Robust Multi-Channel Far-Field Speaker Verification under Different In-Domain Data Availability Scenarios

Xiaoyi Qin, *Student Member, IEEE*, Danwei Cai, *Student Member, IEEE*, Ming Li, *Senior Member, IEEE*

***Abstract*—The popularity and application of smart home devices have made far-field speaker verification an urgent need. However, speaker verification performance is unsatisfactory under far-field environments despite its significant improvements enabled by deep neural networks (DNN). In this paper, we summarize our previous work and propose multiple training strategies and models for the multi-channel far-field speaker verification with different in-domain data availability scenarios. This paper takes the experiments on the FFSVC20 dataset, a far-field multi-channel speaker verification dataset. In the FFSVC20 dataset, we proposed the cross-device and cross-domain trial files, e.g., enrollment data is chosen from single-channel close-talking cellphone audios and the test data comes from multi-channel far-field microphone array audios. We focus on the single-channel and multi-channel speaker verification training based on the dataset. For single-channel speaker verification, considering the size of training data and availability of labels, we introduce three training scenarios and given our proposed training methods, including 1) given zero out-of-domain data and few in-domain labeled data; 2) given large-scale out-of-domain labeled data and few in-domain labeled data; 3) given large-scale out-of-domain labeled data and few in-domain unlabeled data. To this end, we propose a meta-learning approach, refined transfer learning methods, and semi-supervised learning for three scenarios, respectively. For multi-channel speaker verification, we first introduce two types of 3 dimension convolution (3D Conv) residual network (ResNet) model proposed in our previous works, including fully 3D ResNet and incorporating 3D Conv with 2D Conv ResNet (3D2D-ResNet). In this paper, we propose channel-wise 3D squeeze-and-excitation ResNet (C3DSE-ResNet) and spatial-wise 3D SE ResNet (S3DSE-ResNet) to further explore the channel dependencies and improve the 3D ConvNet performance. The results show that the proposed strategies and models can significantly boost the performance under the far-field scenario.**

***Index Terms*—speaker verification, far-field, multi-channel**

I. Introduction

Automatic speaker verification (ASV) is a biometric technology that verifies the speaker identity based on audio signals. Over the past few years, deep speaker framework based on time-delay neural network (TDNN) [1]–[4] or ResNet [5], [6] has significantly improved the performance of ASV systems under the settings of close-talking and clean recording environments. However, the performance of ASV systems is limited by the scenarios of low signal-to-noise ratio (SNR) or cross-domain mismatch conditions, such as far-field [7], [8] and cross-lingual [9]–[11]. For the far-field scenario, the recording quality of the collected audio is affected by energy decay, noise and reverberation, which causes data quality degradation and domain mismatch between the close-talking training data and far-field testing data.

Speaker verification under this challenging far-field setting has attracted a lot research interests. Multiple challenges such as VOiCES [7], [12] and FFSVC [8], [13] were launched to foster research in this field. Different methods, including front-end signal processing [14]–[16], domain adaptation [17]–[20], and joint learning of speech enhancement and speaker representation learning [21], [22], have been proposed for far-field speaker verification. Following these research works, this paper further investigates far-field speaker verification under both single and multiple channel settings.

Far-field speaker verification can be formulated as a domain adaptation task given a large-scale close-talking dataset and a far-field dataset. For single channel data, we apply different training strategies considering the size of the far-field data and the availability of the label. With a relatively small size of far-field dataset, transfer learning is used to adapt the model to in-domain far-field data. When only few far-field data samples is provided for training, a meta-learning method is proposed to perform domain adaptation. With an unlabeled far-field dataset, we propose a semi-supervised approach to generate pseudo-labels for the unlabeled data before transfer learning.

For multi-channel data, existing methods adopt front-end processing of beamforming [14], [15], [23], [24] or joint learning of speech enhancement and ASV [21]. However, front-end speech enhancement may damage the speech quality and thus affect the verification performance of close-talking data [15], [25], [26]. In this paper, we directly model multi-channel signals to learn speaker embeddings. Specifically, we propose to apply 3-dimensional (3D) convolutional network (ConvNet) to the 3D input feature (multi-channel spectrogram) on our previous work [27]. The proposed multi-channel training framework utilizes the information carried out by multiple speech observations at different spatial locations and simultaneously processes the time-, frequency- and channel-information to learn a robust speaker embedding. Considering the large computational cost and memory consumption of full 3D ConvNet model, we propose to incorporate 3D convolution with 2D convolution to reduce the model size. To further explore the channel dependencies, we extend the attention module of squeeze-excitation (SE) [28] from the 2D ConvNet layer to the 3D ConvNet layer for the proposed multi-channel framework in this paper.

Xiaoyi Qin and Ming Li are from School of Computer Science at Wuhan University and Data Science Research Center at Duke Kunshan University. Danwei Cai is from Data Science Research Center at Duke Kunshan University. Corresponding Author: Ming Li, ming.li369@duke.edu

For both single and multiple channel settings, most studies used data simulation to obtain far-field audio due to the lack of real data. However, the simulation can not perfectly match the real data. The simplification of the recording environment's interior structure during the simulation process of the image source method (ISM) [29] leads to the domain gap between the simulated and genuine data. To fill this gap, we conduct the experiments on FFSVC20 dataset[1] [8]: a far-field multi-channel dataset recorded in real world environment. The dataset consists of multiple recording devices with a variety of distances. The recording contents include fixed text and free text. Also, we design the challenging cross-channel testing trials to investigate the conditions of close-talking enrollment and far-field testing.

This paper extends our previous works on far-field speaker verification. In our previous works, we released a large-scale far-field speaker verification dataset (FFSVC20) [8], adopted the transfer learning on far-field ASV field (only discuss the fine-tuning with in-domain data scenario) [25], [30] and proposed multi-channel training frameworks with 3D ConvNet for far-field speaker model [27]. Compared to those studies, we propose multiple training strategies for far-field speaker verification with different in-domain data availability scenarios and improve modeling using new proposed modules with multi-channel far-field data in this paper. To sum up, the main contributions in this paper are:

- Systematically discuss the performance of fine-tuning strategy on the far-field field, including fine-tuning only with in-domain data (FT-domain) and fine-tuning using both in-domain and out-of-domain data (FT-mix).
- Investigate two new training scenarios for far-field data: when few far-field data is provided, meta-learning is applied to far-field speaker verification; when an unlabeled far-field dataset is given, semi-supervised is used.
- Further explore the channel dependencies for far-field multi-channel training framework, squeeze-excitation (SE) attention module is extended to 3D ConvNet. The channel-wise 3D squeeze-and-excitation module and spatial-wise 3D squeeze-and-excitation module are proposed.

The remaining paper is organized as follows. We briefly discuss related works in Section II. Section III details the FFSVC20 dataset. Section IV describes different training strategies for different data size and labeling scenarios. Section V introduces the details of the proposed multi-channel training framework. The experimental setting and results are presented in Section VI and VII. Section VIII gives the conclusion.

II. Related works

As mentioned before, the speech intelligibility and quality of far-field audio is affected by long-range fading, room reverberation, and environmental noises. To compensate the adverse impacts, signal processing methods aim at improving the speech quality or extracting robust acoustic features for far-field audio. Speaker modeling methods aim at learning a robust speaker embedding from far-field data via domain adaptation or advanced DNN architecture.

A. Front-end processing for far-field speaker verification

At the feature level, sub-band Hilbert envelopes based features [31]–[33], warped Minimum Variance Distortionless Response (MVDR) cepstral coefficients [34], Blind Spectral Weighting (BSW) based features [35], Power-Normalized Cepstral Coefficients (PNCC) [36], [37] and DNN bottleneck features [38] have been applied to ASV system to suppress the adverse impacts of reverberation and noise. Also, Liu *et al.* [39] proposed to jointly optimize the parameters of per-channel energy normalization or parameterized cepstral mean normalization with DNN speaker embedding extractor.

