TUSUR (Team 21) Systems Description for the Short-duration Speaker Verification (SdSV) Challenge

Ivan Rakhmanenko, Evgeny Kostyuchenko, Alexander Shelupanov

Tomsk State University of Control Systems and Radioelectronics (TUSUR), Tomsk, Russia
ria@keva.tusur.ru

1. Introduction

In this paper we present our contribution to the Task 2 of the short-duration speaker verification (SdSV) challenge. The main task for this challenge is to find new technologies for text-dependent and text-independent speaker verification in short duration scenario. Some of the approaches used by the authors during participation in the challenge are presented. Described speaker verification systems include x-vector system with PLDA backend, additional audio augmentation, gender recognition and scores normalization, x-vector system with neural PLDA backend and gender recognition, and fusion of both systems.

Our main goal in this challenge was to find out what is the maximum of capabilities for the x-vector frontend system. We found that system with PLDA backend and ZT-normalization method (single system) gives superior performance on Farsi speech, but worse results on English speech. Overall, in terms of minDCF single system performs 46.3% better than baseline x-vector system. In addition, we found that performance of Neural PLDA backend system does not improve after adding augmented enroll data, but PLDA backend is affected and performs significantly better. Single system with ZT-scores normalization and additional audio augmentation performs 14.8% better than Neural PLDA backend system.

2. Systems description

In this section, we present components of speaker verification systems used in SdSV 2020 challenge. This includes complete description of the submitted systems components, including front-end and back-end modules along with their configurations.

2.1. X-vector PLDA (single) system

First system is baseline x-vector [1] with several modifications. It consists of three main components: x-vector frontend, gender recognition classifier and gender-dependent PLDA backend. It is considered as single system, because it has simple linear structure, it does not have scores fusion and only one gender-dependent modeling is used.

X-vector frontend for this system is based on provided baseline Kaldi recipe without any significant modifications. As a features 30 mel-frequency cepstral coefficients (MFCC) are used, with 25 ms frame length, 20-7600 Hzs frequency range. After MFCC extraction we apply sliding window cepstral mean and variance normalization (CMVN) and remove silence frames. Silence frames are removed using Kaldi energy-based voice activity detection script.

X-vector extraction network follows E-TDNN structure [2], but first TDNN-ReLU layer has bigger context $(t-2, t-1, t, t+1, t+2)$, 9-th Dense-ReLU layer is deleted and last dense layer before pooling has size=1536. Size of the last layer is 7323, equals to number of speakers in the training dataset. X-vector network was trained for 6 epochs with batch size = 128 at 1st epoch and batch size = 164 at 2nd-6th epochs. First dense layer after pooling was used for x-vector extraction, giving x-vector size = 512.

Next component of the single system is gender recognition classifier. It is logistic regression classifier that was trained on x-vectors without x-vector mean normalization, with LDA transformation and Kaldi x-vector length normalization. LDA dimension is 200. Output of this classifier was used to select proper gender-dependent PLDA model for every evaluation x-vector.

Last component of the system is PLDA model. It gets mean normalized, LDA-transformed and length-normalized x-vectors as input features. There was two PLDA models trained, one for male and one for female speakers. Scores for this models were normalized using Z- and T-normalization methods [3]. Z-normalization includes normalizing PLDA scores using target speaker averaged x-vector against a set of impostor speakers statistics (1). T-normalization includes normalizing PLDA scores using test segments x-vectors against a set of impostor speakers statistics (2). We selected 10% of the best scores for normalization statistics calculation. Single system's final score is sum of Z- and T-normalized PLDA scores (3).

$$S_{Z-norm} = \frac{S(X,\lambda) - \mu_{\lambda}}{\sigma_{\lambda}} \tag{1}$$

$$S_{T-norm} = \frac{S(X,\lambda) - \mu_{X}}{\sigma_{X}} \tag{2}$$

$$S_{ZT-norm} = S_{Z-norm} + S_{T-norm} \tag{3}$$

2.2. Neural PLDA system

As a second backend for this challenge we decided to use Neural PLDA (N-PLDA) backend [4].This model, operates on pairs of x-vector embeddings (a pair of enrollment and test x-vectors) and outputs a score that allows the decision of target versus non-target hypotheses. We wanted to compare this backend's performance to standard PLDA backend.

