Joint Submission of CUHK and PolyU for NIST SRE 2010

Weiwu Jiang[1], Man-Wai Mak[2], Helen Meng[1]

[1]The Chinese University of Hong Kong, Hong Kong SAR of China

[2]The Hong Kong Polytechnic University , Hong Kong SAR of China

General Overview

Submissions overview

Submission	Sub-system				
	JFA	JSV	JSF	FSH	GSV
hkcupu1 (all systems)	√	√	√	√	√
hkcupu2* (best 1 system)				(√)	(√)
hkcupu3** (best 3 systems)	(√)	(√)		√	√

Descriptions of sub-systems

1.JFA: **Joint Factor Analysis** system (*M=m+Vy* without *D*) is similar to idea proposed by Kenny [1]. But during testing, the estimated channel noise is subtracted at the feature level and scoring based on the log-likelihood ratio (LLR) is applied. **JFA is only for female phonecall-tel test condition

2.JSV: apply **JFA supervectors** (*M = m + Vy*) to construct linear kernels of SVM by using LibSVM

3.JSF: apply **JFA speaker factor** *y* to construct cosine kernel of *SVM*

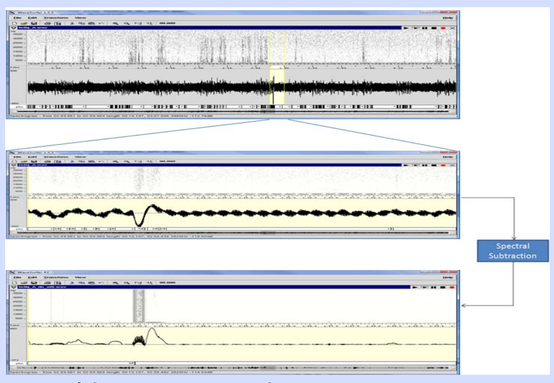
4.FSH: uses **JFA speaker factor** *y* as input vector to create a **Fisher's discriminative projection matrix** [2]. Each target speaker's factor y is projected by this matrix and regarded as accorded target model. Similar to the training stage, each test utterance is also projected via the Fisher's discriminative projection matrix mentioned above. Direct cosine distance is calculated as each trial score. * FSH is only for phonecall-tel test condition

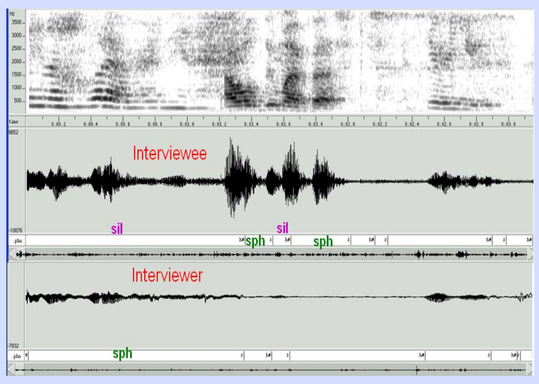
5.GSV: use MAP adapted **GMM mean supervectors** and nuisance attribute projection (NAP) [3]

Front-end Processing

Silence removal

- Phonecall tel speech: ETSI AMR VAD [4]
- Phonecall mic speech: spectral subtraction followed by enhanced energy-based VAD
- Interview speech: use speech from both interviewer and interviewee





Cepstral features extraction

system	feature	frame	mixture
JFA	17 MFCC_0 +Δ+ ΔΔ (51 Dim)	25ms	1024tel+1024mic
JSV	17 MFCC_0 +Δ+ ΔΔ (51 Dim)	25ms	1024tel+1024mic
JSF	12 PLP +Δ+ ΔE+ΔΔ+ΔΔE+ΔΔΔ+ΔΔΔE (52 Dim)	20ms	1024tel+1024mic
FSH	12 PLP + Δ+ ΔE+ΔΔ+ΔΔE+ΔΔΔ+ΔΔΔE (52 Dim)	20ms	1024tel+1024mic
GSV	12 MFCC + Δ (24 Dim, channel-dependent)	25ms	512

Feature Specification/Method

- CMN followed by Gaussian Feature Warping
- Use HTK tools for feature extraction

