# Team22 System Description for SdSV Challenge 2020

## Abstract

This report describes the systems used by the Team22 for Task 2 of the short-duration speaker verification challenge 2020. The challenge is based on DeepMine dataset. DeepMine is a large-scale database in Persian and English. Two systems based on x-vector and ResNet and their scores fusion were investigated for this purpose. PLDA was used as backend for x-vector and cosine similarity for ResNet. The challenge baseline system EER was 10.67 % and the best result in our experiments was 8.20 % EER.

**Index Terms**: speaker verification, x-vector, ResNet, PLDA, scores fusion, SdSV

## 1. Introduction

As mentioned in the abstract, this report describes the Team22 submissions for Task 2 of the short-duration speaker verification (SdSV) challenge 2020. This was the first challenge using DeepMine dataset [1]. The main goal of the challenge is to evaluate new technologies for text-dependent (TD) and text-independent (TI) speaker verification (SV) in a short duration scenario. The challenge has two separate tasks: Task 1 is defined as speaker verification in text-dependent mode and Task 2 is speaker verification in text-independent mode. Both modes consist of trials where the enrollment and test utterances are from the same language (Persian). Unlike Task 1, Task 2 also includes cross-lingual trials where the enrollment utterances are in Persian and the test utterances are in English. There are no cross-gender trials in this challenge [2].

Our systems are based on Deep Neural Network (DNN) embedding: The x-vector [3], which was the challenge baseline, and the ResNet34 [4]. Good results have been reported from both architectures in short duration speaker verification [5, 6]. Source normalized LDA [7] and DNN based methods [8] can be used to reduce the language effects in cross-lingual trials. The structure of this document is as follows: In Section 2, the implementation consideration, including the datasets and the configuration of the systems are described. The results of the experiments are presented and discussed in Section 3.

## 2. Experimental Setup

### 2.1. Datasets

The evaluation dataset used for the challenge is DeepMine dataset. DeepMine is a large-scale database in Persian and English, with its current version containing more than 1850 speakers and 540 thousand recordings overall. It is the first large-scale speaker recognition database in Persian [1]. Each trial in this task contains a test segment of speech along with a model identifier which indicates one to several enrollment utterances. The net enrollment speech for each model is uniformly distributed between 3 to 120 seconds (after applying an energy-based VAD) [2].

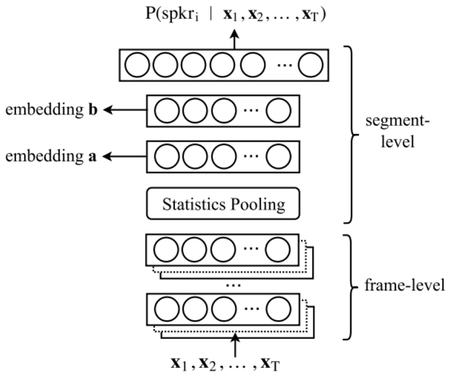
P(spkr$_i$ | $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$)
embedding **b**
embedding **a**
Statistics Pooling
segment-level
frame-level
$\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$

Figure 1: *Diagram of x-vector architecture. Embedding a or b is used as speaker embedding [9].*

The challenge adopted a fixed training condition where the system should only be trained using the following set: VoxCeleb1, VoxCeleb2, LibriSpeech, and DeepMine. For this Challenge, the DeepMine dataset is presented in three parts: Train, Enrollment, and Evaluation. The in-domain training data in Task 2contains text-independent Persian utterances from 588 speakers. Each trial contains a test segment of speech along with a model identifier which indicates one to several enrollment utterances. The net enrollment speech for each model is uniformly distributed between 3 to 120 seconds (after applying an energy-based VAD) [2].

### 2.2. x-vector

The x-vector is a TDNN based speaker embedding extracted from a speech sample [3]. The x-vector architecture are shown in Figure 1. We used the voxceleb recipe[1] from the Kaldi toolkit [10] and a pre-trained model[2] that was trained using VoxCeleb1 and VoxCeleb2 datasets. The training data artificially augmented with noises and reverberation. The voxceleb recipe uses a 7-layer x-vector. 30 MFCC coefficients extracted from a 25ms frame with 4 was form the150-dimensional input of network. An energy-based voice activity detector was used to discard non-speech frames. The 7-th layer (embedding b as shown in Figure 1) was used as speaker embedding and VoxCeleb1, LibriSpeech[3], and DeepMine datasets were used for LDA and PLDA.

Some other modifications such as training model from scratch (with VoxCeleb1, LibriSpeech, and DeepMine datasets) were tested but did not effective. Due to our time and Hardware resources constraints, we were unable to perform further experiments such as training model from scratch with

[1] https://github.com/kaldi-asr/kaldi/tree/master/egs/voxceleb/v2

[2] https://kaldi-asr.org/models/m7

[3] http://www.openslr.org/resources/12/train-clean-100.tar.gz

all training data or source normalized LDA to reduce the language effects in cross-lingual trials.

### 2.3. ResNet34

The ResNet is a well-known convolutional DNN and has good results in speaker recognition [5, 6]. We use Thin ResNet34 architecture as speaker embedding extractor based on [6] and a pre-trained model[1] that was trained with VoxCeleb1 and VoxCeleb2 datasets. A fixed length 2-second temporal segment, extracted randomly from each utterance was used during training. No data augmentation was performed during training, apart from the random sampling. Spectrograms are extracted with a hamming window of width 25ms and step 10ms. The 257-dimensional raw spectrograms are used as the input to the network. Mean and variance normalization (MVN) is performed by applying instance normalization to the network input. Since the VoxCeleb dataset consists mostly of continuous speech, voice activity detection (VAD) is not used in training and testing.