At the signal processing level, DNN based denoising method of single-channel speech enhancement [40]–[43] has been applied for noise robust speaker recognition. Also, Shi *et al.* [21] proposed to jointly optimize speech enhancement with speaker modeling to improve speaker verification performance in various acoustic conditions. For multi-channel signal enhancement, beamforming has been successfully applied in far-field speaker verification [14], [15], [23], [44]. Yang and Chang [45] proposed to jointly optimize the DNN acoustic beamforming and dereverberation with speaker embedding extractor for far-field audio. To reduce the reverberation level in far-field audio, dereverberation method of weighted prediction error (WPE) [46], [47] is also applied for far-field speaker verification [37], [48].

The usage of front-end signal processing modules may potentially increase the model complexity and distort the speaker information. Therefore, we focus on the training strategy and end-to-end modeling in this paper. The conventional operations of beamforming and dereverberation methods will be considered as a comparison in experimental results.

B. Speaker Embedding Extraction for far-field speaker verification

For single channel far-field audio data, Jati *et al.* [49] propose to apply a discriminative model that hybridizes DNN and total variability model for speaker verification. Zhao *et al.* [50] propose a novel DNN architecture of channel-interdependence enhanced Res2Net for single channel far-field speaker verification. To reduce the mismatch between the far-field audio and the close-talking audio, adversarial learning [15], [17], [51] and transfer learning [18], [25], [52] are also applied for single channel far-field speaker verification.

The performance of far-field speaker verification systems is also investigated under the setting of ad-hoc microphone arrays whose spatial arrangement are unknown. DNN attention mechanism has been applied to aggregate speaker embeddings of different recordings captured by different channels of the distributed microphone arrays [53], [54]. Also, Liang *et al.* propose the spatio-temporal processing block at the frame-level to capture the contextual relationship in both channel and time axis.

Compared with related work that pays attention to the single-channel and ad-hoc data modeling, we focus on the

[1]http://2020.ffsvc.org/DataDownload

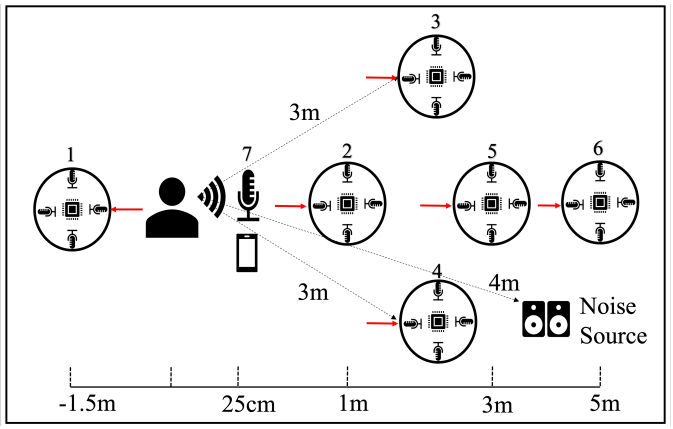


Fig. 1. Room setting of the FFSVC20 dataset. The red arrow points to channel 0 of each microphone array.

multi-channel data modeling and propose new multi-channel attention modules based on the previous work [27].

III. The FFSVC20 dataset

The development of smart home devices gives rise to open research questions for speaker verification in the far-field environments. Most research use data simulation to generate far-field data due to the lack of a far-field dataset collected in real. However, far-field simulation does not perfectly match the real data. To this end, we provided the FFSVC20 dataset to the speaker verification community [8]. FFSVC20 dataset is the part of DMASH dataset[2]. Compared with the HIMIA dataset [30], FFSVC20 dataset provides more recording device and richer text. The HI-MIA dataset provides 340 speaker data consists of one HIFI microphone and multiple microphone arrays. However, the content are "hi,mia" and "ni hao,mi ya" only. The FFSVC 20 dataset, which is recorded with multiple devices in two different room settings, exhibits far-field characteristics in real scenarios. Fig.1 shows the recording setup of the FFSVC20 dataset. The recording devices include one close-talking microphone, one smartphone at a distance of 25 centimeters, and six four-channel microphone arrays at different locations as shown in Fig.1. The shape of the microphone array is circular with a 5 centimeters radius.

The dataset is recorded in Mandarin. The text content includes the fixed text of '*ni hao, mi ya*' and other text-independent utterances. During recording, each speaker visited three times with a time gap of 7 to 15 days to ensure recording diversity. For a single visit of one speaker, the first 30 recordings are text-dependent utterances and the remaining recordings are with free texts. Different recording environments are set for different recording visits: for the first visit, the noise sources include television and electric fan noises; for the second visit, audio data is recorded in a clean environment without noise; for the third visit, a working electric fan is used as the noise source.

The dataset includes 395 speakers: the training set contains 120 speakers, the development set contains 35 speakers and the evaluation set contains 240 speakers. To measure the speaker verification performance in different scenarios, we define three evaluation tasks as follows:

- Task 1: far-field text-dependent speaker verification with single microphone array.
- Task 2: far-field text-independent speaker verification with single microphone array.
- Task 3: far-field text-dependent speaker verification with distributed microphone arrays.

The evaluation set is equally divided into three non-overlapped subsets with 80 speakers. Different subsets are used to construct verification trials for different evaluation tasks.

To match the application in real scenario, the recording from the close-talking cellphone is used for enrollment; and the recording from far-field microphone array is used for testing. For any target trial, the enrollment and the testing utterances are from different visits of the speaker. Under this evaluation protocol, a challenging cross-domain trial list is constructed. In this paper, we focus on task 1 and task 2.

IV. Single-Channel Far-Field Speaker Verification

In real applications, far-field speaker verification sometimes can be regarded as a domain adaptation task given a large-scale close-talking speech dataset and a far-field dataset. This section introduces various training strategies for far-field speaker verification systems according to different in-domain data availability scenarios.

A. Transfer Learning

Transfer learning, also known as domain adaptation [55], is the strategy we used to train a far-field speaker verification model. Under this strategy, the far-field model is fine-tuned from a pre-trained one trained on a large-scale general speaker recognition dataset. Thus we can transfer the discriminative speaker knowledge from the pre-trained model and reduce domain mismatch problems.

Generally, collecting a large-scale labeled far-field dataset is time-consuming and expensive. Thus datasets for far-field speaker verification are usually in small-scale, while models training on a small dataset can be easily overfitted. In this case, transfer learning is widely used to improve the speaker verification performance in data sparse scenarios [25], [56]. Typically, the large-scale general dataset is regarded as the out-of-domain data, while the small far-field dataset is considered as the in-domain data. The adaptation process of the transfer learning strategy is shown in Fig.2, which contains the following steps [25]:

- Pre-train a deep speaker embedding model using the large-scale dataset with sufficient speakers;
- Retain all parameters of the model except for the output speaker classification layer; replace the speaker classifier with respect to the number of speakers in the in-domain far-field training data;
- Finetune and adapt the the new model with the in-domain data until it converges. All parameters, including those

[2]https://www.aishelltech.com/DMASH_Dataset

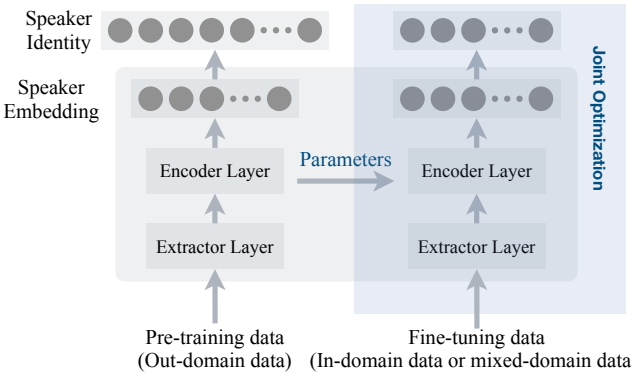


Fig. 2. Adaptation pipeline of the transfer learning strategies.