System Configuration

Data for UBM

- Tel_UBM − NIST04, 05, 06 tel speech Mic_UBM − NIST 05, 06 mic speech

Data for JFA

- Matrix V − NIST04, 05, 06, Switchboard Phase2, Phase 3, Cellular Parts 2 (300 rank)
- Tel Matrix U − NIST04, 05, 06 tel speech (100 rank)
- Mic Matrix U − NIST05. 06 mic speech (75rank)
- Interview Matrix U − NIST08 interview speech (75 rank)

(Totally rank of U = 100 tel +75 mic +75 interview)

Data for Fishervoice

- Projection matrix − NIST04, 05, 06 tel speech (400 gender-dependent speakers, where each one contains 8 different utterances) *

Data for NAP

- Tel_NAP − NIST04, 05, 06 tel speech (Corank = 16)
- Mic/interview_NAP − NIST05, 06 mic speech, NIST08 interview speech (Corank = 128)

Data for SVM background

- JSV,JSF − NIST04, 05, 06, 08, Switchboard Cellular Parts 2
- GSV − NIST 05 tel speech for tel positive-class, − NIST 05, 06 mic speech for mic positive-class

Data for score normalization

- Tnorm for JSV, JSF, GSV − NIST04, 05, 06tel speech
- TZnorm for JFA, FSH − NIST04, 05, 06 tel speech (Tnorm), Switchboard Phase2, Phase 3, Cellular Parts 2 (Znorm)

* Ranks of the 3 projection steps are 299, 298, 295 respectively

Results on NIST SRE 2008

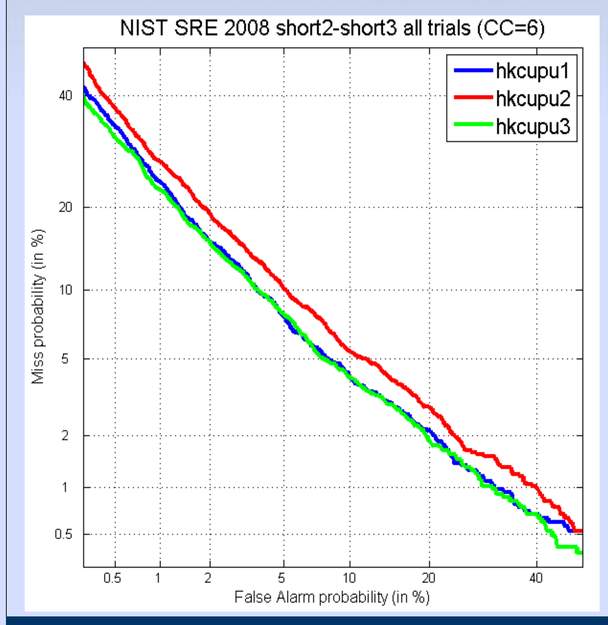
Results on Short2-Short3 (CC=6) : EER and MinDCF

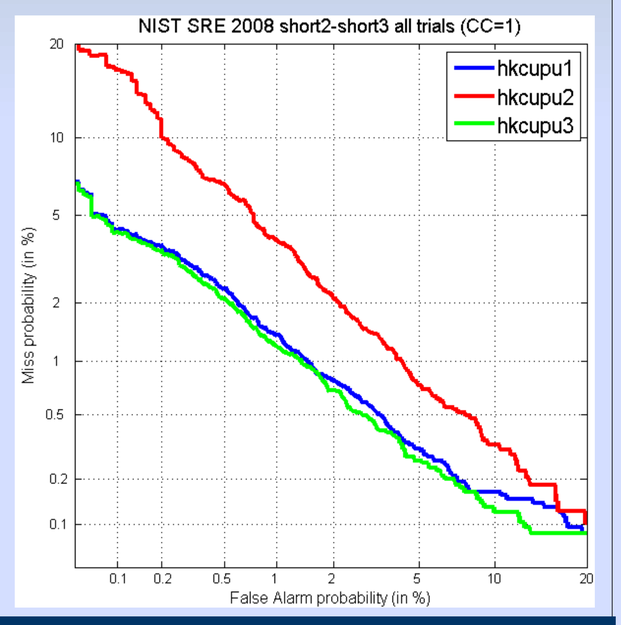
sub-system	male		female		all trials	
	EER(%)	MinDCF	EER(%)	MinDCF	EER(%)	MinDCF
JFA	6.86	0.0337	8.40	0.0446	7.80	0.0411
JSV	7.49	0.0385	8.92	0.0403	8.47	0.0401
JSF	8.81	0.0405	10.02	0.0451	9.56	0.0437
FSH	6.74	0.0354	7.87	0.0374	7.54	0.0370
GSV	6.84	0.0332	9.70	0.0460	8.81	0.0419

