CLUSTERING SIMILAR ACOUSTIC CLASSES IN THE FISHERVOICE FRAMEWORK

Na Li[1,2] *Weiwu Jiang*[3] *Helen Meng*[2,3] *and Zhifeng Li*[2]
[1]College of Marine Engineering, Northwestern Polytechnical University, Xi'an, China,
[2]Shenzhen key lab of CVPR, Shenzhen Institutes of Advanced Technology, CAS, China,
[3]The Chinese University of Hong Kong, Hong Kong SAR
{na.li, zhifeng.li}@siat.ac.cn, {wwjiang, hmmeng}@se.cuhk.edu.hk

ABSTRACT

In the Fishervoice (FSH) based framework, the mean supervectors of the speaker models are divided into several subvectors by mixture index. However, this division strategy cannot capture local acoustic class structure information among similar acoustic classes or discriminative information between different acoustic classes. In order to verify whether or not local structure information can help improve system performance, we develop five different speaker supervector segmentation methods. Experiments on NIST SRE08 prove that clustering similar acoustic classes together improves the system performance. In particular, the proposed method of equal size clustering achieves 5.1% relative decrease on EER compared to FSH[1].

Index Terms— speaker verification, Fishervoice, structure information, subvectors

1. INTRODUCTION

In recent years, many techniques have been proposed in the field of speaker verification. The most popular method is Gaussian mixture model based Joint Factor Analysis (JFA) [1][2]. It achieves significant performance improvement by modeling speaker and channel variability into two low-dimensional subspaces.

Although JFA brings much improvement, the framework has the drawback that it requires the distribution of channel and speaker information to be independent. To address this issue, Dehak proposed a new speaker verification system to compress both speaker and channel information into a low-dimensional space called total variability space [3], then represent each speech utterance by a total factor feature vector (called i-vector). This system achieved much success in the NIST evaluations.

Inspired by the JFA and i-vector systems, we proposed the Fishervoice (FSH) framework and its enhancements [4-6]. These methods make use of the discriminative information [7] of the JFA speaker supervector, as well as project each high-dimensional speaker supervector to a low-dimensional subspace by suppressing intra-speaker variations while emphasizing discriminative information.

If we adopt the view that Gaussian model represent some kind of broad phonetic events, then information about the entire acoustic class structure can be captured by Gaussian Mixture Models (GMM). However, when the number of mixtures increases, some acoustic classes may become close to each other in the modeling process. In previous work, when multiple discriminative projections are applied, some of the close-lying acoustic classes may be distributed into different subvectors, causing partial loss of some discriminative information around local boundaries.

To address this case, we develop several different segmentation strategies to localize and cluster the structural information of Gaussian mixtures. Experimental results confirm that clustering similar acoustic classes together can improve system performance.

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

2. PREVIOUS WORK

2.1. Joint factor analysis (JFA) supervector

The approach of JFA specifies that the speaker and channel noise components, which reside in the speaker-and-channel-dependent supervectors respectively, are assumed to follow Gaussian distributions. Let M_{ih} denote the speaker- and channel-dependent supervector of the mean vectors for the h-th utterance from speaker i. M_{ih} is assumed to be made up of four supervectors as shown below:

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

where m is the mean supervector of the universal background model (UBM) [8], 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 the speaker

[1] The corresponding author is Zhifeng Li. This work was partially conducted when the first author interned in Shenzhen Institutes of Advanced Technology.

vector by grouping the first three terms in Eq. (1):

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

2.2. Fishervoice (FSH) discriminative analysis

The Fishervoice framework includes three projections as illustrated in Eq. (3-5):

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

$$f_1 = W_1^T x, \text{ where } W_1 = \arg\max_W \left\| W^T \Psi W \right\| \quad (3)$$

where x is an any supervector and Ψ is the covariance matrix of all supervectors in the development set.

2) The 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:

$$f_2 = W_2^T f_1, \text{ where } W_2^T S_w W_2 = I, \; W_2 = \Phi \Lambda^{-1/2} \quad (4)$$

where S_w is the standard within-class scatter matrix in [7], Φ is the normalized eigenvector matrix of S_w, and Λ is the eigenvalue matrix of S_w.

