

ZXIC Speaker Verification System for FFSVC 2022 Challenge

Yuan Lei¹, Zhou Cao¹, Dehui Kong^{1,2}, Ke Xu^{1,2}

¹ZTE Sanechips Corporation

²State Key Laboratory of Mobile Network and Mobile Multimedia Technology

lei.yuan1@zte.com.cn, cao.zhou1@zte.com.cn, kong.dehui@sanechips.com.cn,
xu.kevin@sanechips.com.cn

Abstract

This paper presents the development of ZXIC speaker verification system submitted to the task 1 of Interspeech 2022 Far-Field Speaker Verification Challenge (FFSVC2022). Deep neural network based discriminative embeddings, such as x-vectors, have been shown to perform well in speaker verification tasks. In far-field speaker verification system, mismatch between training and testing data and mismatch between enrollment and authentication utterances impact the system performance a lot. To alleviate this mismatch and improve the system performance, in this paper we propose a novel multi-reader domain adaption learning framework based on asymmetric metric learning. In this challenge, we also explore advanced neural network based embedding extractor structures including ECAPA-TDNN and ResNet-SE. A number of experiments on these architectures show that our proposed method is effective and improves the systems performance a lot. The final submitted systems are the fusion of several models. In FFSVC2022, our best system achieves a minimum of the detection cost function (minDCF) of 0.511 and an equal error rate (EER) of 4.409% on the evaluation set.

Index Terms: speech verification, deep learning, domain adaption, metric learning

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 of various lengths into fixed dimensional embedding vectors for back-end scoring and verification [1, 2]. These works have achieved significantly superior results on speaker verification benchmark datasets as well as close-talk scenarios.

However, when ASV systems are deployed in far-field and noisy scenarios, their performance drops obviously. One key factor is the domain mismatch caused by the statistical difference between the training and evaluation datasets. Another critical factor of performance degradation is the mismatch between enrollment and test utterances. It is common for users to enroll their utterances via close-talking devices in quiet conditions but authenticate in the complex far-field environments where 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 compensate these cross-domain mismatches, many strategies have been proposed. One approach for x-vector systems with probabilistic linear discrim-

inative analysis (PLDA) back-end [3] is to apply domain adaption to back-end classifiers. After training of the x-vector network, a transformation of extracted embeddings is learned with the objective of reducing domain mismatch [4]. In [5], domain adaptation is performed by aligning the covariance of labeled out-of-domain and unlabeled in-domain data. A second approach is to design and train domain invariant embeddings. In [6], invariant representation learning method is introduced to improve the system robustness in reverberant and noisy conditions. Alternatively, domain adaptation loss functions have been proved that they can reduce the mismatch effectively [7, 8]. Study from [8] has shown that domain adversarial training with a gradient reversal layer to learn domain-invariant features can bring 0.8% absolute EER reduction in FFSVC2020 challenge [9]. In addition, metric learning to learn effective representations that have small intra-class distance and large inter-class distance is another commonly used approach to reduce domain mismatch which has been investigated in [10, 11, 12].

To promote the development of speaker verification on real application scenarios, Far-Field Speaker Verification Challenge (FFSVC) was first organized in 2020 [9]. In 2022, the specific objective for FFSVC2022 focus on single-channel far-field speaker verification scenarios under noisy conditions [13]. In this paper, we present our submitted system to the fully supervised far-field speaker verification task (task 1) of FFSVC2022. Inspired by previous work, we propose a novel multi-reader domain adaption learning framework based on asymmetric metric learning for deep learning based x-vector embeddings. We build two asymmetric data streams, which interlace to each other can mine considerably richer relationship compared with conventional one stream metric learning approaches. By extracting domain invariant embeddings, the domain adaption learning framework can not only alleviate the mismatch between training and testing data but also the mismatch between enrollment and authentication utterances.

In this work, our main contribution is that we propose a novel multi-reader domain adaption learning. In addition, a complete description of the system components, including front-end, advanced neural network based embedding extractor structures along with their configurations, and back-end are introduced. At last, we conduct a number of experiments to prove the effectiveness of proposed deep learning framework on different architectures.

