The T057 CNSRC2022 System Description

Wei Ju
e-mail: juwei@oppo.com

***Abstract* – This report describes the submission of T057 to first CN-Celeb speaker recognition challenge(CNSRC 2022) at Odyssey 2022.The final submission is a combination of three systems. The System-1 is ECAPA-TDNN, And System-2 and System-3 are the fast Resnet34 with different feature. This report explores several parts, including network structures, large margin fine tuning, and back-end refinement. The MinDCF of our submission is 0.3445, and the corresponding EER is 7.3440%.**

***Keywords* – speaker recognition, speaker verification, CN-Celeb**

I. System Description

For the Task 1 speaker verification, both Fixed Track and Open Track, we submitted the same result.

A. Datasets and Data Augmentation

1) Training Data: This database contains two datasets: CN-Celeb1 and CN-Celeb2. The statistics of the two datasets are shown in CNSRC 2022 Evaluation Plan, and more details can be found in [1]. In conclusion, there are 2753 speakers in total for training. Data augmentation is widely used to improve speaker embedding robustness. We used data augmentation just for System-1 but System-2 and System-3 because of the Sunine [2] author mentioned in https://gitlab.com/csltstu/sunine/-/issues/4. In System-1, Each speech segment was speed perturbed by 0.95 or 1.05 factor, then time domain SpecAugment [3], reverberation and noise [4] also adopted, all the data augmentation methods conducted on the fly. we concatenated different types of augmentation, and in total, the training data size is expanded 5 fold.

2) Features: We extracted 80-dimensional log Mel filter bank for both System-1 and System-3, 256-dimensional spectrogram for System-2. The window size is 25 ms, and the frameshift is 10 ms. Without extra voice activation detection (VAD). Because the datasets have lots of short durations audio, in the first, we concatenated five utterances as a new utterance, and then, the speech segments were sliced to 2 seconds and augmented on the fly in the online training mode mentioned before. These features were extracted based on torchaudio. All features were cepstral mean normalized in a sentence level.

B. Network Structures

1) Backbone: Our backbones include two types of state-of-the-art models:

ECAPA-TDNN: ECAPA-TDNN [5] is a TDNN architecture likes the previous popular x-vector [6], but 1)Incorporating a channel-and context-dependent attention system in the statistics pooling layer 2)Introduces a 1-dimensional variant of Squeeze-Excitation (SE) blocks [7] to inject global information in frame-level layers of the network. 3) multi-layer feature aggregation and feature summation allows the network to efficiently exploit knowledge learned in the preceding layers. The ECAPA-TDNN architecture is depicted in Figure 1, and the configuration is default according to [5].

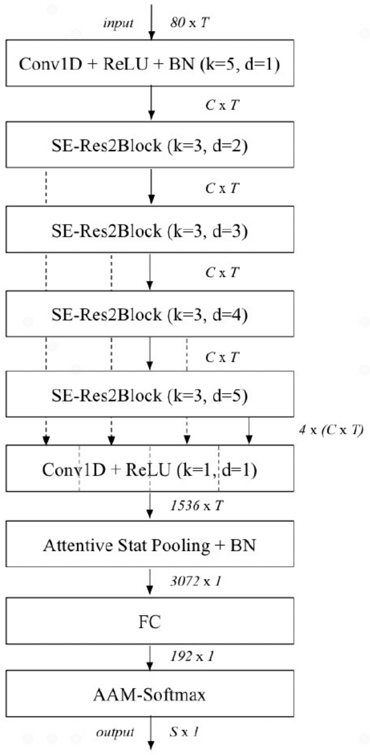

Fig. 1. Network topology of the ECAPA-TDNN baseline system.

ResNet34: ResNet [8] is one of the most classical convolutional neural networks. We adopt the fast ResNet34 according to [9], the number of the channel is small then the classical ResNet34, the size of filter is [32, 64, 128, 256] and the size of kernel is [3, 4, 6, 3].

2) Pooling Method: The traditional pooling method is temporal average pooling and temporal statistics pooling, it aims to aggregate the different frames to an utterance level embedding equally. But different frame has different weight to speaker embedding may be reasonable. Our networks all used attentive statistics pooling [10], the size of attention channel for ECAPA-TDNN and ResNet both 128.

3) Loss Function: Large margin loss function becomes a basic configuration in speaker recognition field recently. The three networks are use AM-Softmax [11] to improve performance.

C. Training Protocol

We used the Sunine and Speechbrain [12] as our training toolkit, both are based on Pytorch [13]. Sunine used for ResNet34 and Speechbrain used for ECAPA-TDNN. The ECAPA-TDNN system was trained through two stages. We used Adam optimizer [14] with a weight decay of 2e-5 and cyclic learning rate scheduler [15]. In the first stage, the max learning rate is 1e-2 and the base is 1e-8, the step size is 65000, we used 4 GPUs V100 with 256 mini-batch with a 200 frames chunk size, the margin of the loss function is 0.2. After 5 epochs training, we entry the large margin fine tuning [16] stage, we increased the training chunk size from 200 frames to 600 frames, and increased the margin from 0.2 to 0.5, and because of the GPU memory, we decreased the batch size to 128. We also decreased the max learning rate in 1e-4. This stage is also trained in 5 epochs. The ResNet34 system training protocol is different from ECAPA-TDNN, we just trained it in on one stage, and most settings is default in the Sunine recipes. We used 1 GPU V100, and because of the GPU memory, the batch size of System-3 is 256 but 32 for System-2.

