THE HKCUPU SYSTEM FOR THE NIST 2010 SPEAKER RECOGNITION EVALUATION

Weiwu Jiang[1], *Man-Wai Mak*[2], *Wei Rao*[2] *and Helen Meng*[1]

[1]The Chinese University of Hong Kong, Hong Kong SAR,
[2]The Hong Kong Polytechnic University, Hong Kong SAR
{wwjiang, hmmeng}@se.cuhk.edu.hk, {enmwmak, 10901332r}@polyu.edu.hk

ABSTRACT

This paper presents the HKCUPU speaker recognition system submitted to NIST 2010 speaker recognition evaluation (SRE). The system comprises five subsystems, each with different acoustic features, session-variability reduction methods, speaker modeling and scoring methods and classifiers. This paper reports the results of individual and fusion systems for the core test and highlights the improvements made by our newly proposed JFA-Fishervoice (FSH) subsystem. Results show that FSH outperforms JFA when its projection matrix is channel-dependent (telephone or microphone) and that FSH is complementary to other state-of-the-art techniques. It was also found that VAD is an important pre-processing step for interview speech.

Index Terms— speaker recognition, factor analysis, Fishervoice, discriminative models, NIST SRE 2010.

1. INTRODUCTION

The NIST 2010 Speaker Recognition Evaluation (SRE) is part of an ongoing series of evaluations conducted by NIST [1]. This year, not only does the evaluation contain more speech materials and longer conversations, but also it comprises speech collected under different conditions (telephone conversations or interviews) and speech produced by different level of vocal efforts. This makes the evaluation this year particular challenging.

In speaker verification, Gaussian mixture model (GMM) [2] based joint factor analysis (JFA) [3] and support vector machine (SVM) [4][5] have become popular methods. More recently, studies [6] have shown that low-dimensional speaker-dependent feature vectors can be obtained from the total factors (also named i-vectors) in JFA. The i-vector of a test speaker can be classified by SVM or by comparison with the i-vector of the target speaker via cosine distance.

This paper presents the HKCUPU submission to NIST 2010 SRE. HKCUPU combines the effort of **C**hinese **U**niversity of Hong Kong and The Hong Kong **P**olytechnic **U**niversity. The paper describes the five subsystems in HKCUPU and reports the performance in terms of EER, minDCF, actual DCF and DET curves. In particular, the paper highlights a new JFA-based Fishervoice [7] approach to speaker verification. The method maps a supervector into a compressed subspace by nonparametric Fisher discriminant analysis [8], which has the effect of suppressing intra-speaker variations and emphasizing the discriminative speaker information. Another advantage of this Fishervoice framework is that it can be applied directly in the testing phase to compute the distance between a test vector and the reference vector of a target speaker. Results suggest that the Fishervoice approach outperforms JFA [3] and GMM-SVM [4] when its projection matrices are channel-dependent. The paper also demonstrates the complementarity between the Fishervoice approach and other state-of-the-art approaches.

2. SYSTEM DESCRIPTION

The HKCUPU system consists of three main modules: (1) feature extraction, (2) a parallel of five classifiers, and (3) system score fusion. We implemented different acoustic features in combination with various speaker modeling and session-variability reduction methods to maximize subsystem diversity. In total, five generative or discriminative subsystems have been built (see Table 1).

Table 1. Speaker modeling and scoring methods, acoustic features, model types, and score normalization techniques used in the HKCUPU system. (*Note*: JSV is short for JFA-supervector with linear SVM; JSF is short for JFA speaker-factor with cosine-kernel SVM; GSV is short for GMM-SVM with NAP; FSH is short for JFA speaker-factor with Fishervoice [7].)

Model	Features	Model Type	Normalization
JFA	MFCC	Generative	**TZnorm**
JSV	MFCC	Discriminative	**Tnorm**
JSF	PLP	Discriminative	**Tnorm**
GSV	MFCC	Discriminative	**Tnorm**
FSH	**PLP**	**Discriminative**	**TZnorm**

2.1. Feature Extraction

The first stage of feature extraction is voice activity detection (VAD). For telephone speech, an energy-based VAD

This work was in part supported by the HKSAR Project No. CUHK 1/02C, PolyU5264/09E and HKPolyU 4-ZZ7W.