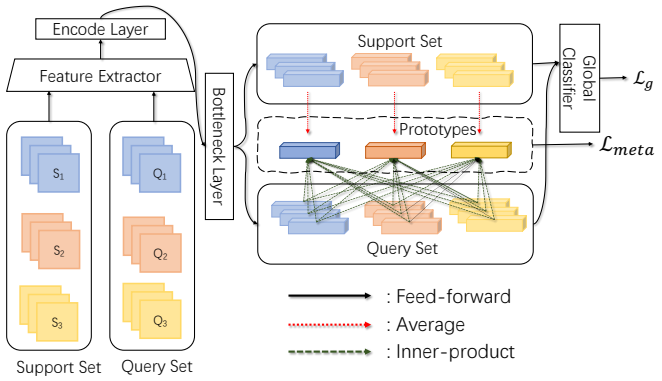


Fig. 3. Pipeline of our proposed meta learning method. Data and features of the same speaker are represented by the same color.

from the pre-trained model and the new speaker classifier, are jointly optimized.

Normally, only in-domain data is used to adapt the pre-trained model from the source domain to our target domain (named **FT-domain**, we proposed in [25]). Since the number of speakers from the in-domain dataset is much smaller, the pre-trained model can be easily over-fitted to such limited data. Thus the discrimination and generalizability of the adapted model may be degraded. To perform domain adaptation with in-domain data while maintaining the discrimination and generalizability of the pre-train model, we propose to use both in-domain and out-of-domain data together (mixed-domain data) to fine-tune the pre-trained model (named **FT-mix**). Compared to the seen data (i.e., out-of-domain data), unseen data (i.e., in-domain data) result in larger losses and gradients in forward and backward propagation. Therefore, we aim to achieve a consistent adaptation of the model regarding of the cross-domain mismatch and reduce the over-fitting by using the mixed-domain data. In addition, the method proposed in [18] adopts two classification heads to fine-tune the model, one for the in-domain data and the other for the out-of-domain data. For FT-mix, the out-of-domain and in-domain speakers share the same classification head. Compared with the [18], the FT-mix is easy to implement.

B. Meta-learning

In some commercial scenarios, where the existed large-scale datasets, e.g., VoxCeleb, are unauthorized, it is difficult to apply the transfer learning strategy. Therefore, we adopt meta-learning to address the performance degradation for the far-field speaker verification with a few in-domain data available. Meta-learning is another domain adaptation strategy when the model lacks data for training. Different from transfer learning where the pre-trained model possesses a wealth of prior knowledge, meta-learning aims to help the model learn new concepts and skills fast with only a few training samples.

Recently, meta learning based metric learning is employed in speaker recognition task [57]–[59] that uses the prototype network [60] to make query set close to the support set. In meta learning scenario, the training set is usually divided into two set, namely the support set and the query set. During the training process, the cost function assesses the performance on the query set for each batch given the corresponding support set. Cai et al. [61] also propose within-sample variability-invariant loss to learn the same embedding among the clean utterance and its noisy copies. Inspired by those works, we introduce a metric learning based meta-learning approach that combines the advantages of prototype network and angular margin softmax [62] to obtain the discriminative embeddings well from a few examples (see Fig. 3).

1) Meta learning with domain gap pairs: Since the enrollment data is close-talking and test data is recorded from any distance, we tackle the problem with mini-batch meta-learning that each mini-batch consists of the support and query sets of speaker classes. For each class, the mean of its support set is taken as the prototype and enforce the query examples to become closer to its own prototype. To simulate the practical scenario, the support sets and the query sets consist of the utterances randomly sampled from close-talking and far-field data.

Specifically, for each mini-batch, we first randomly sample N classes from the given dataset and then sample S and Q examples from each class as the support set and query set, respectively. As a result, we have a support set $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N\times S}$ and a query set $\mathcal{Q} = \{(\hat{\mathbf{x}}_j, \hat{y}_j)\}_{j=1}^{N\times Q}$, where $\mathbf{x}$ and $\hat{\mathbf{x}}$ are input features, y and $\hat{y} = \{1, \cdots, N\}$ are the class labels. The loss calculation for the prototypical network is described as follows.

As with [57], [60], we calculate prototypes of classes by averaging over support sets and forcing examples of queries to become closer to their own prototypes. First, the S_c is defined as the set of support examples in class c and then the prototype of each class $c = 1, \cdots, N$ is computed in a mini-batch:

$$\mathbf{P_c} = \frac{1}{|\mathcal{S}_c|} \sum_{\mathbf{x}\in\mathcal{S}_c} (f_\theta(\mathbf{x})) \tag{1}$$

where $|S_c|$ counts the number of S_c, $\mathbf{P_c}$ is the prototype vector with speaker embedding dimensions of class c and the $f_\theta(\cdot)$ is the model for speaker embedding extraction. Consider the characteristic of ASV, the cosine similarity of each query example with prototypes is compute as the distance metric :

$$d(\hat{\mathbf{x}}_i, \mathbf{P_c}) = \frac{f_\theta(\hat{\mathbf{x}}_i) \cdot \mathbf{P_c}}{\|f_\theta(\hat{\mathbf{x}}_i)\|\|\mathbf{P_c}\|} \tag{2}$$

Each query example is classified as N speakers based on a Softmax function over distances to each speaker prototype, and final loss of meta learning is the combination of the Softmax and cross-entropy loss:

$$\mathcal{L}_{meta} = -\frac{1}{|\mathcal{Q}|} \sum_{(\hat{\mathbf{x}} \in \mathcal{Q})} \log \frac{e^{d(\hat{\mathbf{x}}_i, \mathbf{P_c})}}{\sum_{c'=1}^{N} e^{d(\hat{\mathbf{x}}_i, \mathbf{P_{c'}})}} \tag{3}$$

where the c and c' denote as the class that $\hat{\mathbf{x}}_i$ belong to and any class in a mini-batch, respectively.

2) Global classification: The proposed meta-learning scheme can make the far-field embedding closer to the close-talking embedding. However, if only considering a limited number of speakers N according to the mini-batch composition, it is difficult to exclude the influence brought by speaker-unrelated factors. This makes it difficult to achieve a discriminative embedding space as the centroids of the speaker embeddings are unstable. Inspired by [57], [60], we introduce a global classification strategy to make the speaker embedding distributed on a hypersphere. In this experiment, the ArcFace [62], a modified Softmax with angle margin, is adopted as global classification. The loss function follows,

$$L_g = -\frac{1}{|\mathcal{Q}| + |\mathcal{S}|} \sum_{i=1}^{N} log \frac{e^{s(\cos(\theta_{y_i} + m))}}{e^{s(\cos(\theta_{y_i} + m))} + \sum_{j=1, j \neq y_i}^{n} e^{s \cos \theta_j}} \tag{4}$$

where $cos(\theta_{y_i}) = \frac{\mathbf{W}_{y_i}^T f(\mathbf{x}_i)}{\|\mathbf{W}_{y_i}\| \|f(\mathbf{x}_i)\|}$ and $\{\mathbf{x}_i, y_i\} \in \mathcal{Q} \cup \mathcal{S}$ denotes the speaker embedding feature of the i-th sample from the y_i-th class. m is the angular margin penalty between $\mathbf{x}_i$ and $\mathbf{W}_{y_i}$.