The output of model is a 512-dimensional fixed length embedding that was obtained from self-attentive pooling (SAP). The cosine similarity was used as backend for scoring the trials. For this purpose ten 4-second temporal samples crops at regular intervals from each test segment, and compute the similarities between all possible combinations (10×10 = 100) from every pair of segments. The mean of the 100 similarities is used as the score. Due to our time and Hardware resources constraints, we only use 2 samples as enrollment for each target speaker.

## 3. Results and Discussion

The results of the systems and their fusion on the progress set are displayed in Table 1. Line 1 is the SdSV baseline results that is an 8-layer x-vector topology with a 150-dimensional LDA and a PLDA as backend. The baseline system is trained using VoxCeleb1 and VoxCeleb2 datasets [2]. Line 2 is the x-vector topology result that is trained with VoxCeleb1 and VoxCeleb2. For backend, an LDA with 200 dimensions and a PLDA that was trained with VoxCeleb1, LibriSpeech and DeepMine is used. Although the in-domain Persian DeepMine dataset was used to reduce the language effects in cross-lingual trials but the result was not better than the baseline.

Line 3 is the thin ResNet34 topology result that was explained in Section 2.3. This system's result also was not better than the baseline. Line 4 is the result of the two systems scores fusion based on mean and reduce the baseline EER by 23 %.

In addition to the results announced on the competition website, some good and detailed results were published by the challenge organizers that was displayed in Table 2 for baseline and our best system, i.e., x-vector and ResNet34 fusion. These results are for evaluation set and show that compared to baseline, our system enhancement for male and Persian speakers was more than its enhancement for female and English speakers. For example, our system reduce the baseline EER by 32 % for male speakers' trials while EER reduction for female speakers' trials was 23 %.

Figure 2 compares the performance of three systems, i.e., the challenge baseline, our x-vector system, and the fusion system in a DET plot.

Table 1: *The baseline and our systems results for progress set.*

| | System | minDCF | EER (%) |
|---|---|---|---|
| 1 | baseline | 0.4319 | 10.67 |
| 2 | x-vector | 0.4122 | - |
| 3 | ResNet34 | 0.4777 | - |
| 4 | fusion | 0.3548 | 8.20 |

Table 2: *The baseline and our best system results for evaluation set in terms of EER %.*

| trials | baseline | x-vectror + ResNet34 |
|---|---|---|
| all trials | 10.67 | 8.21 |
| male | 8.26 | 5.62 |
| female | 11.71 | 8.98 |
| EN | 9.58 | 9.12 |
| FA | 6.14 | 5.57 |
| EN-male | 7.11 | 6.63 |
| EN-female | 10.58 | 9.47 |
| FA-male | 4.25 | 3.28 |
| FA-female | 6.53 | 6.10 |

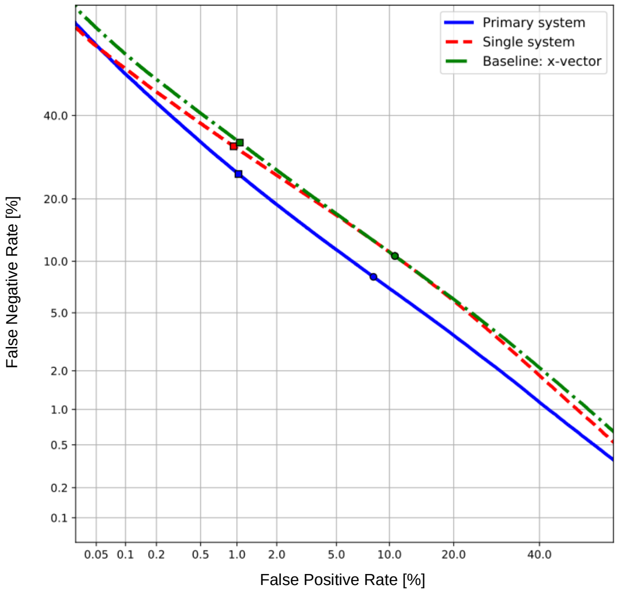

Figure 2: *Comparison of baseline, our x-vector syatem (single) and the fusion system (primary) for evaluation set.*

## 4. References

[1] H. Zeinali, L. Burget, and J. Černocký, "A Multi Purpose and Large Scale Speech Corpus in Persian and English for Speaker and Speech Recognition: the DeepMine Database," arXiv preprint arXiv:1912.03627, 2019.

[2] H. Zeinali, K. A. Lee, J. Alam, and L. Burget, "Short-duration Speaker Verification (SdSV) Challenge 2020: the Challenge Evaluation Plan," arXivpreprint arXiv:1912.06311, 2019.

[3] 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), 2018, pp. 5329-5333.

[1] https://github.com/clovaai/voxceleb_trainer

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770-778.

[5] 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.

[6] J. S. Chung, J. Huh, S. Mun, M. Lee, H. S. Heo, S. Choe, et al., "In defence of metric learning for speaker recognition," arXiv preprint arXiv:2003.11982, 2020.

[7] M. McLaren, M. I. Mandasari, and D. A. van Leeuwen, "Source normalization for language-independent speaker recognition using i-vectors," 2012.

[8] L. Li, D. Wang, A. Rozi, and T. F. Zheng, "Cross-lingual speaker verification with deep feature learning," in 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2017, pp. 1040-1044.

[9] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep Neural Network Embeddings for Text-Independent Speaker Verification," in Interspeech, 2017, pp. 999-1003.

[10] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, et al., "The Kaldi speech recognition toolkit," in IEEE 2011 workshop on automatic speech recognition and understanding, 2011.