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]Dept. of System Engineering and Engineering Management, The Chinese University of Hong Kong

[2]Dept. of Electronic and Information Engineering, The Hong Kong Polytechnic University

Introduction

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.

Subsystems

Sub-System	Description
JFA	Joint Factor Analysis
JSV	Linear SVM on JFA-GMM supervectors
JSF	Cosine-kernel SVM on JFA speaker factors
GSV	Linear SVM on GMM-supervectors with NAP
FSH	Cosine distance on Fishervoice-projected speaker factors

Voice Activity Detection

Type of Speech	VAD Method
Telephone	• Energy-based VAD for GSV subsystem • ETSI Adaptive Multi-Rate (AMR) VAD for other four subsystems
Microphone and Interview	• Spectral subtraction followed by energy-based VAD with crosstalk removal

Acoustic Features

Subsystem	Features and Dimension	Frame Size
JFA	17 MFCC_0+Δ+ΔΔ (51Dim)	25ms
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

- GSV: Cepstral mean normalization followed by feature warping.
- Others: Mean-variance-normalization followed by feature warping.

Subsystem Description

JFA Subsystem:

- 2048-Gaussian GMM-supervector of speaker s:

$$\boldsymbol{m}(s) = \boldsymbol{m} + \boldsymbol{V}\boldsymbol{y}(s) + \boldsymbol{U}\boldsymbol{x}(s, h_s)$$

where $\boldsymbol{m}$ is the UBM supervector, $\boldsymbol{U}$ is the Eigenchannel matrix, $\boldsymbol{V}$ is the Eigenvoice matrix, $\boldsymbol{x}(s,h)$ is the speaker-dependent Eigenchannel factor, $\boldsymbol{y}(s)$ is the session- and speaker-dependent Eigenvoice factor, and h_s represents the enrollment channel.

JSV Subsystem:

- Use GMM-supervectors of target speaker s and claimant c obtained from JFA as feature vectors:

$$\boldsymbol{m}(s) = \boldsymbol{m} + \boldsymbol{V}\boldsymbol{y}(s)$$
$$\boldsymbol{m}(c) = \boldsymbol{m} + \boldsymbol{V}\boldsymbol{y}(c)$$

- Linear SVM Scoring:

$$S_{\mathrm{SVM}}(s,c) = \alpha_{s,0}\langle \boldsymbol{m}(s), \boldsymbol{m}(c)\rangle - \sum_{i \in \mathrm{SV}^-} \alpha_{s,i}\langle \boldsymbol{m}(b_i), \boldsymbol{m}(c)\rangle + d_s$$

where b_i is the i-th background speaker and SV^- contains the support vector indexes of background speakers.

JSF Subsystem:

- Use speaker-factors of target speaker s and claimant c obtained from JFA as feature vectors:

$$\boldsymbol{m}(s) = \boldsymbol{m} + \boldsymbol{V}\boldsymbol{y}(s) + \boldsymbol{U}\boldsymbol{x}(s, h_s)$$
$$\boldsymbol{m}(c) = \boldsymbol{m} + \boldsymbol{V}\boldsymbol{y}(c) + \boldsymbol{U}\boldsymbol{x}(c, h_c)$$

- Cosine-kernel SVM Scoring:

$$S_{\mathrm{CSVM}}(s,c) = \alpha_{s,0}K(\boldsymbol{y}(s), \boldsymbol{y}(c)) - \sum_{i \in \mathrm{SV}^-} \alpha_{s,i}K(\boldsymbol{y}(b_i), \boldsymbol{y}(c)) + d_s$$

GSV Subsystem:

- Use MAP adapted 512-Gaussian GMM-supervectors
- Transformed by NAP matrices (Corank = 16 for tel; Corank = 128 for mic/interview)
- Linear SVM Scoring with T-norm

FSH Subsystem:

- Apply JFA speaker factors $\boldsymbol{y}(s)$ as feature vectors to estimate a nonparametric Fisher discriminant projection matrix $\boldsymbol{W}$:

$$\boldsymbol{W} = \boldsymbol{W}_1\, \boldsymbol{W}_2\, \boldsymbol{W}_3$$

where $\boldsymbol{W}_1$ is PCA projection matrix, $\boldsymbol{W}_2$ is whitened within-class projection matrix, and $\boldsymbol{W}_3$ is nonparametric between-class projection matrix. $\boldsymbol{W}_3$ focuses on the boundaries between speakers without using parametric models to approximate the distribution of $\boldsymbol{y}(s)$.

$$\boldsymbol{W}_1 = \arg\max_{\boldsymbol{W}} \left\| \boldsymbol{W}^T \boldsymbol{\Sigma} \boldsymbol{W} \right\| \qquad \boldsymbol{W}_3 = \arg\max_{\boldsymbol{W}} \left\| \boldsymbol{W}^T \boldsymbol{S}_b \boldsymbol{W} \right\|$$

$$\boldsymbol{W}_2^T \boldsymbol{S}_w \boldsymbol{W}_2 = \boldsymbol{I} \quad \Rightarrow \quad \boldsymbol{W}_2 = \boldsymbol{\Phi}\boldsymbol{\Lambda}^{-\frac{1}{2}}$$