The total loss combines the loss in meta learning in Equation 3 with the global loss in Equation 4:

$$L_{total} = L_{meta} + L_g \tag{5}$$

C. Semi-supervised learning

In this part, we investigate scenarios where the in-domain far-field data is unlabeled. Such situation often occurs in recordings of unregistered participants. To make use of unlabeled data in training, we need to adopt strategies that involves unsupervised learning. One of the strategies is the semi-supervised learning, which generate pseudo-labels using another labeled dataset. Most semi-supervised algorithms, including temporal ensembling [63] and mean-teachers [64], are used under the closed-set protocol – the unlabeled training data belongs to the classes within the labeled training data. In our case, we adopt a large-scale dataset as the out-of-domain labeled data and treat the FFSVC dataset as the in-domain unlabeled data. The labeled data is used to pre-train a model for pseudo-labeling. To apply semi-supervised learning in the transfer learning phase, the speaker embedding from unlabeled data is extracted by using the pre-train model, and the followed clustering algorithms to generate pseudo-labels for the unlabeled data. Pseudo-labels are generated by the following algorithm:

- Step 1. Extract all speaker embeddings $\mathcal{Z} \in \mathbb{R}^{N \times d}$ from the FFSVC20 dataset using the pre-trained speaker model.
- Step 2. Run a clustering algorithm with different number of clusters K to obtain centroid matrix $\mathbf{C} \in \mathbb{R}^d$ for each K.
- Step 3. Calculate the within-class cosine similarity (WCCS) and observe the 'elbow' of the WCCS curve to determine the number of clusters K .
- Step 4. Create the pseudo labels for the FFSVC20 dataset.
- Step 5. Use the pseudo-labels data together with the labeled data into speaker model to fine-tune the model.
- Step 6. Repeat Step 1 with the fine-tuned model from Step 5 as the pre-trained model.

Generating pseudo label by clustering. We adopt the K-means algorithm as the clustering algorithm to generate the pseudo labels. The learning objective of K-means is set to minimize the within-cluster sum-of-squares criterion:

$$\min_{\mathbf{C}} \frac{1}{N} \sum_{i=1}^{N} \min_{k} \|\mathbf{z}_i - \mathbf{C}_k\|^2 \tag{6}$$

where $\mathbf{z}_i \in \mathbb{R}^d$ is the d-dimensional speaker embedding of the i^{th} sample. The cluster with the closest controid to z_i in terms of the L2-norm distance is assigned as the pseudo-label for sample i.

Determine the number of clusters. Inspired by the works of Cai *et al.* [65], [66], we determine the number of clusters by the 'elbow' method. Given $\mathbf{z}_{k,a}$, the assigned a^{th} embedding of the k^{th} cluster. The total WCCS of N elements is:

$$WCCS = \frac{\sum_{k=1}^{K} \sum_{a=1}^{A} cos(\mathbf{z}_{k,a}, \mathbf{C}_k)}{N}, \tag{7}$$

Since the cosine similarity and euclidean similarity are connected linearly for normalized vectors, the WCCS linearly connects with learning objective of K-means. Fig. 4 shows the curve of WCCS results under different Ks. The WCCS monotonically increases as number of clusters K increases. WCCS tends to flatten with some K onwards and forming an 'elbow' of the curve. Such 'elbow'?indicates that the intra-cluster has little variation and increasingly over-fitting. From Fig. 4, the 'elbow' is distributed between 400 and 600.

V. Multi-channel Far-Field Speaker Verification

In this section, we will introduce the far-field multi-channel system. The 3D convolution (3D Conv) operation and 3D Squeeze-and-Excitation module will be applied to replace 2D convolution (2D Conv) operation and learn spatial information in this section. First, we introduce our previous work [27] on Section V-A, V-B and V-C. Then, to further explore the channel dependencies for the far-field multi-channel training framework, the squeeze-excitation (SE) attention module is extended to 3D ConvNet. The channel-wise 3D squeeze-and-excitation ConvNet and spatial-wise 3D SE ResNet are proposed in Section V-D.

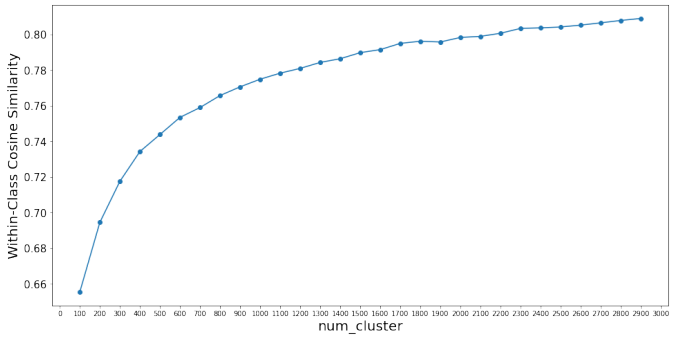


Fig. 4. Within-Cluster Cosine Similarity versus the number of clusters K employed.

Given the microphone array with M microphone channels, the spectro-temporal feature for recording channel m can be represented as $x_m \in R^{F\times T}$, where F is the feature dimension, and T is the number of time frame. The features of multi-channel microphone array utterance can be seen as either a multi-channel 2D feature or a 3D feature representation $X \in R^{M\times F\times T}$. The multi-channel feature representation is then fed into the speaker embedding network.

A. 2D Conv with multi-channel 2D features

Given the multi-channel 2D features, the convolution layer with 2-dimensional kernel takes X as M 2D feature planes and produces the output feature maps of $Y \in R^{C\times H\times W}$, where C denotes the number of convolution output channels, H and W indicate the high and wide of the feature maps. Formally, the c^{th} feature map after the convolution operation can be described as

$$\mathbf{Y}_c = \sum_{m=0}^{M} K(c,m) \otimes \mathbf{X}_m \tag{8}$$

where $K(c,m)$ is the 2D filter weights for input channel m and convolution output channels c, and the $\otimes$ indicates 2D convolution operation. In this study, the first convolution layer of the speaker verification model is designed to receive multiple channel features. With 2D convolution, how multi-channel training works is the same as processing three color channel pictures in computer vision. In this case, the model is denoted as ***2D-ResNet*** *(multi-channel)*.

B. 3D Conv with 3D features

The second scheme for multi-channel training is the use of 3D convolution layers. 3D Conv has been applied for far-field multi-channel speech recognition in [67]. The 3D convolution layer receives 4-dimensional input feature maps with size $C \times D \times H \times W$, where C is the number of feature maps, D, H, W are the depth, height and width of the feature map respectively. For the multi-channel input model, there is more spatial information focused on the recording channels (corresponds to depth dimension in feature maps). The output Y can be defined as

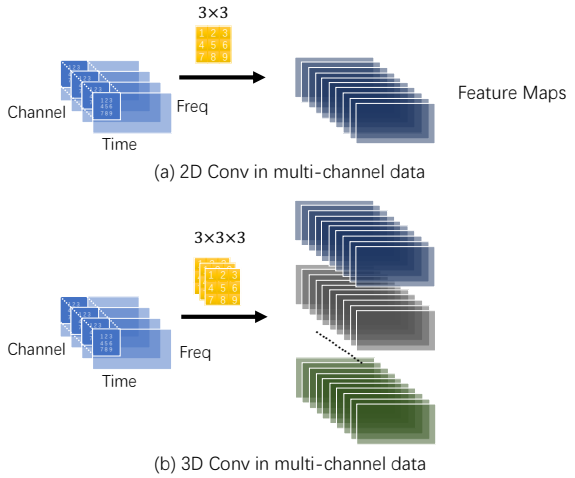


Fig. 5. 2D and 3D convolution operation for multi-channel data.

$$\mathbf{Y}_{\text{out},c} = \sum_{k=0}^{C_{\text{in}}} K(c,k) \otimes \mathbf{Y}_{\text{in},k} \tag{9}$$

where $K(c,k)$ is the 3D filter weights for input channel k and convolution output channels c, the $\otimes$ is the valid 3D cross-correlation operator, $Y_{\text{out},c}$ is the c^{th} feature map of the output feature and $Y_{in,k}$ is the k^{th} feature map of the input feature.

