let C denote the total number of speakers, H_i be the number of samples (or sessions) in the speaker i, ξ_i be the sample mean of the class i and ξ be the sample mean of all training data. The optimal projection W for FDA is calculated as follows:

$$W = \arg\max_W \left(\frac{\left\|W^T S_b W\right\|}{\left\|W^T S_w W\right\|}\right)$$
$$S_w = \sum_{i=1}^{C}\sum_{h=1}^{H_i}(x_{i,h}-\xi_i)(x_{i,h}-\xi_i)^T,\; S_b = \sum_{i=1}^{C} H_i(\mu_i-\xi)(\mu_i-\xi)^T \quad (6)$$

In the proposed Fishervoice framework, we aim to enhance S_w and S_b to extract discriminant information more effectively. First, the space of all SSVs derived from the entire set of training samples undergoes dimension reduction by principal component analysis (PCA) and is projected into a subspace to make S_w nonsingular. Then two procedural steps are applied in solving W:

Step 1: In the PCA projected space, we further enhance S_w by whitening [11], in an attempt to remove intra-speaker variations. This is achieved by a whitening transform matrix T, which is computed as follows:

$$T^T S_w T = I, \;\; T = \Phi\Lambda^{-1/2} \quad (7)$$

where Φ is the normalized eigenvector matrix of S_w, Λ is the eigenvalue matrix of S_w and I is the identity matrix.

Step 2: We enhance S_b by applying nonparametric subspace analysis [11] [12] to obtain a *nonparametric* between-class scatter matrix S_b'. This aims to better characterize inter-speaker variations. For an arbitrary utterance h from speaker i, let $x_{i,h}'$ denote the new SSV that has undergone

[2] In this context, we refer to acoustic feature classes as *structures*.

[3] In practice we take the *log* of the scores.

two projections (PCA and whitening), a process which is consistent with step 1 above. We consider the contribution of $x'_{i,h}$ towards the nonparametric between-class scatter matrix S'_b by focusing on its proximity to the boundary that separates speaker class i and any other class k. S'_b is computed according to the following equations:

$$S'_b = \sum_{i=1}^{C} \sum_{k=1,k\neq i}^{C} \sum_{h=1}^{H_i} g(i,k,h)(x'_{i,h} - m_k(x'_{i,h}))(x'_{i,h} - m_k(x'_{i,h}))^T$$
$$m_k(x'_{i,h}) = \frac{1}{R}\sum_{q=1}^{R} \varphi_{k,q}(x'_{i,h}) \quad (8)$$

where $\varphi_{k,q}(x'_{i,h})$ is q-th sample from speaker k that among the nearest neighbors (projected SSV) of $x'_{i,h}$, R is the total number of such nearest neighbors considered, $m_k(x'_{i,h})$ is the mean of these R nearest neighbors, and $g(i,k,h)$ is a weighting function defined as:

$$g(i,k,h) = \frac{\min\{d^{\alpha}(x'_{i,h}, \varphi_{i,R}(x'_{i,h})), d^{\alpha}(x'_{i,h}, \varphi_{k,R}(x'_{i,h}))\}}{d^{\alpha}(x'_{i,h}, \varphi_{i,R}(x'_{i,h})) + d^{\alpha}(x'_{i,h}, \varphi_{k,R}(x'_{i,h}))} \quad (9)$$

where exponential parameter α controls the variation of the weighting function with respect to the distance $d(o_1, o_2)$, which is the Euclidean distance between two vectors o_1 and o_2. The parameter R is often set as the median of the total sessions for each speaker in the training data [12]. The weighting function $g(i,k,h)$ assesses the proximity of the projected SSV $x'_{i,h}$ to a local speaker class boundary and weights the SSV's contribution towards constructing the matrix S'_b. This weight approaches the highest value of 0.5 if $x'_{i,h}$ is near the classification boundary and decreases as $x'_{i,h}$ moves far away from the classification boundary.

To summarize, subspace projections involved in the overall Fisher's discriminant analysis include the following:

(i) **Subspace projection for dimension reduction:** Compute the PCA projection matrix W_1 from the entire training set and use it to project all SSVs into the PCA subspace. The subspace projection f_1 is obtained by:

$$f_1 = W_1^T x, \text{ where } W_1 = \arg\max_W \left\| W^T \Psi W \right\| \quad (10)$$

where x is an arbitrary SSV from Eq. (5) and Ψ is the covariance matrix of all the SSVs in the training set.

