MULTI-OBJECTIVE PROGRESSIVE CLUSTERING FOR SEMI-SUPERVISED DOMAIN ADAPTATION IN SPEAKER VERIFICATION

Ze Li[1,2],Yuke Lin[1,2], Ning Jiang [3], Xiaoyi Qin[1,2], Guoqing Zhao[3], Haiying Wu[3], Ming Li[1,2]

[1]School of Computer Science, Wuhan University, Wuhan, China
[2]Suzhou Municipal Key Laboratory of Multimodal Intelligent Systems,
Duke Kunshan University, Kunshan, China
[3]Mashang Consumer Finance Co., Ltd, China
ming.li369@dukekunshan.edu.cn

ABSTRACT

Utilizing the pseudo-labeling algorithm with large-scale unlabeled data becomes crucial for semi-supervised domain adaptation in speaker verification tasks. In this paper, we propose a novel pseudo-labeling method named Multi-objective Progressive Clustering (MoPC), specifically designed for semi-supervised domain adaptation. Firstly, we utilize limited labeled data from the target domain to derive domain-specific descriptors based on multiple distinct objectives, namely within-graph denoising, intra-class denoising and inter-class denoising. Then, the Infomap algorithm is adopted for embedding clustering, and the descriptors are leveraged to further refine the target domain's pseudo-labels. Moreover, to further improve the quality of pseudo labels, we introduce the subcenter-purification and progressive-merging strategy for label denoising. Our proposed MoPC method achieves 4.95% EER and ranked the 1st place on the evaluation set of VoxSRC 2023 track 3. We also conduct additional experiments on the FFSVC dataset and yield promising results.

Index Terms— Speaker Recognition, Semi-supervised, Domain Adaptation

1. INTRODUCTION

Automatic speaker verification (ASV) aims to verify the identities of speakers from their voice samples. In recent years, thanks to the advancement of deep neural networks (DNN), deep learning-based SV systems, such as [1–3], have demonstrated impressive performance under different scenarios. Nevertheless, when an ASV system well trained on the source domain is applied to another domain, the performance may degrade significantly. Therefore, domain adaptation is an important and challenging issue in speaker verification.

To address this problem, researchers propose various methods to bridge the domain gap. These methods can be broadly categorized into statistical alignment [4, 5], generative adversarial [6, 7], self-supervised learning [8] and pseudo-labeling [9, 10]. Statistical alignment and generative adversarial approaches mainly focus on leveraging domain distribution information to bridge the domain gap. However, this often leads to overlooking finer-grained details and class-specific information. Self-supervised learning based methods utilize the correlation within target domain data, but the system performance still needs to be improved. Pseudo-labeling based methods achieve competitive results in ASV domain adaptation tasks. It iteratively utilizes clustering algorithms to assign pseudo-labels to unlabeled target domain data and subsequently uses these pseudo-labels

---

Corresponding Author: Ming Li.

for supervised learning. However, the accuracy of the pseudo-labels largely depends on the robustness of the clustering algorithm and the reliability of the source domain's pre-trained model.

Many existing works [4–13] focus on unsupervised domain adaptation, assuming the absence of labels in the target domain. However, considering that manually annotating a small amount of target data is easy to achieve, semi-supervised domain adaptation emerges as a more realistic challenge, which aims to learn target models from a few labeled target domain data and a large amount of unlabeled target domain data. For pseudo-labeling approaches, a straightforward method is to merge limited labeled data with large amount of pseudo-labeled data together for training. However, we believe that the labeled target domain data contains valuable domain-specific information, which aids in filtering out noisy labels from the pseudo-labeled data, thereby enhancing the performance of domain adaptation.

In this study, we propose a novel pseudo-labeling method named Multi-objective Progressive Clustering (MoPC) for semi-supervised domain adaptation. To begin with, we define multiple objectives, including within-graph denoising, intra-class denoising and inter-class denoising. These objectives can be quantified from a small amount of labeled data to improve the embedding clustering quality of unlabeled data. Furthermore, we also introduce the subcenter-purification strategy to remove noisy classes and adopt the progressive-merging strategy to further enhance the quality of pseudo-labels.

