XMUSPEECH SYSTEM FOR NIST 2021 SPEAKER RECOGNITION EVALUATION

Qingyang Hong[*1], *Yiming Zhi*[1], *Haodong Zhou*[2], *Fuchuan Tong*[2], *Jie Wang*[2], *Lin Li*[*2]

[1]School of Informatics, Xiamen University, China
[2]School of Electronic Science and Engineering, Xiamen University, China
{qyhong, lilin}@xmu.edu.cn

ABSTRACT

In this paper, we present our submitted XMUSPEECH system for NIST SRE20 CTS Challenge and SRE21 Evaluation. With the large amounts of data assimilated into training set, the diversity of training data sources inevitably leads to domain mismatch, which becomes a key factor affecting the system performance. In order to solve this problem, we have made a lot of attempts. Based on the x-vector framework, we used different network structures, and tried to modify the performance of factorized time delay deep neural network (F-TDNN) and residual network (ResNet). In addition, in the back-end classifier, we used domain adaption to eliminate the impact of domain mismatch. We also employed Adaptive Symmetric Score Normalization (AS-Norm) for score normalization to adjust the fractional distribution space. These attempts have enriched the diversity of our systems, enabling the fusion system to complement each subsystem and improve the final submission performance. In addition, we describe the processing of video-only track based on the development set.

Index Terms— SRE21, x-vector, ResNet, AS-Norm, domain mismatch, video

1. INTRODUCTION

The Speaker Recognition Evaluation, sponsored by the US National Institute of Standards and Technology (NIST), has been one of the most representative contests in speaker recognition since 1996. Research teams from all over the world constantly explore new algorithms and state-of-the-art technologies for speaker recognition. SRE21 is organized similar to SRE19, focusing on speaker detection over conversational telephone speech (CTS) and audio from video (AfV) [1]. In addition to the audio-only track, SRE21 also features a visual-only track and an audio-visual track involving automatic person detection using audio, image, and video material.

SRE21 offers both fixed and open training conditions to allow uniform cross-system comparisons and to understand the effect of additional and unconstrained amounts of training data on system performance. For the fixed training condition, the baseline speaker recognition system is developed using the NIST SRE CTS Superset (LDC2021E08) [2] and VoxCeleb2 [3] datasets.

Since this evaluation provides options for open training data, it will inevitably lead to the introduction of large-scale publicly available data sets for system development. It is conceivable that the domain mismatch between individual data sets and test data will arise due to the different collection environment of data sets. We started the system development work for this challenge and tried to eliminate the performance degradation caused by the domain mismatch.

The first thing we thought of is to increase the diversity of subsystems, and it is most convenient to extract different acoustic features for training. In our experiments, two types of features (MFCC and FBank) have been employed for training. And it is necessary to find a robust training system based on x-vector [4]. In terms of network structure, we mainly explored F-TDNN [5] and ResNet [6]. F-TDNN factorizes the parameter matrices into smaller matrices, which makes the training more efficient. And ResNet can learn a lot of detailed temporal information.

Following the extraction of x-vector, we used probabilistic linear discriminant analysis (PLDA) [7] for the back-end scoring. We also employed centering, whitening, LDA, domain adaption and length normalization on x-vector before scoring. These have played an important role in eliminating domain mismatches. After the scoring, we also tried AS-Norm [8] to optimize the distribution of scores.

On the other hand, in the video-only track, we use the officially recommended InsightFace based on ResNet101 model [16] for our face recognition. We obtain the results of equal error rate (EER) and minimal detection cost function (min_DCF) based on the development set of multimedia corpus.

The rest of the paper is organized as follow: Section 2 gives the description of datasets and acoustic feature extraction. In Section 3, we described the details of the subsystems we developed for SRE21. Section 4 illustrates the back-end and score normalization. In Section 5, we report the result of our subsystems for SRE20 CTS Challenge. Section 6 describes the video-only track developed for face recognition and its data processing step. Finally, we conclude our work in Section 7.

2. DATA PREPARATION

2.1. Datasets

The fixed training condition designates a common set to facilitate a uniform algorithmic comparison of systems. For this condition, we only use NIST SRE CTS Superset (LDC2021E08) for the common training data, which is named as Train-fixed.

For the open training condition, we use the corpuses of NIST SRE04, 05, 08, 10 and SRE12-tel, which is named as Train-open.

We also employ additive noises and reverberation (i.e., Babble, Noise, Music and Reverb from MUSAN [9] and reverberation [10]) as described in [4] to augment the training data. This operation can make the systems more robust, and alleviate the problem of training data domain mismatch.

