# TEAM50 SYSTEM DESCRIPTION FOR NIST 2020 CTS CHALLENGE

*Team 50*

## ABSTRACT

We briefly describe the primary systems submitted by Team-50 to NIST 2020 CTS Challenge. In this paper, we report the performance on the progress set.

***Index Terms***— x-vector, CTS

## 1. INTRODUCTION

The 2020 CTS Challenge focuses on the speaker verification over conversational telephone speech. We submitted a fusion result with two single systems. For each system, we introduce the datasets used, features extraction and their configurations.

## 2. DATA SETS

The 2020 CTS is the open training condition, thus we used SRE04-10, SwithcBoard, Mixer6 telephone data, Vox-celeb 1&2 [1, 2] and Librispeech. All data is downsampled to 8KHz. Further, we made data augmentation with MUSAN [3] and RIR [4] with the twice size of the original data. The SRE18 and SRE19 [5] are used for adaptation in the backend strategies.

## 3. SINGLE SYSTEM

The two systems are training with Pytorch, and their acoustic feature are Fbank with Kaldi [6]. The first system includes a 14-layer FTDNN structure [7, 8] and an AM-Softmax loss function. The x-vector [9] is extracted from the 512-dimensional affine component of penultimate layer. The model is trained with SGD optimizer and an exponential decay scheduler. The details of FTDNN are as Table 1.

The second system includes a 101-layer Resnet structure and the AM-Softmax loss function. The relative configures are the same as the FTDNN model.

## 4. CONFIGURATIONS

The acoustic feature is a 64-dimensional Fbank with energy. We made the voiced activity detection (VAD) and cepstral mean normalization (CMN) firstly. And then we extracted all chunks based on the chunk length of 200 frames. Each chunk does not overlap each other. After extracting the x-vectors, we firstly trained the linear discriminant analysis (LDA) with the normalized embeddings to reduce the dimensions of the x-vectors. Subsequently, the probability linear discriminant analysis (PLDA) [10] was trained with the 150-dimensional embedding. Finally, the raw scores from the likelihood comparison of the PLDA are enhanced with the adaptive symmetric normalization (AS-Norm) [11].

## 5. RESULTS

The results of the above two systems are as Table 2. The final submission is adapted from their average fusion.

**Table 2**. All submissions of Team 50.

| system | EER(%) | min cost | act cost |
|---|---|---|---|
| FTDNN | 3.33 | 0.147 | 0.590 |
| ResNet | 3.02 | 0.162 | 0.334 |
| Fusion | 2.66 | 0.133 | 0.258 |

## 6. REFERENCES

[1] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: a large-scale speaker identification dataset," in *INTERSPEECH*, 2017.

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

[3] David Snyder, Guoguo Chen, and Daniel Povey, "MUSAN: A Music, Speech, and Noise Corpus," *arXiv:1510.08484v1*, 2015.

[4] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5220–5224.

[5] Seyed Omid Sadjadi, Craig Greenberg, Elliot Singer, Douglas Reynolds, Lisa Mason, Jaime Hernandez-Cordero, et al., "The 2019 nist speaker recognition evaluation cts challenge," in *Speaker Odyssey*, 2020, vol. 2020, pp. 266–272.

**Table 1**. 14-layer Factorized TDNN Framework

| | Layer Type | Context Factor 1 | Context Factor 2 | Skip Conn. from Layer | Size | Inner Size |
|---|---|---|---|---|---|---|
| 1 | TDNN-ReLU-BN | t-2:t+2 | | | 512 | |
| 2 | FTDNN-ReLU-BN | t-2,t | t,t+2 | | 512 | 256 |
| 3 | FTDNN-ReLU-BN | t | t | | 512 | 256 |
| 4 | FTDNN-ReLU-BN | t-3,t | t,t+3 | | 512 | 256 |
| 5 | FTDNN-ReLU-BN | t | t | 3 | 512 | 256 |
| 6 | FTDNN-ReLU-BN | t-3,t | t,t+3 | | 512 | 256 |
| 7 | FTDNN-ReLU-BN | t-3,t | t,t+3 | 2,4,6 | 512 | 256 |
| 8 | FTDNN-ReLU-BN | t-3,t | t,t+3 | | 512 | 256 |
| 9 | FTDNN-ReLU-BN | t | t | 3,5,7 | 512 | 256 |
| 10 | FTDNN-ReLU-BN | t-3,t | t,t+3 | | 512 | 256 |
| 11 | FTDNN-ReLU-BN | t | t | 6,8,10 | 512 | 256 |
| 12 | FTDNN-ReLU-BN | t-3,t | t,t+3 | | 512 | 256 |
| 13 | FTDNN-ReLU-BN | t | t | 7,9,11 | 512 | 256 |
| 14 | Dense-ReLU-BN | t | t | | 1024 | |
| 15 | Pooling(mean+stddev) | Full Seq. | | | 2*1024 | |
| 16 | Dense(x-vector)-ReLU-BN | [0,T] | | | 512 | |
| 17 | Dense-ReLU-BN | [0,T] | | | 512 | |
| 18 | Dense-AMSoftmax | [0,T] | | | Num.Spks. | |


[6] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The Kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.

[7] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[8] Daniel Povey, Gaofeng Cheng, Yiming Wang, Ke Li, Hainan Xu, Mahsa Yarmohammadi, and Sanjeev Khudanpur, "Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks," in *Interspeech*, 2018, pp. 3743–3747.

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

[10] Sergey Ioffe, "Probabilistic linear discriminant analysis," in *European Conference on Computer Vision*. Springer, 2006, pp. 531–542.

[11] Sandro Cumani, Pier Domenico Batzu, Daniele Colibro, Claudio Vair, Pietro Laface, and Vasileios Vasilakakis, "Comparison of speaker recognition approaches for real applications," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.