The bilibili VoxCeleb Speaker Recognition Challenge 2023 System Description

Xingui Zeng[1]*, Zhuo Yang*[2]*, Shiyi Wan*[3]*, Wei Deng*[1]*, XiangCao*[1]

[1]bilibili, China
[2]East China University of Science and Technology, China
[3]Shanghai International Studies University, China

zengxingui@bilibili.com, y30211040@mail.ecust.edu.cn, 1541524161@qq.com,
xuanwu@bilibili.com, caoxiang@bilibili.com

Abstract

This technical report describes our bilibili submissions for the VoxCeleb Speaker Recognition Challenge 2023 (VoxSRC-23) in the supervised speaker verification tracks (Track 1). For the supervised verification track, we trained four Resnet-based systems with architectural and feature variations. All models were trained with AAMSoftmax (Additive Angular Margin Softmax) loss and then fine-tuned with a large margin. Additionally, we utilized quality-aware score calibration, which incorporates quality metrics in the calibration system to generate more consistent scores across varying levels of utterance conditions. Finally, we applied fusion of all systems with both enhancements. The minDCF of our submission is 0.1048, and the corresponding EER is 1.7810%.

Index Terms: speech recognition, speaker verification

1. System Description

For the supervised verification track, we trained four Resnet-based systems. This section will focus on the approach we employed in this challenge.

1.1. Datasets and Data Augmentaion

1.1.1. Training Data

Data augmentation is also quite important in training speaker verification models. To generate additional fourth speakers, we first utilized a 5-fold speed augmentation based on the Sox speed function. Each segment was perturbed by a factor of 0.8, 0.9, 1.1, and 1.2. Since the VoxCeleb 2-dev[1] dataset comprises 1,092,009 utterances and a total of 5,994 speakers, we obtained 5,460,045 utterances and 29,975 speakers. Afterward, we implemented on-the-fly chain-like augmentation with a probability of 0.6. The effect chain is as follows:

- Noise addition augment with MUSAN[2] dataset.
- RIR reverberation with RIRs[3] dataset.
- gain augment

1.1.2. Developing Set

To assess the performance of our models, we employed four test sets as our development sets:

- VoxCeleb 1-O[4]: This test dataset comprises only 40 speakers, and we sampled 37,720 trials from it.
- VoxCeleb 1-E: This is an expanded version of VoxCeleb 1-O and consists of 581,480 trials from 1251 speakers.
- VoxCeleb 1-H: This set has 552,536 trials and is more challenging since each pair shares the same nationality and gender.
- VoxSRC 23-val: It is the validation set of VoxSRC 2023 and includes 49,987 trials.

1.2. Features

We utilized Kaldi[5] to extract 81-dimensional, 96-dimensional, and 120-dimensional log Mel filter bank energies. The window size was 25ms, and the frame shift was 10ms. We extracted 200 frames of features without additional voice activation detection (VAD). The speech segments were sliced to 2 seconds and augmented on the fly. All features were cepstral mean normalization.

1.3. Network Structures

Our systems utilized the bottleneck-block-based ResNet[6], which is one of the most popular convolutional neural networks for speaker verification tasks.

The pooling layer's objective is to aggregate the variable sequence to an utterance level embedding, and we adopted the multi-query multi-head attention pooling mechanism (MQMHA)[7] in our system.

Recently, margin based softmax methods have been widely used in speaker recognition works.Our system employed the AAMSoftmax[8] loss with a subcenter method and introduced the Inter-TopK penalty[7].

1.4. Training Protocol

We conducted our experiments using wespeaker[9] and trained all of our models in two stages.

In the first stage, we utilized the SGD optimizer with a momentum of 0.9 and weight decay of 1e-4. We adopted 200 frames of each sample in one batch to prevent over fitting and accelerate training. We also used an exponential decrease scheduler, with the minimum learning rate set to 5e-5, and the initial learning rate set to 0.2. Additionally, we gradually increased the margin from 0 to 0.2[10].

In the large-margin-based fine-tuning[11] stage, we made some adjustments to the settings used in the first stage. Firstly, we removed the speed augmented part from the training set to avoid domain mismatch, leaving only 5,994 classes. Secondly, we increased the frame size from 200 to 600 and exponentially increased the margin from 0.2 to 0.5. To ensure training stability, we removed the Inter-TopK penalty. The learning rate scheduler was almost the same, with the initial learning rate set to 1e-4 and the final learning rate set to 2.5e-5.

1.5. Back End

Speaker embeddings were extracted from the fully connected (FC) layer, and the score was computed using cosine similar-

ity after model training. Afterward, we utilized AS-Norm[12], QMF[11] (Quality Measure Functions), and score fusion.

