

Cross-Domain ArcFace: Learning Robust Speaker Representation Under the Far-Field Speaker Verification

Yuke Lin¹, Xiaoyi Qin¹, Ming Li^{1,2}

¹School of Computer Science, Wuhan University, Wuhan, China

²Data Science Research Center, Duke Kunshan University, Kunshan, China

linyuke0609@gmail.com

Abstract

The system of speaker verification system shows outstanding performance with the assistance of different types of loss functions with angular margin penalty, which can enforce the intra-class compactness and inter-class discrepancy. However, the power of classification may degrade largely when encountering the cross-domain problems, especially in far-field scenes. Thus, we propose a novel Cross-Domain ArcFace(CD-ArcFace) loss function. By adopting distinct margin penalty in different domain when conducting mix-data fine-tuning, the performance of various speaker verification system can be further improved. This experiment is carried on FFSVC2022. The final score level of our fusion system for the task1 achieves 4.028% and 4.368% EER on the development set and evaluation set.

Index Terms: Far-Field, Speaker verification, Cross-Domain.

1. Introduction

Automatic speaker verification (ASV) is a bio-metric technology that helps to judge whether a pair of utterance belongs to the same speaker or not. With the development of computing power, deep learning based speaker verification system presents outstanding results and gradually becomes the mainstream. Nowadays, impressive performance has been achieved by constructing deep speaker embedding with large neural network scale like ECAPA-TDNN[1] and ResNet[2]. However, those system would degrade severely when employed in a domain-mismatch scenario. Several efforts have been paid to improve the cross-domain robustness in various field like far-field(FFSVC2020, VOICES) and cross-lingual(VoxSRC2021, SdSV)scenario,etc. Among all of them the far-field problem is the most noteworthy. The energy decay and reverberation of audios may mislead the optimization direction and degrade the speaker verification performance.

To enforce higher similarity for intra-class samples and diversity for inter-class samples, margin-based softmax methods would be a better choice when compared with traditional softmax loss function. As a variant of softmax loss, angular softmax(A-softmax)[3] maps features into hypersphere space and shows promising results. More recently, depending on more discriminative embeddings and stronger geometric interpretability, AM-softmax[4] and AAM-softmax(ArcFace)[5] generally become preferred on the ASV system. However, those loss functions merely focus on single-domain scenario. Thus, this article raise a novel cross-domain ArcFace(CD-ArcFace).By allocating different margin to different domain data in mix-data fine-tuning, the domain-gap can be further alleviated.

This paper introduces our proposed our system for FFSVC task 1, which incorporates the mentioned CD-ArcFace. The rest of this paper is organized as follows. In section 2, we describe our novel proposed CD-ArcFace. Section 3 describes our mix-data training pipeline and the models we used. Experiment settings are presented in section 4, While 5 discusses the results based on our experiments. Conclusions are provided in Section 6.

2. Cross-Domain ArcFace

2.1. ArcFace(AAM-Softmax)

ArcFace(also named AAM-softmax)[5] is one of the most popular and fine-grained loss function in speaker verification problems.It is an extension of traditional softmax loss function which introduce an L2-normalization step on the embeddings and an angular margin penalty is added to the penalty when estimating the log-likelihood during the training process. Additionally, it has a clear geometric interpretation due to the exact correspondence to the geodesic distance on the hypersphere, which can be of great help to enforce the model increase inter-speaker distances and ensure intra-speaker compactness. The ArcFace loss function is given by:

$$L = -\frac{1}{n} \sum_{i=1}^n \log \frac{e^{(s \cos(\theta_{y_i} + m))}}{e^{s(\cos(\theta_{y_i} + m))} + \sum_{j=1, j \neq y_i}^N e^{s \cos(\theta_j)}} \quad (1)$$

with n indicates the batch size, θ_{y_i} represents the angle between the current speaker embedding x_i and the AAM-softmax class prototype with speaker identity y_i . The margin penalty is indicated with m , which can be interpreted as a metric of compactness of classification. A scaling factor s is applied to increase the range of the output log-likelihoods.

