System Description for Hi-Fi System

Ruida Li

Ant Group

Abstract

We follow the rule of Task 1 SV fixed track and only the CN-Celeb data are used as the development set. We use ECAPA-TDNN as the architecture and a weighted average of Circle loss and AM-Softmax loss is applied as the training criterion. The experimental result shows that our method achieves 6.8260% EER and 0.3603 minDCF in Task 1 SV fixed track.

1. Data

Following the rule of Task 1 SV fixed track, we only use the CN-Celeb1[1] and CN-Celeb2[1] data as the development set. And we note that the the data in development set and evaluation do not overlap.

2. Models

For data preprocessing, we generate 80-dimensional log Melfilterbanks(Fbanks) from 25ms windows with10ms frameshift. The length of variable-length training samples varies from 2s to 4s, corresponding to 200 frames to 400 frames. And we exploit the cepstral mean normalization on the spectrogram. The feature extraction step is handled with Kaldi toolkit[2].

For the embedding network, we employ the widely used ECAPA-TDNN[3] architecture. ECAPA-TDNN consists of a 1-dimensional Squeeze-Excitation Res2Blocks, with a multi-layer feature aggregation to concatenate the information of different hierarchical levels, and a channel- and context-dependent statistics pooling layer is employed to extract statistic features, followed by a fully-connect layer to compute logits. In our experiment, we use 1024 channels in the convolutional frame layers; and set the nodes of the final fully-connected layer as 192, namely the final embedding is192-dimensional.

The loss function we used is a weighted average of Circle loss and AM-Softmax loss. An annealing strategy is introduced to improve the stability of the training process and boost the convergence.

Our model is trained with Adam optimizer[4] and the batch size is set to 64. The learning rate is started with 1e-3 and continuously decays by 0.95 epoch by epoch. We finish our experiment after 80 epochs. The network training and embedding extraction are implemented by an existing PyTorch tookit[5].

3. Results

The evaluation performance is measured by Equal Error Rate(EER) and the minimum normalized detection cost(minDCF) with $P_{target} = 1e - 2$. Experimental result on Cn-Celeb is shown in Table 1.

As illustrated in table1, our method achieves 6.8260% EER and 0.3603 minDCF in Task 1 SV fixed track.

Table 1: The experimental result on CN-Celeb

	EER(%)	**minDCF**
Our Method	6.8260	0.3603

4. References

[1] Lantian Li, Ruiqi Liu, Jiawen Kang, Yue Fan, Hao Cui, Yunqi Cai, Ravichander Vipperla, Thomas Fang Zheng, and Dong Wang, "CN-Celeb: multi-genre speaker recognition," *Speech Communication*, 2022.

[2] 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, number CONF.

[3] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.

[4] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[5] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer, "Automatic differentiation in pytorch," 2017.