2. METHODS

2.1. Domain-specific Descriptor Extraction

To enhance the quality of graph clustering, we define multiple objectives: within-graph denoising, intra-class denoising and inter-class denoising. Then, we utilized the limited labeled data from the target domain to derive the domain-specific descriptors (denoted as Noise-Edge Descriptor(NED), Intra-Class Descriptor(ICD) and Class-Merging Descriptor(CMD), respectively) based on these objectives. The methods of deriving these descriptors are described as follows:

Descriptor NED . We compute the pairwise cosine similarity between each embedding and those from different classes, then select the maximum value as NED:

$$NED = \max_{\substack{1 \le i \ne j \le K \\ 1 \le a \le N_i, 1 \le b \le N_j}} cosine\left(\mathbf{z}_{i,a}, \mathbf{z}_{j,b}\right) \tag{1}$$
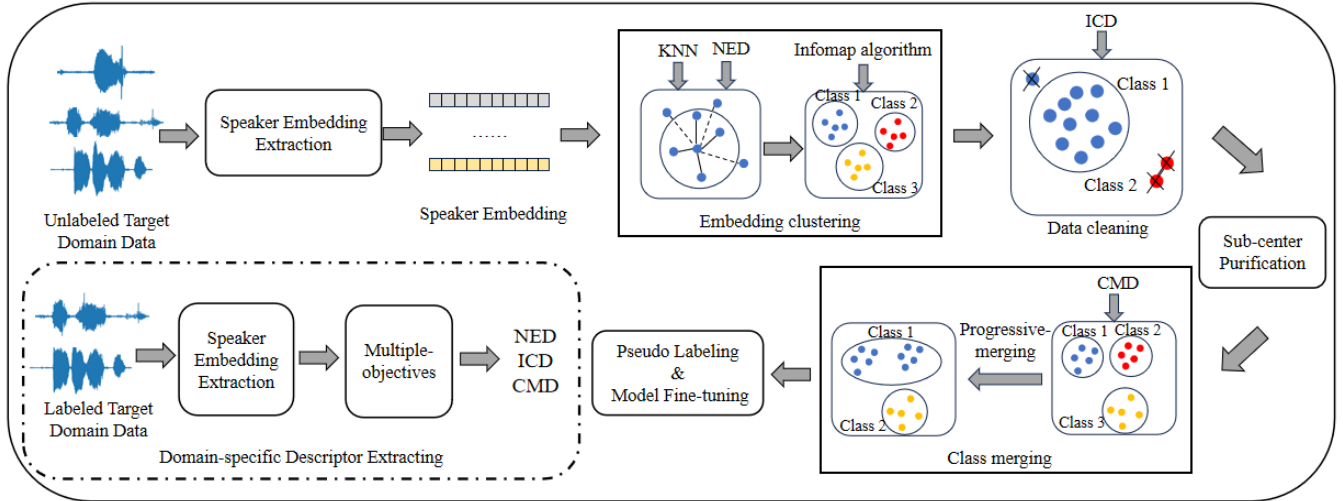
Fig. 1. System framework of our MoPC method.

where $\mathbf{z}_{i,a}$ is the ath embedding of the ith class, K represents the total number of classes and N_i denotes the total number samples of the ith class. The embeddings are extracted by the model trained on the source domain.

Descriptor ICD . We calculate the cosine similarity between the embedding of each class and its respective centroid vector and select the maximum one from the minimum values of each class as ICD:

$$ICD = \max_{1 \le i \le k} \min_{1 \le a \le N_i} cosine\left(\mathbf{z}_{i,a}, \mathbf{C}_i\right) \qquad (2)$$

where $\mathbf{C}_i$ is the centroid vector of the ith class.

