# SYSTERM DESCRIPTION BY TEAM HHU-LB

## 1. INTRODUCTION

In this document, we provide a brief description for our speaker recognition system for FFSVC Challenge. We begin by giving a summary of the data used to train the various components of the system, followed by a description of the system components along with their hyper-parameter configurations. Finally, we report experimental results obtained with this system on our dev sets. Figure 1 shows a block diagram of our x-vector system. It's built using Kaldi (for x-vector extractor training) and the PLDA back-end scoring.

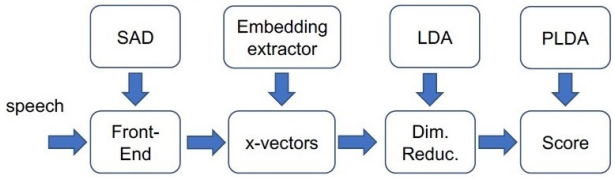


**Fig. 1**. block diagram of the FFSVC challenge x-vector system

## 2. DATA

The training sets which we use in our experiments are VoxCeleb-1+2[1][2], AIshell[3][4], HIMIA[5] and the dataset that FFSVC2020[6] provides.The challenge provides a training set with 120 speakers, and a development set with 35 speakers. The HIMIA dataset has 254 speakers in the training set and 42 speakers in the development set.

In order to expand the total amount of training data, and to deal with the noise problem during the recording of the original data set, we add noise using Kaldi to expand the training data. Then, we remove utterances that less than 60 frames after removing nonspeech frames, and throw out speakers with fewer than 10 utterances. The PLDA-based backends are trained on the same datasets.

## 3. CONFIGURATION

The features are 30 dimensional MFCCs from 25 ms frames every 10 ms spanning the frequency range 20Hz-7600Hz. Before dropping the non-speech frames using an energy based SAD, a short-time cepstral mean subtraction is applied over a 0.5-second sliding window.

For x-vector extraction, the network consists of layers that operate on speech frames, a statistics pooling layers that operate at segment-level and finally a softmax output layer. The nonlinearities are rectified linear units (ReLUs). The first 5 layers of the network at the frame level, with a time-delay architecture. Layers vary in size, from 512 to 1536, depending on the splicing context used. The statistics pooling layer receives the output of the final frame-level layer as input, aggregates over the input segment, and computes its mean and standard deviation. These segment-level statistics are concatenated together and passed to two additional hidden layers with dimension 512 and 300 (either of which may be used to compute embeddings) and finally the softmax output layer.

Prior to dimensionality reduction through LDA (to 180), x-vectors are unit-length normalized.For backend scoring, a Gaussian PLDA model with a full-rank subspace is trained using the x-vectors extracted from all the speech segments from the training sets, as well as one corrupted version randomly selected from babble, noise, music, reverb.

In this challenge, we use two main training methods. Firstly, we train a base model using the three data sets: Vox1, Vox2 and AISHELL together with their augmented data, and then using the data set provided by the FFSVC data set and HIMIA to finetune. In the process of fine-tuning, we fixed the parameters of the first five layers and train only the final classifier. The other is that, we use HIMIA data set and FFSVC data set to train the model directly. The optimal results of the two methods are different, which is reflected in the results

## 4. RESULT

In this section, we present the experimental results on dev set, and the final result on the eval set. Results are reported in terms of the equal error rate (EER) as well as minDCF. Table 1 summarizes the baseline results on the fine-tuning and training directly. Table 2 shows that the final result on eval set.

**Table 1**. Results of systerm about train and finetune

| system | minDCF | EER |
|---|---|---|
| Train directly | 0.521 | 5.57 |
| Finetune | 0.480 | 4.97 |

**Table 2**. Results on dev and eval set

| system | minDCF | EER |
|---|---|---|
| dev | 0.480 | 4.97 |
| eval | 0.591 | 5.84 |

## 5. REFERENCES


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

[2] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[3] Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*. IEEE, 2017, pp. 1–5.

[4] Jiayu Du, Xingyu Na, Xuechen Liu, and Hui Bu, "Aishell-2: transforming mandarin asr research into industrial scale," *arXiv preprint arXiv:1808.10583*, 2018.

[5] Xiaoyi Qin, Hui Bu, and Ming Li, "Hi-mia: A far-field text-dependent speaker verification database and the baselines," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7609–7614.

[6] Xiaoyi Qin, Ming Li, Hui Bu, Wei Rao, Rohan Kumar Das, Shrikanth Narayanan, and Haizhou Li, "The interspeech 2020 far-field speaker verification challenge," in *Interspeech 2020*, 2020.