(ii) **Subspace projection to reduce intra-speaker variations:** In the PCA subspace above, compute the whitened subspace according to Eq. (7) and adjust the dimension of the whitened subspace to reduce intra-speaker variability. The subspace projection f_2 is obtained by:

$$f_2 = W_2^T f_1, \text{ where } W_2^T S_w W_2 = I, \; W_2 = \Phi \Lambda^{-1/2} \quad (11)$$

where W_2 is the whitening transformation matrix applied to the within-class scatter matrix S_w via Eq. (10), Φ is the normalized eigenvector matrix of S_w, Λ is the eigenvalue matrix of S_w.

(iii) **Subspace projection to extract discriminant speaker class boundary information:** In the projected subspace above (after PCA and whitening), compute the matrix S'_b according to Eq. (8-9). Perform PCA on the nonparametric between-class scatter matrix S'_b and choose dominant eigenvectors to form the PCA projection matrix W_3. The subspace projection f_3 is obtained by:

$$f_3 = W_3^T f_2, \text{ where } W_3 = \arg\max_W \left\| W^T S'_b W \right\| \quad (12)$$

(iv) Finally, the overall subspace transformation matrix W_N is denoted as:

$$W_N = W_1 W_2 W_3 \quad (13)$$

W_N is computed from all training utterances. For each speaker in the training set, we calculate the mean of the subspace projections from all of his/her training utterances to form the *reference vector* of that speaker in the projected SSV space.

2.3 Testing procedure

Based on the training data set, we compute the overall Fisher's discriminant subspace projection matrix W_N in Eq. (13). During testing, we compute the SSV of the input test utterance and perform the subspace projections in W_N. This projected testing sample (an SSV) is compared with the *reference vector* for each speaker (see the subsection above), in terms of distance metrics [1] such as the Euclidean distance (EUC) and the normalized correlation (COR), in order to perform speaker recognition. The two distance metrics are shown in Eq. (14):

$$D_{EUC}(o_1, o_2) = \sqrt{(o_1 - o_2)^T (o_1 - o_2)}$$
$$D_{COR}(o_1, o_2) = \left\| o_1^T o_2 \right\| \Big/ \sqrt{o_1^T o_1 o_2^T o_2} \quad (14)$$

where o_1 and o_2 are two projected SSVs.

3. EXPERIMENTAL RESULTS AND ANALYSIS

We experimented with the XM2VTS database [14], which is comprised of 295 speakers, recorded from four different sessions. Each speaker reads two numeric sequences with audio and face recordings. The first three sessions are used for training and the last for testing. We have previously used this corpus for face recognition and currently extend our work to speaker recognition. The corpus offers multiple sessions with a relatively large number of subjects, and supports future work in multimodal person recognition.

Front-end processing includes: (i) Down-sampling the audio data from 32 kHz to 8 kHz, to follow the common experiment setup with reference to previous works [2] [3]. (ii) Extracting thirteen MFCCs and their time derivatives (delta and delta delta MFCCs) using a 30ms Hamming window, with a step of 15 ms between successive windows.

We reference the other (common) approaches to speaker recognition such as the GMM-UBM [2] and Eigenvoice GMM [3]. The UBM combines two gender-dependent GMMs trained respectively on male and female training data. Each speaker model is adapted (mean-only) with one iteration and the relevance factor is 16. As for the Fishervoice approach, the UBM is used as the unified GMM. We compute the SSVs with M=1024. The parameter

R is set to 2 (as explained earlier) and α is set to 2 (in order to empirically balance the rate of decrease of the weighting function g). The Fisher's discriminant subspace projection matrices, W_1, W_2 and W_3, have the dimensions of 884, 590 and 294 respectively, corresponding to the upper limit of their matrix ranks.

The first experiment investigates the sensitivity of speaker recognition performance with regards to the number of key frames (K) used in the SSV. As mentioned, we apply the Fishervoice framework along with the two distance metrics – Euclidean distance (EUC) and normalized correlation (COR). Results are shown in Figure 2. We observe that the speaker recognition performance remains largely stable across a range of K values above 6. This suggests that a relatively small number of key frames can contribute significantly towards modeling speaker identities.

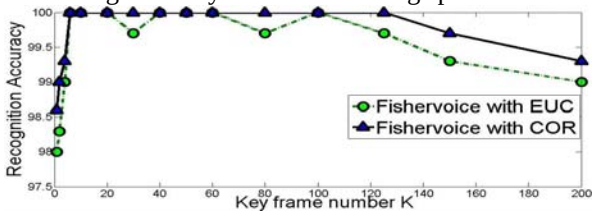

Figure 2. Speaker recognition accuracy (%) based on the Fishervoice framework with two distance metrics – Euclidean distance (EUC) and normalized correlation (COR).