Descriptor CMD . We compute the cosine similarity between the centroid vectors of each class, then select the maximum one as CMD:

$$CMD = \max_{1 \le i \ne j \le k} cosine\left(\mathbf{C}_i, \mathbf{C}_j\right) \qquad (3)$$

2.2. Overall Description of the MoPC Framework

This section describes the proposed overall framework of our MoPC pseudo-labeling method for semi-supervised domain adaptation. The framework is shown in Fig 1, and the steps are described as follows:

Step 1: Speaker embedding extraction. Firstly, we utilize the model trained on the source domain as a feature extractor, to obtain speaker embeddings from target domain data. To maintain stability, utterances less than one second are discarded.

Step 2: Embedding clustering. We generate the graph with speaker embedding using the K-Nearest Neighbors (KNN) algorithm; the parameter K is determined by the 'elbow' method [14, 15]. Descriptor NED represents the highest similarity among embedding that do not belong to the same class. Therefore, it is highly probable that embeddings with similarity higher than NED belong to the same class. Hence, we retain edge weights greater than NED to eliminate within-graph noise. Then, the Infomap [16] algorithm is employed for clustering based on the graph.

Step 3: Data cleaning. Descriptor ICD represents the highest similarity of edge embeddings to their respective class centers. There-

fore, embeddings with similarity higher than ICD to their class center are highly likely to belong to that class. Hence, we retain embeddings with similarity higher than ICD to their class center to eliminate intra-class noise. Additionally, eliminate classes with deficient numbers of instances.

Step 4: Sub-center purification. To further improve the quality of pseudo labels, we introduce a subcenter-purification strategy to remove noisy classes by the soft labels output from the Sub-Center head. For more detailed information, please refer to Section 2.3.

Step 5: Progressive class merging. To reduce inter-class noise, we employ a progressive-merging strategy to merge classes with the same ground-truth label while making every effort to avoid the increase in intra-class noise caused by erroneous merges. For more detailed information, please refer to Section 2.4.

Step 6: Pseudo labeling and model fine-tuning. After completing the aforementioned steps, we assign pseudo-labels to unlabeled target domain data based on the clustering results. Then, both the unlabeled data with pseudo-labels and the labeled data are used to fine-tune the pre-trained speaker model.

2.3. Sub-center Purification

To address the issue of class impurity, inspired by the sub-center ArcFace [17] loss, we introduce the subcenter-purification strategy. In this process, we begin by assigning pseudo-labels to the unlabeled data after Step 3, using these as the input to train a Sub-center Arc-Face classifier. After convergence, we stop training and pass all the data through the classifier to compute the selection probability for each class's sub-center. As shown in Fig 2, most of the highly pure class data will tend to choose one specific sub-center, while the class with lower purity will exhibit multiple sub-centers, and we could consider removing it.

