POLYU SUBMISSION OF NIST 2018 SPEAKER RECOGNITION EVALUATION

Man-Wai MAK, Youzhi TU and Weiwei LIN

Dept. of Electronic and Information Engineering
The Hong Kong Polytechnic University, Hong Kong SAR
25 Oct. 2018
http://www.eie.polyu.edu.hk/∼mwmak/papers/polyu-sre18-sysdesc.pdf

ABSTRACT

This report describes the systems submitted by The Hong Kong Polytechnic University. The submitted systems are the fusion of a GMM i-vector system and three DNN x-vector systems. In one of the x-vector systems, we applied our recently proposed maximum-mean-discrepancy autoencoder [1, 2] for domain adaptation. In another two x-vector systems, we adapted the PLDA model using the unlabeled data in SRE18 development data and the SITW data with and without S-norm. The most up-to-date description can be found in the URL above.

***Index Terms*—** Speaker verification; i-vectors; x-vectors; probabilistic LDA; domain adaptation; maximum mean discrepancy

1. SYSTEM DESCRIPTION

1.1. Acoustic Features and VAD

For the i-vector system, we used Kaldi[1] (`v1/conf/mfcc.conf` in SRE16 receipe) to extract 20-dimensional MFCCs plus their delta and double delta coefficients, followed by energy-based voice activity detection (VAD) [`v1/conf/vad.conf`]. For the x-vector systems, we extracted 23-dimensional MFCC based on `v2/conf/mfcc.conf` in Kaldi's SRE16 recipe, followed by energy-based VAD.

we down-sampled the `.flac` files in SRE18-dev and SITW to 8kHz using Sox.[2] For the VAST enrollment utterances, we replaced the Kaldi's VAD decisions by the diarization labels in the SRE18-dev set. No diarization was applied to the test segments.

1.2. I-vector Extraction

The training procedure of the i-vector extractor was derived from Kaldi's SRE16 recipe. Specifically, a gender-independent UBM with 2048 Gaussians was trained using utterances from SRE18-unlabeled, SRE16-dev-test, SRE16-eval-test, and SRE16-minor. Using this UBM, a total variability matrix with 600 factors was trained using Switchboard 2 Phases I-III, Switchboard Cellular Parts 1-2, SRE04-12, and Mixer 6. For SRE12, we only used telephone segments under the `tel_phn` directory of the corpus.

This work was in part supported by The RGC of Hong Kong SAR, Grant No. PolyU 152137/17E.

[1] http://kaldi-asr.org/

[2] https://sourceforge.net/projects/sox/

1.3. Data Augmentation

We used Kaldi's SRE16 recipe to create the augmented data. Specifically, we considered the noise, music and speech from the MUSAN database as noise sources and digitally added noise to the waveform files of Switchboard 2 Phases I-III, Switchboard Cellular Parts 1-2, SRE04-10, and Mixer 6 at an SNR from 0 dB to 20 dB. We also applied various reverberation effects to the waveform files in these datasets. Then, we selected 128,000 files from the noise-contaminated waveform files and reverberated waveform files. The resulting augmented dataset was added to the original Switchboard (Phases I–III and Cellular 1-2), SRE04-10, and Mixer 6 for training an x-vector extractor and PLDA models.

For the i-vector system, we also included the telephone segments of SRE12 in the data augmentation process, i.e., the SRE04-10 for the x-vector systems is replaced by SRE04-12.

1.4. X-vector Extraction

We used two DNNs for extracting x-vectors. For Systems S_3 and S_4 in Table 1, we used the pre-trained DNN available from the Kaldi repository.[3] For System S_2 in Table 1, we retrained the DNN using the augmented data described in Section 1.3.

1.5. PLDA and Adapted PLDA

For both i-vector and x-vector systems, we used the augmented data described in Section 1.3 to compute an LDA projection matrix with a rank of 300. The mean of the LDA-projected vectors were then removed, followed by length normalization. The processed vectors were then used for training PLDA models with 300 latent factors.

As shown in Fig. 1(a), the x-vectors from SRE04-10+MX6 and Switchboard are significantly different from the x-vectors of SRE18-eval. But utterances from SRE04-10+MX6 and Switchboard were used for training the PLDA models. To reduce domain mismatch, we used SRE18-dev data to adapt the PLDA model for CMN2 data and used SITW-dev-enroll data to adapt the PLDA model for VAST data. Fig. 1(b) shows that the domain mismatch between PSTN and VoIP utterances is negligible. Therefore, we did not separate the CMN2 into PSTN and VoIP.

In System S_4 in Table 1, we also used our recently proposed maximum-mean-discrepancy autoencoder [1, 2] to transform the x-vectors to a space with less domain variability before carrying out the LDA and PLDA training.