The remainder of this 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. Experiments are presented in Section 4, followed by results and conclusions.

Table 1: *ResNet34-SE architecture configuration*. K denotes kernel size. C denotes channels.

Layer Name	Configurations
Conv1	$[K = 3 \times 3, C = 32] \times 1$
Res-SE-Block 1	$\begin{bmatrix} K = 3 \times 3, C = 32 \\ K = 3 \times 3, C = 32 \\ SE - Block \end{bmatrix} \times 3$
Res-SE-Block 2	$\begin{bmatrix} K = 3 \times 3, C = 64 \\ K = 3 \times 3, C = 64 \\ SE - Block \end{bmatrix} \times 4$
Res-SE-Block 3	$\begin{bmatrix} K = 3 \times 3, C = 128 \\ K = 3 \times 3, C = 128 \\ SE - Block \end{bmatrix} \times 6$
Res-SE-Block 4	$\begin{bmatrix} K = 3 \times 3, C = 256 \\ K = 3 \times 3, C = 256 \\ SE - Block \end{bmatrix} \times 3$
Statistics pooling	ASP
Linear	5120×256

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 with the entire utterance are selected for feature extraction. 80 dimensional logarithm Mel filter bank energies are generated with a hamming window of 25ms width and 10ms step. All features are cepstral mean normalized without extra voice activity detection (VAD).

2.2. Speaker embedding extractors

In total, we train 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 [14]. We introduce them in the following.

2.2.1. ResNet-SE

Residual networks [14], which are widely used in image recognition have recently been implemented in speaker recognition system [15]. Squeeze-and-excitation residual network (ResNet-SE) is a variant of ResNet that employs squeeze-and-excitation blocks to enable the network to perform dynamic channel-wise feature recalibration [16, 17]. 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 frame layers are followed by an attentive statistics pooling layer (ASP) [18] that calculates the mean and standard deviations of the final frame-level features to aggregate frame-level features into utterance-level features.

2.2.2. ECAPA-TDNN

ECAPA-TDNN is one of the state-of-the-art speaker verification models [19]. As a enhancement of original time delay neural network (TDNN) architecture, ECAPA-TDNN model intro-

Table 2: *ECAPA-TDNN1024 architecture configuration*. K denotes kernel size. C denotes channels.

Layer Name	Configurations
Conv1	$[K = 5, C = 1024]$
SE-Res2Block 1	$\begin{bmatrix} K = 1, C = 1024 \\ [K = 3, C = 128] \times 8 \\ K = 1, C = 1024 \\ SE - Block \end{bmatrix}$
SE-Res2Block 2	$\begin{bmatrix} K = 1, C = 1024 \\ [K = 3, C = 128] \times 8 \\ K = 1, C = 1024 \\ SE - Block \end{bmatrix}$
SE-Res2Block 3	$\begin{bmatrix} K = 1, C = 1024 \\ [K = 3, C = 128] \times 8 \\ K = 1, C = 1024 \\ SE - Block \end{bmatrix}$
Conv2	$[K = 1, C = 1536]$
Statistics pooling	ASP
Linear	3072×192

duces additional skip connections to propagate and aggregate channels throughout the system. Table 2 shows the model architecture we have used in this challenge. The number of SE-Res2Blocks is set to 3 with dilation values 2, 3 and 4. The number of channels is set to 1024. Attention statistic pooling (ASP) is used and 192 dimensional speaker embedding vectors are extracted.

2.3. Back-end

In this work, we use L2-normalization to converts extracted embedding vectors to unit vectors. We want to force learned embeddings to lie on a sphere and make the system focus on the angle. According experiment results [6, 20] 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 each development and evaluation trial, embedding vectors (e, t) of the entire enroll and test utterances are extracted and whole utterance similarity score is computed. We also extract N embedding vectors ($e_i, t_j, 1 \leq i, j \leq N$) from N equally-spaced segments in enroll and test utterances, compute the score matrix and get the average as matrix average score. As shown in (1), the final score is the sum of whole utterance similarity score and matrix average score.

