Dezhafzar System Description for 2020 SdSV Challenge

Fatemeh Arabnejad, Abbas Khosravani

Dezhafzar Co., Iran

info@dezhafzar.com

Abstract

This report describes our system submitted to 2020 Short-duration Speaker Verification (SdSV) challenge [1].
Index Terms: speech verification, x-vector, PLDA

1. Introduction

Different biometrics are used the claimed identity of a person, including fingerprints, face or iris. Such biometrics need complicated hardwares and also the person must be presented physically. Speech is the most convenient way to communicate with each other. It can be captured by a simple microphone which is available in most of devices. Research on speaker verification has started since 1990s [2]. Different techniques have been developed, from Hidden Markov Models(HMM) to DNN models. Recently, x-vector which is based on Time Delay Neural Network (TDNN) achieves superior performance on most of evalaution datasets. This article describes the system developed by Dezhafzar company to improve the baseline system introduced by SdSV challenge. The submitted system has EER of 5.68% and minDCF 0.23 on progress data.

2. System Description

2.1. Acoustic Features

In this work we extracted 30 MFCC features with 25ms window size using Hamming window from length variation audio signals.

2.2. x-vector Extraction

Most of the state-of-the-art automatic speaker verification systems are developed based on x-vectors [3, 4, 5] . The x-vectors are extracted from the affine component of a TDNN layer. The properties of a TDNN model is to capture time invariant features. In our experiments we use kaldi toolkit for training the TDNN model [6]. Table 1 shows the structure of the TDNN model used in our experiment.

Table 1: *Struct of TDNN architecture*

Layer	Layer Context	Total Context
tdnn 1	[t-2, t+2]	5
tdnn 2	{t-2,t,t+2}	9
tdnn 3	{t-3,t,t+3}	15
tdnn 4	{t-4,t,t+4}	23
tdnn 5	{t}	23
tdnn 6	{t}	23
stat pooling	[0,T)	T
FC	{0}	T
FC	{0}	T
Softmax	{0}	T

After x-vector length normalization, x-vectors are centered and projected to 200 dimension vectors by using Linear Discriminant Analysis (LDA).

2.3. PLDA Scoring

To score a trial, cosine dissimilarity function or Probabilistic Linear Discriminant Analysis(PLDA) could be used. We used PLDA which has shown superior performance in our experiment.

We train two PLDA models on LDA projected x-vectors of training dataset based on speaker genders, one for male and one for female. On the evaluation phase, we evaluate all utterances on both PLDA models and then the average of scores is submitted to challenge.

3. Results

Three datasets were used in this method: VoxCeleb1 [7], VoxCeleb2 [8] and DeepMine dataset [9, 10]. The proposed x-vector is trained on the whole VoxCeleb data and DeepMine training data.

The equal error rate of the submitted system on leaderboard (30 percent of evaluation dataset) is 5.69 and the minimum discriminant cost function is 0.2369. Table 3 shows the results on the progress and evaluation dataset.

Table 2: *minDCF and EER of submitted system on progress data*

		Proposed Method	Baseline
minDCF	Farsi	**0.1500**	0.2949
	English	**0.3874**	0.4111
	Total	**0.2369**	0.4319
EER(%)	Farsi	**3.42**	6.17
	English	9.14	9.61
	Total	**5.69**	10.67

Table 3: *minDCF and EER of submitted system on evaluation data*

		Proposed Method	Baseline
minDCF	Farsi	**0.1499**	0.2962
	English	**0.3893**	0.0.4118
	Total	**0.2374**	0.4324
EER(%)	Farsi	**3.40**	6.14
	English	9.06	9.58
	Total	**5.67**	10.67

Figure 1a and 1b plot the DET curve of progress and evaluation set of DeepMine dataset.

4. Conclusions

We described our submitted model SdSV challenge. In the proposed method we tried to explore the effect of speaker's gender on speaker verification system. It has been shown that training two PLDA models, one for male and one for female, and averaging the scores of these models per utterance, improves the baseline system from 10.67% and 0.4324 down to 5.69% and 0.2374 in terms of EER and minDCF, respectively.