2.2. Acoustic feature extraction

2.2.1. MFCC

For the Mel frequency cepstral coefficient (MFCC) feature extraction, all audios were converted to the cepstral features of 23-dimensional MFCC with a frame-length of 25ms and a frame shift of 10ms. The cepstral filter banks were selected within the range of 20 to 3700 Hz. Then, a frame level energy-based voice activity detector (VAD) selection was conducted to the features. This was followed by local cepstral mean and variance normalization (CMVN) over a 3-second sliding window. All operations of feature extraction were based on Kaldi toolkit [11].

2.2.2. FBank

The other subsystems were based on the filter bank (FBank) feature. The FBank feature retains a lot of raw information, which makes it possible for the neural network to learn more useful information. Of course, this also requires the neural network itself to have strong modeling capability. The FBank feature vectors include 80 dimensional FBanks and energy value extracted from the raw signal with a 25ms frame-length. Similar to MFCC, VAD and CMVN were also used for FBank features.

3. SUBSYSTEMS

The final submitted system is based on the fusion of several x-vector systems with different datasets and features. In this section we will introduce the details of each subsystem.

3.1. Factorized TDNN x-vector systems

The core trick of F-TDNN is factorizing matrices with a semiorthogonal constraint. This obviously reduces the amount of parameters and proves that there is no loss of modeling capability through the singular value decomposition (SVD). The configuration of the first two factorized TDNN x-vector systems could be found in [12].

- F-TDNN-v1: Factorized TDNN x-vector trained on 5-fold Train-open with 23-dimension MFCC features.
- F-TDNN-v2: Factorized TDNN x-vector trained on 5-fold Train-open with 81-dimention FBank features.

3.2. ResNet x-vector systems

ResNet models are optimized based on the AM-Softmax loss.

- ResNet-34: ResNet-34 trained on 5-fold Train-open with 81-dimention FBank features.
- ResNet-50: ResNet-50 trained on 5-fold Train-open with 81-dimention FBank features.
- ResNet-34-SE: ResNet-34-SE trained on 5-fold Train-open with 81-dimention FBank features.
- ResNet-50-SE: ResNet-34-SE trained on 5-fold Train-open with 81-dimention FBank features.

For ResNet-34 and ResNet-50, we extract the far and near embeddings (corresponding to the output layer) from the first and second layer after the statistic pooling layer. For the far embedding, PLDA scoring is used. And for the near embedding, Cosine scoring is adopted. These two kinds of scores are fused for that model. In our experiments, all the subsystems were implemented on ASV-Subtools [14].

4. BACK-END

4.1. Scoring

For all the systems above, the PLDA of the system was trained using embeddings of the 5-fold training data since the PLDA is sensitive to the domain. For the post-processing of the embeddings extracted from the embedding extractors, length normalization, centering, whitening and LDA transformation for feature dimensionality reduction have been applied to the embeddings in sequence, finally followed by the PLDA training. Furthermore, the PLDA parameters are adapted on the in-domain data. All scores of subsystems were estimated using the adapted PLDA (APLDA).

4.2. Score normalization and fusion

We also applied the AS-Norm [13] to compare the performance. However, it only helps the Cosine scoring to reduce the act_C result but still can't surpass the APLDA scoring in our experiments.

5. RESULTS OF AUDIO TRACK

We present the experimental results on the progress set of SRE20 CTS challenge, since we can't obtain the results on the test set of SRE21. The results of all subsystems are shown in Table 1. The fusion rate of APLDA and Cosine scoring for ResNet-34 and ResNet-50 are 7:3 and 5:5 respectively. ResNet-50-SE-r4 adopts the reduced learning late (0.0025), which is different with ResNet-50-SE.

Table 1. The results of subsystems on the progress set of NIST SRE 20 CTS Challenge

System	SRE20 progress set		
	EER (%)	min_C	act_C
F-TDNN-v1	5.48	0.220	0.223
F-TDNN-v2	4.67	0.222	0.227
ResNet-34-SE	4.41	0.196	0.199
ResNet-50-SE	4.47	0.187	0.189
ResNet-34	3.67	0.178	0.194
ResNet-50	3.89	0.180	0.188
ResNet-50-SE-r4	4.50	0.188	0.193
Fusion	3.11	0.154	0.160

From the act_C results, we can see the best subsystem is ResNet-50. Most the ResNet-50 subsystems are better than ResNet-34, which shows that Bottleneck-Block is superior compared with Basic-Block. We also find that adding SE-block can't lead to further improvement. Finally, we fused all seven subsystems with the same weight and obtained the act_C of 0.160. We submitted this fusion system for the SRE21-Open task.

For the SRE21-Fixed task, we only trained the ResNet-34-SE on 5-fold Train-open with 81-dimention FBank features, and submitted the result of this single system.

