Intema System Description for the VoxCeleb Speaker Recognition Challenge 2023

Ali Aliyev[1]

[1]Intema SARL, Luxembourg

a.aliyev@intema.ai

Abstract

This report describes our solution for the VoxCeleb Speaker Recognition Challenge 2023 (VoxSRC-23). In our solution, we fuse various ResNet-based models trained on VoxCeleb2 dev and different languages from the Common Voice dataset. Our best submission for track 2 achieves 0.095 in minDCF and 1.787% in EER on the VoxSRC-22 evaluation set.

Index Terms: speech recognition, speaker verification, VoxSRC-23

1. Introduction

The VoxSRC-23 challenge contains two full supervised speaker verification tracks (track 1 and track 2). However, we decided to participate only in the second track, because it allows us to train on other public datasets, and we already had some models and pipelines prepared. We used 9 different Resnet based models and then applied AS-Norm and Score Calibration. And at the end we used fusion of all our systems output to improve our final submission.

2. System description

2.1. Datasets

We used datasets such as, VoxCeleb 2[1] and Mozilla Common Voice[2] to train the models. The standard dataset for training speaker verification models is VoxCeleb 2, but training on a single dataset can lead to a bit of overfitting of our models, to avoid this, and for our other purposes in addition, we used some data from Mozilla Common Voice. By adding another dataset, we allow our model to prepare for the greater diversity that may be encountered in this challenge. It is worth considering that Common Voice has no ground truth labels for speaker verification task and the training pairs were generated by us. Therefore, the results of our experiments may differ from those of other researchers because the quality of the dataset depends on the script that generates the labels.

We also used Room Impulse Response (RIR)[1] and MUSAN [3] datasets for augmentation during the training.

2.2. Model architecture

We used a modification of ResNet, which was proposed by the winning team of the CN-Celeb Speaker Recognition Challenge 2022 [4], as the basis for all our models.

As we can see in the Table 2, we change the filter size in the first layer from 3×3 to 1×1, as well as we add another layer with 1×1 filter with the number of filters 32×4. All other blocks have been modified in the same way, more details on the architecture are described in the original paper.

[1]https://www.openslr.org/28/

ResNetBlock-1	Output
$\begin{bmatrix} 3\times3, 32 \\ 3\times3, 32 \end{bmatrix} \times N$	$80 \times T \times 32$

Table 1: *Original ResNetBlock structure*

ResNetBlock-1	Output
$\begin{bmatrix} 1\times1, 32 \\ 3\times3, 32 \\ 1\times1, 128 \end{bmatrix} \times N$	$80 \times T \times 128$

Table 2: *Modified ResNetBlock structure*

2.3. Training process

During the training of our models, we used different approaches and techniques to improve and diversify the results of our models.

2.3.1. Data Augmentation

To augment the data during training, we used RIR's and MUSAN's datasets to add extraneous sounds and reverberation on the original signals. Also, to increase the number of speakers we added speed augmentation, which helped us to increase the number of speakers by three times. We also tried to apply SpecAugment, but unfortunately it only worsened the results of our model. At each iteration, one of the augmentation was applied with a probability of 0.6

2.3.2. Training strategy

We trained our models on 2 s segments for 130–160 epochs using AAM-Softmax[5] loss and SGD. Most models were trained using TSTP, but two of them were trained using MQHASTP [6]. Then after the main training phase, we applied Large Margin Fine-tuning(LMF)[7] and as stated in the original paper input data was increased to 6 s.

2.4. Scoring

Our scoring pipeline consists of the following steps:

1. Calculating the cosine similarity between audio files in pairs.

#	System Description	Dataset	VoxSRC23-val	
			EER	DCF
1	ResNet-293-TSTP-AAM	VoxCeleb 2	2.927	0.1502
2	ResNet-293-TSTP-AAM-LM	VoxCeleb 2	2.609	0.1421
3	ResNet-221-TSTP-AAM	VoxCeleb 2	2.874	0.1484
4	ResNet-221-TSTP-AAM-LM	VoxCeleb 2	2.575	0.1435
5	ResNet-101-TSTP-AAM	VoxCeleb 2 + Common Voice 13.0	2.763	0.1693
6	ResNet-101-TSTP-AAM-LM	VoxCeleb 2 + Common Voice 13.0	2.571	0.1369
7	ResNet-152-TSTP-AAM	VoxCeleb 2 + Common Voice 13.0	2.590	0.1600
8	ResNet-152-TSTP-AAM-LM	VoxCeleb 2 + Common Voice 13.0	2.428	0.1324
9	ResNet-221-TSTP-AAM	VoxCeleb 2 + Common Voice 13.0	2.339	0.1428
10	ResNet-221-TSTP-AAM-LM	VoxCeleb 2 + Common Voice 13.0	2.208	0.1045
11	ResNet-34-TSTP-AAM	VoxCeleb 2 + Common Voice 9.0	3.917	0.2572
12	ResNet-34-TSTP-AAM-LM	VoxCeleb 2 + Common Voice 9.0	3.651	0.2091
15	ResNet-101-MQMHASTP-Inter-TopK	VoxCeleb 2 + Common Voice 9.0	2.825	0.1695
16	ResNet-101-MQMHASTP-Inter-TopK-LM	VoxCeleb 2 + Common Voice 9.0	2.691	0.1520
17	ResNet-152-MQMHASTP-Inter-TopK	VoxCeleb 2 + Common Voice 13.0	2.727	0.1605
18	ResNet-152-MQMHASTP-Inter-TopK-LM	VoxCeleb 2 + Common Voice 13.0	2.587	0.1448
19	**Fusion (Cosine Score + AS-Norm Score + Calibrated Score)**		**1.853**	**0.8204**

Table 3: *Results of our experiments. Note, all metrics are calculated after AS-Norm and Score Calibration.*

2. Next, we applied adaptive symmetric normalization (AS-Norm)[8]. For AS-Norm we took the average embeddings of the speakers of the desired training dataset, depending on which dataset a particular model was trained on. Then the cosine scoring was calibrated based on the top-300 imposter scores.
3. Then we applied Score Calibration[9], where a logistic regression was trained based on various statistics and score after AS-Norm.
4. Finally, we trained a logistic regression to combine scorers from all of our models. Initially we used scores after score calibration as training data, but we noticed that using all three scores (Cosine Score, AS-Norm Score and Calibrated Score) improves the results of logistic regression quite a bit.

VoxCeleb 2 was chosen as the training dataset for all logistic regressions.

3. Experiments results

As you can see from the Table 3, in our final submission, we decided to use a fusion of 9 different models, each of them being fine-tuned with LM, we ended up with 18 models for our fusion. On the VoxSRC23-val dataset, we achieved 1.853 EER and 0.8204 minDCF. Our best single model, ResNet-221-TSTP-AAM-LM, which was trained on VoxCeleb2 + Common Voice 13.0, achieved 2.208 EER and 0.1045 minDCF.

4. Conclusions

In this report, we have described our approach for the VoxSRC-23. We tried a combination of different methods and datasets during training, used LM finetuning, applied post-processing techniques such as AS-Norm and Score Calibration which helped to improve our single model results, and then built a fusion system based on all the obtained data which worked based on linear regression. Since the primary metric this year was DCF, we focused on it when building our solution. Due to hardware problems and lack of sufficient time, it was not possible to fully complete the experiments and test all possible combinations. Our proposed method achieved 0.095 in minDCF and 1.787% in EER on the VoxSRC-23 evaluation set.

5. References

[1] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Interspeech 2017*. ISCA, aug 2017. [Online]. Available: https://doi.org/10.21437%2Finterspeech.2017-950

[2] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *CoRR*, vol. abs/1912.06670, 2019. [Online]. Available: http://arxiv.org/abs/1912.06670

[3] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," 2015.

[4] Z. Chen, B. Liu, B. Han, L. Zhang, and Y. Qian, "The sjtu x-lance lab system for cnsrc 2022," 2023.

[5] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4685–4694.

[6] 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," *CoRR*, vol. abs/2110.05042, 2021. [Online]. Available: https://arxiv.org/abs/2110.05042

[7] 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," *CoRR*, vol. abs/2010.11255, 2020. [Online]. Available: https://arxiv.org/abs/2010.11255

[8] S. Cumani, P. D. Batzu, D. Colibro, C. Vair, P. Laface, and V. Vasilakakis, "Comparison of speaker recognition approaches for real applications," in *Interspeech*, 2011. [Online]. Available: https://api.semanticscholar.org/CorpusID:97395

[9] G. Liu, T. Zhou, Y. Zhao, Y. Wu, Z. Chen, Y. Qian, and J. Wu, "The microsoft system for voxceleb speaker recognition challenge 2022," 2022.