

ChinaTelecom System Description to VoxCeleb Speaker Recognition Challenge 2023

Mengjie Du, Xiang Fang, Jie Li

China Telecom Corporation Ltd. Data&AI Technology Company, Beijing, China

{dumj2, fangx1, lij86}@chinatelecom.cn

Abstract

This technical report describes ChinaTelecom system for Track 1 (closed) of the VoxCeleb2023 Speaker Recognition Challenge (VoxSRC 2023). Our system consists of several ResNet variants trained only on VoxCeleb2, which were fused for better performance later. Score calibration was also applied for each variant and the fused system. The final submission achieved *minDCF* of 0.1066 and *EER* of 1.980%.

Index Terms: speaker recognition, speaker verification, ResNet

1. Fully Supervised Speaker Verification

The track 1 of VoxSRC 2023 in this year is a fully supervised speaker verification task, where participants can use only VoxCeleb2 development set [1] for training. For this track, we trained 7 ResNet variants with varying sizes. Details are as followed.

1.1. Data augmentation

We applied the similar data augmentation method from the SJTU online system [2].

- Do speed perturbation with the ratio from 0.9, 1.0, 1.1 randomly, expanding the number speaker classes by a factor of 3.
- Randomly decide whether to do noise augmentation with the ratio of 0.6. If do, randomly select noise from MUSAN [3] and RIR [4] datasets.
- Randomly select a fixed length segment from the current utterance.

For all models, 80-dimension Mel filter banks (fbank) of a 2s segment were taken as input, with a 25ms frame length and 10ms frame shift.

1.2. Model architectures

ResNet has achieved state-of-the-art performance in speaker recognition within recent years. Thus, we chose ResNet and its variant, the Res2Net-based architectures, as the foundational backbone networks. Specifically, we incorporated Res2Net [5] and ERes2Net [6] to leverage both local and global speaker information for enhanced discriminative speaker embeddings. To elevate the complexity of architectures, we expanded the depths of these three frameworks (ResNet, Res2Net, ERes2Net) to 152 and 293, respectively. Furthermore, We also integrated SimAm-ResNet293 [7] as an auxiliary model for performance improvement.

1.3. Pooling layer

Pooling methods coupled with attention mechanisms have been proven effective in speaker verification, which assign larger weights to more discriminative speaker characteristics. In this case, we employed multi-query multi-head attention pooling method (MQMHA) [8] to alleviate model sticking in some certain patterns. We set the head number to 8, the query number to 2, the scale factor to 2, and the final speaker embedding dimension to 256.

1.4. Loss functions

We employed AAM [9], K-subcenter [10] with Inter-TopK [8] as loss function to train all single backbone networks, where the scale and margin were set to 32 and 0.2 in AAM loss, the sub-center number K was set to 3, and the penalty and TopK were set to 0.06 and 5, respectively.

We employed Sphereface2 [11, 12] as loss function for models with depth of 293. The weight of positive and negative pairs was set to 0.7, and the margin was set to 0.2 working as C-type.

Table 1: Network architectures of system. The base channel number of all models is 32.

Network	ID	Step 1		Step 2
		AAM+K-Subcenter+Inter-TopK	Sphereface2	AAM+K-subcenter
ResNet152	1	✓		✓
Res2Net152	2	✓		✓
ERes2Net152	3	✓		✓
ResNet293	4a	✓		✓
	4b		✓	✓
Res2Net293	5a	✓		✓
	5b		✓	✓
ERes2Net293	6a	✓		✓
	6b		✓	✓
Simam-ResNet293	7a	✓		✓
	7b		✓	✓

1.5. Training protocol

Our speaker verification system was implemented with WeSpeaker toolkit [13]. All our single models were trained on Nvidia A100 and V100 GPUs. Our training protocol encompassed two distinct steps.

1.5.1. Step 1: initial training

For all models, SGD is used as the optimizer, initialized with a learning rate of 0.1 and decayed to 1e-5 at the end. The learning rate decreased exponentially with a ratio of 1e-4. The training batch size for models with a depth of 152 was 32, and for models with a depth of 293 was adjusted to 16. All models were trained for a total of 150 epochs to ensure model convergence.

Table 2: Performance of models. Results already include AS-Norm and calibration with QMFs.

ID	Network	VoxSRC2023 val		VoxCeleb1-O		VoxCeleb1-E		VoxCeleb1-H	
		EER	minDCF	EER	minDCF	EER	minDCF	EER	minDCF
1	ResNet152	2.827	0.147	0.420	0.025	0.578	0.032	1.038	0.057
2	Res2Net152	3.163	0.170	0.532	0.028	0.665	0.040	1.187	0.068
3	ERes2Net152	2.999	0.153	0.415	0.026	0.601	0.034	1.106	0.060
4a	ResNet293	2.913	0.156	0.362	0.023	0.553	0.033	1.058	0.058
4b		2.929	0.172	0.457	0.027	0.627	0.036	1.154	0.066
5a	Res2Net293	2.772	0.146	0.362	0.024	0.543	0.031	0.990	0.054
5b		2.706	0.144	0.425	0.026	0.556	0.033	1.011	0.055
6a	ERes2Net293	2.772	0.149	0.298	0.019	0.559	0.031	1.029	0.056
6b		2.847	0.143	0.351	0.020	0.545	0.032	1.020	0.055
7a	Simam-ResNet293	2.948	0.165	0.372	0.024	0.608	0.037	1.199	0.067
7b		2.863	0.177	0.415	0.022	0.629	0.038	1.154	0.065
	Fusion	2.136	0.1260	-	-	-	-	-	-
	Fusion (test set)	1.980	0.1066	-	-	-	-	-	-

1.5.2. Step 2: large margin fine-tuning

Large margin fine-tuning [14] has been widely used for further increasing discriminative ability of speaker embeddings. Some changes were made that speed perturbation and Inter-TopK loss were removed, when doing large margin fine-tuning. For all models trained on AMM+K-subcenter+Inter-TopK or Sphereface2, AAM+K-subcenter was the only loss function at the step 2. The margin was set to 0.5 from 0.2. The length 2s of training segments in step 1 was adjusted to 6s as well.