3) The 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) in [4] from the whitened subspace above as:

$$f_3 = W_3^T f_2, \text{ where } W_3 = \arg\max_W \left\| W^T S_b' W \right\| \quad (5)$$

Finally, to extract discriminative information from the scatter matrices S_w and S_b effectively, the overall transformation matrix W_{NF} for nonparametric Fisher discriminative analysis is given by:

$$W_{NF} = W_1 W_2 W_3 \quad (6)$$

Details about this framework can be found in [4].

3. PROPOSED SYSTEMS

In this paper, we investigate several strategies of dividing the original supervector into subvectors for parallel training and projection, and then concatenate these projected subvectors together as the projected speaker vector.

3.1. Training and testing stage

The training procedure of the Fishervoice-based framework is described as follows:

1) Extract the input speaker supervector according to Eq. (2) from each utterance.

2) Divide each speaker supervector into K subvectors using one of the proposed strategies.

3) Apply Fishervoice discriminative analysis on each subvector in parallel according to Eq. (3-6) and obtain the transformation matrix W_k for the k-th subvector

$$W_k = W_{k1} W_{k2} W_{k3} \quad (7)$$

where W_{k1} ,W_{k2} ,W_{k3} denotes the projection matrices described in section 2.2.

4) Concatenate all the transformation matrices W_k (k=1,2,…,K) to form the total projection matrix W_{Total} as follows:

$$W_{Total} = \left[W_1 \cdots W_k \cdots W_K \right] \quad (8)$$

5) For target speaker enrollment, each speaker's supervector is projected into a low-dimensional training reference vector R_{train} by the total projection matrix W_{Total}.

In the testing stage, we extract the supervector from the test speaker, similar to the training procedure. Then each supervector is projected into a testing reference vector R_{test} by the total projection matrix W_{Total}. We calculate the distance score between R_{trainl} and R_{test} in terms of the normalized correlation (COR) which is shown as:

$$D\left(R_{train}, R_{test}\right) = \frac{\left\| R_{train}{}^T R_{test} \right\|}{\sqrt{R_{train}{}^T R_{train} R_{test}{}^T R_{test}}} \quad (9)$$

3.2. Subvector division strategies

We propose five different strategies to divide the high-dimensional speaker supervector into subvectors.

3.2.1 Random Gaussian mixture index selection (R-FSH)

In this method, we first collect all Gaussian mixture index labels of the UBM model. Then all these index labels are randomly divided into K classes with equal class size. As a result, each subvector is concatenated by the Gaussian mixture mean (speaker mean) vectors with all corresponding index labels in a class. In our previous work [5][6], we simply divide all Gaussian mixture index labels into K classes in order, which can be considered as a special case of R-FSH.

3.2.2 Non-equal size clustering (NE-FSH)

a. Given a UBM model with M mixtures, we consider the M mean vectors m_j (j=1,2,…,M) as input data points. Then we create a GMM with K mixtures using the M data points.

b. For each mean vector m_j, we calculate its posterior probabilities with the K Gaussian components.

c. We classify each mean vector m_j, into the k-th class (k=1…K) if the posterior probability of m_j with the k-th Gaussian component has the largest value P_{max_j}.

d. We arrange the indexes of all mean vectors in each class in descending order and concatenate the corresponding mean vectors as k-th subvector.

3.2.3 Equal size clustering based on GMM (E-FSH)

a. Same steps (a-c) as in the NE-FSH.

b. For the mean vector m_j, if the number of vectors in class k exceeds the average number for each class, we compare P_{max_j} of m_j with the smallest value in class k. We place m_j into class k and move the original smallest value corresponding vector into other class to meet requirement if P_{max_j} is the larger one. Otherwise, we continue to relocate vector m_j to other class till requirement is achieved.

3.2.4 Feature dimension alignment clustering (F-FSH)

Suppose the input mean vector m_j is $[x_j, \triangle x_j, \triangle\triangle x_j]^T$, ($j$=1…$M$), where $x_j \in R^N$ is extracted from the $3N$ dimensional MFCC feature. We concatenate all k-th dimensional component of mean vectors m_j to generate the k-th subvector (k=1…$3N$). The dimension of each subvector is M.