Results on Short2-Short3 (CC=1) : EER and MinDCF

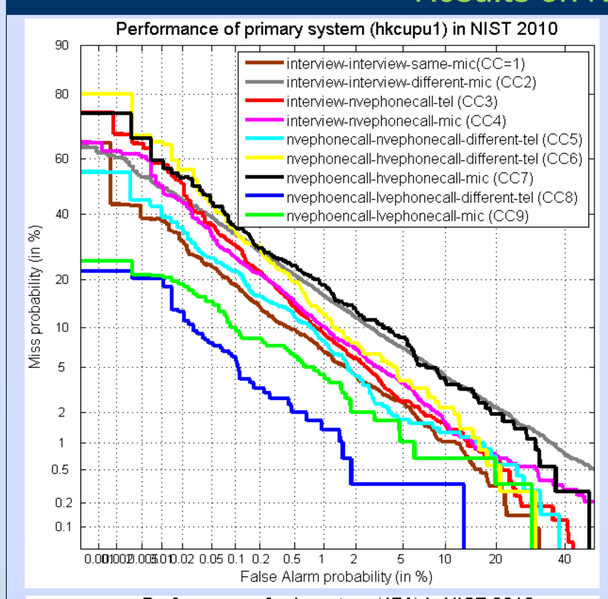
sub-system	male		female		all trials	
	EER(%)	MinDCF	EER(%)	MinDCF	EER(%)	MinDCF
JFA	4.13	0.0207	7.49	0.0434	6.37	0.0342
JSV	2.33	0.099	3.48	0.0174	3.57	0.0179
JSF	3.40	0.0135	4.91	0.0220	4.29	0.0186
FSH	2.78	0.0131	3.44	0.0158	3.15	0.0147
GSV	2.07	0.0107	2.24	0.0114	2.14	0.0113

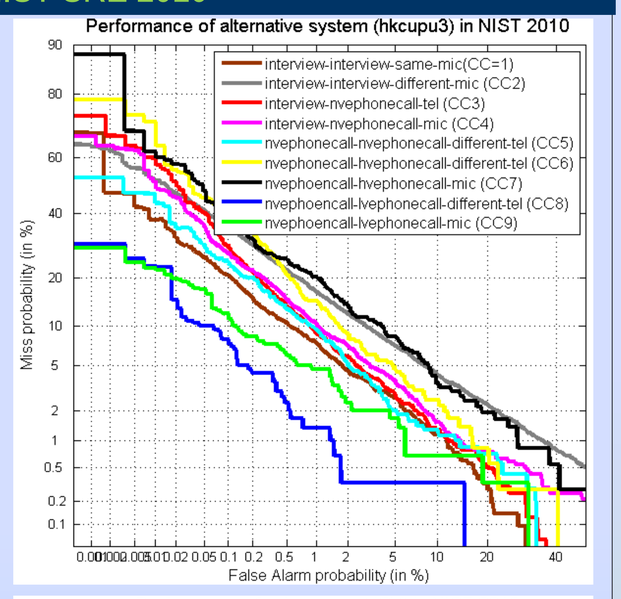
Fusion results on NIST SRE 2008: DET curves

