System Description for CNSRC 2022

Team 126

Affiliation and Department

E-mail address

Abstract

This paper gives a brief description of our submission systems for the three tracks in CN-Celeb Speaker Recognition Challenge 2022 (CNSRC 2022), i.e., Task 1. Speaker Verification (SV) Fixed Track and Open Track, Task 2. Speaker Retrieval (SR) Open Track.

1. Data

1.1. Task 1 SV *fixed track*

In the fixed track, only the development set of *CN-Celeb v1* and *v2* are allowed for training the systems. For data augmentation, publicly available [1] MUSAN [1] and Reverberation datasets are utilized, in which the latter involves convolving room impulse responses (RIR) with audio [2].

1.2. Task 1 SV and Task 2 SR *open track*

Vox-Celeb v1 [3] and *v2* [4] are utilized to pretrain the models and the pretrained models are fine-tuned on the development set of *CN-Celeb*.

2. Models

The architecture of our model is similar to [5], which integrates several 2D convolutional layers with ECAPA-TDNN [6]. The architecture is shown in Fig. 1. The model input is 64-dimensional Log Mel filterbank (*MFbank*). The Conv2D layers ahead ECAPA-TDNN process the input *MFbank* spectrum as an one channel image. There is an instance normalization layer (IN) [7] at the top of the model, which normalizes each frequency bin of the spectrum across time dimension.

Different 2D convolutional structures are tried for the "Conv2D block" in our experiments, including ResBaseBlock, ResBottleBlock [8], ResNextBlock [9], Res2Block [10], SKblock [11]. The number of output channel of the Conv2D is represented as C_1. The output of Conv2D is flatten first and processed by a Conv1D layer, giving C_2 channels. In our experiments, C_1 and C_2 are set as 32 and 512 respectively.

3. Experimental settings

3.1. Training

In the training process, the model is trained to classify the speaker identities from the development set of *CN-Celeb v1* and *v2* using AM-softmax [12] or AAM-softmax [13] loss. The hyperparameter scale and margin for the angular loss are 30 and 0.2, respectively. The speaker embedding dimension used in our experiments is 128. A dropout layer is added before the speaker embedding layer with ratio 0.3.

[1] http://www.openslr.org/

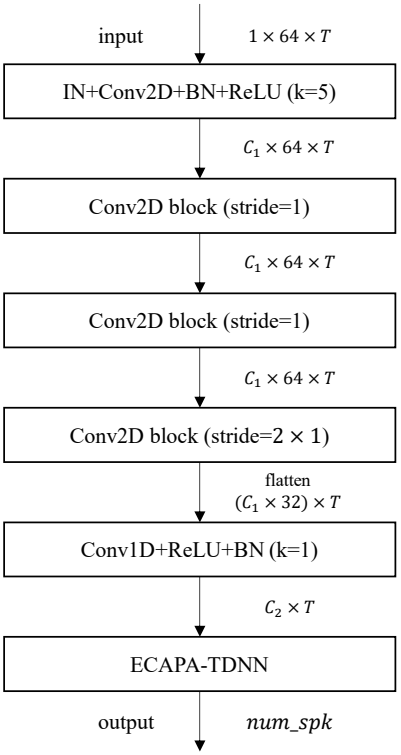

Figure 1: The structure of our model.

A 2-second duration segment is randomly cropped from each input speech waveform and the cropped input is augmented by adding noise with a probability of p_{aug}. The noise is randomly selected from MUSAN or RIR with probabilities 0.7 and 0.3 respectively. The noise is added on the original signal with SNR 0.01-0.5. Frequency and time masks [14] are applied on the transferred *MFbank* spectrum before feeding it into the model. In the masking process, 0-0.2 ratio of the spectrum is masked as 0 in time or frequency dimension.

3.1.1. Fixed track

In this track, only the development set of *CN-Celeb v1* and *v2* are allowed for model training. The model is trained with Adam optimizer [15] for 30 epochs in total. The learning rate is initialized as 0.001, and it is decayed by 10 at the epoch 15 and 25.