3.2.5 Feature dimension alignment clustering with derivative information (FD-FSH)

For each input mean vector $m_j = [x_j, \triangle x_j, \triangle\triangle x_j]^T$, ($j$=1…$M$), $x_j \in R^N$, we concatenate the k-th dimensional component of $x_j, \triangle x_j$ and $\triangle\triangle x_j$ to generate the k-th subvector. The dimension of each subvetor is 3M.

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.

4.2. Feature extraction

ETSI Adaptive Multi-Rate (AMR) GSM VAD [9] is applied to prune out silence. Then the speech is segmented into 25ms Hamming window frames 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-dimensional vector. Finally, the feature warping process [10] is applied to all the MFCCs.

4.3. Subspace training

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 utterances 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.

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 ranks (800, 799 and 550) respectively. The number of subvectors in each speaker supervector is set to 16 by cross validation in our previous experiments [11]. The parameter R which controls the number of nearest neighbors for constructing S_b' in was set to 4, according to the median number of sessions for each speaker.

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. 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. Results are given in terms of equal error rate (EER) and minDCF.

5.1. Random selection of Gaussian mixtures

The first experiment investigates the sensitivity of R-FSH system with regards to the different selection of Gaussian mixtures. We randomly create five R-FSH systems for training and compare these results with those of the standard JFA and our previous Fishervoice (FSH) based framework.

Table 1. Comparison among the results of R-FSH, FSH, and JFA on NIST 2008 male core task (tel-tel condition)

System Type	EER (%)	minDCF (×100)
R-FSH	4.25	**2.15**
	4.37	2.18
	4.36	2.18
	4.28	2.26
	4.17	2.17
FSH	4.34	2.16
JFA	4.65	2.50

Table 2. Experimental results of NE-FSH, E-FSH and FSH on the NIST 2008 male core task (tel-tel condition)

System Type	EER (%)	minDCF (×100)
NE-FSH	4.36	2.30
	4.44	2.30
	4.52	2.30
	4.18	2.29
	4.12	2.27
E-FSH	4.21	2.24
	4.28	2.30
	4.23	2.20
	4.28	2.27
	4.28	2.26
FSH	4.34	**2.16**

Table 1 suggests that the R-FSH method outperforms the standard JFA system for both EER and minDCF metrics. However, the performance of the randomly created systems is not stable. FSH is a special case of the R-FSH method

since each Gaussian mixture is independent from others.

5.2. Comparison with GMM-based clustering

In the second experiment, five different Gaussian models are trained to cluster mean vectors. Accordingly, five NE-FSH and E-FSH systems are created.

Table 2 summarizes the results of the proposed NE-FSH and E-FSH systems. Key observations include: First, both the NE-FSH and E-FSH systems perform better than FSH system on average. This is because GMM-based clustering can select similar acoustic classes together to enhance location boundary information for discriminative training which improves the performance of the system. Second, E-FSH system performs more stable than NE-FSH system. Third, NE-FSH system obtains the lower EER than that of E-FSH system on average. Lastly, in some cases, NE-FSH may perform worse than E-FSH. The possible reason is that the number of mean vectors (2048) may not be enough to train a stable GMM with 16 mixtures.

5.3. MFCC feature alignment clustering

In this section, we further investigate the sensitivity of the FSH framework without using any local class boundary information between Gaussian mixtures. Table 3 shows the results of F-FSH and FD-FSH. We observe that when we discard the information of acoustic class structure in the speaker supervector, the performance of F-FSH and FD-FSH degrades significantly compared to FSH. However, compared to the JFA system, the F-FSH and FD-FSH systems perform slightly better in both EER and minDCF metrics. This indicates that the discriminative classifier improves the performance of the generative model.