2.4. Progressive Class Merging

During the clustering process, there may be cases where multiple classes share the same ground-truth label. In such instances, it is necessary to merge these classes; however, we also need to avoid the increase in intra-class noise caused by incorrect merges. To achieve
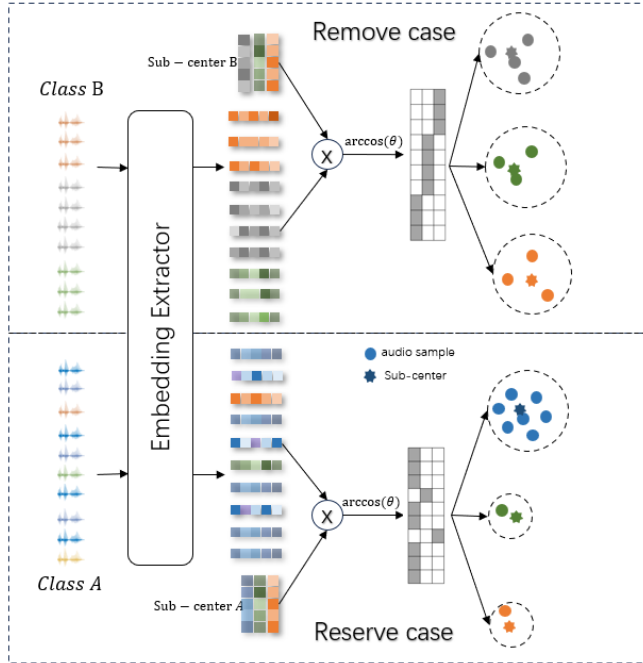
Fig. 2. Sub-center Purification. Noisy data in one class tends to be uniformly distributed into various sub-centers (Top), while purer class data is more likely to select a distinct sub-center (Bottom).
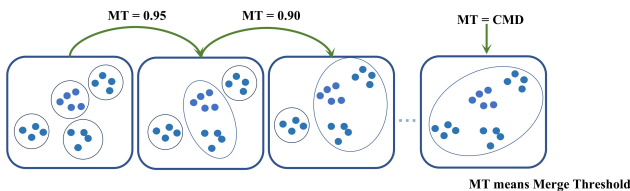
Fig. 3. Progressive class merging. We set a series of descending threshold values to merge classes with high center similarities iteratively.

this, inspired by the Agglomerative Hierarchical Clustering, we introduce the progerssive-merging strategy. As shown in Fig 3. Firstly, we set a series of merge thresholds ranging from higher value down to the descriptor CMD. Then, for each merge threshold, we merge those classes whose class center similarities satisfy the threshold condition and are each other's closest neighbors.

3. EXPERIMENTS

3.1. Dataset

We employ two datasets from different languages or scenarios in our experiments to perform domain adaptation. The VoxCeleb2 [18] dev set with speaker labels is used as the labeled data from the source domain. For unlabeled target domain data, we use the following two datasets: a subset of the CNCeleb2 [19] dev set provided by VoxSRC23 challenge [1], the FFSVC2020 supplementary set provided

---

[1] http://mm.kaist.ac.kr/datasets/voxceleb/voxsrc/competition2023.html

---

by FFSVC2022 challenge [20]. To perform semi-supervise domain adaptation, we further select the small subset of CNCeleb (1,000 utterances from 50 speakers) provided by VoxSRC23 challenge as the labeled target domain data for CNCeleb2, and randomly select 1,000 utterances from 10 speakers in the FFSVC2020 supplementary set as the labeled target domain data for FFSVC.

For evaluation, we use the official validation and test trial list of the VoxSRC23 Track 3 task and the official development trial list of FFSVC2022 to evaluate the systems trained in the semi-supervised manner.

3.2. Model Usage

We utilize two ResNet-Based systems for model training, including the ResNet100-based one [21] and the ResNet152-based one [22]. In addition, we also adopt the SimAM [23] modules in the backbone block. The acoustic features are 80-dimensional log Mel-filterbank energies with a frame length of 25ms and a hop size of 10ms. The extracted features are mean-normalized before feeding into the deep speaker network. We also introduce the temporal statistic pooling (TSP) and the attentive statistic pooling (ASP) [24] to obtain fix-dimension vectors.

3.3. Training Details

Our training process can be divided into two stages, including (i) Supervised learning on the source domain and (ii) Semi-supervised domain adaptation learning on the target domain.

3.3.1. Supervised Learning Settings

Since the source domain data is labeled, it can be used to train a feature extractor. In this phase, we adopt the on-the-fly data augmentation [25] to add additive background noise or convolutional reverberation noise for the time-domain waveform. The speed perturbation [26], which speeds up or down each utterance by a factor of 0.9 or 1.1, is applied to yield shifted pitch utterances that are considered from new speakers. The SGD optimizer with a momentum of 0.95 and weight decay of 1e-4 is used. We adopt the StepLR scheduler with 15 epochs decay. The initial learning rate starts from 0.1, the minimum learning rate is 1.0e-4, and the decay factor is 0.1. The margin and scale of ArcFace are set as 0.2 and 32, respectively. We perform a linear warm-up learning rate schedule at the first 5 epochs to prevent model vibration and speed model training. The input frame length is fixed at 200 frames.