After training for 30 epochs, the model is fine-tuned for five epochs with a larger margin as described in [16]. The input duration is increased to 4-second and the margin in the loss

Table 1: Results of the Task 1. SV fixed track. In the original ECAPA-TDNN, three SE-Res2Blocks are used. We add the number of the blocks to 6, and concatenate the 2,4,6th blocks' outputs. The model using modified version of ECAPA-TDNN is marked with *.

Conv2D block	p_{aug}	Loss	minDCF	EER
SK	0.7	AAM	0.4491	8.05
SK	0.8	AAM	0.4503	8.34
ResNext	0.7	AAM	0.4648	8.25
ResNext	0.8	AAM	0.4559	8.06
ResNext*	0.7	AM	0.4685	8.66
SK*	0.7	AM	0.4484	8.35
SK*	0.7	AAM	0.4436	8.12
model fuse			0.4263	7.59

function (AM or AAM-softmax) is increased to 0.4.

3.1.2. Open track

The model is pre-trained on *Vox-Celeb v1* and *v2* first, and then trained on *CN-Celeb v1* and *v2* following the process described above. Due to the time limitation, we did not conduct the large margin fine-tuning in this track.

3.2. Evaluation

3.2.1. SV

The utterances in each trial is divided into 4-second duration segments, with 2-second overlap between each adjacent segments. The averaged embeddings' cosine similarities between the test segments and enrollment segments is normalized in 0-1, which can represent the probability that the test utterance has the same speaker as the enrollment utterance.

4000 utterances are randomly sampled from the training set and utilized as an imposter cohort set for score normalization. Adaptive S-Norm [17] is utilized in our experiments with $topk = 300$.

3.2.2. SR

The cosine similarities between the embeddings of enrollment utterances and utterances in the pool are calculated, and we select the utterances with top 10 scores as the retrieval outputs.

4. Results

The results of the speaker verification task for the fixed tack and open track are shown in Table 1 and 2, respectively. The model fusing result is given by averaging the scores from each model, and the fused results are submitted to CNSRC 2022.

The result of speaker retrieval task is achieved by the fused score using models in Table 2, and gives mAP 0.3563.

5. Conclusion

The report gives a brief description of our submission to CNSRC 2022 and we believe there is a considerable improvement that can be achieved in future research.

Table 2: Results of the Task 2. SV open track. The model using modified version of ECAPA-TDNN is marked with *.

Conv2D block	p_{aug}	Loss	minDCF	EER
ResBase*	0.7	AAM	0.4336	7.34
ResBottle*	0.7	AAM	0.4405	7.37
ResNext*	0.7	AAM	0.4250	7.31
Res2*	0.7	AAM	0.4279	7.27
SK*	0.7	AAM	0.4396	7.31
model fuse			0.4125	6.77

6. References

[1] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[2] 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.

[3] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," *Proc. Interspeech 2017*, pp. 2616–2620, 2017.

[4] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *Proc. Interspeech 2018*, pp. 1086–1090, 2018.

[5] J. Thienpondt, B. Desplanques, and K. Demuynck, "Integrating frequency translational invariance in tdnns and frequency positional information in 2d resnets to enhance speaker verification," in *Proc. Interspeech2021*, 2021, pp. 2302–2306.

[6] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech 2020*, 2020, pp. 3830–3834.

[7] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016.

[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] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[10] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 2, pp. 652–662, 2019.

[11] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 510–519.

[12] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[13] 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.

[14] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *Proc. Interspeech 2019*, pp. 2613–2617, 2019.

[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015.

[16] J. Thienpondt, B. Desplanques, and K. Demuynck, "The idlab voxsrc-20 submission: Large margin fine-tuning and quality-aware score calibration in dnn based speaker verification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5814–5818.

[17] Z. N. Karam, W. M. Campbell, and N. Dehak, "Towards reduced false-alarms using cohorts," in *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2011, pp. 4512–4515.