5. References

[1] K. A. Zeinali, Hossein nad Lee, J. Alam, and L. Burget, "Short-duration speaker verification (SdSV) challenge 2020: the challenge evaluation plan." arXiv preprint arXiv:1912.06311, Tech. Rep., 2020.

[2] J. M. Naik, "Speaker verification: a tutorial," *IEEE Communications Magazine*, vol. 28, no. 1, pp. 42–48, 1990.

[3] H. Zeinali, S. Wang, A. Silnova, P. Matějka, and O. Plchot, "But system description to voxceleb speaker recognition challenge 2019," *arXiv preprint arXiv:1910.12592*, 2019.

[4] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and 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.

[5] S. Ramoji, P. Krishnan, and S. Ganapathy, "Nplda: A deep neural plda model for speaker verification," *arXiv preprint arXiv:2002.03562*, 2020.

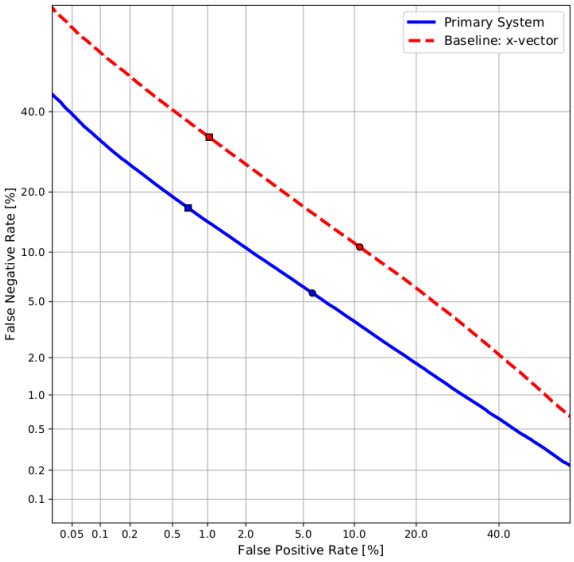
[6] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[7] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Proc. Interspeech 2017*, 2017, pp. 2616–2620.

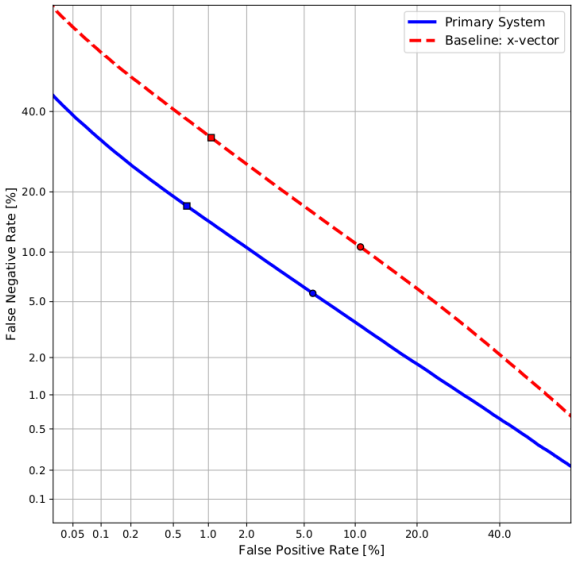
[8] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. Interspeech 2018*, 2018, pp. 1086–1090.

[9] H. Zeinali, H. Sameti, and T. Stafylakis, "DeepMine speech processing database: Text-dependent and independent speaker verification and speech recognition in Persian and English." in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 386–392.

[10] H. Zeinali, L. Burget, and J. Cernocky, "A multi purpose and large scale speech corpus in Persian and English for speaker and speech recognition: the DeepMine database," in *Proc. ASRU 2019 The 2019 IEEE Automatic Speech Recognition and Understanding Workshop*, 2019.



(a) *Progress set*



(b) *Evaluation set*

Figure 1: *DET curve plots indicating the baseline system vs our submited system on both the progress and evaluation dataset.*