Fig. 5 shows the difference between the 2D and 3D convolution layer. Multi-Channel 2D Conv only learns the information between channels on the first layer. Since the 2D convolution aggregates the channel information together into the 2D feature maps after the first convolution layer, the feature map after passing through the first layer does not contain the information between channels. But the 3D convolution retains the channel axis and keeps the channel information in the whole ConvNet. On the other hand, since the kernel size of multi-channel 2D Conv must equal the number of the input feature map, multi-channel 2D Conv can not slide along the channel axis. Therefore, only by fixing the sound source and microphone array position, the first 2D Conv layer could effectively learn the channel information. But in real scenarios, the location of the sound source is dynamic. Conversely, 3D Conv could slide along the channel axis, thus 3D could learn the information between adjacent microphone channels. We think 3D Conv could capture the relationship of time, frequency and channel, and implement beamformer performance implicitly by learning 3D information.

In this case, all the 2D convolution operations in the original ResNet are replaced by the 3D convolution layers. We thus modify the global statistical pooling (GSP) layer to aggregate the mean and std statistics along with the time-, frequency- and depth-axis.

According to [68], the 3D Convolution Network (ConvNet) is more suitable for learning multi-channel spatio-temporal features than the 2D ConvNet. In terms of the far-field multi-channel data, convolution and pooling operations are performed in a spatio-spectral-temporal manner using 3D ConvNet, while 2D ConvNet operates only along the temporal

and frequency dimensions. Fig. 5 illustrates the difference, 2D Conv applied on an image will output an image, 3D Conv applied on multiple images (treating them as different channels) also results in an image. Hence, 2D Conv loses information of the input signal spatially right after every convolution operation. In contrast, 3D Conv preserves the spatial information of the input signals resulting in an output volume. The model is denoted as ***3D-ResNet*** in this paper.

C. Incorporate 3D Conv with 2D Conv

As stated before, the 3D convolution retains the channel axis in the whole ConvNet, while the 2D convolution drops the channel axis after the first convolution layer. However, using 3D convolution layers in ResNet may greatly increase the model size. This motivates us to incorporate 3D Conv with 2D Conv. To match the dimension between the 3D convolution feature maps (4D tensor) and 2D convolution feature maps (3D tensor), a 3D Conv layer with kernel size of $D_{in} \times 1 \times 1$ is adapted to convert the 4-dimensional feature maps into a 3 dimensional feature maps, where C_{in} is designed to match the channel size of the input feature maps. In this way, the channel axis of the 4-dimensional feature maps output has length 1, and it is then reshaped to 3-dimensional feature maps and fed into the 2D Conv layers. In this case, the model is denoted as ***3D2D-ResNet***. Fig. 7 show the structures of the *3D-ResNet* and *3D2D-ResNet* models.

D. 3D Squeeze-and-Excitation module

The application of the attention mechanism in speaker verification has achieved significant success, especially the usage of Squeeze-and-Excitation Networks (SENet) [28]. Inspired by SENet, we introduce two modules, the channel-wise 3D squeeze-and-excitation (C3D-SE) module and the spatial-wise 3D squeeze-and-excitation (S3D-SE) module. The attention mechanism acts on the channel dimension and depth dimension of feature maps, respectively.

1) Channel-wise 3D squeeze-and-excitation module: Inspired by the success of SENet in the single-channel ASV task [4], [65], we adopt the C3D-SE module for the far-field speaker verification. Fig. 6 shows the normal modified SE module that is learning the channel weight for each convolution channels of 3D Conv. Fig 6 (a) present the common SE module [28]. Since the SE operation is performed along the channel level on the 2D feature map, the attention weight $\mathbf{e} \in \mathbb{R}^{\mathbf{C}}$ of conventional SE is 1D vectors, and this SE module is named the channel-wise 2D SE (C2D-SE) in this paper. The C3D-SE is similar to C2D-SE, we also adopt the global average pooling at the channel level and generates a 1D weight vector. We calculate a vector $e \in \mathbb{R}^{C}$ containing the mean descriptor for each convolution channel of the intermediate feature maps in the following manner:

$$e^{c} = \frac{1}{D \times H \times W} \sum_{i=1}^{D} \sum_{j=1}^{H} \sum_{k=1}^{W} x_{i,j,k}^{c} \tag{10}$$

with $x_{i,j,k}^{c}$ the elements of $\mathbf{X}^{c} \in \mathbb{R}^{D \times H \times W}$, the component of input feature map $\mathbf{X} \in \mathbb{R}^{C \times D \times F \times T}$ corresponding with convolution channel position c. After the squeeze operation, the followed excitation operation is:

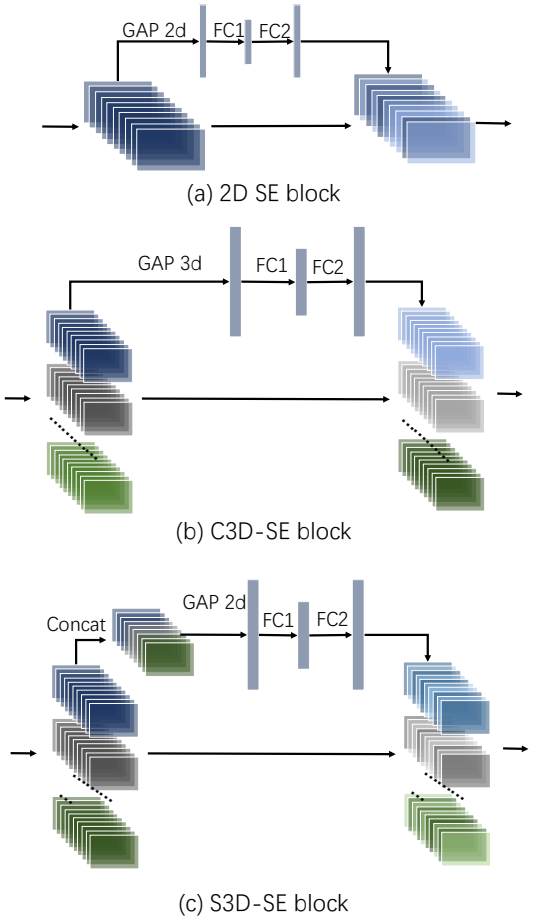


Fig. 6. 2D and 3D Squeeze-and-Excitation operations. Fig. 6 (a) describes the 2D-SE, Fig. 6 (b) describes the C3D-SE module, Fig.6 (c) describes the S3D-SE module.

$$\mathbf{s} = \sigma(\mathbf{W}_2 f(\mathbf{W}_1 \mathbf{e} + \mathbf{b}_1) + \mathbf{b_2}) \tag{11}$$

where $\mathbf{W}$ and $\mathbf{b}$ indicating the weights and bias of a linear layer, $f(.)$ the ReLU activation function and $\sigma(\cdot)$ the sigmoid function. Finally, $\mathbf{X}^{c}$ is scaled with the corresponding scalar in $\mathbf{s}$. The proposed C3D-SE blocks are inserted at the end of each residual module before the additive skip connection. The model is denoted as ***C3DSE-ResNet***.

2) Spatial 3D squeeze-and-excitation module: The existing attention modules are usually operated on the channel or frequency dimensions [28], [69]. Those methods generate 1-D weights following the channel dimension, which means the various recording channels in one feature map channel share the same weight. However, the quality of multiple recording channel is different, attention weights may need more refinement and we could assign different weights for different recording channels. Therefore, we introduce S3D-SE for 3D ConvNet, which performs SE with 2D weights. Since the C3D-SE with 2D weight (channel-wise and deep-wise attention weights) will additionally increase the parameters,

TABLE I
THE DATA USAGE OF SINGLE-CHANNEL AND MULTI-CHANNEL MODEL. K DENOTES THE SPEAKER NUMBER AFTER CLUSTERING OF SEMI-SUPERVISED LEARNING.