[9] was used in the GSV subsystem, while the VAD in the ETSI Adaptive Multi-Rate (AMR) coder [10] was used for other four subsystems. For microphone and interview speech, the AMR coder either fails to detect any speech or considers the whole speech file as speech. Therefore, spectral subtraction followed by energy-based VAD was applied [9]. Note that spectral subtraction was only used for speech/non-speech segmentation; acoustic features were extracted from the speech segments of the original signals. Except for the GSV subsystem, we used the NIST10 ASR transcripts of the interviewers to remove the interviewer's speech segments that appear in the interviewee's channel (crosstalk). For the GSV subsystem, we applied VAD [9] to both interviewer and interviewee channels in order to remove the crosstalk appeared in the interviewee channel. After VAD, speech segments were converted to sequences of feature vectors using HTK. Three types of cepstral features were used and they are detailed in Table 2. All feature vectors were processed by mean-variance-normalization (MVN) followed by feature warping [11].

Table 2. Acoustic features used in HKCUPU.

Subsystem	**Features and dimension**	**Frame Size**
JFA	17 MFCC_0+Δ+ΔΔ (51Dim)	**25 ms**
JSV	17 MFCC_0+Δ+ΔΔ (51Dim)	**25ms**
JSF	12 PLP_E+Δ+ΔΔ+ΔΔΔ (52 Dim)	**20ms**
GSV	12 MFCC+Δ (24 Dim)	**25ms**
FSH	**12 PLP_E+Δ+ΔΔ+ΔΔΔ (52 Dim)**	**20ms**

2.2. Subsystem Descriptions

JFA Subsystem. The training of the JFA subsystem mainly follows [3], with the assumption that the speaker- and channel-dependent GMM supervector *M* can be expressed as the sum of four supervectors:

$$M = m + Vy + Dz + Ux \qquad (1)$$

In Eq. 1, *m* is the UBM supervector, *U* is the Eigenchannel matrix, *V* is the Eigenvoice matrix, *D* is the diagonal residual scaling matrix, *x* is the speaker-dependent Eigenchannel factor, *y* is the session- and speaker-dependent Eigenvoice factor, and *z* is the session- and speaker-dependent speaker-residual. The JFA subsystem is different from [3] in that it does not have matrix *D*. Furthermore, log-likelihood ratio (LLR) based scoring similar to [12] was used during verification. This scoring approach aims to reduce the session variation at the feature level.

In the training phase, two 2048-Gaussian gender-dependent UBMs were created by combining the mixture components of the UBM of telephone speech and microphone speech, each comprising 1024 Gaussians. We used NIST SRE04, SRE05 and SRE06 telephone data to train the telephone UBMs and used NIST SRE05 and SRE06 microphone data to train the microphone UBMs. The gender-dependent Eigenvoice matrix *V* (speaker space with rank = 300) was trained by using 893 male speakers (11,204 utterances) and 1,365 female speakers (16,556 utterances) from Switchboard II Phase 2 and Phase 3, Switchboard Cellular Parts 2, and NIST SRE04, SRE05, and SRE06. We trained 3 Eigenchannel matrixes *U*, one for each channel type. Specifically, we used (1) telephone data in NIST SRE04, SRE05, and SRE06 to train a telephone-Eigenchannel matrix with 100 channel factors; (2) microphone data in NIST SRE05 and SRE06 to train a microphone-Eigenchannel matrix with 75 channel factors; and (3) interview data in NIST SRE08 to train an interview-Eigenchannel matrix with 75 channel factors. We combined these three subspaces to obtain a full channel space of 250 channel factors. Both *U* and *V* were trained using 15 iterations of expectation maximization. For training speaker factors, we used a relevance factor of 14.