For AS-Norm, we based it on the VoxCeleb 2-dev set and employed speaker-wise adaptive score normalization. We utilized the original VoxCeleb 2-dev dataset without any augmentation and averaged all the embeddings speaker-wise, resulting in 5,994 cohorts. Then, we calibrated the scores using top 300 imposter scores with this speaker-wise AS-Norm.

For QMF, we used four qualities to form QMF, including speech duration computed by an energy-based VAD, score, normed score, and magnitude of non-normalized embeddings. We also selected 100k trials from the original VoxCeleb 2-dev as the QMF training set. Finally, we trained an XGBoost to serve as our QMF model.

Lastly, we adopted score fusion to further enhance the system's performance.

2. Results

2.1. Sub-Systems

We trained four ResNet-based models, and their details are presented in Table 1. To increase the diversity of the models, we make small architectural changes across all four models:

- Changing input feature dimension
- Changing model channels
- Changing model depths

We have also tried other model structures such as RepVGG, Cam++, ECAPA_TDNN, but we couldn't achieve better results than Resnet.

Table 1: *Resnet variant*

Name	Features	Resnet Channels	Resnet Depth
$R1$	fbank96	32	$3 \times 8 \times 36 \times 3$
$R2$	fbank120	32	$3 \times 8 \times 36 \times 3$
$R3$	fbank120	64	$3 \times 8 \times 36 \times 3$
$R4$	fbank80	32	$10 \times 20 \times 64 \times 3$

2.2. Ablation Study

We conducted an ablation study on our baseline system in this subsection. Our baseline system, R1, utilized a ResNet-152 backbone followed by MQMHA pooling and AAM-Softmax. We evaluated the system's performance using the Equal Error Rate (EER) and the minimum Decision Cost Function (DCF) calculated with $C_{FA} = 1$, $C_M = 1$, and $P_{target} = 0.05$ for different trials. We took the performance of VoxSRC 23-val as our benchmark, as shown in Table 2.

We first conducted ablation studies by using large-margin fine-tuning, which improved the EER from 3.221% to 3.073% and the minDCF from 0.182 to 0.162. After applying AS-Norm, the EER was further improved to 2.753%, and the minDCF was reduced to 0.151. Finally, the QMF process achieved an EER of 2.387% and a minDCF of 0.141.

We followed the same procedure for all of our models, with the sole variable being the backbone.

2.3. Sub-Systems and Fusion Performance

We used four different backbones to generate distinct representations, and Table 3 displays some of our submissions to VoxSRC 2023, along with our fusion system's final result. We fine-tuned the fusion weights of all models based on the results of VoxCeleb 1-H and VoxSRC 23-val. In the VoxSRC 2023 challenge, our final fusion achieved a 0.1048 minDCF and a 1.7810% EER. Compared to our R4 model, the fusion result improved by 14.10% relatively in minDCF and 16.81% relatively in EER.

Table 2: *Ablation Study on the VOXSRC23-val set*

Methods	EER	$\mathbf{MinDCF_{0.05}}$
$R1$	3.221%	0.182
+Large Magin Fintuning	3.073%	0.162
+AS-Norm	2.753%	0.151
+QMF	2.387%	0.141

Table 3: *Our Submissions to VoxSRC23-test*

System	Voxsrc23-val		Voxsrc23-test	
	EER	$\mathbf{MinDCF_{0.05}}$	EER	$\mathbf{MinDCF_{0.05}}$
$R1$	2.387%	0.141	2.263%	0.1364
$R2$	2.238%	0.123	-	-
$R3$	2.436%	0.136	-	-
$R4$	2.141%	0.122	-	-
Fusion				
$R1 \sim R4$	1.835%	0.107	1.7810 %	0.1048

3. Conclusions

For this challenge, we utilized ResNet as our backbone and applied MQMHA pooling layer, Inter-TopK loss, and domain-based large margin fine-tuning methods. Additionally, we adopted AS-Norm and QMF. All of these methods, along with the large backbones, contributed significantly to enhancing the system's performance. As a result, our system achieved a final result of 0.1048 minDCF and 1.781%.

4. References

[1] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

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

[3] 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.

[4] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[5] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[6] 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.

[7] M. Zhao, Y. Ma, M. Liu, and M. Xu, "The speakin system for voxceleb speaker recognition challange 2021," *arXiv preprint arXiv:2109.01989*, 2021.

[8] 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.

[9] 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," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[10] Y. Liu, L. He, and J. Liu, "Large margin softmax loss for speaker verification," *arXiv preprint arXiv:1904.03479*, 2019.

[11] 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.

[12] 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.