3.3.2. Semi-Supervised Learning Settings

In this stage, labeled and pseudo-labeled target domain data are used to fine-tune the pre-trained model. Only speaker augmentation with speed perturbation is retained for data augmentation, and the ArcFace is replaced by the Sub-center ArcFace. The learning rate starts from 1.0e-3 and gradually drops till convergence.

Considering the size of the target domain data, we employ distinct fine-tuning strategies for VoxSRC Track3 and FFSVC. For VoxSRC Track3, the source model is directly fine-tuned using the target domain data. In contrast, the FFSVC dataset, with its fewer speakers, demands a cautious approach to avoid overfitting. Hence, we set the initial learning rate to 1.0e-5 and apply the Mix-FT strategy[27]. Additionally, for experiments on the VoxSRC23 dataset, AS-norm and QMF mentioned in [22] are used for score calibration.

Table 1. The results of different backbone systems using our MoPC based semi-supervised learning framework on the VoxSRC 22/23 Track 3 datasets.

ID & Model	VoxSRC23 val		VoxSRC23 test		VoxSRC22 test	
	EER[%]	mDCF$_{0.05}$	EER[%]	mDCF$_{0.05}$	EER[%]	mDCF$_{0.05}$
1 SimAM-ResNet100-ASP	7.490	0.342	5.287	0.3037	6.927	0.409
2 ResNet100-TSP	7.350	0.360	-	-	-	-
3 SimAM-ResNet100-ASP(v2)	7.525	0.335	-	-	-	-
4 ResNet152-ASP	7.240	0.347	-	-	-	-
5 ResNet152-TSP	7.535	0.358	-	-	-	-
Fusion(1+2+3+4+5)	6.725	0.311	4.952	0.2777	6.584	0.374
Fusion [10] *	-	-	-	-	7.030	0.388
Fusion [9] †	-	-	-	-	7.153	0.389

* PGMVG clustering method, 1st place in VoxSRC 2022 track 3
† k-means clustering method, 2st place in VoxSRC 2022 track 3

4. RESULTS

We utilize the ground-truth labels of the target domain data and the following metrics to access the quality of pseudo-labels: intra-class noise rate, inter-class noise rate, number of pseudo-label classes, number of ground-truth classes, and NMI [15]. For a class, if the frequency of a particular label is the highest, that label can be considered the primary label for that class. Intra-class noise rate refers to the proportion of samples occupied by labels other than the primary label within the class. Inter-class noise rate is the proportion of samples occupied by classes with the same primary label.

Table 2 displays the effectiveness of each step of our MoPC method. As we can see, in the scenario where only the Infomap algorithm is applied as the primary clustering method, both the intra-class noise and inter-class noise rates are quite high, reaching 11% and 34.4%, respectively. After we introduce the descriptor NED to eliminate within-graph noise, we can observe a significant decrease in both intra-class noise rate and inter-class noise rate, and the NMI has shown some improvement. Furthermore, due to the fact that some classes have very few samples, after removing the within-graph noise, the embeddings of these classes cannot form edges with embeddings from other classes and are ultimately treated as isolated points and removed. Consequently, the number of ground-truth classes decreases by over one hundred. After introducing the descriptor ICD and subcenter-purification strategy to improve the quality of pseudo-labels further, we can see an improvement in all metrics. After introducing the descriptor CMD with the progressive-merging strategy to eliminate inter-class noise, we see a significant decrease in inter-class noise rate, which has dropped from 28.5% to 17.1%. Due to some inevitable erroneous merges, the metric intra-class noise rate has increased from 1.71% to 7.7%. However, at this point, the number of pseudo-label classes is closer to the number of ground-truth label classes, the improvement in the performance of the inter-class noise rate far outweighs that of the intra-class noise rate, and the NMI metric has not decreased significantly. Therefore, we believe that the quality of pseudo-labels at this point is superior to the previous state.