JSV Subsystem. This subsystem uses supervectors ($M' = m + Vy$) determined by JFA as feature vectors for classification by SVM [13]. Specifically, given a test utterance, one iteration of EM was applied to estimate a speaker factor *y* (taking *Ux* into account) from which a 104448-dim GMM-supervector M' was obtained. The matrices *U* and *V* are the same as those in the JFA subsystem. A special background data set was constructed by selecting utterances (including non-English) from NIST SRE04, SRE05, SRE06, SRE08, and Switchboard Cellular Parts 2 training set, which amounts to 3,000 male speakers and 3,500 female background speakers.

JSF Subsystem. This subsystem uses JFA speaker-factor *y* to construct kernels for SVM. Its training procedure is similar to that of the JFA subsystem, with the following differences: (1) 52 PLP features were used, (2) 20 iterations of EM was used to estimate *U* and *V*. Given a test utterance, a JFA-based GMM is estimated by a single iteration of EM, followed by extracting a 300-dim speaker factor *y* for SVM classification with a cosine kernel [6]. The background speaker set was identical to that of the JSV subsystem.

GSV Subsystem. We created two 512-Gaussian UBMs – one from a subset (totally 5,077 utterances) of microphone speech in NIST SRE05 and SRE06 and another one from a subset (totally 5,162 utterances) of telephone speech in NIST SRE04, SRE05, and SRE06. For each target speaker in NIST SRE10, we created two channel-dependent (microphone and telephone) speaker models by applying MAP adaptation with a relevance factor of 16, to form two 12288-dim GMM-supervectors [4]. Similarly, 300 gender- and channel-dependent background GMM-supervectors were obtained. NAPs with 64 co-ranks for telephone speech and 128 co-ranks for microphone/interview speech were applied to the GMM-supervectors. The gender-dependent projection matrix for telephone speech was obtained from 517 male and 934 female speakers from NIST SRE04, SRE05, SRE06 and SRE08. For microphone/interview speech, 143 male and 178 female speakers were selected from NIST SRE05, SRE06 and SRE08.

FSH Subsystem. This subsystem extends Fishervoice [7] to speaker verification. Specifically, it uses JFA speaker factors y as feature vectors to estimate a nonparametric Fisher discriminant projection matrix W [8]:

$$W = W_1 \; W_2 \; W_3 \qquad (2)$$

Given a test utterance, the matrix W is used to project the corresponding JFA speaker-factor vector y to a low-dimensional discriminant subspace that better represents speaker characteristics. Then direct cosine distance is calculated to obtain a trial score. In (2), W is a combination of three subspace projection matrixes: PCA projection matrix W_1, whitened within-class projection matrix W_2, and nonparametric between-class projection matrix W_3. Matrices W_2 and W_3 aim to minimize the distance between the projected vectors of the same speaker while maximizing the distance among different speakers. Unlike classical LDA, it is not necessary to use parametric models to approximate the distribution of y. This characteristic leads to a W_3 that focuses on the boundaries between speakers. As will be shown in Section 4, focusing on the boundary allows us to exploit the discriminative features of speakers, leading to better verification performance.

During training, telephone utterances from NIST SRE04, SRE05 and SRE06 were used to train the gender-dependent Fishervoice matrices (W_1, W_2 and W_3). This amounts to 400 male and 400 female speakers, each has 8 different utterances. The projection matrices, W_1, W_2 and W_3, have dimensions 300×299, 299×298, and 298×295, respectively. These correspond to the upper limit of their matrix ranks. The parameter R in [7] was set to 4, according to the median number of sessions for each speaker. The JFA parameters U and V are the same as the JSF subsystem.