1.6. Score procedure and fusion

Cosine distance was used as the score metric. Adaptive score normalization [15] was also applied after score computation, where the imposter cohort size was set to 300. The cohort was estimated from the development set of VoxCeleb2, mirroring the methodology employed by the SJTU system.

Besides, we constructed 30k trials denoted as *Vox2QmfsDev* from the development set for score calibration, following the strategy [14]. We utilized 6 QMF values the same in the ID R&D system [16]:

- a) speech length of the enrollment utterance;
- b) speech length of the test utterance;
- c) logarithm of the sum of *a* and *b*;
- d) logarithm of the sum of test and enrollment utterance lengths;
- e) SNR of the test utterance;
- f) SNR of the enrollment utterance;

The final calibrated fusion scores was calculated as

$$S' = v_0 \cdot \mathbf{W}^T \mathbf{S} + \mathbf{V}^T \mathbf{Q} + b \quad (1)$$

where v_0 , b and $\mathbf{V} \in \mathbb{R}^{6 \times 1}$ are learnable weights trained on *Vox2QmfsDev*, $\mathbf{W} \in \mathbb{R}^{n \times 1}$ is fixed weight, and $\mathbf{S} \in \mathbb{R}^{n \times 1}$ and $\mathbf{Q} \in \mathbb{R}^{6 \times 1}$ are normalized score matrix and QMF values, respectively.

1.7. Evaluation metric

For track 1, the speaker verification task, there are two evaluation metrics as followed:

- Equal Error Rate (*EER*): the error rate when False Acceptance (FA) and False Rejection (FR) error rates are equal.

- minimum detection cost function *minDCF*: the cost considering that achieving a low false positive rate is more important than achieving a low false negative rate. The following parameters were used to compute the cost: $C_{miss} = 1$, $C_{FA} = 1$, and $P_{Target} = 0.05$

2. Results

Table 2 shows the performance of single models and the fusion system on VoxSRC2023 val, VoxCeleb1-O, VoxCeleb-E and VoxCeleb-H. It can be seen that for single model, Res2Net293 and ERes2Net293 achieved the best results, with their respective performance really close to each other. For example, Res2Net293 trained with Sphereface2 loss (ID 5b) obtained the lowest EER of 2.706% on VoxSRC2023 validation set. It is observed that the performance of ResNet293 and Simam-ResNet293 fell short of initial expectations, which seemed to be attributed to the inadequacy of the training data volume, potentially leading to overfitting. Score fusion yielded a substantial enhancement in system performance, resulting in a remarkable reduction of EER to 2.136%, coupled with a decrease in minDCF to 0.1260 on the validation set. Our final submission achieved EER of 1.980% and minDCF of 0.1066, underscoring the effectiveness of our fusion strategy.

3. Conclusions

In this report, we make a detailed description of our solution for Track 1 of VoxSRC 2023. We employed ResNet-based variants with different depths and loss functions. It is suggested that score fusion of these variants plays a significant role for speaker verification, which brings impressive performance improvement.

4. References

- [1] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.
- [2] Z. Chen, B. Han, X. Xiang, H. Huang, B. Liu, and Y. Qian, "Sjtu-aispeech system for voxceleb speaker recognition challenge 2022," *arXiv preprint arXiv:2209.09076*, 2022.
- [3] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," 2015.
- [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] 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.
- [6] S. Z. Yafeng Chen, L. C. Hui Wang, and J. Q. Qian Chen, “An enhanced res2net with local and global feature fusion for speaker verification,” in *Interspeech 2023*, 2023.
- [7] X. Qin, N. Li, C. Weng, D. Su, and M. Li, “Simple attention module based speaker verification with iterative noisy label detection,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6722–6726.
- [8] M. Zhao, Y. Ma, Y. Ding, Y. Zheng, M. Liu, and M. Xu, “Multi-query multi-head attention pooling and inter-topk penalty for speaker verification,” in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6737–6741.
- [9] 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.
- [10] J. Deng, J. Guo, T. Liu, M. Gong, and S. Zafeiriou, “Sub-center arcface: Boosting face recognition by large-scale noisy web faces,” in *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer, 2020, pp. 741–757.
- [11] Y. Wen, W. Liu, A. Weller, B. Raj, and R. Singh, “Sphereface2: Binary classification is all you need for deep face recognition,” *arXiv preprint arXiv:2108.01513*, 2021.
- [12] B. Han, Z. Chen, and Y. Qian, “Exploring binary classification loss for speaker verification,” in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.
- [13] H. Wang, C. Liang, S. Wang, Z. Chen, B. Zhang, X. Xiang, Y. Deng, and Y. Qian, “Wespeaker: A research and production oriented speaker embedding learning toolkit,” *arXiv preprint arXiv:2210.17016*, 2022.
- [14] 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.
- [15] S. Cumani, P. D. Batzu, D. Colibro, C. Vair, P. Laface, V. Vasilakakis *et al.*, “Comparison of speaker recognition approaches for real applications,” in *Interspeech*, 2011, pp. 2365–2368.
- [16] R. Makarov, N. Torgashov, A. Alenin, I. Yakovlev, and A. Okhotnikov, “Id r&d system description to voxceleb speaker recognition challenge 2022,” 2022.