ZXIC Speaker Verification System for FFSVC 2022 Challenge

Yuan Lei

ZTE Corporation

lei.yuan1@zte.com.cn

Abstract

This paper describes the ZXIC speaker verification system submitted to the task 1 of Interspeech 2022 Far-Field Speaker Verification Challenge (FFSVC 2022). We focus on solve the performance degradation problem caused by mismatch between enrollment utterances and far-field test utterances. We introduce a novel multi-reader domain adaption learning method to alleviate this mismatch impact. In this challenge, we explore 2 kinds of advanced neural network structures ECAPA-TDNN and ResNet-SE with different loss functions. The submitted systems are the fusion of different models. Finally, the best system achieves a minimum of the detection cost function (minDCF) of 0.511and an equal error rate (EER) of 4.409% on the evaluation set of the challenge.

Index Terms: speech verification, fully supervised, deep learning, domain adaption

1. Introduction

Speaker verification is the process of verifying a person from characteristic of voices. Recently, due to the development of deep learning technology and the availability of large-scale speech datasets, automatic speaker verification (ASV) has become one of the most promising biometric authentication methods in smart speakers and smartphones. Pioneering works on speaker verification based on embedding extracted by deep neural network can transform speaker utterances into fixed dimensional embedding vectors for back-end scoring [1, 2]. These works have achieved significantly superior results on speaker verification benchmark datasets and close-talk scenarios. However, when ASV systems are deployed in real word, their performance drop obviously. The performance degradation is caused by the combination of many factors. One factor is the mismatch between domains on real scenarios and domains in which the systems are trained. Another critical factor is the mismatch between enrollment utterances and test utterances which are collected on real scenarios. Unexpected cross-domain problems, such as cross-distances, cross-channels, cross-devices and cross-time problems will damage the system's performance a lot. To promote the development of speaker verification on real application scenarios, Far-Field Speaker Verification Challenge (FFSVC) was first organized in 2020 [3]. In this year, the specific objective for FFSVC 2022 focus on single-channel far-field speaker verification scenarios under noisy conditions [4].

In this paper, we presented our submitted system to the fully supervised far-field speaker verification task (task1) of FFSVC 2022. In this challenge as well as common real scenarios, enrollment data are usually collected via close-talking telephones while test utterances are collected in complex far-field home/office environments. To solve performance degradation caused by the mismatch. We propose a novel multi-reader domain adaption learning method for deep learning based x-vector embeddings. We introduce this training technique with strong focus on mismatch between enrollment and test speeches. In addition, we conduct a number of experiments on deep learning based embedding extractors. Classification objectives loss functions and metric learning based loss functions are also explored in this work.

The rest of the paper is organized as follows: Section 2 describes the system components of our system including front-end, feature extraction, model extractors and back-end strategies. Section 3 introduces the proposed novel multi-reader domain adaption learning method. Experimental results are presented in Section 4, followed by conclusions.

2. System components

2.1. Feature extraction

All raw input signals are resampled to 16kHz, normalized and pre-emphasized before feature extraction. During training, we randomly extract a fixed length 2-seconds segment from each utterance. While during testing, 5 equally-spaced 2-seconds segments and the entire utterance are selected for feature extraction. 80 dimensional logarithm Mel filter bank energies are generated with a hamming window of width 25ms and step 10ms. All features were cepstral mean normalized without extra voice activity detection (VAD).

2.2. Speaker embedding extractors

In total, we trained two advanced neural network architectures to extract speaker embedding from acoustic features. One is variant of x-vector [2] and the other is variant of ResNet [5]. We introduce them in the following.

2.2.1. ResNet-SE

Residual networks [5], which are widely used in image recognition have recently been used in speaker recognition system [6]. Squeeze-and-excitation residual network (ResNet-SE) is a variant of a ResNet that employs squeeze-and-excitation blocks to enable the network to perform dynamic channel-wise feature recalibration [7, 8]. In this work, we implement ResNet-SE with 34 layers to extract speaker embeddings. As shown in Table 1 which describes the ResNet34-SE architecture, 256 dimensional speaker embedding vectors are extracted. The statistics pooling layer can aggregates all frame-level outputs to integrate information across time dimension so that subsequent layers operate on the entire segment. In this work, we use attentive statistics pooling (ASP) [9] to aggregate frame-level features into utterance-level features.

Table 1: *ResNet34-SE architecture configuration*

Layer name	Configurations
Conv1	$[3\times 3 \quad 32]$
Res-SE 1	$\begin{bmatrix}3\times 3 & 32\\ 3\times 3 & 32\end{bmatrix}\times 3$
Res-SE 2	$\begin{bmatrix}3\times 3 & 64\\ 3\times 3 & 64\end{bmatrix}\times 4$
Res-SE 3	$\begin{bmatrix}3\times 3 & 128\\ 3\times 3 & 128\end{bmatrix}\times 6$
Res-SE 4	$\begin{bmatrix}3\times 3 & 256\\ 3\times 3 & 256\end{bmatrix}\times 3$
Statistics pooling	ASP
Linear 1	512×256

2.2.2. ECAPA-TDNN

ECAPA-TDNN is one of the state-of-the-art speaker verification models [10]. ECAPA models consists of blocks of time delay neural blocks (TDNNs) and squeeze-and-excitation layers unified with blocks of Res2Block layers. The model we used is based on ECAPA-TDNN architecture with the following parameters: the number of SE-Res2Net Blocks is set to 3 with dilation values 2, 3 and 4. The number of channels is set to 1024. The pooling layer of ECAPATDNN-1024 is attention statistic pooling (ASP) [9].

