PAU SUBMISSION OF NIST 2018 SPEAKER RECOGNITION EVALUATION

Youzhi TU[*], *Yingke ZHU*[#], *Man Wai MAK*[*], *Dongpeng CHEN*[†], *Weiwei LIN*[*], *Brian K.W. MAK*[#], *Zhuxin CHEN*[†] *and Weibin ZHANG*[†]

[*]Dept. of Electronic and Information Engineering,
The Hong Kong Polytechnic University, Hong Kong SAR of China
[#]Department of Computer Science and Engineering,
Hong Kong University of Science and Technology, Hong Kong SAR of China
[†]Voice AI, China
28 Oct. 2018
http://www.eie.polyu.edu.hk/∼mwmak/papers/pau-sre18-sysdesc.pdf

ABSTRACT

This report describes the systems submitted by the PAU team, which is composed of The Hong Kong Polytechnic University, The Hong Kong University of Science and Technology, and VoiceAI, China. The submitted systems are the fusion of a GMM i-vector system and five DNN x-vector systems. Specifically, we have two x-vector systems trained on 16kHz speech data for VAST data, and the other three x-vector systems were trained on 8kHz speech utterances. In one of the 8kHz x-vector systems, we incorporated a self-attention layer into the x-vector extractor. 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; self attention

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, except for System S_4 for VAST data (referred as S_4-VAST in the sequel) in Table 1, we extracted 23-dimensional MFCC based on `v2/conf/mfcc.conf` in Kaldi's SRE16 recipe, followed by energy-based VAD.

For System S_4-VAST, the input acoustic features are 30-dimensional MFCCs with a frame-length of 25ms extracted using a 30-channel mel-scale filterbank spanning the frequency range 20Hz-7,600Hz. Cepstral mean subtraction is applied over a 3-second sliding window and an energy-based VAD is employed to filter out non-speech frames from the utterances. During evaluation stage, a statistics-based VAD was applied only on the test utterances to extract speech frames. System S_5 uses a different VAD [1] based on a bidirectional long short-term memory (BLSTM) network.

For Systems S_4-VAST and S_5 in Table 1, we down-sampled the `.flac` files in SRE18-dev, SITW, Voxceleb1 (excluding speakers that overlap with SITW) and Voxceleb2 to 16kHz using Sox.[2] For other systems, these utterances were down-sampled to 8kHz. We replaced the Kaldi's VAD decisions by the diarization labels for the VAST enrollment utterances in the SRE18-dev set. No diarization was applied to the test segments.

1.2. 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 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 91,102 files from the noise-contaminated waveform files and reverberated waveform files. In the sequel, we refer to the i-vectors and x-vectors of these files as "aug". The resulting augmented data were added to the original SRE04-10 and Mixer 6 for training an x-vector extractor, LDA projection matrix, and PLDA models. We refer to this augmented data set as "sre04-10+mx6+aug". This dataset comprises 154,157 segments spoken by 4,407 speakers.

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. We selected 64,000 noise contaminated or reverberated files and added them to SRE04-12 and Mixer 6, which results in an augmented dataset "sre04-12+mx6+aug" with 133,516 segments spoken by 4,408 speakers.

For the S_4-VAST system in Table 1, Voxceleb1 and Voxceleb2 were augmented using the same technique as above for training both the x-vector extractor and the PLDA backend. We refer to this dataset as "voxceleb1-2+aug", which amounts to 1,236,567 segments from 7,185 speakers.

1.3. 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

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/

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.

1.4. X-vector Extraction

We used two DNNs to extract 512-dimensional x-vectors for Systems S_2 and S_3 in Table 1. Specifically, for the former system we used the pre-trained DNN available from the Kaldi repository.[3] While for the latter, we retrained the DNN using Switchboard 2 Phases I–III, Switchboard Cellular Parts 1–2, SRE04–10, Mixer 6, and the augmented data described in Section 1.2. Totally, the training data comprise 208,807 speech files spoken by 5,009 speakers.

System S_4-VAST differs from Systems S_2 and S_3 in that the statistics pooling layer in the DNN is replaced by a self-attention layer [2]. The self-attention layer derives a weighted mean vector and a weighted deviation vector from the outputs of the previous hidden layer over each speech segment. The remaining components are the same as the conventional x-vector extractor. System S_5 has a larger network architecture than Systems S_2 and S_3 with an x-vector dimension of 1024 rather than 512.

1.5. LDA, PLDA and Adapted PLDA

We used the original plus augmented data described in Section 1.2 to compute an LDA projection matrix with a rank of 300 for Systems S_1 and S_2 in Table 1. 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. For Systems S_3 and S_4, the LDA-projected vectors and the latent dimension of PLDA models are 150.