2.2. Cross-Domain ArcFace

Previous research[6] has shown that higher values of m will result in more compact classes with large inter-class distance, which allow network to capture more abstract features and improve the ability to classify different identities. However, large margin initialization brings much difficulties may confuse the network, which makes it hard to converge. In our transfer learning process, the target domain data is put into the model together with source domain data. Thus, speaker embedding from source domain would be easier for classifier to discriminate when compared with target domain data in the fine-tuning progress. In other words, large margin fine-tuning still works when it comes to the source domain data. As for the target domain, a small initialization of margin would be beneficial.

Here we conduct a toy experiment to prove that the margin of data from different domain should be differentiated when

Corresponding Author: Ming Li.

carrying on mix-data training. We randomly select 5 speaker from source domain and 5 speakers from target domain. Each speaker has approximately 500 samples. The features extracted from the pre-trained encoder pass through the t-SNE and the dimension reduces from 128 to 2. As is shown on Fig.2, data from the source domain is relatively compact compared to target domain, which indicates that different decision boundary can be applied. Meanwhile, the number of utterance per speaker of target domain is basically large, so it should start with a smaller margin to make it easier to converge.

$$L = -\frac{1}{n} \sum_{i=1}^n \log \frac{e^{(s \cos(\theta_{y_i} + m_i))}}{e^{(s \cos(\theta_{y_i} + m_i))} + \sum_{j=1, j \neq y_i}^N e^{s \cos(\theta_j)}} \quad (2)$$

$$m_i = \begin{cases} m_s & x_i \in \mathcal{A}_s \\ m_t & x_i \in \mathcal{A}_t \end{cases} \quad (3)$$

The proposed Cross-Domain ArcFace is presented above. Here the \mathcal{A} represents mixed training samples which consists of samples from source domain and target domain. m_s and m_t are the hyper-parameters based on the degree of difference between source domain and target domain. The rest parameters refer to Equation (1).

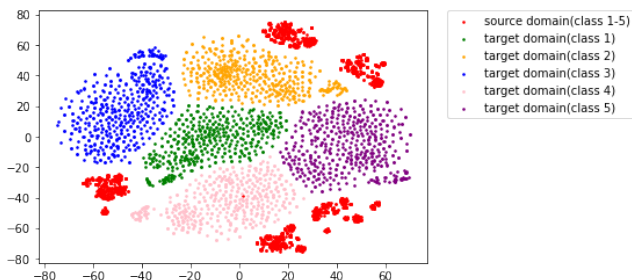


Figure 1: The scatter diagram 2D-speaker representations through T-SNE. Source domain and target domain share half of the samples respectively.

3. Training Framework

3.1. Transfer learning pipeline

Transfer learning is a common strategy to transfer the knowledge learned from a general scenario to a specific scenario. By freezing a certain number of parameters or layers of network, the ability of extracting abstract information of system can be partially reserved. Transfer learning, can be also interpreted as domain adaptation. A simple method is by the means of finetuning pre-trained model from source domain data with target domain data, the domain gap can be diminished accordingly. The tactic is proved to be feasible in some previous research[7],[8] in speaker verification problem.

As is shown in Figure1, the front-end feature extractor is pre-trained with large-scaled source domain data. In our experiment, there is an imbalance between the number of categories of source domain data and target domain data, which makes it easier to overfit on the target domain. Therefore, in the fine-tuning process, the target domain data is input into model together with source domain data. Speed perturbation is used to increase the number of input audios and speakers. Additionally, all parameters are jointly optimized till the convergence

under a small number of epochs and learning rate, which can be helpful to avoid over-fitting as well. After the speaker embedding is collected, the cross-domain ArcFace assists to make those from same speaker more compact and make those from different speaker looser.