Model	Strategy	Pre-train		Fine-tuning	
		Dataset	Num. Spk & Num. Utt	Dataset	Num. Spk & Num. Utt
Single-Channel	Baseline	FFSVC20+HIMIA	447 & 208300	-	-
	Meta-Learning	FFSVC20+HIMIA	447 & 208300	-	-
	Mix	VoxCeleb 1&2 + OpenSLR + FFSVC20+HIMIA	11001 & 805285	-	-
	FT-Mix	VoxCeleb 1&2 + OpenSLR	10554 & 596985	VoxCeleb 1&2+OpenSLR+ FFSVC20+HIMIA	11001 & 805285
	FT-domain	VoxCeleb 1&2 + OpenSLR	10554 & 596985	FFSVC20+HIMIA	447 & 208300
	SSL	VoxCeleb 1&2 + OpenSLR	10554 & 596985	VoxCeleb 1&2+OpenSLR+ FFSVC20+HIMIA	10554+K & 805285
Multi-Channel	FT-Mix	VoxCeleb 1&2 + OpenSLR	10554 & 596985	VoxCeleb 1&2+OpenSLR+ FFSVC20+HIMIA	11001 & 805285
	FT-domain	VoxCeleb 1&2 + OpenSLR	10554 & 596985	FFSVC20+HIMIA	447 & 208300

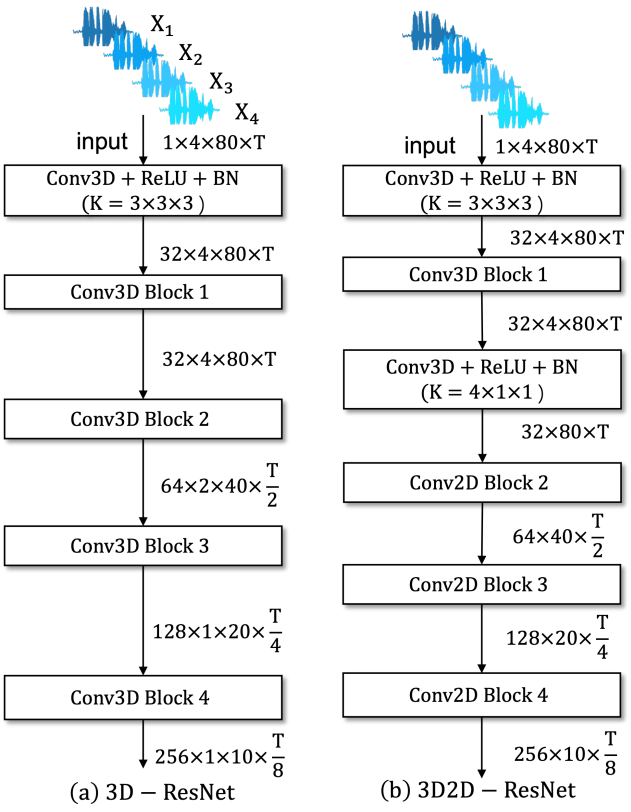


Fig. 7. Model structure of 3D-ResNet and 3D2D-ResNet.

we further simplify the attention generative mechanism: the 4D feature maps at various channel levels are flattened and concatenated to transform into 3D feature maps as shown in Fig. 6 (c). Therefore, the final S3D-SE function is given by:

$$e^{c,d} = \frac{1}{H \times W} \sum_{j=1}^{H} \sum_{k=1}^{W} x_{j,k}^{c,d} \tag{12}$$

To further improve the performance and reduce the model size, we propose the jointly S3D-SE attention and C2D-SE attention model (denoted ***S3C2SE-ResNet***) for far-field speaker verification. The backbone is structured to follow the 3D2D-ResNet framework. We adopt the S3D-SE module in the first block to learn spatial information, and the C2D-SE module is integrated into the remaining blocks.

VI. EXPERIMENTAL SETTING

A. Dataset and data usage

Table.I shows the details of the data usage for our experiments. The training phases are divided into two parts: pre-train and fine-tuning.

1) large-scale out-of-domain training dataset: **VoxCeleb 1&2**. The VoxCeleb 1&2 [70], [71] are widely used large-scale speaker verification datasets including 7323 speakers. Those audios include multiple languages (mainly English) under different kinds of realistic scenarios.

OpenSLR. The *OpenSLR* platform is a free and open speech resource website. We add some Chinese datasets listed on OpenSLR into the training set to reduce the language mismatch and obtain better performance for ASV. The aforementioned Chinese databases listed on OpenSLR are AISHELL-1 (SLR33) [72], Free ST Chinese Mandarin Corpus (SLR38) [73],Primewords Chinese Corpus Set (SLR47) [74], aidatatang_200zh (SLR62) [75] and MAGICDATA (SLR68) [76], with in total 3231 speakers.

The whole VoxCeleb dataset with 7323 speakers and selective datasets from OpenSLR are employed to train the out-of-domain model. To balance the data and avoid biasing the model towards any dataset, finally we randomly select 60 utterances from each speaker to refine the large-scale out-of-domain datasets.

2) in-domain dataset: Since the HIMIA[3] [30] database shares the same recording environment as FFSVC20, we consider both HIMIA and FFSVC20 as the in-domain datasets. The HI-MIA database includes two sub-datasets, which are

[3]https://openslr.org/85/

the AISHELL-wakeup1 with 254 speakers and the AISHELL-2019B-eval with 86 speakers. Since, there is some overlapping between the HIMIA dataset and the FFSVC20 training data, we also release the list[4] of overlapped speaker with its corresponding relationship. Finally, we have in-domain data from 447 speakers for training, and for each device and each speaker, 50 single or multi-channel audio files are randomly selected.

B. Model training and setting

1) Data Augmentation: We perform offline data augmentation using the MUSAN dataset [77] and the *pyroomacoustic* toolkit [78]. *pyroomacoustic* is adopted to generate the reverberated and noisy data as a simulation of the far-field condition. We randomly set the width and length of the acoustic room between 3 meters and 8 meters, and the height of the room is 3 meters. The distance between the source speaker and the recorded microphone array is randomly selected between 0.5 and 8 meters. The distance between the noisy source loudspeaker and the recorded microphone array is 4 meters. The microphone array we designed is a circular device with a 5cm radius and four channels (same with the setting of the FFSVC20 microphone array).

2) Single channel training setting: 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. In this experiment, a strong baseline system for speaker embedding extraction, namely SE-ResNet34, is adopted. For the SE-ResNet34 module, we adopt the same structure as the one in [5]. The network structure contains three main components: a front-end pattern extractor, an encoder layer, and a back-end classifier. The ResNet34 [79] model with Squeeze-and-Excitation Module (SE) [28], and different residual blocks $[32, 64, 128, 256]$, is employed as the front-end pattern extractor, the 256-dimensional fully connected layer following the global statistic pooling (GSP) based encoder layer is adopted as the speaker embedding layer. The ArcFace [62] ($s = 32, m = 0.2$) is used as the classifier.

Adam optimizer is used to update model parameters, and we adopt the MultiStepLR as the learning rate (LR) decays strategy that decays the learning rate of each parameter group by 0.1 once the number of epoch reaches one of the milestones. The milestone epochs are 10, 20, and 30. In the pre-train stage, LR decreases from the initialized 0.001 to 0.00001 until its performance no longer decreases on the development set. In the fine-tuning stage, the initialized LR is set to a fixed constant of 0.00001.

We divided our experiments into three scenarios based on the availability of in-domain and out-of-domain data.

- Scenario 1. Given zero out-of-domain data and few in-domain labeled data;
- Scenario 2. Given large-scale out-of-domain labeled data and few in-domain labeled data;
- Scenario 3. Given large-scale out-of-domain labeled data and few in-domain unlabeled data.