Table 3. Experimental results of F-FSH, FD-FSH, FSH

System Type	EER (%)	minDCF(×100)
F-FSH	4.61	2.41
FD-FSH	4.65	2.42
FSH	**4.34**	**2.16**
JFA	4.65	2.50

5.4. Different amount of information retained strategy

In the fourth experiment, we aim to apply dynamic dimensions for projection matrixes based on the energy reserved. Table 4 shows that the performance of NE-FSH improves slightly with increasing variance retained in W_{k1}. NE-FSH achieves its lowest EER value when retained-to-total information ratio is 90% in W_{k1}.

5.5. System fusion

In the last experiment, we select the best and the worst system from the above systems and fuse them together. Results in table 5 indicate that the fused systems can achieve better and more stable performance compared to each best individual system.

Table 4. Experimental results of different amounts of variance preserved in the projection matrix

System	Variance Preserved in (W_{k1} ,W_{k2} ,W_{k3})	EER (%)	minDCF (×100)
NE-FSH	(80%, 99%, 99%)	4.04	**2.34**
	(85%, 99%, 99%)	4.04	2.35
	(90%, 99%, 99%)	**4.01**	2.37
	(800, 799, 550)[2]	4.12	2.27

Table 5. Fusion results between R-FSH, NE-FSH and E-FSH. EER (%), minDCF (×100)

Fusion Scheme	Fusion of the best individual system		Fusion of the worst individual system	
	EER	minDCF	EER	minDCF
R-FSH+NE-FSH	4.08	**2.20**	4.28	2.20
R-FSH+E-FSH	4.11	2.21	4.27	2.30
NE-FSH+E-FSH	4.11	2.27	4.28	2.30
R-FSH+NE-FSH +E-FSH	**4.03**	2.25	4.27	2.30

6. CONCLUSION

This paper develops five different systems to investigate the influence of different methods of speaker supervector segmentation on the speaker verification task. Experimental results on the NIST SRE08 male core task indicate that both NE-FSH and E-FSH methods improve over the previous approach using Fishervoice. Hence we can conclude that clustering similar acoustic classes together can enhance local class boundary information between Gaussian mixtures for discriminative training, leading to better and more robust performance.

7. ACKNOWLEDGEMENT

This work is affiliated with the CUHK MoE-Microsoft Key Laboratory of Human-centric Computing and Interface Technologies. Also, this work is partly supported by National Natural Science Foundation of China (61103164, 61002042), Shenzhen Basic Research Program (JC2010052 70350A, JCYJ20120903092050890, JCYJ2012061711461 4438), Foshan Research (2011BY100082) and Guangdong Innovative Research Team Program (No.201001D010464 8280). The authors are grateful to Dr. Frank Soong for helpful and informative discussions on this research topic.

[2] The combination of this retained-to-total information ratio is around (77%, 99%, 99%).

8. REFERENCES

[1] P. Kenny, “Joint factor analysis of speaker and session variability: Theory and algorithms, Tech. Report CRIM-06/08-13,” 2005.

[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, P. Dumouchel and P. Ouellet, “Front-End Factor Analysis for Speaker Verification IEEE Transactions on Audio, Speech and Language Processing, November 2009.

[4] Z. Li, W. Jiang and H.Meng “Fishervoice: a discriminant subspace framework for speaker recognition,” ICASSP2010.

[5] W. Jiang, H.Meng and Z. Li, “An enhanced Fishervoice subspace framework for text-independent speaker verification,” ISCSLP 2010.

[6] W. Jiang, Z. Li and H.Meng, “An analysis framework based on random subspace sampling for speaker verification,” Interspeech 2011.

[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. Reynolds, T. Quatieri and R. Dunn, “Speaker verification using adapted Gaussian mixture models,” Digital Signal Processing, vol. 10, no. 1-3, pp. 1941, 2000.

[9] 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.

[10] J. Pelecanos and S. Sridharan “Feature warping for robust speaker verification.” In Odyssey: The Speaker and Language Recognition Workshop, 2001.

[11] Weiwu Jiang, Helen Meng and Zhifeng Li, "An Enhanced Fishervoice Subspace Framework for Text-independent Speaker Verification," in the Proceedings of the 7th International Symposium on Chinese Spoken Language Processing (ISCSLP), 2010