2.3. Back-end

According experiment results [11] and previous experience, PLDA will not enhance the system performance if the model is trained with margin-based loss functions. So in our work, we use cosine similarity to calculate back-end score.

2.3.1. Score Integration

For development and evaluation trials, embedding vectors of the entire enroll and test utterances are extracted and whole utterance similarity score is computed. We also extract 5 embedding vectors from 5 equally-spaced segments in enroll and test utterances, compute the score matrix and get the average as matrix average score. The final score is the mean of whole utterance similarity score and matrix average score.

3. Multi-Reader domain adaption learning

In this section, we propose a multi-reader domain adaption transfer learning method to solve mismatch problems between enrollment and test utterances in far-field speaker verification.

As shown in Figure1, in fine-tuning stage, we generate a batch that contains N speakers, and one close-talking enrollment utterance and M far-field test utterances from each speaker.

We feed these utterances into our system described in Section 2. During training, in each batch, the extracted normalized enrollment speaker embedding vectors are ee_i and test speaker embedding vectors are $et_{i,j}$ where $1 \le i \le N$ and $1 \le j \le M$.

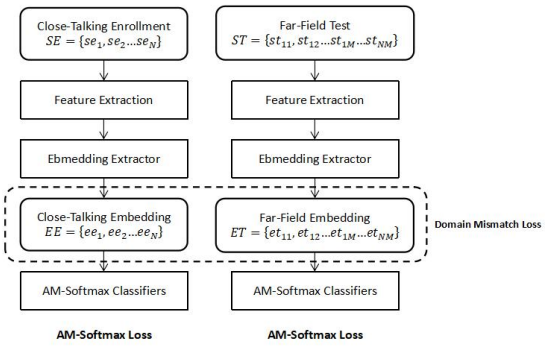

Figure 1: *Multi-Reader domain adaption learning*

3.1. Domain mismatch loss

If the embedding extractor is trained well, the embedding vectors extracted from enrollment speech and test speech should have similar distribution in the embedding space. The similarity matrix between each enrollment utterance embedding and test utterance embedding is defined as:

$$S_{i,j,k} = \cos\left(et_{i,j}, ee_k\right) + b \tag{1}$$

where b is learnable bias, $1 \le i,k \le N, 1 \le j \le M$. As shown in Figure 2, for a well trained system, the similarity should be large in gray area and be small in white areas. We define the target labels for similarity matrix are positive when $i = k$ and negative when $i \ne k$. Cross-entropy loss is selected to calculate the loss between softmax of similarity matrix and target labels. The final domain mismatch loss is defined as:

$$L_{DM} = -\frac{1}{M \times N}\sum_{i=1}^{N}\sum_{j=1}^{M}\log\frac{e^{S_{i,j,i}}}{\sum_{k=1}^{N} e^{S_{i,j,k}}} \tag{2}$$

3.2. Multi-Reader domain adaption learning

Figure 1 illustrates the Multi-Reader domain adaption learning process. Besides the domain mismatch loss, we also calculate the additive margin softmax [12] loss $L_{AMS}(SE)$ of enrollment utterances and $L_{AMS}(ST)$ of test utterances separately. And in this work, the combined loss is defined as:

$$L = L_{AMS}(SE) + \alpha L_{AMS}(ST) + \beta L_{DM} \tag{3}$$

where the parameters α and β control the contributions from the these losses.

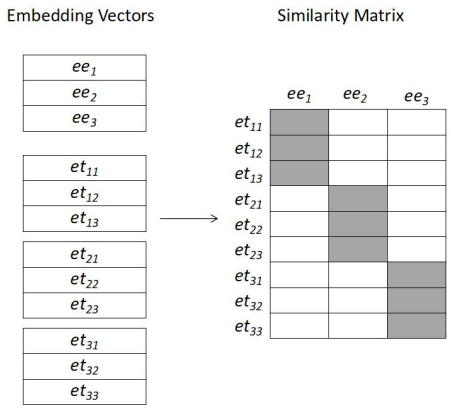

Figure 2: *Similarity matrix of domain mismatch loss*

4. Experiments and Results

4.1. Training strategy

For FFSVC2022 task 1, only Voxceleb 1&2 dataset[13, 14], FFSVC2020 dataset and supplementary set can be used for training. We use Voxceleb2 corpus to pre-train all our models and Voxceleb1 to evaluate the performance of pre-trained models. Adam optimizer with the learning rate decreases from 0.001 to 0.0005 linearly is used.