Table 1 reports the results on different VoxSRC 2022 and 2023 Track 3 evaluation sets. As we can see, compared to both PGMVG and k-means methods, our approach is significantly ahead. Even on the SimAM-ResNet100-ASP single-system, our model outperforms the results achieved by their fused model on the VoxSRC 2022 Track 3 evaluation set. Furthermore, our fused system achieved 4.952% EER and ranked the 1st place on the VoxSRC 2023 Track 3 evaluation set.

Additionally, we conducted experiments on the FFSVC dataset to assess the universality of our method. It can be observed in Table 3 that our method is equally effective on the FFSVC dataset. Compared to the k-means method, our method exhibits an improvement of nearly 20% on the FFSVC 2022 development set.

Table 2. NR$_1$ and NR$_2$ indicates the Intra-class and Inter-class noise rate, respectively. P and GT are Pseudo label and ground truth label.

Method	NR$_1$[%]	NR$_2$[%]	Spk. of P	Spk. of GT	NMI
k-means	12.3	42.6	1800	1807	0.9179
Based (Infomap)	11.0	34.4	2128	1807	0.9412
+ NED	2.4	29.9	2080	1642	0.9783
++ ICD	2.2	29.7	2079	1636	0.9794
+++ subcenter	1.7	28.5	2070	1626	0.9827
++++ CMD	7.7	17.1	1705	1626	0.9811

Table 3. The results of the SimAM-ResNet100-ASP single system on the FFSVC22 development set.

Method	FFSVC-dev	
	EER[%]	mDCF$_{0.01}$
Based	10.097	0.736
+ k-means	5.147	0.523
+ MoPC(ours)	4.125	0.461

5. CONCLUSION

In this paper, we proposed a novel pseudo-labeling method named Multi-objective Progressive Clustering (MoPC) for semi-supervised domain adaptation. The differences between our semi-supervised approach and previous methods lies in the more extensive utilization of the limited amount of target domain labeled data. The subcenter-purification and progressive-merging strategy are also introduced to improve the quality of pseudo labels further. Experimental results on CNCeleb2, FFSVC show that our proposed method is effective on different datasets.

6. ACKNOWLEDGEMENT

References

[1] Weicheng Cai, Jinkun Chen, and Ming Li, "Exploring the encoding layer and loss function in end-to-end speaker and language recognition system," *arXiv preprint arXiv:1804.05160*, 2018.

[2] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *Proc.ICASSP*. IEEE, 2018, pp. 5329–5333.

[3] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.

[4] Baochen Sun, Jiashi Feng, and Kate Saenko, "Correlation alignment for unsupervised domain adaptation," *Domain adaptation in computer vision applications*, pp. 153–171, 2017.

[5] Kong Aik Lee, Qiongqiong Wang, and Takafumi Koshinaka, "The coral+ algorithm for unsupervised domain adaptation of plda," in *Proc. ICASSP*. IEEE, 2019, pp. 5821–5825.

[6] Gautam Bhattacharya, Joao Monteiro, Jahangir Alam, and Patrick Kenny, "Generative adversarial speaker embedding networks for domain robust end-to-end speaker verification," in *Proc. ICASSP*. IEEE, 2019, pp. 6226–6230.

[7] Johan Rohdin, Themos Stafylakis, Anna Silnova, Hossein Zeinali, Lukáš Burget, and Oldřich Plchot, "Speaker verification using end-to-end adversarial language adaptation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6006–6010.

[8] Zhengyang Chen, Shuai Wang, and Yanmin Qian, "Self-supervised learning based domain adaptation for robust speaker verification," in *Proc.ICASSP*. IEEE, 2021, pp. 5834–5838.