3.2. Deep Speaker Embedding Model

In this part, three different speaker verification systems are introduced, which are consist of ResNet-SimAM[9], ResNet-SE[10] and the ECAPA-TDNN[1]. The acoustic features are 80-dimensional log Mel-filterbank energies with a frame length of 25ms and hop size of 10ms. The extracted features are mean normalized before feeding into the deep speaker network.

3.2.1. ResNet-SE

In this experiment, the ResNet34 structure is utilized as the front pattern extractor, which can transform acoustic features into frame level features. The width of the residual blocks is 64, 128, 256, 512 basically. The Squeeze and Excitation block is placed after each residual block, which can capture global channel information. Then, the global statistic pooling(GSP) concatenates the calculated mean and deviation of the output feature map, which integrates variable-length features into fix-length features. Finally, the bottleneck linear layer convert the high-dimension vector to the low-dimension utterance-level vector we expect.

3.2.2. ResNet-SimAM

Simple attention module(SimAM) has been proved to be effective in both computer vision field[11] and speaker verification field. Different from other attention module, the SimAM is designed based on some well-known neuroscience theories, which is more interpretable. The pooling layer we adopt here is attentive statistics pooling(ASP)[12]. the classifier is the same as the ResNet-SE System.

3.2.3. ECAPA-TDNN

The ECAPA-TDNN Network achieves great success in the speaker verification task and provides the start-of-the-art performance. In this experiment, 1024 feature channels are used to scale up the network. The dimension of the bottleneck in the SE-Block is set to 128. The front-end feature extractor is followed by an attentive statistics pooling (ASP) layer that calculates the mean and standard deviations of the final frame-level features.

4. Experiment Settings

4.1. Data usage

The experiment is conducted with following datasets:

- VoxCeleb 2.
- FFSVC2020 supplementary set & dev set.
- FFSVC2022 development set

VoxCeleb2[13] contains 1,092,009 utterances from 5,994 speakers, which is treated as the source domain data. The FFSVC2020 supplementary with dev set comprises 1,213,766 utterances from 154 different speakers. The latter dataset is considered as target domain data for fine-tuning. As for evaluation, we adopt the FFSVC2022 development set for hyperparameters tuning.

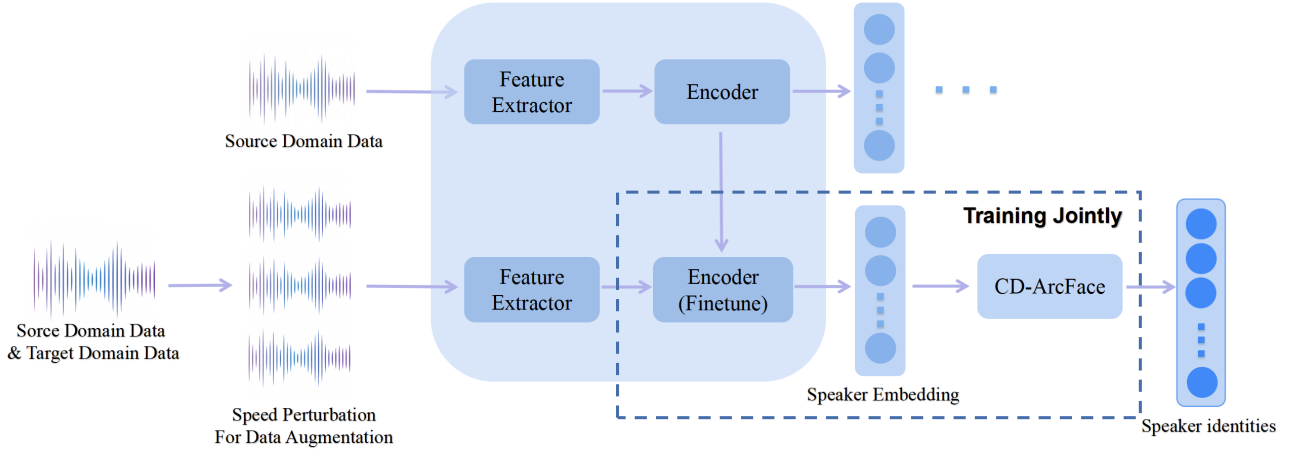


Figure 2: The brief illustration of the proposed system. The encoder optimized from source domain data shares its parameter to the encoder below. The back-end classification process of pre-trained model is omitted here.