In fine-tune stage 1, firstly we combine voxceleb2 dataset and FFSVC2022 supplementary dataset to fine-tune all models using Adam optimizer with 0.0001 learning rate with 0.95 decay each epoch.

Then in fine-tune stage 2, we select close-talk iphone dataset as enrollment dataset and others in FFSVC2022 supplementary data as test dataset. Then we use proposed multi-reader domain adaption learning method to future tuning all models.

All training process are processing on 2 NVIDIA Tesla T4 GPUs with Intel Xeon Gold CPU at 2.30 GHz.

4.2. Data augmentation

In pre-train stage, music, noise and speech part from MUSAN dataset [15] is used as additive noise with random SNR setting from 5db to 20db.

In addition, we also use SpecAugment method [16] to enhance the system's robust in pre-train and fine-tune stage.

At last, to close the real scenarios, we collect office and home background noise which is mingled with different kinds of noise such as air conditioner noise, keyboard noise and television noise. We randomly add this noise to the test utterances in fine-tune stage.

4.3. Experimental results

In this section, we report the results of the speaker embedding-based systems as well as the fusion system on the FFSVC2022 development data and evaluation data. The primary metric to evaluate system performance is the Minimum Detection Cost(mDCF). In addition, Equal Error Rate (EER) is selected as performance criteria.

As shown in Table 2, data augmentation and score integration method both contribute the system improvement. Multi-reader domain adaption learning in fine-tune stage 2 aiming to solve the authentication mismatch problem , boost the system performance a lot. System performance results on evaluation dataset is shown in Table 3. The scores from the models o1 to 8 are linearly weighted into one fusion score.

Table 2: *Performance of our systems on the FFSVC2022 development set*

ID	System	EER (%)	min-DCF
1	ResNet34-SE + fine-tune1	7.321	0.590
2	System 1 + data augmentation	6.925	0.560
3	System 2 + fine-tune2	6.194	0.505
4	System 3 + score integration	5.922	0.507
5	ECAPA-TDNN1024 + fine-tune1	6.200	0.514
6	System 5 + data augmentation	6.083	0.517
7	System 6 + fine-tune2	5.497	0.496
8	System 7 + score integration	5.322	0.483

Table 3: *Performance of our systems on the FFSVC2022 evaluation set*

System	EER(%)	min-DCF
Baseline[4]	7.021	0.681
System 4	6.971	0.630
System 8	6.091	0.573
Fusion 1	5.556	0.524
Fusion 2	4.409	0.511

5. Conclusions

This paper describes the ZXIC speaker verification system submitted to the task 1 of Interspeech 2022 Far-Field Speaker Verification Challenge (FFSVC 2022). ECAPA-TDNN and ResNet-SE are used to extract speaker embeddings. In this paper, we put forward a novel muti-reader domain adaption training method to solve the problem of mismatch between close-talk enrollment and far-field test speech. According to the experiment result, this method improve the systems performance a lot.

6. References

[1] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *Proc. Interspeech*, 2017, pp. 999–1003.

[2] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[3] X. Qin, M. Li, H. Bu, W. Rao, R. K. Das, S. Narayanan, and H. Li, "The interspeech 2020 far-field speaker verification challenge," in *Proc. Interspeech*, 2020, pp. 3456–3460.

[4] X. Qin, M. Li, H. Bu, S. Narayanan, and H. Li, "Far-field speaker verification challenge (FFSVC) 2022: challenge evaluation plan," *https://ffsvc.github.io.*

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[6] W. Xie, A. Nagrani, J. S. Chung, and A. Zisserman, "Utterance-level aggregation for speaker recognition in the wild," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[7] J. Thienpondt, B. Desplanques, and K. Demuynck, "The idlab voxsrc-20 submission: Large margin fine-tuning and qualityaware score calibration in dnn based speaker verification, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5814–5818.

[8] L. Zhang, Q. Wang, K. A. Lee, L. Xie, and H. Li, "Multi-Level Transfer Learning from Near-Field to Far-Field Speaker Verification," in *Proc. Interspeech*, 2021, pp. 1094–1098

[9] W. Cai, J. Chen, and M. Li, "Exploring the encoding layer and loss function in end-to-end speaker and language recognition system," in *Speaker Odyssey*, 2018, pp.74–81.

[10] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPATDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.

[11] Y. Tong, W. Xue, S. Huang, L. Fan, C. Zhang, G. Ding and X. He, "The JD AI speaker verification system for the FFSVC 2020 challenge," in *Proc. Interspeech*, 2020, pp. 3476–3480.

[12] F. Wang, W. Liu, H. Liu, and J. Cheng, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[13] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A LargeScale Speaker Identification Dataset, " in *Proc. Interspeech*, 2017, pp. 2616–2620.

[14] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep Speaker Recognition," in *Proc. Interspeech*, 2018.

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

[16] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," *in Proc. INTERSPEECH*, 2019, pp. 2613–2617.