Since the x-vectors extracted from SRE04-10+MX6 are significantly different from the x-vectors of SRE18-eval, there is severe domain mismatch between these datasets for all systems except for S_4-VAST. To reduce domain mismatch, we used SRE18-dev data to adapt the PLDA model for CMN2 and used SITW-dev-enroll or voxceleb1-enroll data to adapt the PLDA model for VAST in these systems. Note that for System S_4-VAST, the PLDA model was trained on a 100k subset of the "voxceleb1-2+aug" set. As the domain of "voxceleb1-2+aug" is similar to that of VAST, no PLDA adaptation was performed in this case.

1.6. PLDA Scoring and Score Normalization

For each trial, we averaged 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.

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

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

2. PERFORMANCE AND COMPUTATION TIME

Table 2 shows the performance (in terms of EER, minimum DCF and actual DCF) of different systems and their fusions on 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 scoring times and percentage of total scoring time for scoring a pair of i-vectors and a pair of x-vectors. Table 5 shows another CPU execution time as well as memory requirements for computing the score of one verification trial.

3. REFERENCES

[1] Carlos Busso Fei Tao, "Bimodal recurrent neural network for audiovisual voice activity detection," *Proc. Interspeech 2017*, pp. 1938–1942, 2017.

[2] Yingke Zhu, Tom Ko, David Snyder, Brian Mak, and Daniel Povey, "Self-attentive speaker embeddings for text-independent speaker verification," *Proc. Interspeech 2018*, pp. 3573–3577, 2018.

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+mx6+aug	sre18-unlabeled	Yes	sre16-major	sre16-major	sitw-eval-enroll	sitw-eval-test	sitw-eval-test	sre18-unlabeled	sre18-unlabeled	sre18-dev-enroll	sre18-dev-test	sre18-unlabeled
	x-vector16k	VAST	voxceleb1-2+aug	–	No	–	–	sitw-dev-enroll	sitw-dev-test	sre18-unlabeledU	–	–			sitw-eval-test
S_5	–	CMN2	–	–	–	–	–	–	–	–	–	–	–	–	–
	x-vector16k	VAST	voxceleb1-2	–	No	–	–	–	–	–	–	–	sre18-dev-enroll	sre18-dev-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*: unlabeled data in SRE18. *sre18-unlabeledU*: 16kHz upsampled version of *sre18-unlabeled*. *i-vector* and *x-vector*: speaker embedding based on 8kHz speech data. *x-vector16k*: speaker embedding based on 8kHz speech data.

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	7.39	0.553	0.565
	VAST	4.53	0.416	0.671
	Both	–	–	0.618
S_5	CMN2	–	–	–
	VAST	4.94	0.523	0.576
	Both	–	–	–
$S_1 + S_2$	CMN2	8.66	0.554	0.612
	VAST	7.82	0.519	0.601
	Both	–	–	0.607
$S_1 + S_2 + S_4$	CMN2	6.72	0.499	0.509
	VAST	5.35	0.407	0.407
	Both	–	–	0.458
$S_1 + S_2 + S_3 + S_4$	CMN2	6.57	0.496	0.510
	VAST	4.94	0.412	0.486
	Both	–	–	0.498
$S_1 + S_2 + S_4 + S_5$	CMN2	6.78	0.502	0.514
	VAST	5.35	0.449	0.572
	Both	–	–	0.543

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_5, refer to Table 1.

Submission	Score Fusion Equation	Sub-Task	EER (%)	minDCF	ActualDCF
Primary	$0.5(w_0 + w_1 S_1 + w_2 S_2) + 0.5 S_4$	CMN2	6.72	0.499	**0.509**
	$0.8(w'_0 + w'_1 S_1 + w'_2 S_2) + 0.2 S_4$	VAST	5.35	**0.407**	**0.407**
		Both	–	–	0.458
Contrastive 1	$0.9(0.5(w_0 + w_1 S_1 + w_2 S_2) + 0.5 S_4) + 0.1 S_3$	CMN2	**6.57**	**0.496**	0.510
	$0.9(0.8(w'_0 + w'_1 S_1 + w'_2 S_2) + 0.2 S_4) + 0.1 S_3$	VAST	**4.94**	0.412	0.486
		Both	–	–	0.498
Contrastive 2		CMN2	–	–	–
	$w_{b0} + w_{b1}(0.8(w'_0 + w'_1 S_1 + w'_2 S_2) + 0.2 S_4) + w_{b2} S_5$	VAST	5.35	0.449	0.572
		Both	–	–	–

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 and S_5 are uncalibrated scores. 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 w_{bi} were determined by Bosaris using SITW-eval trial scores. For the configuration of Systems S_1 to S_5, 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.

Task	CPU Time(sec.)	Memory(MB)
I-vector system	11.5	462
X-vector system	11.75	394
Attentive x-vector system (1 head)	13.32	421
Attentive x-vector system (2 head)	15.14	593

Table 5. Computation time and memory consumption of various systems to produce the score of one verification trial. All tasks were performed on a 64-bit Linux server with 128G Ram and an Intel Xeon E5-2650 running at 2.20GHz. All CPU times are based on one core of the processor.