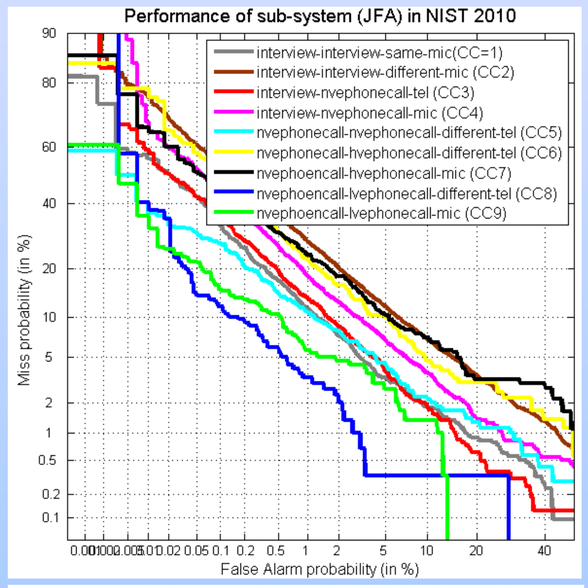


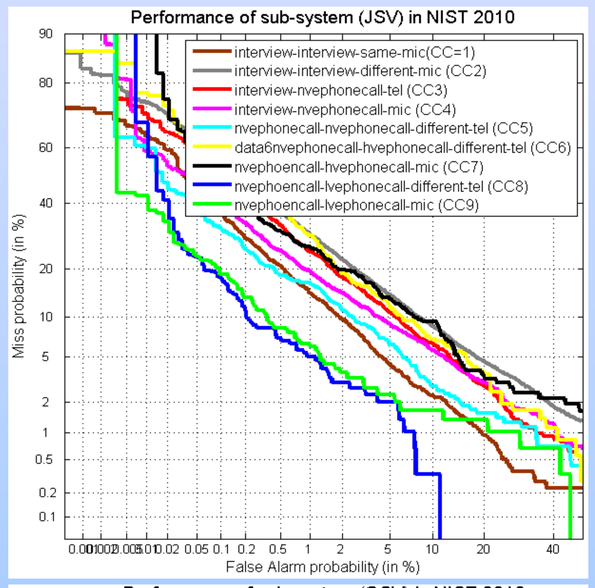


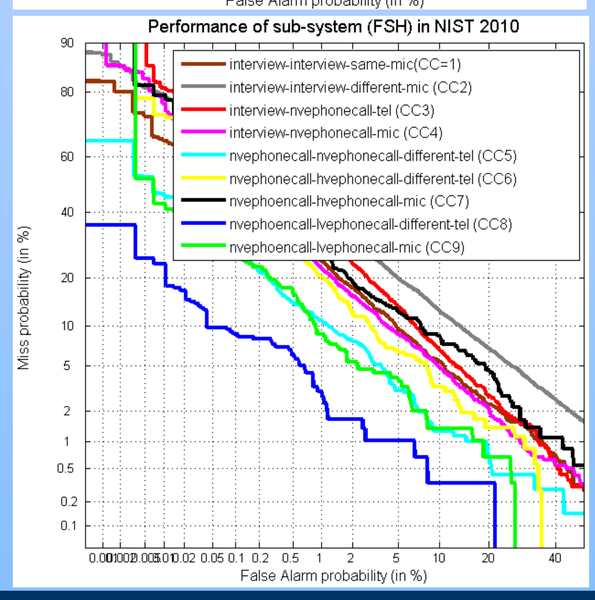
Results on NIST SRE 2010

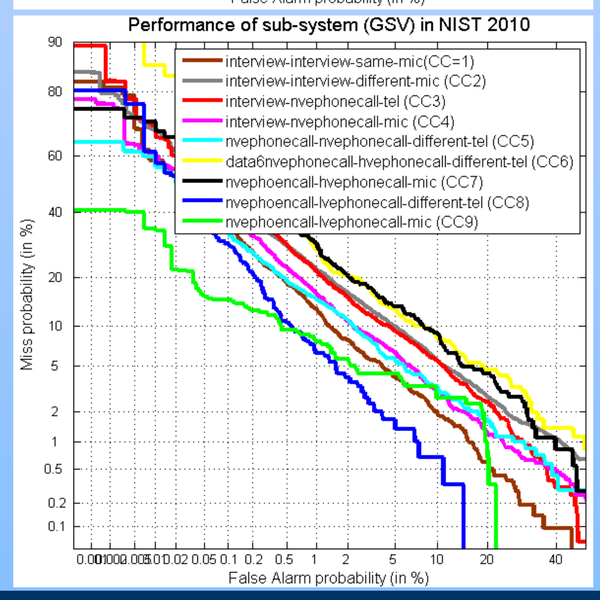












Conclusions

1. Results show that fusion with side information works reasonably well. FSH, JFA, JSV and GSV sub-systems provide significant contribution in overall performance.
2. FSH shows good performance for the tel condition but poor performance for the mic/interview conditions (where training was based on tel speech). This suggests that the distribution of the speakers' tel supervector *M* processed by eigenchannel is differs significantly (possibly due to channel information) from the ones needed for the mic/interview.

Reference

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

[2] Z. Li, W. Jiang and H. Meng "FISHERVIOCE: A discriminant subspace framework for speaker recognition," ICASSP 2010

[3] 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.

[4] 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, 1999.