- Cosine distance scoring:

$$S_{\mathrm{FSH}}(s,c) = \frac{\langle \boldsymbol{W}\boldsymbol{y}(s), \boldsymbol{W}\boldsymbol{y}(c)\rangle}{\left\|\boldsymbol{W}\boldsymbol{y}(s)\right\| \left\|\boldsymbol{W}\boldsymbol{y}(c)\right\|}$$

Fusion System	CC5	CC6	CC8
JFA+JSV	3.94%	6.64%	1.89%
JFA+JSV	3.49%	5.49%	**1.23%**
FSH+JFA	3.21%	5.26%	1.34%
FSH+GSV	**2.93%**	**4.06%**	1.34%
FSH+JSV	3.21%	4.71%	1.34%
GSV+JSV	3.60%	4.97%	1.34%

EER in % (1st line) and Minimum DCF (2nd line) for CC1 to CC9.

Com. Cond.	JFA	JSV	JSF	GSV	FSH	Fusion	Rel. Imp.
1	**3.88** **0.63**	4.55 0.73	7.51 0.83	4.40 0.67	7.10 0.73	2.69 0.41	31% 35%
2	8.04 0.84	9.11 0.81	13.55 0.89	**7.39** **0.77**	11.32 0.86	5.70 0.57	23% 27%
3	**4.53** **0.67**	7.59 0.79	11.32 0.89	6.28 0.70	8.26 0.90	2.94 0.51	35% 23%
4	5.79 0.76	7.14 0.68	11.11 0.83	**5.58** **0.68**	6.96 0.85	3.59 0.54	36% 20%
5	4.52 **0.47**	5.73 0.65	5.77 0.60	4.76 0.57	**4.09** 0.54	2.36 0.39	42% 18%
6	7.17 0.82	8.31 0.84	9.14 0.83	7.75 **0.79**	**6.09** 0.81	3.87 0.68	36% 14%
7	**7.52** **0.74**	8.79 0.91	8.63 0.78	9.15 0.75	8.35 0.85	5.00 0.55	34% 26%
8	2.01 0.46	2.68 0.55	3.69 0.44	2.57 0.48	**1.68** **0.28**	1.00 0.22	40% 24%
9	3.45 0.40	**2.76** 0.46	3.79 0.41	4.13 **0.39**	4.14 0.49	1.69 0.17	39% 56%

Training Data

UBM:

- Tel – NIST04, 05, 06 tel speech
- Mic – NIST 05, 06 mic speech

JFA:

- Matrix $\boldsymbol{V}$ – NIST04, 05, 06, Switchboard Phase2, Phase 3, Cellular Parts 2 (300 speaker factors)
- Tel Matrix $\boldsymbol{U}$ – NIST04, 05, 06 tel speech (100 Ch factors)
- Mic Matrix $\boldsymbol{U}$ – NIST05, 06 mic speech (75 Ch factors)
- Interview Matrix $\boldsymbol{U}$ – NIST08 interview speech (75 Ch factors)
- Totally rank of $\boldsymbol{U}$ = 100 tel +75 mic +75 interview

Fishervoice:

- Projection matrix $\boldsymbol{W}$ – NIST04, 05, 06 tel speech (400 gender-dependent speakers with 8 different utterances)

NAP:

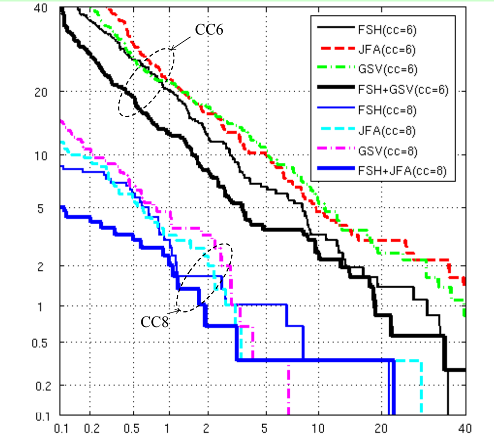
- Tel – NIST04, 05, 06 tel speech (Corank = 16)
- Mic/interview – NIST05, 06 mic speech, NIST08 interview speech (Corank = 128)

SVM Imposter-Class:

- JSV and JSF – NIST04, 05, 06, 08, Switchboard Cellular Parts 2
- GSV – NIST 06 tel speech and NIST 06 mic speech

Score normalization:

- Tnorm for JSV, JSF and GSV – NIST04-06 tel/mic speech
- TZnorm for JFA and FSH – NIST04-06 tel/mic speech



Conclusions

The HKCUPU system submitted to NIST 2010 SRE is composed of 5 subsystems. 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.