D. Back-end

After the ECAPA-TDNN and the two ResNet34 Systems training completed, the classifiers were removed. There are three speaker embeddings were extracted from these networks, 192 dimensional from ECAPA-TDNN and two 256-dimensional from both ResNet34. Firstly, we use the αQE to take full advantage of multi enroll segment, like [17] mentioned. Secondly, we used AS-Norm [18] to calibrate our scores, for the imposter, we randomly select one utterance from every training speaker, and resulted in 2753 cohorts total, and the top 800 imposter scores used to calculate the mean and variance for every trial.

E. Results

1) Baseline System Ablation Study: The performance was evaluated by the Equal Error Rate (EER) and the minimum Decision Cost Function (MinDCF) as required by CNSRC2022. The results of our baseline system ablation study were showed in the three tables. For convenience, we took ECAPA-TDNN System to introduce. The large margin fine tuning improve the MinDCF from 0.5237 to 0.4863, and the EER decrease slightly. It is surprised that the αQE improve the speaker verification performance drastically, the MinDCF from 0.4863 to 0.4177, and the EER from 9.83% to 8.52%, and it also works in another systems, both the ResNet34 System performance improved obviously.

TABLE 1
Ablation Study on ECAPA-TDNN System

	EER(%)	MinDCF
ECAPA-TDNN	10.01	0.5237
+Fine-tuning	9.83	0.4863
++αQE	8.52	0.4177
+++AS-Norm	**8.26**	**0.3964**

TABLE 2
Ablation Study on ResNet34(Spec) System

	EER(%)	MinDCF
ResNet34(Spec)	10.56	0.5242
+αQE	9.30	0.4338
++AS-Norm	**8.77**	**0.4066**

TABLE 3
Ablation Study on ResNet34(Fbank) System

	EER(%)	MinDCF
ResNet34(Fbank)	9.94	0.5261
+αQE	8.80	0.4323
++AS-Norm	**8.31**	**0.4041**

To explore the best value of α, we done some experiments shows in Table4, we found that α=3 for ECAPA-TDNN embedding, α=5 for ResNet34(Spec) embedding, α=4 for ResNet34(Fbank) embedding are the best choice.

TABLE 4
Exploration of the best value of α

	α	EER(%)	MinDCF
ECAPA-TDNN	2	**8.44**	0.4206
	3	8.52	**0.4117**
	4	8.62	0.4180
	5	8.76	0.4193
	6	8.84	0.4208
ResNet34(Spec)	3	**9.19**	0.4364
	4	9.23	0.4347
	5	9.30	**0.4338**
	6	9.36	0.4339
ResNet34(Fbank)	3	**8.75**	0.4332
	4	8.80	**0.4323**
	5	8.89	0.4334

2) Fusion Performance: Table 5 shows the best subsystems of ours and the final result of fusion system. It is worth mentioning that the ResNet34 is obviously worse than ECAPA-TDNN at the beginning, but after αQE and AS-Norm in the back-end, the three systems have comparable performance. After score fusion, the MinDCF and EER decreased into 0.3445 and 7.34%, it demonstrates the complementarity of these systems,and we submitted the fusion result to the CNSRC2022.

TABLE 5
Our Submissions to CNSRC2022

	EER(%)	MinDCF
S1	8.26	0.3964
S2	8.77	0.4066
S3	8.31	0.4041
Fusion	**7.34**	**0.3445**

II. Conclusions

In the CNSRC2022, due to time constraints, some other methods were not explored, and we submitted the some result both in fixed Track and open Track. The final MinDCF and ERR is 0.3445 and 7.34%, the performance is not very competitive, and should be improved even more in the future.

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," arXiv preprint arXiv:2012.12468, 2020.

[2] Lantian Li, Yang Zhang, Dong Wang, "Sunine: THU-CSLT Speaker Recognition Toolkit," 2022.

[3] Park D S, Chan W, Zhang Y, et al. Specaugment: A simple data augmentation method for automatic speech recognition[J]. arXiv preprint arXiv:1904.08779, 2019.

[4] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. 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] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN:Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in Proc. Interspeech,2020.

[6] Snyder D, Garcia-Romero D, Sell G, et al. X-vectors: Robust dnn embeddings for speaker recognition[C]//2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2018: 5329-5333.

[7] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation networks," in Proc. IEEE/CVF CVPR, 2018, pp. 7132–7141.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[9] Chung J S, Huh J, Mun S, et al. In defence of metric learning for speaker recognition[J]. arXiv preprint arXiv:2003.11982, 2020.

[10] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," arXiv preprint arXiv:1803.10963, 2018.

[11] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2019, pp. 4690–4699

[12] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "Speechbrain: A general-purpose speech toolkit," 2021.

[13] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga et al., "Pytorch:An imperative style, high-performance deep learning library," Advances in neural information processing systems, vol. 32, pp. 8026–8037, 2019.

[14] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," Proc. ICLR, 2014.

[15] L. N. Smith, "Cyclical learning rates for training neural networks," in IEEE WACV, 2017, pp. 464–472.

[16] Thienpondt J, Desplanques B, Demuynck K. The IDLab VoxSRC-20 submission: Large margin fine-tuning and quality-aware score calibration in DNN based speaker verification[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 5814-5818.

[17] Gusev A, Vinogradova A, Novoselov S, et al. SdSVC Challenge 2021: Tips and Tricks to Boost the Short-Duration Speaker Verification System Performance[C]//Proceedings of the Interspeech. 2021: 2021-1737.

[18] W. Wang, D. Cai, X. Qin, and M. Li, "The dku-dukeece systems for voxceleb speaker recognition challenge 2020," arXiv preprint arXiv:2010.12731, 2020.