For N-PLDA network initialization we used gender-dependent PLDA models train for X-vector PLDA system. We used the same network parameters as in [4], except batch size = 4096 and beta = 9.9. N-PLDA model was trained for 20 epochs with learning rate = 0.0001, Adam optimizer and learning rate halving after error on the validation set increased for 2 epochs.

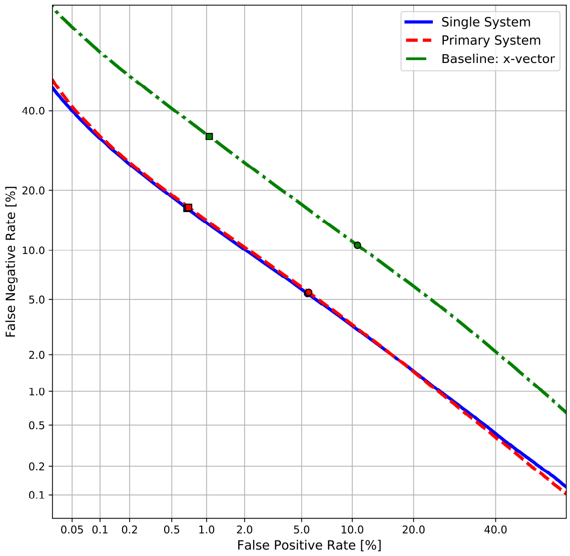

Figure 1: *Detection error trade-off (DET) curves for presented systems on evaluation dataset*

2.3. Fusion (primary) system

Primary system consists of two components described earlier: X-vector PLDA and Neural PLDA systems. Primary system score was computed as sum of two subsystems normalized scores. We thought that this fusion could give better performance stability on full set and could reduce overfitting on the training data.

3. Data usage

For x-vector network training we used Kaldi scripts of the x-vector baseline system. There was only two datasets used for network training: full VoxCeleb1 and VoxCeleb2 [5]. Audio augmentation methods provided with the Kaldi were used. This includes adding reverberation, music, noise and babble from RIRS noises and MUSAN corpuses to the training audio. Two millions of augmented audio was selected in addition to the original VoxCeleb datasets.

For gender recognition classifier training we used part of the VoxCeleb dataset (650k audio segments) and whole SdSVC task 2 train dataset.

For PLDA and Neural PLDA models training we used xvectors extracted from VoxCeleb1 and VoxCeleb2 datasets with 2 million augmented audio, SdSVC train dataset and LibriSpeech (clean-100, clean-360, other-500) datasets adding audio augmentations (reverberation, music, noise and babble).

Enroll dataset was expanded with augmented audio using the same audio augmentations techniques as for the x-vector network's training.

For scores normalization whole SdSVC train set was used as gender-dependent impostor speakers data.

4. Experimental evaluation

Performance of the three presented systems on the progress and evaluation subsets is shown in Table 1. Despite our suggestion that fusion of the PLDA and N-PLDA x-vector systems could improve verification results, primary system showed worse performance than our single system. Single system had better results on both the progress and the evaluation subset. Overall, results for primary and single system are pretty similar (Figure 1).

In Table 2, we could see detailed results for our best submission. We could see that there is no overfitting on the progress subset, as performance on progress and evaluation subsets are pretty similar. In addition, our single system performs better on male and Farsi speech (Figure 2). It seems that our system recognized cross-lingual speakers much worse than speakers with native Farsi language. One of the main reasons is that there was no English speech in SdSVC Train and Enroll datasets and system overfitted to Farsi speech.

We tried to apply audio augmentations to test data, but it gave us much worse results than after applying enroll data augmentation (Table 3). For N-PLDA backend enroll data augmentation gave worse results than without this augmentation.

After scores submission deadline we tried to vary fraction of data used for scores normalization and found that ZT-scores normalization with selecting top 40% of the highest impostor scores gave us a little boost to our single system's performance (Table 3). In addition, it could be seen that applying Z- and T-normalization gives worse performance than applying ZT-scores normalization simultaneously.