The second experiment compares the Fishervoice framework (with K=6) with two other approaches, namely, GMM-UBM [2] and Eigenvoice [3]. In the Eigenvoice approach, each speaker's GMM mean vectors are reconstructed by a linear weighted combination of eigenvectors. A maximum likelihood solution called MLED (Maximum Likelihood Eigen-Decomposition) is used to estimate the weights. The number of Gaussian mixtures is set at 1024 and 2048. The number of eigenvectors used in Eigenvoice method is set to 270 with 98% of the variational energy retained in eigenspace. Results are shown in Table 1. They suggest that the integration of SSV with nonparametric Fisher's discriminant analysis in the Fishervoice framework leads to superior performance (100%) in the XM2VTSDB test set with 295 utterances.

Table 1. Comparison among three approaches to speaker recognition accuracy (%): GMM-UBM, Eigenvoice and the proposed Fishervoice framework

Gaussian Mixtures	GMM-UBM	Eigenvoice	Fishervoice (EUC)	Fishervoice (COR)
1024	98.0	96.3	100	100
2048	98.9	96.9	100	100

4. CONCLUSIONS

This paper proposes a new framework for speaker recognition, referred as Fishervoice. It includes the design of a feature representation known as the structured score vector (SSV), which relates acoustic structures (obtained from a Gaussian mixture model) with key frames in an input utterance in capturing relevant speaker characteristics. The framework also includes the application of nonparametric Fisher's discriminant analysis to map the SSVs into a discriminant subspace, where matching is performed between a test sample and reference speaker samples. 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. Experiments based on the XM2VTSDB shows that the Fishervoice framework gave superior performance, compared with other commonly used approaches, e.g. GMM-UBM and Eigenvoice. Future work includes experimentation with NIST database and SVM-based methods.

5. REFERENCES

[1] D. Reynolds and R. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. SAP*, vol.3, no.1, pp.72-83, 1995.

[2] D. Reynolds, F. Thomas, and B. Robert, "Speaker verification using adapted Gaussian Mixture Models," *DSP,* vol.10, no.1-3, pp.19–41, 2000.

[3] R. Kuhn, P. Nguyen, and J. Junqua, "Eigenvoices for speaker adaptation," *ICSLP*, pp.1771–1774, 1998.

[4] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Factor analysis simplified," *ICASSP* 2005

[5] P. Kenny, et al., "Improvements in factor analysis based speaker verification," *ICASSP* 2006.

[6] W. Campbell, et al., "Phonetic speaker recognition with support vector machines," *NIPS*, 2004.

[7] W. Campbell, D. Sturim, D. Reynolds and A. Solomonoff, "SVM based speaker verification using a GMM super vector kernel and NAP variability compensation," *ICASSP,* 2006.

[8] A. Stolcke, S. Kajarekar, and L. Ferrer, "Nonparametric feature normalization for SVM-based speaker verification," *Proc. ICASSP,* pp.1577-1580, 2008.

[9] M. Turk and A. Pentland, "Face recognition using eigenfaces," *Proc. CVPR*, pp. 586-591, 1991.

[10] P. Belhumeur, J. Hespanda, and D. Kiregeman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *PAMI,* vol.19, pp.711-720, 1997.

[11] K. Fukunaga, *Statistical Pattern Recognition*, Academic Press, 1990.

[12] Z. Li, D. Lin, X. Tang, "Nonparametric discriminant analysis for face recognition," *IEEE Trans. on PAMI*, vol. 31, no. 4, pp. 755-761, 2008

[13] J. Kittler, Y. Li, and J. Matas. "On matching scores for LDA-based face verification," *Proc. British Machine Vision Conference*, pp. 42–51, 2000.

[14] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Matitre, "XM2VTSDB:The extended M2VTS database," *Second International Conference on Audio- and Video-Based Biometric Person Authentication*, 1999.

[15] T. Kinnunen, H. Li, "An overview of text-independent speaker recognition: from features to supervectors", *Speech Communication* 2009

[16] J.Kharroubi, D.Petrovska-Delacretaz, and G.Chollet, "Combining GMM's with suport vector machines for text-independent speaker verification", *Eurospeech 2001*, 1761-1764.