[4]http://2020.ffsvc.org/HIMIA_FFSVC2020_overlap.txt

Scenario 1. Given zero out-of-domain data and few in-domain labeled data, the speaker model directly trained with in-domain data (FFSVC and HIMIA) is viewed as the baseline system. For the proposed meta-learning method, 80 speakers are randomly sampled in each mini-batch. $S = 2$ examples and $Q = 1$ example are randomly selected for each speaker as the support set and the query set, respectively.

Scenario 2. Given large-scale out-of-domain labeled data and few in-domain labeled data, the mix-training (MIX) strategy that integrates both out-of-domain and in-domain data together to train the model is used as the baseline.

Scenario 3. Given large-scale out-of-domain labeled data and few in-domain unlabeled data, the pre-trained model is adopted to extract speaker embeddings, perform clustering, and generate pseudo-labels. Those in-domain data with pseudo-labels are employed to fine-tune the pre-trained model in a semi-supervised manner.

3) Multi-channel training setting: 80-dimensional log Mel-filterbank energies extracted with a frame length of 25ms and a hop size of 10ms are adopted as the acoustic features. We adopt the ResNet34 model as a basic backbone network for all multi-channel experiments. Considering the extensive computational consumption and memory requirement, the residual block channels are scaled down to [32,64,128,256], the kernel size of the 3D Conv is $(3 \times 3 \times 3)$ and the kernel size of the 2D Conv is (3×3). For the *C3SE-ResNet34SE* and *S3C2SE-ResNet34* model, the reduction ratio of SE-module is 4. Finally, we expand the channel of *S3C2SE-ResNet34* to [64,128,256,512] to explore the best single system performance.

Considering the channel number difference between single-channel close-talking enrollment data and multi-channel far-field test data, we adopt the following approach to feed our acoustic features into a multi-channel model:

- For single-channel data, the 2D acoustic feature map is replicated three times and viewed a 3D input feature map.
- For both simulated and real multi-channel data, the 3D feature map is constructed by the 2D acoustic features of all channels according to the channel order.

Other training settings are the same as the single-channel experiments, including the optimizer, the LR schedule, and the classifier.

We also conduct three scenario experiments for multi-channel training. Results of scenario 2 will be reported first. The best speaker embedding model in Scenario 2 will be selected to perform the remaining scenarios experiment.

C. Evaluation metrics

The speaker verification systems are measured by Equal Error Rate (EER) and minimum normalized Detection Cost Function (mDCF) with $C_{Miss} = C_{FA} = 1$ and $P_{target} = 0.01$.

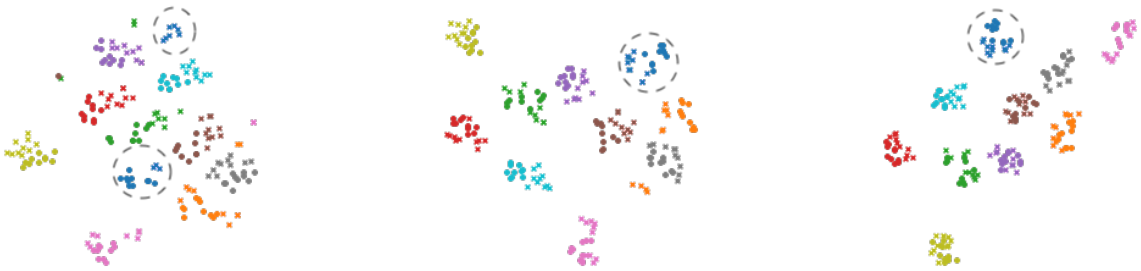
VII. Experimental Results

A. Single-channel training

This section presents the experimental results of three different training scenarios and strategies. Since the test data were

TABLE II
THE PERFORMANCE OF VARIOUS SINGLE-CHANNEL SPEAKER EMBEDDING SYSTEMS ON THE FFSVC20 EVALUATION SET. SINGLE- MEAN±STD INDICATES THAT THE MEAN AND STANDARD DEVIATION OF ALL RECORDING CHANNELS, AND MULTI- FUSION INDICATES THAT MULTI-CHANNEL EMBEDDING-LEVEL FUSION.

Strategy	**Task 1 Single- mean±std**		**Task 1 Multi- fusion**		**Task 2 Single- mean±std**		**Task 2 Multi- fusion**	
	EER[%]	**mDCF$_{0.01}$**	**EER[%]**	**mDCF$_{0.01}$**	**EER[%]**	**mDCF$_{0.01}$**	**EER[%]**	**mDCF$_{0.01}$**
Scenario 1. with SE-ResNet34 (C=32)								
Baseline	11.87±0.17	0.94±0.00	10.86	0.92	14.31±0.13	0.989±0.00	13.56	0.99
+ WPE	11.32±0.16	0.93±0.00	10.31	0.92	13.71±0.14	0.987±0.00	12.73	0.98
+ Delay-and-Sum	-	-	10.46	0.93	-	-	13.32	0.99
+ MVDR	-	-	10.24	0.93	-	-	12.57	0.98
+ GEV	-	-	10.23	0.93	-	-	12.49	0.98
Meta-Learning (S=2,Q=1)	9.36±0.05	0.88±0.00	**8.84**	**0.86**	11.76±0.06	0.97±0.00	**10.63**	**0.97**
Scenario 2. with SE-ResNet34 (C=32)								
MIX	6.78±0.06	0.62±0.01	6.18	0.58	7.44±0.06	0.66±0.01	6.84	0.62
Pre-train	11.49±0.05	0.84±0.00	10.68	0.82	15.34±0.06	0.97±0.00	14.33	0.94
+ FT-domain	6.71±0.10	0.66±0.01	6.07	0.62	7.41±0.06	0.78±0.01	6.96	0.75
+ FT-Mix	**6.35±0.09**	**0.60±0.01**	**5.77**	**0.56**	**6.27±0.11**	**0.67±0.01**	**5.76**	**0.63**



(a) Pre-train (b) Semi-supervised Learning (c) FT-Mix

Fig. 8. The visualization of t-SNE for different training strategies under scenario 3. The figure 8(b) describes the embedding description of semi-supervised learning under the best performance. Each color corresponds to a different speaker. The × and • are indicated that far-field data and close-talking data, respectively.

TABLE III
THE PERFORMANCE OF METRIC-LEARNING AND META-LEARNING ON THE VOXCELEB 1 ORIGINAL EVALUATION SET (CLEANED)

Training set	Loss	EER[%]	$mDCF_{0.01}$
Vox1 Dev	Arcface	2.56	0.219
Vox1 Dev	Prototype + Arcface	**1.89**	**0.140**
Vox2 Dev	Arcface	**0.79**	**0.084**
Vox2 Dev	Prototype + Arcface	1.80	0.110

collected by the microphone array with multiple recording channels, we report the performance on a) the mean and standard deviation of all recording channels and b) multi-channel embedding-level fusion in Table II.

1) Scenario 1. zero out-of-domain data and few in-domain labeled data: The result of the baseline system directly trained by the in-domain data without any training strategy is shown in Table II. Other than that, speech enhancement technologies based on signal processing are employed in training and test phase for multi-channel data. In this scenario, Delay-and-Sum (DS), MVDR [80] and GEV [81] beamformer are adopted to enhance the speech quality, and WPE [46] technology is used for de-reverberation[5]. Since the multi-channel signal is converted to single-channel data after beamforming, we only reported the beamformer results in the column of *Multi- fusion*. From Table II, we can observe that the performances of speech enhanced signals has little improvement than the original signal. Therefore, we will not report the speech enhancement results in the following experiments.

For the proposed meta-learning method, we first conduct experiments on VoxCeleb to verify the effectiveness. Accordingly, the performances of models trained by the VoxCeleb1 dev set (Vox1Dev) and those trained by the VoxCeleb2 dev (Vox2Dev) set are shown in Table III. In the limited training data (Vox1 Dev) scenario, the meta-learning with global classification (with Prototype and ArcFace loss) approach shows significant improvements over the performance of Metric-Learning baseline method (with ArcFace loss). Since the role of meta-learning is to find a good initialized model under the limited data scenarios, meta-learning does not improve the performance with large-scale training datasets (Vox2 Dev).