6. VISUAL-ONLY SYSTEM

The 2021 NIST multimedia SRE recognizes a person through the fusion of audio and video. Therefore, each video provides personal voice segment and face information at the same time. The visual system is to detect whether the target person exists in another test video. The baseline face recognition system is built using Pytorch based on the InsightFace with a face detector termed RetinaFace [17], and a face embedding extractor using a ResNet101 architecture, and utilizes a pre-trained model which has been trained on MS-Celeb-1M dataset [18]. In this section, we will show our processing of visual-only track.

6.1. Face detection

We use ffmpeg to extract one frame per second to process the development set test video. Then, on the extracted frames, we apply RetinaFace face detector to select the frames with faces, and get the bounding boxes of all faces in each frame. Next, we cut out the face image through the bounding box, align it with the 5-point facial landmark model, and finally resize the image to 112×112 pixels and normalized.

6.2. Face embedding

The face embeddings are extracted using InsightFace, and the pre-trained ResNet101 model is used to extract face encodings from the cropped, aligned,and normalized resized images.

6.3. Score

We use kmean++ algorithm [19] to classify per video and get its center vector. Then, in the enrollment process, the cosine similarity between the enrollment image and the center vector of each video is calculated as the output score. Finally, we obtained the EER of 2.08% and min_DCF of 0.036 for the development set. The results are compared with the baseline system in Table 2.

Table 2. The results of development set of Visual Track

System	EER (%)	min_DCF
Baseline	1.82	0.035
Ours	2.08	0.036

Due to time constraint, we did not submit the results on Visual and Audio-Visual Track.

7. CONCLUSION

We have presented the description of the XMUSPEECH submission to SRE 20 CTS Challenge and SRE 21 Evaluation. In view of the large amount of training data and the domain mismatch problem, we have made various attempts in network structures, back-end scoring and score normalization. Different network structures greatly enhance the diversity and complementarity of our systems. These attempts have eliminated the impact of domain mismatch to some extent from different stages, allowing our final fusion system to achieve great improvement in comparison with the subsystems. And our results on the visual-only track is close to the NIST baseline system.

8. ACKNOWLEDGEMENT

This work is supported by the National Natural Science Foundation of China (Grant No.61876160 and No. 62001405).

12. REFERENCES

[1] S. O. Sadjadi, C. Greenberg, E. Singer, L. Mason, and D. Reynolds, "NIST 2021 Speaker Recognition Evaluation Plan," https://www.nist.gov/document/nist-2021-speaker-recognition-evaluation-plan, July 2021, [Online; accessed 6-October-2021].

[2] S. O. Sadjadi, "NIST SRE CTS Superset: A large-scale dataset for telephony speaker recognition," 2021.

[3] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," arXiv preprint arXiv:1806.05622, 2018.

[4] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE ICASSP*, April 2018.

[5] Daniel Povey, Gaofeng Cheng, Yiming Wang, Ke Li, Hainan Xu, Mahsa Yarmohammadi, and Sanjeev Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks.," in *INTERSPEECH*, 2018, pp. 3743–3747.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[7] Simon JD Prince and James H Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.

[8] Sandro Cumani, Pier Domenico Batzu, Daniele Colibro, Claudio Vair, Pietro Laface, and Vasileios Vasilakakis, "Comparison of speaker recognition approaches for real applications," in *INTERSPEECH*, 2011.

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

[10] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. IEEE ICASSP*, 2017.

[11] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The Kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, 2011.

[12] Jesus Villalba, Nanxin Chen, David Snyder, Daniel Garcia-Romero, Alan McCree, Gregory Sell, Jonas Borgstrom, Fred Richardson, Suwon Shon, Franc¸ois Grondin, et al., "The JHU-MIT system description for NIST SRE18," Johns Hopkins University, Baltimore, MD, Tech. Rep, 2018.

[13] H. Lu, J. Zhou, M. Zhao, W. Lei, Q. Hong, L. Li, "XMU-TS systems for NIST SRE19 CTC challenge", in *Proc. IEEE ICASSP*, 2020.

[14] F. Tong, M. Zhao, J. Zhou, H. Lu, Z. Li, L. Li, and Q. Hong, "ASV-Subtools: Open source toolkit for automatic speaker verification," in *Proc. IEEE ICASSP*, 2021.

[15] F. Tong, Y. Liu, S. Li, J. Wang, L. Li, Q. Hong, "Automatic error correction for speaker embedding learning with noisy labels," in *INTERSPEECH*, 2021.

[16] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in Proc. the Thirty-First AAAI Conference on Artificial Intelligence, ser. AAAI'17. AAAI Press, 2017, pp. 4278 – 4284.

[17] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-Shot Multi Level Face Localisation in the Wild," in CVPR, 2020.

[18] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in European conference on computer vision. Springer,2016, pp. 87-102.

[19] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in Proc. ACM-SIAM Symposium on Discrete Algorithms, 2007, p. 1027-1035.