System Description of Fixed Track for T113

Hang-Rui Hu, Jian-Tao Zhang
{hhr, zhangjiantao}@mail.ustc.edu.cn

May 30, 2022

Abstract

This report describes our submission to the Fixed Track of the CN-Celeb Speaker Recognition Challenge 2022 (CNSRC 2022). CN-Celeb is the most real-world complex speaker recognition dataset and suffers from label noise, speaking style variations, cross-channel problems and long-short test scenarios. We will describe our system from several parts, including data processing, network structures, label noise filtering and score normalization. Due to time constraints, we only trained a few small models for this competition. The minDCF of our final submission is 0.3999, and our final system is mainly a fusion of 2 models, of which the best single model is ResNet34_32 with a minDCF of 0.4123.

1 System Description

1.1 Data Processing

1.1.1 Training data

We used the updated CN-Celeb1 and CN-Celeb2 as our dataset. It covers 11 genres and the total amount of speech waveforms is 1356 hours. The entire dataset was split into two parts: the first part CN-Celeb(T) involves 632, 736 utterances from 2793 speakers and was used as the training set; the second part CN-Celeb(E) involves 18, 579 utterances from 200 speakers and was used as the evaluation set. In our experiments,we used the traditional Kaldi-based method(offline augmentation) to extract acoustic features.

1.1.2 Data Augmentation

We augmented the CN-Celeb(T) data with reverberation,noise, music, and babble,which based on the Kaldi CNCeleb recipe and combined them with the clean data. After the augmentation,3,163,680 utterances from 2793 speakers were generated to extract acoustic features

1.1.3 Features

We extracted both 41-dimensional and 81-dimensional log Mel filter bank energies based on Kaldi. The window size is 25 ms,and the frameshift is 10ms. We also applied voice activation detection(VAD) to remove nonspeech frames. Besides,we removed features that was less than 5s (500 frames) per utterance after removing silence frames. ,and throwed out speakers with fewer than 8 utterances. All features were cepstral mean normalized in our training modes.

1.2 Network Structures

We mainly uses **ResNet** as our backbone as in (cite abp). All the early experiments were conducted on ResNet18_16, where 18 denotes the layers and 16 denotes the base channels. After the training strategy is determined, we increase the depth and width of the network to get better performance. Since our models were not large enough to significantly overfit the training data, we did not apply additional regularization strategies to them.

We also tried two other backbones: **ECAPA-TDNN** and **RepVgg**. Note that we do not down-sample the temporal dimension in the convolution layers of RepVgg. However, due to the lack of time,

the training of these two models may be insufficient, so their performance is much worse than the ResNet system with comparable parameters.

1.3 Label Noise Filtering

We observed many speech segments in CN-Celeb do not contain the current speaker's voice, which may be composed of real-world noise or other speaker's voice, etc.

To improve the robustness to label noise, we first trained a ResNet18_16 with sub-center AAM loss (s=30, m=0.1, K=3 or 10), to automatically cluster within each class such that hard samples and noisy samples are separated away from the dominant clean samples. After convergence, we treated the samples in the sub-center with too small proportions (less than 10% or 20%) of the current class and too large angles (more than 40° or 80°) from the dominant sub-center as noisy samples.

Then, we directly dropped these high-confidence noisy samples, and all models in our systems were trained with the remaining clean samples, to further enhance intra-class compactness.

1.4 Training Protocol

All networks are implemented using the PyTorch framework. The mini-batch for training is set to 128 or 256. All networks are optimized using Adam optimizer and CosineAnnealing schedule with a maximum learning rate of 1e-3 and a cycle size of 20k iterations. A weight decay of 2e-5 is applied to all parameters except the norm layers and bias layers. All networks are trained with AAM-Softmax loss with a scale of 15˜30 and a margin of 0˜0.3 for 5˜10 cycles.

1.5 Scoring

Length-normalized 256-dimensional speaker embeddings are extracted for all systems, and the trial scores are produced using the simple cosine distance. All scores are normalized with AS-Norm using top 400 imposter scores, to improve the robust to scenarios changes. To save time, we select the length-normalized AAM weights as the imposter cohort, and we observe that these classifier weights are almost the same as the speaker-wise average embeddings.

2 Results

As shown in Tabel.1, the workhorse model in our system is ResNet. RepVgg and ECAPATDNN were not sufficiently trained due to lack of time, but they still participated in the final model fusion with small weights.

Table 1: main results in CN-CELEB

ID	Model	Features	minDCF	EER(%)	Remark
R1	ResNet18_16	fbank_41	0.488	8.9	18_16 denotes the layers and base channels
R2	ResNet34_32	fbank_41	0.450	8.5	
R3	ResNet34_**64**	fbank_41	0.419	8.0	
R4	ResNet34_32	fbank_**81**	0.412	7.8	
V1	RepVgg_b1_32	fbank_41	0.441	8.7	
V2	RepVgg_b2_32	fbank_41	0.448	8.8	
E1	ECAPA_TDNN_512	fbank_41	0.481	9.1	512 denotes the base channels
E2	ECAPA_TDNN_1024	fbank_41	0.472	10.2	
	R3+R4		0.402	7.53	fusion weights are 0.8 and 1
	R3+R4+V1+E2		0.399	7.56	fusion weights are 0.8, 1, 0.3 and 0.2