$$score = \cos(e, t) + \frac{1}{N^2} \times \sum_{i=1}^N \sum_{j=1}^N \cos(e_i, t_j) \quad (1)$$

3. Multi-Reader domain adaption learning

In this section, we propose a novel multi-reader domain adaption learning framework based on asymmetric metric learning to solve system degradation problem caused by mismatch between enrollment and authentication utterances. For the following proposed learning method, the weights are initialized with the weights of pre-train models which are trained with Voxceleb

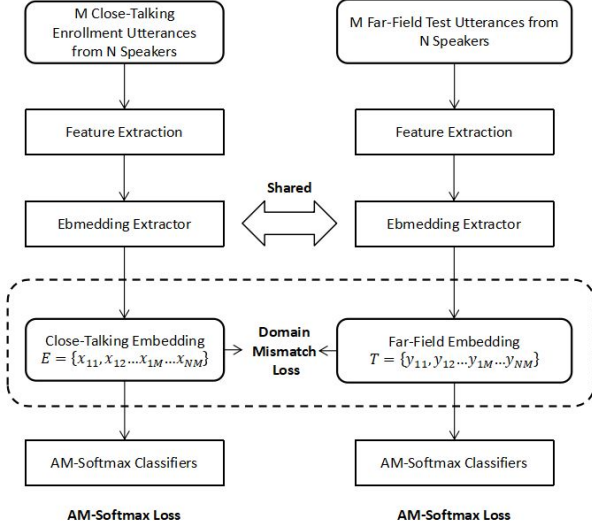


Figure 1: Domain adaption learning framework builds two asymmetric data streams of enrollment and test utterances in each training iteration to train a single network. Domain mismatch loss functions are proposed to align the embeddings.

dataset [21, 22].

3.1. Domain mismatch loss

Additive margin variant of Softmax loss function, AM-Softmax [23, 24], uses a cosine margin penalty to the target to increase inter-class variance and has achieved good performance in many tasks [11, 15]. However, it is sensitive to the parameters of scale and margin. Metric learning loss functions are alternatives to classification loss functions to learn embeddings directly with small intra-class and large inter-class distance [25, 26]. Inspired by prototypical loss function [27], we propose a domain mismatch loss function.

As shown in Figure 1, in fine-tuning stage, we generate a batch that contains N speakers, and M close-talking enrollment utterances and M far-field test utterances from each speaker. We feed these utterances into our systems described in Section 2. During training, in each batch, the extracted normalized enrollment speaker embedding vectors are denoted as $x_{i,j}$ and test speaker embedding vectors are denoted as $y_{i,j}$ where $1 \leq i \leq N$ and $1 \leq j \leq M$. The center of enrollment embeddings $\{x_{1,1}, x_{1,2}, \dots, x_{1,M}, \dots, x_{N,M}\}$ is denoted as $\{x_{c1}, x_{c2}, \dots, x_{cN}\}$ where x_{ck} is defined as

$$x_{ck} = \frac{1}{M} \sum_{j=1}^M x_{k,j} \quad (2)$$

The similarity matrix between each enrollment utterance embeddings and test utterance embeddings is defined as

$$s_{i,j,k} = \cos(y_{i,j}, x_{ck}) + b \quad (3)$$

where b is a learnable bias, $1 \leq i, k \leq N$ and $1 \leq j \leq M$. As shown in Figure 2, for a well trained system, the embedding vectors extracted from enrollment speech and test speech should have similar distribution in the embedding space. Therefore, the similarity should be large in gray areas and small in white areas. We define the target labels for similarity matrix are positive

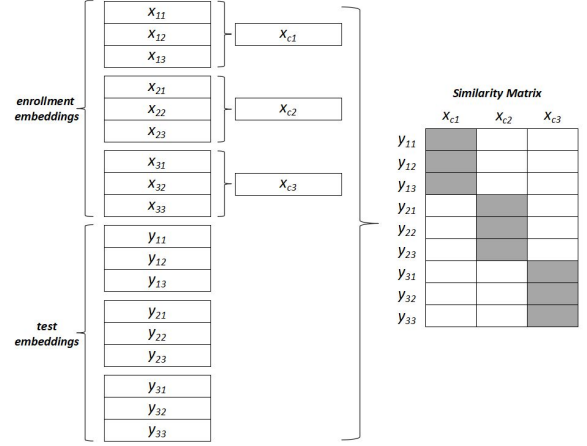


Figure 2: Similarity matrix of domain mismatch loss. The batch contains $N = 3$ speakers and each speaker has $M = 3$ close-talking and far-field test utterances.

