POLYU SUBMISSION OF NIST 2016 SPEAKER RECOGNITION EVALUATION

Man-Wai MAK and Weiwei LIN

Dept. of Electronic and Information Engineering
The Hong Kong Polytechnic University, Hong Kong SAR

enmwmak@polyu.edu.hk

ABSTRACT

This report describes the details of the submission of NIST 2016 Speaker Recognition Evaluation by The Hong Kong Polytechnic University

Index Terms— Speaker verification; i-vectors; probabilistic LDA; domain adaptation; spectral clustering.

1. SYSTEM DESCRIPTION

1.1. Acoustic Features

Speech regions in the speech files were extracted by using a two-channel VAD [1]. For each speech frame, 19 MFCCs together with energy plus their 1st and 2nd derivatives were computed, followed by cepstral mean normalization and feature warping [2] with a window size of 3 seconds. A 60-dim acoustic vector was extracted every 10ms, using a Hamming window of 25ms.

1.2. I-vector Extraction and PLDA Model Training

The i-vector/PLDA system is based on a gender-independent UBM with 512 mixtures and a gender-independent total variability matrix with 300 total factors. Non-test utterances from CallMyNet development data were used for training the UBM and total variability (TV) matrix. The TV matrix and UBM were used for extracting i-vectors from the speech files (both gender) in Switchboard-2 Phase I to Phase III, Switchboard Cellular Part 1 and Part II, and NIST 2004–2010 SREs. Utterances with bad recordings (e.g., without speech or contain tone only) as detected by the VAD and utterances with speech frames less than 10s were excluded. Speaker-to-utterance mappings were determined from the key files of these corpora, with the identical speaker IDs across multiple speech corpora considered to be the same speakers. Speakers with less than 4 speech segments were excluded. This amounts to 66,505 speech segments (i-vectors) spoken by 4,959 speakers.

This work was in part supported by The RGC of Hong Kong SAR, Grant No. PolyU 152068/15E.

Following [3], within-class covariance normalization (WCCN) [4] and i-vector length normalization [5] were applied to the 300-dimensional i-vectors. Then, linear discriminant analysis (LDA) [6] and WCCN were applied to reduce the dimension to 200 before training an unadapted gender-independent PLDA model with 200 latent variables.

1.3. Domain Adapation

To make the PLDA model amenable to CallMyNet data, the following domain adaptation procedures were applied. First, pairwise PLDA scores of non-test utterances in the CallMyNet development set were computed. Then, spectral clustering [7] was applied to the resulting pairwise scoring matrix to cluster the i-vectors in CallMyNet into 300 clusters.[1] The i-vectors of these 300 hypothesized speakers were than added to the pool of training i-vectors to retrain the PLDA model.

The following whitening step was also applied to make the i-vectors of target-speakers and test utterances better reflecting the acoustic characteristics of CallMyNet data. Specifically, the mean of the non-test CallMyNet i-vectors was subtracted from each of the target-speakers' and test i-vectors before applying i-vector pre-processing (WCCN whitening, length-normalization, and LDA-WCCN projection).

1.4. PLDA Scoring and Score Normalization

According to the evaluation protocol, for each evaluation trial, a test utterance was tested against the target-speaker's i-vectors representing the Model ID of that trial, which produces one or multiple PLDA scores. The average of these scores is considered as the trial score.

To reduce the actual DCF, the PLDA scores were normalized by T-norm and S-norm [8]. 400 utterances (i-vectors) from the non-test segments in CallMyNet were used for T-

[1] While the number of speakers in the development data of CallMyNet is much smaller than this value, we found that performance is better if we set this value higher than the actual number of speakers.

norm, and another 400 were used as impostor utterances for Z-norm.

2. PERFORMANCE AND COMPUTATION TIME

Table 2 shows the performance (in terms of EER, minDCF and actual DCF) of three systems in the development set of SRE16. In the table, Sys A, Sys B, and Sys C represent the system without score normalization, with T-norm and with S-norm, respectively.

Table 3 shows the CPU time and memory requirements for computing the score of one verification trial. Tasks 1–3 were implemented in C and Tasks 4–6 were implemented in Matlab. The memory consumption in the Matlab tasks includes the memory of the Matlab shell without the GUI.

3. REFERENCES

[1] M. W. Mak and H. B. Yu, "A study of voice activity detection techniques for NIST speaker recognition evaluations," *Computer, Speech and Language*, vol. 28, no. 1, pp. 295–313, Jan 2014.

[2] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, Crete, Greece, Jun. 2001, pp. 213–218.

[3] M. McLaren, M.I. Mandasari, and D.A. Leeuwen, "Source normalization for language-independent speaker recognition using i-vectors," in *Proc. Odyssey*, 2012, pp. 55–61.

[4] A. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *Proc. ICSLP*, 2006, pp. 1471–1474.

[5] D. Garcia-Romero and C.Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Interspeech'2011*, 2011, pp. 249–252.

[6] C.M. Bishop, *Pattern recognition and machine learning*, springer, New York, 2006.

[7] W. Y. Chen, Y. Q. Song, H. J. Bai, C. J. Lin, and E. Y. Chang, "Parallel spectral clustering in distributed systems," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 33, no. 3, pp. 568–586, 2011.

[8] Stephen Shum, Najim Dehak, Reda Dehak, and James R Glass, "Unsupervised speaker adaptation based on the cosine similarity for text-independent speaker verification.," in *Odyssey*, 2010, p. 16.

Sys	**Mandarin (CMN)**						**Cebuano (CEB)**						**CMN+CEB**		
	Male			Female			Male			Female			Male+Female		
	EER(%)	mDCF	aDCF	EER(%)	mDCF	aDCF	EER(%)	mDCF	aDCF	EER(%)	mDCF	aDCF	EER(%)	mDCF	aDCF
A	5.36	0.458	0.714	17.34	0.819	4.905	22.68	0.866	1.806	23.29	0.956	4.734	17.49	0.775	3.040
B	7.39	0.442	0.674	19.01	0.797	0.842	25.30	0.811	0.955	24.15	0.935	0.975	20.89	0.746	0.862
C	6.03	0.461	0.748	17.00	0.781	0.875	23.36	0.823	0.975	23.61	0.945	0.980	18.93	0.752	0.895

Table 1. Performance of three systems in the development set of SRE16. *Sys A*: No score normalization; *Sys B*: T-norm; *Sys C*: S-norm. *mDCF*: Minimum DCF; *aDCF*: Actual DCF

Sys	Score Norm	Equalized			Unequalized		
		EER(%)	minDCF	actDCF	EER(%)	minDCF	actDCF
A	None	18.14	0.869	2.496	17.65	0.844	3.015
B	T-norm	22.54	0.810	0.875	21.06	0.793	0.860
C	S-norm	20.72	0.824	0.908	19.03	0.796	0.893

Table 2. Performance of three systems in the development set of SRE16.

Task	Task Name	CPU Time (sec.) per Utt.	% of Real Time	Memory Consumption (MB)
1	Voice Activity Detection	0.659	11.70	8.3
2	MFCC Extraction	0.095	1.70	4.2
3	Feature Warping	3.127	55.40	7.3
4	Computing Sufficient statistics	0.933	16.50	329
5	I-vector Estimation	0.817	14.50	543
6	PLDA Scoring + Tnorm/Snorm	0.012	0.20	286
	Overall	5.643	100.00	–

Table 3. Computation time and memory consumption of various part of the system to produce the score of one verification trial. All tasks were performed on a 64-bit Linux server with 8G Ram and an Intel Q9550 running at 2.83GHz. All CPU times are based on one core of the processor.