The xx205 VoxCeleb Speaker Recognition Challenge 2023 system description

Xu Xiang

chinoiserie@sjtu.edu.cn

Abstract

This report describes the system submitted to the third track of the VoxCeleb Speaker Recognition Challenge (VoxSRC) 2023, which is based on the VoxSRC 2020 submission [1]. Codes are available at `https://github.com/xx205/voxsrc2020_speaker_verification`.

1. System

1.1. Architecture

In this submission, a Res2Net model is used as the speaker model. Res2Net [2] exploits a multi-scale structure to enhance the expressive power of residual block. Figure 1 shows the structure of a res2net block.

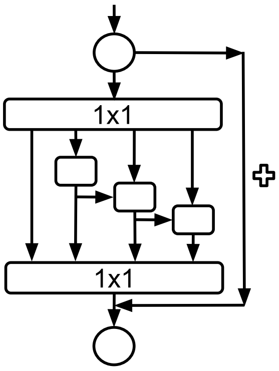

Figure 1: *Single Res2Net block.*

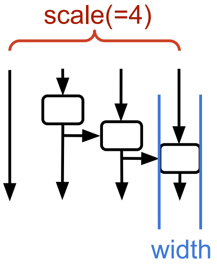

Figure 2: *Scale and width of a Res2Net block.*

1.2. Pooling function

Channel- and context-dependent statistics pooling function that introduced in [3] is used to aggregate the frame level information into segment level fixed dimension presentation.

1.3. Loss function

The additive margin softmax (AM-Softmax) [4] and additive angular margin softmax (AAM-Softmax) [5] have been widely applied to speaker recognition. Composite margin softmax (CM-Softmax) loss function [1] generalize AAM-softmax and AM-softmax, which can be expressed as

$$L_{\text{CM}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{s(\cos(\theta_{y_i,i}+m_1)-m_2)}}{e^{s(\cos(\theta_{y_i,i}+m_1)-m_2)} + \sum_{j\neq i} e^{s\cos(\theta_{j,i})}}$$

where $\theta_{j,i}$ is the angle between the class prototype vector (a column vector of the projection weight matrix) and the input vector, y_i is the ground truth class, s is a scaling factor, and m_1, m_2 are two positive margins. In this submission, a K-subcenter (K=3) version [6] of CM-softmax is used to train the speaker model.

2. Experiments

2.1. Dataset

The original training data, VoxCeleb2 development set, is augmented using Kaldi's VoxCeleb recipe[1] to create 4 additional copies. 80-dimensional FBANKs are extracted as the input features for training.

2.2. Training schedule

In this submission, the training schedule is nearly the same as the VoxSRC 2020 submission [1]:

- Total number of epochs is 23. A full sweep of the training samples by 8 GPUs is defined as one epoch.
- Learning rate is linearly increased to the maximum value in the first 3 epochs, then keep the same in the following 10 epochs and decayed exponentially in the final 10 epochs.
- The tuple of the composite margins (m_1, m_2) stays $(0.0, 0.0)$ in the first 3 epochs, then increased to their maximums in the following 10 epochs and keeps the same in the final 10 epochs.

To further boost the performance of the speaker model, there is an additional epoch for large margin fine-tuning (LMFT) [3].

2.3. Scoring

First for each trial, cosine score is computed with length normalized embeddings extracted by the speaker model. The cosine scores are then normalized by adaptive symmetric adaptive normalization [7]. For adaptive s-norm, the corresponding cohort set consists of the speaker-wise averages of the original VoxCeleb2 development set and top 400 cohorts are selected to compute the corresponding mean and standard deviation.

[1] https://github.com/kaldi-asr/kaldi/blob/master/egs/voxceleb/v2/run.sh

Table 1: *The experimental results on the VoxCeleb1 test sets and the VoxSRC 2020 validation set.*

#Params	**Model**	**VoxCeleb1**		**VoxCeleb1-E**		**VoxCeleb1-H**		**VoxSRC-20 Val**	
		EER(%)	**MinDCF$_{0.01}$**	**EER(%)**	**MinDCF$_{0.01}$**	**EER(%)**	**MinDCF$_{0.01}$**	**EER(%)**	**MinDCF$_{0.05}$**
4.8M	res2net50_w8_s6_c16	0.7869	0.0821	0.8903	0.0917	1.5370	0.1141		
17.7M	res2net50_w24_s4_c32	0.5529	0.0625	0.8003	0.0820	1.3751	0.1271		
29.3M	res2net101_w24_s4_c32	0.5210	0.0534	0.6081	0.0618	1.0940	0.0999		
32.9M	res2net152_w24_s4_c32	0.4572	0.0463	0.6099	0.0587	1.0737	0.0965		
35.5M	res2net200_w24_s4_c32	0.3668	0.0388	0.5930	0.0581	1.0330	0.0912	1.5017	0.0974

2.4. Results

Table 1 reports the system performance on the VoxCeleb1 test sets and the VoxSRC 2020 validation set. The model res2net200_w24_s4_c32[2] performs the best on the VoxCeleb1 test set. Table 2 shows the best model performance on VoxSRC 2023 track3.

Table 2: *Performance of the res2net200_w24_s4_c32 systems on VoxSRC 2023 track3.*

Model	**VoxSRC-23 Test**	
	EER(%)	minDCF$_{0.05}$
res2net200_w24_s4_c32	8.1330	0.3564

3. References

[1] X. Xiang, "The xx205 system for the voxceleb speaker recognition challenge 2020," *arXiv preprint arXiv:2011.00200*, 2020.

[2] S. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. H. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[3] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.

[4] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," 2018, pp. 5265–5274.

[5] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," 2019, pp. 4690–4699.

[6] 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.

[7] P. Matejka, O. Novotnỳ, O. Plchot, L. Burget, M. D. Sánchez, and J. Cernockỳ, "Analysis of score normalization in multilingual speaker recognition." in *INTERSPEECH*, 2017, pp. 1567–1571.

[2] This network has 200 layers in total, for its first res2net block, width=24, scale=4 and #channels=32.