4.2. Data augmentation

In our experiment, the pre-training and fine-tuning progress adopt different augmentation strategy. The model pre-trained from source domain data adopts different types of noises to improve the generalization ability. Specifically, the MUSAN[14] and RIR Noise are applied as the additive source of noise and room impulse response functions respectively. We also amplified or change the speed of audio signals to further improve the diversification of data.

To reduce the domain gap, only the noise of RIR is introduced in the fine-tuning process. Mean while, speed perturbation has been verified to make a difference on the performance of SV system[15].Hence, We speed up or down each utterance by 0.8,0.9,1.1,1.2 times during fine-tuning, and the utterances with different speeds are considered from new speaker. Finally the number of utterances increase from 2,305,775(1,092,009 from source domain and 1,213,766 from target domain) to 11,528,875 with the speaker number increase from 6,148(5,994 from source domain and 154 from target domain) to 30,740.

4.3. Model setup

4.3.1. pre-trained settings

VoxCeleb2 is treated as the source domain data. For feature extraction, logarithmical Mel-spectrogram is extracted by applying 80 Mel filters on the spectrogram computed over Hamming windows of 20ms shifted by 10ms. After 4-round warm-up epochs, the learning rate(LR) falls every 15 epochs from initial 0.1 to 0.0001 until its performance no longer decreases on the development set. The SGD optimizer is adopted to update the model parameters and batch size is set to 128.The ArcFace(s = 32,m = 0.2) is used as the classifier.

4.3.2. fine-tune settings

FFSVC supplement set is used as the target domain data. The characteristic of input features are same as that of pre-trained input features.The m_s is set 0.3 and the m_t is set 0.1. The batch size is 64 at this phase. In case the model overfits on target

domain, the learning rate is limited to 0.001 at the initial stage and drops to 0.00001 within 3 epochs.

4.4. Score Normalization and System Fusion

The cosine similarity is the back-end scoring method. After scoring, results from all trials are subject to score normalization. We utilize Adaptive Symmetric Score Normalization (AS-Norm) in our systems. Specifically, we utilize the AS-Norm1, which is defined as Eq.4. Specifically, only top-30 of the target training data is utilized as cohort set to compute mean and deviation for normalization.

$$s(e, t)_{as-norm1} = \frac{1}{2} \left(\frac{s(e, t) - \mu(S_e(\varepsilon_e^{top}))}{\sigma(S_e(\varepsilon_e^{top}))} + \frac{s(e, t) - \mu(S_t(\varepsilon_t^{top}))}{\sigma(S_t(\varepsilon_t^{top}))} \right) \quad (4)$$

In the system fusion stage, we adopt manual calibration and automatic calibration. According to the system performance in the development data set, we adopt the score level fusion that assigns weights to different models. Considering that the model may overfit on the development set with manual calibration, we use the BOSARIS Toolkit[16] for calibrating.

5. Results and discussion

5.1. Performance on the Original Dataset

Table 1: The performances of different speaker verification systems on the VoxCeleb1 original test set.

| Model | Vox-O | |
|--------------|--------|--------|
| | EER[%] | minDCF |
| ResNet34-SE | 0.956 | 0.104 |
| ResNet-SimAM | 0.845 | 0.085 |
| ECAPA-TDNN | 1.127 | 0.121 |

Table 2: The performances of different speaker verification systems and fusion system on the FFSVC 2022 development and evaluation set.

