The CUHK Systems for NIST SRE 2016

Jinghua Zhong and Helen Meng

Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong

General Overview

➢ Submission Overview

Submission	Sub-system			
	GMM i-vector		DNN i-vector	
	PLDA-LLR	PLDA-SVM	PLDA-LLR	PLDA-SVM
Primary	√	√	√	√
Contrastive 1	√	√		
Contrastive 2			√	√

- Descriptions of sub-systems
  - PLDA-LLR: Gaussian based PLDA with log-likelihood ratio scoring
  - PLDA-SVM: Gaussian based PLDA with SVM scoring
- Score calibration and fusion: Bosaris toolkit

Feature Extraction

➢ Acoustic feature for speaker modeling
- 60-dimensional MFCC features: the first 19 Mel frequency cepstral coefficients and log energy, together with their first and second derivatives
- The frame length was 25ms
- The energy-based voice-activity detection (VAD) and sliding-window cepstral mean and variance normalization (CMVN)

➢ Feature vector for the DNN
- 40-dimensional MFCC features without cepstral truncation
- The frame length was 25ms
- The sliding-window cepstral mean and variance normalization (CMVN)

Classification

➢ GMM i-vector
- Training sets:
  - In-domain: Call My Net Speech Collection, 2,472 utterances with no speaker label
  - Out-of-domain: select from Switchboard II Phase 2, NIST SRE2004-2012, total 57,548 utterances from 4,583 speakers
- GMM i-vector
  - UBM: in-domain training set (2048 mixtures)
  - i-vector extractor: out-of-domain training set (600 rank)
- PLDA-LLR scoring :
  - LDA and PLDA with whitening and length normalization: out-of-domain training set
  - Rank of LDA: 300
- PLDA-SVM scoring :
  - PLDA with whitening and length normalization: out-of-domain training set
  - SVM: both in-domain and out-of-domain training set using LIBSVM

➢ DNN i-vector
- Training sets:
  - In-domain: Call My Net Speech Collection, 2,472 utterances with no speaker label
  - Out-of-domain: a subset of the out-of-domain training set in GMM i-vector part, total 31,882 utterances from 4,231 speakers
- DNN i-vector
  - DNN: multisplice time delay DNN using Kaldi toolkit
    - ✓ About 100 hours speech from Fisher data set
    - ✓ A 6-hidden-layer p-norm neural networks with power p=2
    - ✓ P-norm input/output dimensions: 3500/350
    - ✓ A narrow temporal context of only 2 frames before and after: 200 input nodes
    - ✓ The softmax output layer: 3,820 senones
  - The ancillary UBM and i-vector extractor: out-of-domain training set (400 rank)
- PLDA-LLR scoring :
  - LDA and PLDA with whitening and length normalization: out-of-domain training set
  - Rank of LDA: 300
- PLDA-SVM scoring :
  - PLDA with whitening and length normalization: out-of-domain training set
  - SVM: both in-domain and out-of-domain training set using LIBSVM

Results on NIST SRE 2016 Development Set

➢ The performance comparison of four sub-systems

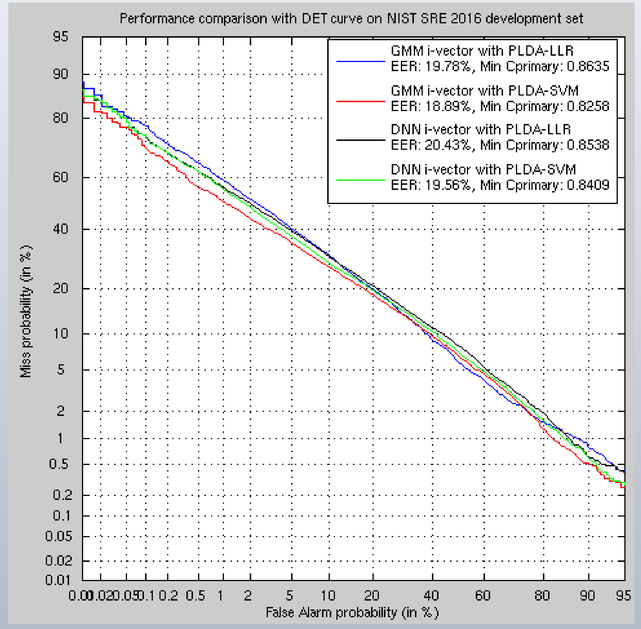
Sub-system		EER (%)	Min $C_{primary}$
GMM i-vector	**PLDA-LLR**	19.78	0.8635
	PLDA-SVM	**18.89**	**0.8258**
DNN i-vector	**PLDA-LLR**	20.43	0.8538
	PLDA-SVM	**19.56**	**0.8409**

➢ Observations:
- SVM scoring improved the performance of PLDA with log-likelihood scoring, because SVM model training could make use of the unlabeled in-domain training set
- DNN i-vector based model performed worse than GMM i-vector based model, possibly because of insufficient training data for DNN (only about 100h data)

➢ Performance of CUHK systems on development set

Submission	EER (%)	Min $C_{primary}$	Act $C_{primary}$
Unequalized, Primary	17.58	0.7955	0.807348
Unequalized, Contrastive 1	17.75	0.8000	0.799995
Unequalized, Contrastive 2	17.69	0.7879	0.795878
Equalized, Primary	17.29	0.8098	0.811102
Equalized, Contrastive 1	17.53	0.8079	0.829106
Equalized, Contrastive 2	17.66	0.8119	0.815431

➢ Performance comparison of four sub-systems on DET curve



Results on NIST SRE 2016 Evaluation Set

➢ Performance of CUHK systems on evaluation set

Submission	EER (%)	Min $C_{primary}$	Act $C_{primary}$
Unequalized, Primary	12.92	0.8121	0.873583
Unequalized, Contrastive 1	12.86	0.8128	0.907490
Unequalized, Contrastive 2	12.80	0.8030	0.879793
Equalized, Primary	13.46	0.8094	0.870625
Equalized, Contrastive 1	13.48	0.8121	0.942227
Equalized, Contrastive 2	13.44	0.8035	0.879927

➢ Equalized performance of primary system based on different catalogues

Catalogue		EER (%)	Min $C_{primary}$	Act $C_{primary}$
Gender	**Male**	13.56	0.8021	0.870796
	Female	13.37	0.8137	0.870455
Language	**Tagalog**	18.28	0.8743	1.079041
	Cantonese	**8.77**	**0.6500**	**0.662210**
Number of enrollments segments	**1**	15.98	0.8544	0.872697
	3	**11.03**	**0.6865**	**0.868553**
Phone number for enrollment and test	**Same**	**11.69**	**0.7558**	**0.811491**
	Different	17.51	0.9381	1.016879

➢ Observations:
- Language, enrollment segments No. and the phone difference highly influenced the performance
- Cantonese trials performed much better than Tagalog trials
  - Mandarin trials performed much better than Cebuano trials in the development set