when $i = k$ and negative when $i \neq k$. 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

$$L_{DM} = -\frac{1}{M \times N} \sum_{i=1}^N \sum_{j=1}^M \log \frac{e^{s_{i,j,i}}}{\sum_{k=1}^N e^{s_{i,j,k}}} \quad (4)$$

3.2. Multi-Reader transfer learning

Figure 1 illustrates the multi-Reader domain adaption learning process. Besides the domain mismatch loss $L_{DM}(E, T)$, we also calculate the additive margin Softmax loss of enrollment utterances $L_{AMS}(E)$ and of test utterances $L_{AMS}(T)$ separately. Finally, in this work, the combined loss is defined as

$$L = L_{DM}(E, T) + \alpha L_{AMS}(E) + \beta L_{AMS}(T) \quad (5)$$

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

4. Experiments

4.1. Training strategy

For FFSVC2022 task 1, only Voxceleb 1 and 2 dataset [21, 22], FFSVC2020 dataset and supplementary set [9, 13] can be used for training. We use Voxceleb2 corpus to pre-train all our models with AM-Softmax loss 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.

4.1.1. Fine-tune stage 1

In fine-tune stage 1, all models are fine-tuned by FFSVC2022 supplementary dataset. However, the number of speaker IDs in FFSVC2022 supplementary dataset is not enough, which may cause over-fitting problem. In this work, based on the multi-reader method in [28], we use two data streams. One contains small number of speaker IDs from in-domain FFSVC2022 data D_1 and the other is intended to classify a large number of speaker IDs from out-of-domain voxceleb data D_2 . The total

Table 3: Performance of our systems on the FFSVC2022 development set and evaluation set.

ID	Configurations	EER(%) (Dev)	minDCF (Dev)	EER(%) (Eval)	minDCF (Eval)
0	Baseline System	-	-	7.021	0.681
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	6.971	0.630
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	6.091	0.573

Table 4: Performance of fusion systems on the FFSVC2022 evaluation set.

System	EER(%) (Eval)	minDCF (Eval)
Baseline System	7.021	0.681
Fusion1	5.556	0.524
Fusion2	4.409	0.511

loss is calculated as

$$L(D_1, D_2) = L(D_1) + \lambda L(D_2) \quad (6)$$

where λ is the regularization parameters. In this stage, adam optimizer is set to 0.0001 learning rate with 0.95 decay each epoch.

4.1.2. Fine-tune stage 2

Then in fine-tune stage 2, according to trial case settings in FFSVC2022 evaluation plan [13], close-talking and iphone recorded audios are selected as enrollment dataset and other audios in FFSVC2022 supplementary data are selected as test dataset. We utilize proposed multi-reader domain adaption learning framework described in Section 3 to future tuning all models.

4.2. Data augmentation

In pre-train stage, music, noise and speech part from MUSAN dataset [29] are used as additive noise with random SNR setting from 5db to 20db. In addition, we also use SpecAugment method described in [30] to enhance the system’s robust in both 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, keyboard and television noise. We randomly add this noise to the test utterances in fine-tune stage.

5. Results Analysis

Table 3 and 4 display results of experiments on systems as well as fusion systems on FFSVC2022 development dataset and final evaluation dataset. The primary metric to evaluate system performance is the Minimum Detection Cost(minDCF) with $P_{tar} = 0.01$ and both C_{miss} and C_{fa} are equal to 1.0. In addition, Equal Error Rate (EER) is selected as performance criteria.

As shown in Table 3, firstly, ECAPA-TDNN based systems perform better than and ResNet-SE based system. As a cost, parameter size and amount of computation of ECAPA-TDNN based systems is also larger.