[3] http://kaldi-asr.org/models/3/0003_sre16_v2_1a.tar.gz

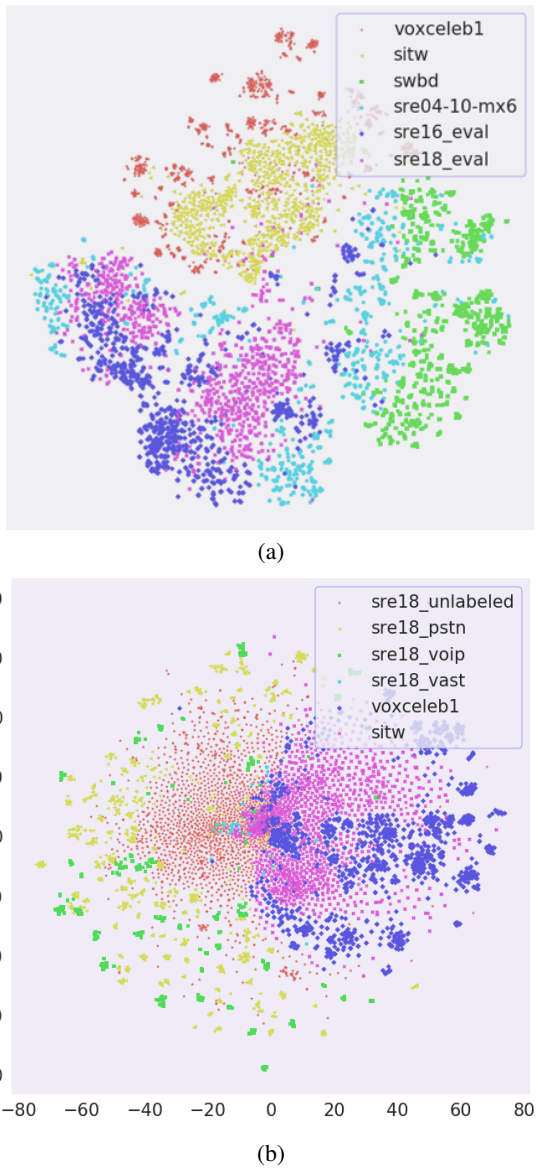

Fig. 1. t-SNE [3] plot of x-vectors from various datasets. *SWBD*: Switchboard 2 Phases I-III and Switchboard Cellular Parts 1-2.

1.6. PLDA Scoring and Score Normalization

For each trial, we average multiple i-vectors/x-vectors of the target speaker so that each target speaker only have one i-vector/x-vector for scoring. A mean vector computed during the training stage is subtract from this enrollment vector, followed by LDA projection. The test vectors were also subject to the same mean-subtraction (centering in Table 1) and LDA projection. The datasets used for mean subtraction depend on the the sub-tasks (CMN2 or VAST). For the exact usages, see Table 1.

1.7. Score Calibration

We used the Bosaris toolkit[4] to calibrate the scores produced by the Kaldi program `ivector-plda-scoring`. This means that we used utterances in the development datasets shown in Table 1 as enrollment and test data to obtain a set of target and non-target scores. Then, we presented these scores to the function `linear_calibrate_scores` in Bosaris to find the calibration weights. The data for centering the i-vectors/x-vectors are different for CMN2 and VAST. Also, systems with S-norm during scoring require S-norm during calibration. The datasets for computing the S-norm parameters for calibration and for scoring are shown in Table 1.

2. PERFORMANCE AND COMPUTATION TIME

Table 2 shows the performance (in terms of EER, minimum DCF and actual DCF) of different systems and their fusion in the development set of SRE18. Table 3 shows the fusion weights of the three submitted systems and their performance on SRE18-dev.

Table 4 shows the computation times and % of real-time for the i-vector and x-vector systems.

3. REFERENCES

[1] W. W. Lin, M. W. Mak, and J. T. Chien, "Multi-source i-vectors domain adaptation using maximum mean discrepancy based autoencoders," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 12, pp. 2412–2422, 2018.

[2] W. W. Lin, M. W. Mak, L. X. Li, and J. T. Chien, "Reducing domain mismatch by maximum mean discrepancy based autoencoders," in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 162–167.

[3] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[4] https://sites.google.com/site/bosaristoolkit/

Sys	Embedding	Sub-Task	PLDA Training	PLDA Adaptation	Snorm	Score Calibration					PLDA Scoring				
						Adapt Train	Adapt Test	Train Vector	Test Vector	Centering	Adapt Train	Adapt Test	Train Vector	Test Vector	Centering
S_1	i-vector	CMN2	sre04-12+	sre18-unlabeled	Yes	sre16-major	sre16-major	sre16-eval-enroll	sre16-eval-test	sre16-major	sre16-eval-test	sre16-eval-test	sre18-dev-enroll	sre18-dev-test	sre18-unlabeled
		VAST	mx6+aug	sitw-dev-enroll		sre16-major	sre16-major	sitw-eval-enroll	sitw-eval-test	sre18-unlabeled	sitw-eval-enroll	sitw-eval-test			
S_2	x-vector	CMN2	sre04-10+	sre18-unlabeled	Yes	sre16-major	sre16-major	sre16-eval-enroll	sre16-eval-test	sre16-major	sre16-eval-test	sre16-eval-test	sre18-dev-enroll	sre18-dev-test	sre18-unlabeled
		VAST	mx6+aug	sitw-dev-enroll		sre16-major	sre16-major	sitw-eval-enroll	sitw-eval-test	sre18-unlabeled	sitw-eval-enroll	sitw-eval-test			
S_3	x-vector	CMN2	sre04-10+	sre18-unlabeled	No	–	–	sitw-eval-enroll	sitw-eval-test	sitw-eval-test	–	–	sre18-dev-enroll	sre18-dev-test	sre18-unlabeled
		VAST	mx6+aug	voxceleb1-enroll		–	–	sitw-eval-enroll	sitw-eval-test	voxceleb1-enroll	–	–			sre04-10+mx6+aug
S_4	x-vector	CMN2	sre04-10+	None	Yes	sre16-major	sre16-major	sre16-eval-enroll	sre16-eval-test	sre16-major	sre16-eval-test	sre16-eval-test	sre18-dev-enroll	sre18-dev-test	sre18-unlabeled
		VAST	mx6+aug	None		sre16-major	sre16-major	sre18-dev-enroll	sre18-dev-test	sre18-unlabeled	sitw-eval-enroll	sitw-eval-test			