| ID | Model | strategy | Dev | | Eval | |
|-------|---------------------|--------------|--------|--------------|--------|--------|
| | | | EER[%] | minDCF | EER[%] | minDCF |
| | BaseLine | - | - | - | 7.021 | 0.681 |
| 1 | ECAPA-TDNN | pre-train | 11.675 | 0.824 | - | - |
| | | +ft-mix | 6.643 | 0.535 | - | - |
| | | +CD-ArcFace | 6.394 | 0.512 | 6.760 | 0.600 |
| | | +As-Norm | 6.133 | 0.497 | 6.712 | 0.598 |
| 2 | ResNet34-SE | pre-train | 9.978 | 0.772 | - | - |
| | | +ft-mix | 4.947 | 0.532 | - | - |
| | | +CD-ArcFace | 4.700 | 0.515 | 5.067 | 0.511 |
| | | +As-Norm | 4.675 | 0.488 | 4.969 | 0.503 |
| 3 | ResNet-SimAM | pre-train | 9.235 | 0.727 | - | - |
| | | +ft-mix | 4.789 | 0.493 | - | - |
| | | +CD-ArcFace | 4.369 | 0.476 | 4.643 | 0.491 |
| | | +As-Norm | 4.354 | 0.469 | 4.575 | 0.486 |
| | fusion | 2+3 | 4.042 | 0.463 | 4.488 | 0.466 |
| 1+2+3 | | 4.028 | 0.456 | 4.368 | 0.458 | |

5.2. Performance on the FFSVC challenge

Table 2 shows our different system for FFSVC 2022 task 1. The pre-train strategy indicates that the model optimized from source domain is evaluated on the target domain directly. The difference between ft-mix and CD-ArcFace is the former employs ArcFace as classifier, while the latter uses the cross-domain ArcFace, which is introduced in Sec2. It should be noted that the AS-Norm is composed with CD-ArcFace system to present the best performance. "Fusion" represents the result of fusing all the listed single systems with score weighted $\{0.4, 0.4, 0.2\}$ for "1+2+3" and $\{0.5, 0.5\}$ for "2+3" respectively.

6. Conclusion

From table 2 our proposed CD-ArcFace is capable of diminish the domain gap generally when conducting mix-dataset transfer learning. No matter on which model, CD-ArcFace can further optimize the system performance more and less. Therefore, CD-ArcFace can be proved to learn representation with more robustness when encountering far-field scene.

7. References

- [1] 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.
- [2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [3] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6738–6746.
- [4] 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.
- [5] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.
- [6] 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.
- [7] X. Qin, D. Cai, and M. Li, "Far-Field End-to-End Text-Dependent Speaker Verification Based on Mixed Training Data with Transfer Learning and Enrollment Data Augmentation," in *Proc. Interspeech 2019*, 2019, pp. 4045–4049.
- [8] X. Qin, C. Wang, Y. Ma, M. Liu, S. Zhang, and M. Li, "Our Learned Lessons from Cross-Lingual Speaker Verification: The CRMI-DKU System Description for the Short-Duration Speaker Verification Challenge 2021," in *Proc. Interspeech 2021*, 2021, pp. 2317–2321.
- [9] X. Qin, N. Li, C. Weng, D. Su, and M. Li, "Simple attention module based speaker verification with iterative noisy label detection," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6722–6726.
- [10] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.
- [11] L. Yang, R.-Y. Zhang, L. Li, and X. Xie, "Simam: A simple, parameter-free attention module for convolutional neural networks," in *International conference on machine learning*. PMLR, 2021, pp. 11 863–11 874.
- [12] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive Statistics Pooling for Deep Speaker Embedding," in *Proc. Interspeech 2018*, 2018, pp. 2252–2256.
- [13] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Proc. Interspeech 2018*, 2018, pp. 1086–1090.
- [14] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," 2015, arXiv:1510.08484v1.
- [15] H. Yamamoto, K. A. Lee, K. Okabe, and T. Koshinaka, "Speaker augmentation and bandwidth extension for deep speaker embedding," in *Interspeech*, 2019, pp. 406–410.
- [16] N. Brümmer and E. De Villiers, "The bosaris toolkit: Theory, algorithms and code for surviving the new dcf," *arXiv preprint arXiv:1304.2865*, 2013.