[9] Xiaoyi Qin, Na Li, Yuke Lin, Yiwei Ding, Chao Weng, Dan Su, and Ming Li, "The dku-tencent system for the voxceleb speaker recognition challenge 2022," *arXiv preprint arXiv:2210.05092*, 2022.

[10] Zhenduo Zhao, Zhuo Li, Wenchao Wang, and Pengyuan Zhang, "The hccl system for voxceleb speaker recognition challenge 2022," *arXiv preprint arXiv:2305.12642*, 2023.

[11] Zhenyu Wang, Wei Xia, and John HL Hansen, "Cross-domain adaptation with discrepancy minimization for text-independent forensic speaker verification," *arXiv preprint arXiv:2009.02444*, 2020.

[12] Garrett Wilson and Diane J Cook, "A survey of unsupervised deep domain adaptation," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 11, no. 5, pp. 1–46, 2020.

[13] Zhengyang Chen, Shuai Wang, and Yanmin Qian, "Adversarial domain adaptation for speaker verification using partially shared network.," in *Interspeech*, 2020, pp. 3017–3021.

[14] Robert L Thorndike, "Who belongs in the family?," *Psychometrika*, vol. 18, no. 4, pp. 267–276, 1953.

[15] Danwei Cai, Weiqing Wang, and Ming Li, "Incorporating visual information in audio based self-supervised speaker recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1422–1435, 2022.

[16] Usha Nandini Raghavan, Réka Albert, and Soundar Kumara, "Near linear time algorithm to detect community structures in large-scale networks," 2007, vol. 76, p. 036106, APS.

[17] Jiankang Deng, Jia Guo, Tongliang Liu, Mingming Gong, and Stefanos Zafeiriou, "Sub-center arcface: Boosting face recognition by large-scale noisy web faces," in *Proc. ECCV*, 2020, pp. 741–757.

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

[19] Lantian Li, Ruiqi Liu, Jiawen Kang, Yue Fan, Hao Cui, Yunqi Cai, Ravichander Vipperla, Thomas Fang Zheng, and Dong Wang, "Cn-celeb: multi-genre speaker recognition," *Speech Communication*, vol. 137, pp. 77–91, 2022.

[20] Xiaoyi Qin, Ming Li, Hui Bu, Shrikanth Narayanan, and Haizhou Li, "The 2022 Far-field Speaker Verification Challenge: Exploring domain mismatch and semi-supervised learning under the far-field scenario," in *Proc. The 2022 Far-field Speaker Verification Challenge (FFSVC2022)*, 2022, pp. 10–14.

[21] Rostislav Makarov, Nikita Torgashov, Alexander Alenin, Ivan Yakovlev, and Anton Okhotnikov, "ID R&D System Description to VoxCeleb Speaker Recognition Challenge 2022," .

[22] Ze Li, Yuke Lin, Xiaoyi Qin, Ning Jiang, Guoqing Zhao, and Ming Li, "The dku-msxf speaker verification system for the voxceleb speaker recognition challenge 2023," *arXiv preprint arXiv:2308.08766*, 2023.

[23] Xiaoyi Qin, Na Li, Chao Weng, Dan Su, and Ming Li, "Simple attention module based speaker verification with iterative noisy label detection," in *Proc. ICASSP*, 2022, pp. 6722–6726.

[24] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive Statistics Pooling for Deep Speaker Embedding," *Proc. Interspeech*, 2018.

[25] W. Cai, J. Chen, J. Zhang, and M. Li, "On-the-Fly Data Loader and Utterance-Level Aggregation for Speaker and Language Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1038–1051, 2020.

[26] Weiqing Wang, Danwei Cai, Xiaoyi Qin, and Ming Li, "The DKU-DukeECE Systems for VoxCeleb Speaker Recognition Challenge 2020," *arXiv.2010.12731*.

[27] Xiaoyi Qin, Danwei Cai, and Ming Li, "Robust multi-channel far-field speaker verification under different in-domain data availability scenarios," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 71–85, 2023.