Table 1. Datasets used by various systems for PLDA training, PLDA adaptation, calibration, and PLDA scoring. The columns "Adapt Train" and "Adapt Test" indicate the data sources for computing the S-norm parameters. *sre18-unlabeled* stands for unlabeled data in the development set of SRE18.

System	Sub-Task	EER (%)	minDCF	ActualDCF
S_1	CMN2	13.14	0.684	0.778
	VAST	7.82	0.490	0.819
	Both	–	–	0.798
S_2	CMN2	8.86	0.567	0.644
	VAST	9.47	0.481	0.642
	Both	–	–	0.643
S_3	CMN2	8.60	0.546	0.556
	VAST	7.41	0.498	0.535
	Both	–	–	0.545
S_4	CMN2	09.17	0.572	0.580
	VAST	07.41	0.412	0.412
	Both	–	–	**0.496**
$S_1 + S_2$	CMN2	8.66	0.554	0.612
	VAST	7.82	0.519	0.601
	Both	–	–	0.607
$S_1 + S_2 + S_3$	CMN2	7.75	0.531	0.539
	VAST	7.41	0.424	0.572
	Both	–	–	0.555
$S_1 + S_2 + S_4$	CMN2	8.11	0.532	0.540
	VAST	7.41	0.374	0.523
	Both	–	–	0.531
$S_1 + S_2 + S_3 + S_4$	CMN2	7.51	0.526	0.547
	VAST	7.41	0.412	0.490
	Both	–	–	0.518

Table 2. Performance of various systems and their fusions in the development set of SRE18. The symbol '+' denotes fusion of scores from the respective systems. For the configuration of Systems S_1 to S_4, refer to Table 1.

Submission	Score Fusion Equation	Sub-Task	EER (%)	minDCF	ActualDCF
Primary	$0.8(w_0 + w_1S_1 + w_2S_2) + 0.2S_3$	CMN2	7.75	0.531	**0.539**
	$0.2(w'_0 + w'_1S_1 + w'_2S_2) + 0.8S_3$	VAST	**7.41**	0.424	0.572
		Both	–	–	0.555
Contrastive 1	$0.2(w_0 + w_1S_1 + w_2S_2 + w_4S_4) + 0.8S_3$	CMN2	**7.51**	**0.526**	0.547
	$0.6(w'_0S_1 + w'_2S_2 + w'_4S_4) + 0.4S_3$	VAST	**7.41**	0.412	**0.490**
		Both	–	–	0.518
Contrastive 2	$w_0 + w_1S_1 + w_2S_2 + w_4S_4$	CMN2	8.11	0.532	0.540
	$w'_0 + w'_1S_1 + w'_2S_2 + w'_4S_4$	VAST	**7.41**	**0.374**	0.523
		Both	–	–	0.531

Table 3. Performance of submitted systems in SRE18-dev. In the second column, S_1 to S_4 are the calibrated scores from the respective systems. For Primary and Contrastive 1, the fusion weights w_i and w'_i were determined by the `linear_fusion_scores` function of Bosaris using SRE16-eval trial scores and SITW-eval trial scores, respectively. For Contrastive 2, the fusion weights were determined by Bosaris using SRE18-dev trial scores. For the configuration of Systems S_1 to S_4, refer to Table 1.

Task	Task Name	CPU Time (sec.) per Utt.	% of Total Time
1	Voice Activity Detection	0.039	2.32
2	MFCC Extraction	0.406	24.20
3	I-vector Extraction	1.232	73.42
4	PLDA Scoring	0.001	0.06
	Overall	1.677	100.00

(a)

Task	Task Name	CPU Time (sec.) per Utt.	% of Total Time
1	Voice Activity Detection	0.039	1.57
2	MFCC Extraction	0.406	16.32
3	X-vector Estimation	2.042	82.07
4	PLDA Scoring	0.001	0.04
	Overall	2.527	100.00

(b)

Table 4. Computation time of various part of the (a) i-vector system and (b) x-vector systems to produce the score of one verification trial. All tasks were performed on a 64-bit Linux server with 32G Ram equipped with an Intel i7-5820K running at 3.30GHz. All CPU times are based on one core of the processor.