2.3. Score Normalization

The scores of the JFA and Fishervoice subsystems were normalized by TZnorm, whereas the scores of subsystems based on SVM were normalized by Tnorm. For the JFA, JSV, JSF and FSH subsystems, NIST SRE04, SRE05 and SRE06 training data was used for training cohort models for Tnorm (300 speakers for each gender). For the GSV subsystem, 261 male utterances and 277 female utterances from NIST SRE05 were used to create the GMM-SVM telephone Tnorm models; for microphone/interview speech, 300 male and 300 female utterances extracted from NIST SRE05 and SRE06 were used. For the Znorm, we used 800 speakers for each gender from the Switchboard II Phase 2, Phase 3 and Switchboard Cellular Parts 2 training data.

3. SYSTEM FUSION

The scores obtained from five different systems, including JFA, JSV, JSF, GSV, FSH were fused using a set of linear fusion weights that achieve the best fusion performance (in terms of minimum DCF) in NIST SRE10. A 5-dimensional grid search was performed to determine the fusion weights.

4. RESULTS

Table 3 shows the fusion of two subsystems under common conditions 5, 6 and 8 (cc5, cc6, and cc8). These common conditions involve telephone speech only. Here, the focus is on the non-interview data conditions because the FSH subsystem was trained on telephone data only. We chose the best four subsystems for fusion, which amount to six fusion systems. The results show that in 4 out of 6 cases, the fusion systems involving FSH achieve either the lowest EER or the lowest minDCF, suggesting that FSH is complementary to JFA, GSV, and JSV. Fig. 1 provides further evidences on this complementarity property. Fig. 1 shows the DET curves of JFA, GSV, FSH, and the fusion of FSH with either JFA or GSV under common conditions 6 and 8. Evidently, when FSH is fused with either JFA or GSV, significant performance gain can be achieved across a wide range of decision thresholds.

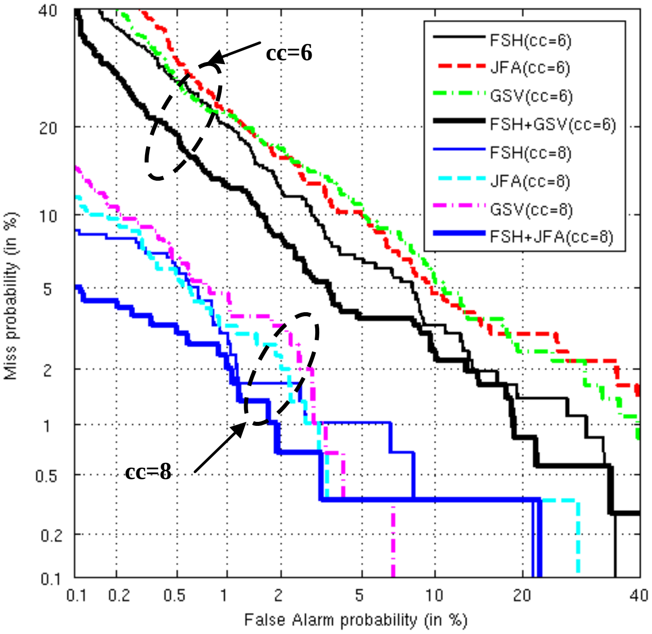

Figure 1. DET curves of JFA, GSV, FSH, and the fusion of FSH with another best performing subsystem under common conditions 6 and 8.

Table 4. Performance of individual subsystems and the fusion of 5 subsystems. For each common condition, the first line represents EER (%) and the second line represents minDCFx1000. The relative improvement is the improvement of the fusion system over the best individual system. As a reference for comparison, the actual DCF of the fusion system under CC5 is 0.585.