Table 1: *Submitted systems evaluation results.*

System	progress		evaluation	
	m-DCF	**EER**	**m-DCF**	**EER**
Baseline	0.4319	10.67	0.4324	10.67
PLDA (single)	**0.2313**	**5.47**	**0.2324**	**5.46**
N-PLDA	0.2706	6.19	0.2727	6.21
Fusion (primary)	0.2352	5.55	0.2361	5.55

Table 2: *Single system evaluation results.*

System	progress		evaluation	
	m-DCF	**EER**	**m-DCF**	**EER**
Male	0.1825	4.16	0.1845	4.17
Female	0.2606	6.19	0.2610	6.17
English	0.2468	5.46	0.2442	5.46
Farsi	0.1423	2.98	0.1435	2.96
EN-male	0.1908	4.21	0.1898	4.23
EN-female	0.2797	6.19	0.2764	6.17
FA-male	0.1083	1.99	0.1087	2.00
FA-female	0.1627	3.55	0.1640	3.52
Farsi-TC-vs-IC-subset	0.1334	3.21	0.1362	3.20

Table 3: *Additional experiments results.*

System	Aug.	Norm.	evaluation	
			m-DCF	**EER**
PLDA (single)	Enroll	ZT	0.2324	5.46
N-PLDA	-	-	0.2727	6.21
N-PLDA (enr aug)	Enroll	-	0.2802	6.40
PLDA (top 40% norm)	**Enroll**	**ZT**	**0.2306**	**5.45**
PLDA (Z-norm)	Enroll	Z	0.2371	5.32
PLDA (T-norm)	Enroll	T	0.2439	5.90
PLDA (without norm)	-	-	0.2962	7.38
PLDA (test aug)	Test	ZT	0.3174	7.51

We could conclude, that ZT-scores normalization and enroll data augmentation could significantly improve x-vector PLDA speaker verification system's performance.

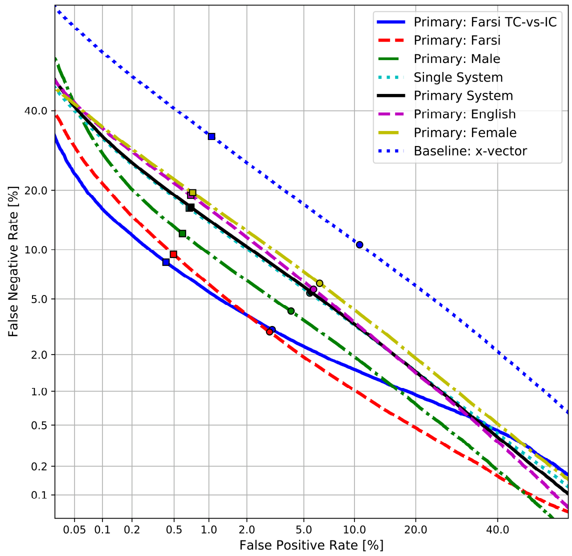

Figure 2: *Detailed DET-curves for primary system*

5. Conclusions

In this paper we presented our contribution to the Task 2 of the short-duration speaker verification challenge. Described speaker verification systems include baseline x-vector system with PLDA backend and scores normalization, x-vector system with neural PLDA backend and fusion of both systems.

We found that system with PLDA backend and ZT-normalization method (single system) gives superior performance on Farsi speech, but worse results on English speech. Overall, in terms of minDCF single system performs 46.3% better than baseline x-vector system. In addition, we found that performance of Neural PLDA backend system does not improve after adding augmented enroll data. Single system with ZT-scores normalization and additional audio augmentation performs 14.8% better than Neural PLDA backend system.

6. Acknowledgements

The results were obtained as part of the implementation of the basic part of the state task of the Russian Federation Education and Science Ministry, project 8.9628.2017/8.9.

7. References

[1] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, S. Khudanpur, X-vectors: Robust dnn embeddings for speaker recognition, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, pp. 5329–5333.

[2] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, S. Khudanpur, Speaker recognition for multi-speaker conversations using x-vectors, in: ICASSP 2019-2019 IEEE 13 International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 5796–5800.

[3] C. Barras and J.-L. Gauvain, "Feature and score normalization for speaker verification of cellular data," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., vol. II, Hong Kong, China, Apr. 2003, pp. 49–52.

[4] Ramoji, Shreyas, Prashant Krishnan, and Sriram Ganapathy. "NPLDA: A Deep Neural PLDA Model for Speaker Verification." arXiv preprint arXiv:2002.03562 (2020).

[5] Nagrani, Arsha, Joon Son Chung, and Andrew Zisserman. "Voxceleb: a large-scale speaker identification dataset." arXiv preprint arXiv:1706.08612 (2017).