SPEAKER VERIFICATION SYSTEM FOR FAR-FIELD SPEAKER VERIFICATION CHALLENGE BY TEAM XD-RTB

1. INTRODUCTION

In This paper, we describe the systems submitted by team XD-RTB to the Far-Field Speaker Verification Challenge. In this challenge, we focus on constructing deep Neural Network architectures based on ResNet. The challenge aims to benchmark the state-of-the-art speaker verification technology under far-field and noisy conditions. The challenge has three tasks, far-field text-dependent speaker verification from single microphone array, far-field text-independent speaker verification from single microphone array and far-field text-dependent speaker verification from distributed microphone array. Our team mainly participate in task1 and task3. Our final system which consist of Neural Network embeddings are applied with PLDA backend. Also, we explore the use of AM-softmax loss function in this challenge.

2. SYSTEM DESCRIPTION

This section describes the system we develope for the FFSVC 2020 challenge. Firstly, we introduce the data sets, data augmentation and spectral feature used in model training. Secondly, we introduce our model ResNet[1][2][3][4] architecture. Thirdly, we explore the use of AM-softmax(the Additive Margin Softmax)[5] loss to improve the performance. Finally, backend method, such as PLDA(Probabilistic Linear Discriminant Analysis)[6][7]are described.

2.1. Training data

The training sets used in our experiments are AIshell[8][9], HIMIA[10] and the data set that FFSVC 2020[11] provides.

The challenge officially provides a training set with 120 speakers, and a development set with 35 speakers. The HIMIA data set has 254 speakers in the training set and 42 speakers in the development set. There are 405 speakers in total after excluding duplicates in two data sets. For the purpose of increasing the amount and the diversity of the training data, all training data is augmented by using the freely available MUSAN[12] and RIRs datasets, creating four corrupted copies of the original recordings with Kaldi recipe.

2.2. Feature

All training datasets are resampled to 16kHz and pre-emphasized before feature extraction.64-dimensional MFB (Log Mel-filter Bank Energies) from 25ms frames with 10ms overlop, with frequency limits 0-8000Hz are used in this challenge.

2.3. Loss Function

We also explore AM-softmax loss in this challenge. Recent studies have shown that AM-softmax loss has greatly improved performance in the field of speaker verification[**?**]which is formulated as:

$$L_{AMS} = \frac{1}{N} \sum_{i} -log \frac{e^{s*(cos\theta_{y_i} - m)}}{e^{s*(cos\theta_{y_i} - m)} + \sum_{j \neq y_i} e^{s*cos(\theta_{y_j}, i)}}$$

where s is a scale factor and m is the margin factor.

2.4. Backend

In this work, both CS(cosine similarity) and PLDA are used for scoring.

3. RESULT

We conduct two training strategies, the first is to train a base 101-layer ResNet with AIshell data, then finetune with 405 speakers, and the second is to train a 101-layer ResNet directly with 405 speakers. The result shows that fine-tuning is better than training directly.

3.1. Submitted result

In this section, we present the experimental results on dev set, and the final result on the task1 and task3 set. Results are reported in terms of the equal error rate (EER) as well as minDCF. Table 1 summarizes the baseline results on the fine-tuning and training directly. Table 2 shows that the submitted result on task1 and task3 set.

Table 1. Train directly and fine-tuning

	cosine similarity		PLDA	
system	minDCF	EER	minDCF	EER
dev Train directly	0.488	4.96	0.462	4.53
dev Finetune	0.435	4.66	0.425	4.47

Table 2. Submitted results on Task1 and Task3

system	minDCF	EER
Task1	0.576	5.41
Task3	0.626	6.88

4. REFERENCES

[1] 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.

[2] Weicheng Cai, Jinkun Chen, and Ming Li, "Exploring the encoding layer and loss function in end-to-end speaker and language recognition system.," in *Proc. of Odyssey*, 2018, pp. 74–81.

[3] Jesús Villalba, Nanxin Chen, David Snyder, Daniel Garcia-Romero, Alan McCree, Gregory Sell, Jonas Borgstrom, Leibny Paola García-Perera, Fred Richardson, Réda Dehak, et al., "State-of-the-art speaker recognition with neural network embeddings in nist sre18 and speakers in the wild evaluations," *Computer Speech & Language*, vol. 60, pp. 101026, 2020.

[4] Aleksei Gusev, Vladimir Volokhov, Tseren Andzhukaev, Sergey Novoselov, Galina Lavrentyeva, Marina Volkova, Alice Gazizullina, Andrey Shulipa, Artem Gorlanov, Anastasia Avdeeva, et al., "Deep speaker embeddings for far-field speaker recognition on short utterances," *arXiv preprint arXiv:2002.06033*, 2020.

[5] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[6] Sergey Ioffe, "Probabilistic linear discriminant analysis," in *European Conference on Computer Vision*. Springer, 2006, pp. 531–542.

[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] Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*. IEEE, 2017, pp. 1–5.

[9] Jiayu Du, Xingyu Na, Xuechen Liu, and Hui Bu, "Aishell-2: transforming mandarin asr research into industrial scale," *arXiv preprint arXiv:1808.10583*, 2018.

[10] Xiaoyi Qin, Hui Bu, and Ming Li, "Hi-mia: A far-field text-dependent speaker verification database and the baselines," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7609–7614.

[11] Xiaoyi Qin, Ming Li, Hui Bu, Wei Rao, Rohan Kumar Das, Shrikanth Narayanan, and Haizhou Li, "The interspeech 2020 far-field speaker verification challenge," in *Interspeech 2020*, 2020.

[12] David Snyder, Guoguo Chen, and Daniel Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.