System Description for T067

Jianchen Li

School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

lijianchen_hit@outlook.com

Abstract

This report describes our submission to the *Fixed Track* of speaker verification task. We elaborate the system input, model structure, loss function, training details, and back-end scoring.

1. Data

We use *CN-Celeb.T* to train the network, which consist of 632,740 utterances from 2,793 speakers. The following strategies are used to augment the training data:

- Reverberated [1]
- Augment with Musan noise [2]
- Augment with Musan music [2]
- Augment with Musan speech [2]
- SpecAugment [3]

where the ratios of frequency and time masking in SpecAugment are 0.05 and 0.2, respectively. After these augmentations, 3,163,680 utterances from 2,793 speakers were generated to extract acoustic features.

2. Models

2.1. Acoustic Features

For the acoustic features, 81-dimensional log-mel filter banks are extracted within a 25ms sliding window for every 10ms. Cepstral mean normalization (CMN) is performed within a 3-second sliding window. Voice activity detection (VAD) is used to remove the silent segments. During training, each utterance is cut into 200-frame chunks to keep the same input length.

2.2. Backbone Network

ResNet34 [4] is adopted as the speaker embedding netowrk $f(\cdot)$. This network uses 2-dimensional CNN layers and several shortcut connections. Multi-head attention pooling layer [5] is used to aggregate the frame-level embeddings. The details of the used ResNet34 is shown in Table 2.

2.3. Loss Function

AM-softmax [6] loss is used to optimize the embedding network. It introduces the additive margin into the softmax function to further suppress intra-speaker variability. The formulation of AM-softmax loss is:

$$\mathcal{L}_{\text{ams}} = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{e^{s\cdot\left(\cos\delta_{y_i,i}-m\right)}}{e^{s\cdot\left(\cos\delta_{y_i,i}-m\right)}+\sum_{c=1,c\neq y_i}^{C}e^{s\cdot\cos\delta_{c,i}}}, \tag{1}$$

where s denotes the scaling factor, n denotes the batch size, m denotes the margin, C denotes the total number of speakers of the training set, y_i is the ground-truth speaker label of $\mathbf{x}_i^s$, and $\delta_{c,i}$ is the angle between the weight vector $\mathbf{w}_c$ and the embeddings $f\left(\mathbf{x}_i\right)$.

2.4. Training Details

The stochastic gradient descent (SGD) optimizer with a momentum of 0.9 and weight decay of 1e-3 is used to optimize the network. The learning rate is initialized to 0.02 and updated by using ReduceLROnPlateau scheduler with a frequency of validating every 8,000 iterations. The batch size is 256. The scaling factor s and margin m of AM-softmax loss are 30 and 0.2, respectively. Our experiments are conducted using ASV-subtools [7].

2.5. Backend

After training, the 256-dimensional embedding of each utterance is extracted by the network. The embeddings are subtracted from their mean on the training set. For simplicity, we use cosine similarity for scoring. Moreover, we use speaker-wise adaptive score normalization (AS-Norm) to calibrate the scores, which greatly enhanced the performance.

3. Results

Table 1 shows the performance of the ResNet34 model on the official *CN-Celeb.E* evaluation set.

4. References

[1] 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.

[2] David Snyder, Guoguo Chen, and Daniel Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[3] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[5] India Massana M À, Safari P, and Hernando Pericás F

Table 1: Results of the ResNet34 model.

Model	EER (%)	minDCF	Rank
ResNet34	8.7580	0.4279	16

Table 2: The ResNet34 architecture, $\mathbf{C}\,(\text{kernal size}, \text{stride})$ denotes the convolutional layer, $\mathbf{S}\,(\text{kernal size}, \text{stride})$ denotes the shortcut convolutional layer, $[\cdot]$ denotes the residual block.

Layer Name	**Output**	**Structure**
Conv1	$32 \times 81 \times L$	$\mathbf{C}\,(3 \times 3, 1)$
Residual Layer 1	$32 \times 81 \times L$	$\begin{bmatrix} \mathbf{C}\,(3 \times 3, 1) \\ \mathbf{C}\,(3 \times 3, 1) \end{bmatrix} \times 3$
Residual Layer 2	$64 \times 41 \times \frac{L}{2}$	$\begin{bmatrix} \mathbf{C}\,(3 \times 3, 1) \\ \mathbf{C}\,(3 \times 3, 1) \\ \mathbf{S}\,(1 \times 1, 2) \end{bmatrix} \begin{bmatrix} \mathbf{C}\,(3 \times 3, 1) \\ \mathbf{C}\,(3 \times 3, 1) \end{bmatrix} \times 4$
Residual Layer 3	$128 \times 21 \times \frac{L}{4}$	$\begin{bmatrix} \mathbf{C}\,(3 \times 3, 1) \\ \mathbf{C}\,(3 \times 3, 1) \\ \mathbf{S}\,(1 \times 1, 2) \end{bmatrix} \begin{bmatrix} \mathbf{C}\,(3 \times 3, 1) \\ \mathbf{C}\,(3 \times 3, 1) \end{bmatrix} \times 6$
Residual Layer 4	$256 \times 11 \times \frac{L}{8}$	$\begin{bmatrix} \mathbf{C}\,(3 \times 3, 1) \\ \mathbf{C}\,(3 \times 3, 1) \\ \mathbf{S}\,(1 \times 1, 2) \end{bmatrix} \begin{bmatrix} \mathbf{C}\,(3 \times 3, 1) \\ \mathbf{C}\,(3 \times 3, 1) \end{bmatrix} \times 3$
Pooling	256	Multi-head Attention Pooling
FC 1	256	Fully Connected Layer
FC 2	256	Fully Connected Layer
AM-softmax	2775	Fully Connected Layer

J, "Self multi-head attention for speaker recognition," in *Proc. Interspeech 2019*, 2019, pp. 4305–4309.

[6] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[7] Fuchuan Tong, Miao Zhao, Jianfeng Zhou, Hao Lu, Zheng Li, Lin Li, and Qingyang Hong, "Asv-subtools: Open source toolkit for automatic speaker verification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6184–6188.