Common Condition	JFA	JSV	JSF	GSV	FSH	Fusion	Relative Improvement
cc=1	**3.88** **0.628**	4.55 0.725	7.51 0.825	4.40 0.669	7.10 0.734	2.69 0.408	31% 35%
cc=2	8.04 0.843	9.11 0.813	13.55 0.893	**7.39** **0.774**	11.32 0.857	5.70 0.566	23% 27%
cc=3	**4.53** **0.665**	7.59 0.792	11.32 0.893	6.28 0.696	8.26 0.896	2.94 0.509	35% 23%
cc=4	5.79 0.760	7.14 0.681	11.11 0.831	**5.58** **0.678**	6.96 0.853	3.59 0.545	36% 20%
cc=5	4.52 **0.467**	5.73 0.646	5.77 0.603	4.76 0.574	**4.09** 0.539	2.36 0.385	42% 18%
cc=6	7.17 0.819	8.31 0.843	9.14 0.830	7.75 **0.786**	**6.09** 0.807	3.87 0.675	36% 14%
cc=7	**7.52** **0.740**	8.79 0.905	8.63 0.775	9.15 0.747	8.35 0.852	5.00 0.548	34% 26%
cc=8	2.01 0.457	2.68 0.549	3.69 0.443	2.57 0.475	**1.68** **0.284**	1.00 0.215	40% 24%
cc=9	3.45 0.395	**2.76** 0.464	3.79 0.406	4.13 **0.390**	4.14 0.494	1.69 0.172	39% 56%

Table 3. EER and (minDCFx1000) of the fusion of best performing subsystems.

Fusion System	cc5	cc6	cc8
JFA + JSV	3.94 (0.50)	6.64 (0.78)	1.89 (0.42)
JFA + GSV	3.49 (**0.37**)	5.49 (0.72)	**1.23** (0.23)
FSH + JFA	3.21 (0.46)	5.26 (0.78)	1.34 (**0.17**)
FSH + GSV	**2.93** (0.45)	**4.06** (0.73)	1.34 (0.26)
FSH + JSV	3.21 (0.49)	4.71 (**0.69**)	1.34 (0.20)
GSV + JSV	3.60 (0.41)	4.97 (0.69)	1.34 (0.40)

Table 4 shows the performance of the 5 subsystems and their fusions. The results show that fusion of five subsystems reduces both EER and minDCF of individual subsystems significantly. In particular, FSH shows superior performance for all conditions (cc5, cc6, and cc8) that involve telephone speech only. On the other hand, FSH shows no improvement on microphone/interview or cross-channel conditions because its projection matrices were trained by telephone data only.

5. DISCUSSIONS AND CONCLUSIONS

The HKCUPU system submitted to NIST 2010 SRE is composed of 5 subsystems. Different acoustic features, speaker modeling techniques, session-variability reduction methods, and VAD schemes have been used for individual systems. This strategy has led to a significant performance gain when the subsystems were fused. Specifically, the fusion system reduces the EER by 42% and minDCF by 56% when compared with the best individual subsystems. It was also found that the newly proposed FSH subsystem is complementary to JFA and performs significantly better than JFA when its projection matrices were trained by the type of speech that matches the evaluation conditions.

7. REFERENCES

[1] http://www.itl.nist.gov/iad/mig//tests/sre/2010/index.html

[2] D. Reynolds, T. Quatieri and R. Dunn, “Speaker Verification using Adapted Gaussian Mixture Models,” Digita Signal Processing, 2000.

[3] 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.

[4] W. M. Campbell and D. E. Sturim and D. A. Reynolds, "Support Vector Machines Using GMM Supervectors for Speaker Verification", IEEE Signal Processing, 2006.

[5] 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,” Proceedings of ICASSP 2009.

[6] 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.

[7] Z. Li, W. Jiang and H.Meng “Fishervoice: A Discriminant Speaker Recognition,” ICASSP2010.

[8] Z. Li, D. Lin, X. Tang, "Nonparametric Discriminant Analysis for Face Recognition," *PAMI*, 2009.

[9] M.W. Mak and H.B. Yu, "Robust Voice Activity Detection for Interview Speech in NIST Speaker Recognition Evaluation, "Proc. APSIPA ASC 2010, Singapore.

[10] 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.

[11] J. Pelecanos and S. Sridharan, .Feature Warping for Robust Speaker Verification,. in *Proc. A Speaker Odyssey*, 2001.

[12] D. Matrouf, N. Scheffer, B. Fauve, "A Straightforward and Efficient Implementation of the Factor Analysis Model for Speaker Verification", Interspeech, 2007.

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