In addition, comparing to the system without any augmentation, it is clearly data augmentation in fine-tune stage can upgrade system performance. For ResNet-SE and ECAPA-TDNN systems, data augmentation get about 5% and 2% EER improvement. Score integration method describe in Section 2 also contributes to the system improvement compared with single cosine similarity scoring method.

Finally, Proposed multi-reader domain adaption learning framework in fine-tune stage 2 boosts the system performance a lot. For EER criteria, the system performances are improved by about 11% and 10% respectively.

Fusion system performance results on evaluation dataset are shown in Table 4. For Fusion 1, scores of systems ID5 to ID8 are linearly weighted into one fusion score. For Fusion 2, all scores of systems in Table 3 are linear combined together.

6. 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 multi-reader domain adaption training framework based on asymmetric metric learning to solve the system degradation caused by mismatch between close-talk enrollment and far-field test speech. In FFSVC2022, our best system achieves minDCF of 0.511 and EER of 4.409% on the evaluation phase.

7. 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*, 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 *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.
- [3] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *2007 IEEE 11th International Conference on Computer Vision*, 2007, pp. 1–8.
- [4] P.-M. Bousquet and M. Rouvier, "On Robustness of Unsupervised Domain Adaptation for Speaker Recognition," in *Proc. Interspeech 2019*, 2019, pp. 2958–2962.
- [5] M. J. Alam, G. Bhattacharya, and P. Kenny, "Speaker Verification in Mismatched Conditions with Frustratingly Easy Domain Adaptation," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2018)*, 2018, pp. 176–180.
- [6] W. Chen, J. Huang, and T. Bocklet, "Length- and Noise-Aware Training Techniques for Short-Utterance Speaker Recognition," in *Proc. Interspeech 2020*, 2020, pp. 3835–3839.
- [7] G. Bhattacharya, M. J. Alam, and P. Kenny, "Adapting end-to-end neural speaker verification to new languages and recording conditions with adversarial training," 05 2019, pp. 6041–6045.
- [8] L. Zhang, J. Wu, and L. Xie, "NPU Speaker Verification System for INTERSPEECH 2020 Far-Field Speaker Verification Challenge," in *Proc. Interspeech 2020*, 2020, pp. 3471–3475.
- [9] 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*, 2020, pp. 3456–3460.
- [10] R. Duroselle, D. Jouviet, and I. Illina, "Metric Learning Loss Functions to Reduce Domain Mismatch in the x-Vector Space for Language Recognition," in *Proc. Interspeech 2020*, 2020, pp. 447–451.
- [11] J. S. Chung, J. Huh, S. Mun, M. Lee, H.-S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, and I. Han, "In Defence of Metric Learning for Speaker Recognition," in *Proc. Interspeech 2020*, 2020, pp. 2977–2981.
- [12] Y. Lei, X. Huo, Y. Jiao, and Y. K. Li, "Deep Metric Learning for Replay Attack Detection," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 42–46.
- [13] X. Qin, M. Li, H. Bu, S. Narayanan, and H. Li, "Far-field speaker verification challenge (ffsvc) 2022 : Challenge evaluation plan," 2022.
- [14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [15] W. Xie, A. Nagrani, J. S. Chung, and A. Zisserman, "Utterance-level aggregation for speaker recognition in the wild," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5791–5795.
- [16] 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," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5814–5818.
- [17] 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*, 2021, pp. 1094–1098.
- [18] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive Statistics Pooling for Deep Speaker Embedding," in *Proc. Interspeech 2018*, 2018, pp. 2252–2256.
- [19] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech 2020*, 2020, pp. 3830–3834.
- [20] 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*, 2020, pp. 3476–3480.
- [21] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," in *Proc. Interspeech 2017*, 2017, pp. 2616–2620.
- [22] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Proc. Interspeech 2018*, 2018, pp. 1086–1090.
- [23] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.
- [24] 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," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [25] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, 2005, pp. 539–546 vol. 1.
- [26] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 815–823.
- [27] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.
- [28] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4879–4883.
- [29] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," <http://arxiv.org/abs/1510.08484>, 2015.
- [30] 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*, 2019, pp. 2613–2617.