As shown in Table II, the proposed meta-learning training strategy achieves a 20% relative improvement compared to the baseline system for far-field speaker verification. Therefore, the proposed meta-learning method, which uses far-field and close-talking prototypes and makes cross-channel embedding closer together, enables domain adaptation.

[5]We implement DS, MVDR, GEV and WPE methods following the https://github.com/funcwj/setk toolkit.

TABLE IV
THE PERFORMANCE OF THE SEMI-SUPERVISED LEARNING APPROACH ON THE FFSVC20 EVALUATION SET UNDER SCENARIO 3.

Model	Cluster	Task 1		Task 2	
		EER[%]	mDCF	EER[%]	mDCF
Pre-train	-	10.675	0.817	14.325	0.940
Round 1	K=400	7.858	0.730	9.350	1.000
	K=500	7.867	0.702	11.223	1.000
	K=600	7.892	0.707	10.124	1.000
Round 2 (K=400)	K=400	7.218	0.691	8.010	0.812
	K=450	**7.192**	**0.683**	**7.968**	**0.801**
	K=500	7.201	0.688	7.997	0.809
Round 2 (K=500)	K=450	7.190	0.695	8.367	0.833
	K=500	7.213	0.701	8.456	0.823
	K=550	7.224	0.701	8.478	0.835
Round 2 (K=600)	K=550	7.214	0.712	8.356	0.835
	K=600	7.242	0.732	8.468	0.843
	K=650	7.301	0.738	8.432	0.838
Round 3	K=450	7.193	0.699	7.674	0.767
Round 4	K=450	7.448	0.690	7.817	0.757
Pre-train + FT-Mix	-	5.768	0.555	5.760	0.630

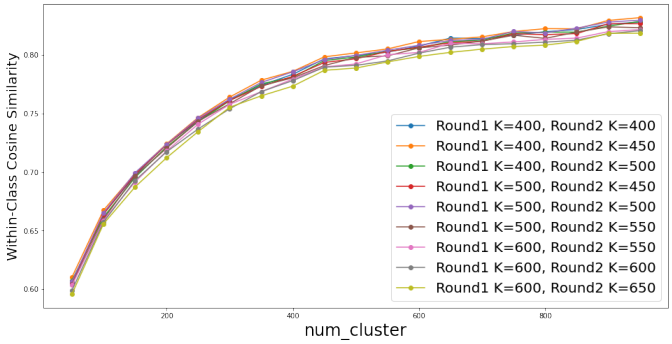


Fig. 9. Within-Cluster Cosine Similarity versus the number of cluster K employed in Round 2 under scenario 3.

2) Scenario 2. large-scale out-of-domain labeled data and few in-domain labeled data: From Table II, we can observe that adopting transfer learning strategy achieves better results compared to the MIX training strategy. The FT-Mix system has 15% relative improvement against the MIX baseline.

Moreover, compared with the lightweight model (ResNet18) we adopted previously [25], the SE-ResNet34 structure has stronger modeling abilities but is also easy to overfit. Therefore, the FT-Mix performance is better than FT-domain in this experiment. In addition, by comparing the performances of the mean and standard deviation of single-channel and embedding-level fusion of multi-channel, the embedding fusion with average weight has 10% relative improvement over the scoring with single-channel only. The result indicates that multi-channel embedding have certain complementary characteristics.

3) Scenario 3. large-scale out-of-domain labeled data and few in-domain unlabeled data: Finally, we discuss the performance of semi-supervised learning. Table IV reports the result of our semi-supervised learning method. In this experiment, the performance of the fully supervised FT-Mix strategy is also provided for reference.

TABLE VI
THE PERFORMANCE OF DIFFERENT MULTI-CHANNEL DATA AVAILABILITY SCENARIOS ON THE FFSVC20 EVALUATION SET BASED ON THE S3C2SE-RESNET34(C=32) MODEL.

Strategy	Task 1		Task 2	
	EER[%]	$\text{mDCF}_{0.01}$	EER[%]	$\text{mDCF}_{0.01}$
Scenario 1.				
Baseline	9.734	0.896	11.323	0.969
Meta-Learning (S=2,Q=1)	**8.012**	**0.813**	**9.125**	**0.876**
Scenario 3.				
Pre-train	9.358	0.770	15.200	0.998
Round1 (K=500)	7.313	0.697	7.418	0.701
Round2 (K=500)	**6.572**	**0.588**	6.636	**0.653**
Round3 (K=500)	6.632	0.596	**6.598**	0.657
Pre-train + FT+mix	5.298	0.494	5.017	0.602

After the first round of clustering by the *K-means* algorithm, the curve of within-cluster cosine similarity is shown in Fig. 4, where the 'elbow' is around 400 to 600. Therefore, we chose the 400, 500, 600 as the number of centroids in the first semi-supervised learning training. In fact, there are 447 speakers in the in-domain training set. The verification results of the first round are close in Task 1. All EER of K=400, K=500, and K=600 achieve about 30% relative improvement compared with the pre-train model. The WCCS curve of the second-round clustering is presented in Fig. 9, the curves of 'elbow' have the same trend. In this case, we select 3 clusters with 50 intervals based on the centroid of the first round. There is about 8% relative improvement compared to the first round, and the system achieves the best performance in task 1 and task 2 when K=450. In addition, we also observe that the verification results with different centroids are relatively stable. Therefore, we keep the number of centroids unchanged for the third round and the forth round while using the best model of the second round for the next clustering. Finally, after multiple iterations of semi-supervised learning, the final performance of semi-supervised learning has 30% relative improvement than the pre-trained model.

By observing Fig. 8, we can find that the main challenge of the far-field scenario is the domain mismatch, as the close-talking data and far-field data are naturally divided into two categories (Fig. 8(a)). However, as shown from the embedding distributions obtained by the semi-supervised learning model (Fig. 8(b)) and the fully supervised learning model (Fig. 8(c)), the embeddings of far-field and close-talking are aggregated more compact than the pre-trained one. On the other hand, since the number of clusters cannot accurately match the real speaker quantity, each cluster inevitably contains noisy data that affects the performance.

B. Multi-channel training

1) Scenario 2. large-scale out-of-domain labeled data and few in-domain labeled data: Table V reports the performance of different far-field models and where the training condition indicates that training model with single-channel/multi-

TABLE V
THE PERFORMANCE OF VARIOUS MULTI-CHANNEL MODEL UNDER FFSVC20 EVAL SET. MULTI-FUSION INDICATES THAT MULTI-CHANNEL EMBEDDING-LEVEL FUSION.

Model ID	Training Condition	Model	Strategy	Task 1 Eval		Task 2 Eval	
				EER[%]	**mDCF$_{0.01}$**	**EER[%]**	**mDCF$_{0.01}$**
1	**Single-Channel**	**2D ResNet34 (C=32)** (Multi- fusion)	Pre-train	12.238	0.956	13.531	0.971
			+FT-mix	**6.212**	**0.627**	**6.534**	**0.638**
			+FT-domain	7.172	0.690	7.441	0.815
2	**Single-Channel**	**2D ResNet34 (C=32)** (MVDR based Beamforming)	Pre-train	11.778	0.901	12.113	0.963
			+FT-mix	**6.113**	**0.619**	**6.411**	**0.637**
			+FT-domain	7.076	0.676	7.304	0.808
3	**Single-Channel**	**2D SE-ResNet34 (C=32)** (Multi- fusion)	Pre-train	10.675	0.817	14.325	0.940
			+FT-mix	**5.768**	**0.555**	**5.760**	**0.630**
			+FT-domain	6.073	0.617	6.957	0.750
4	**Multi-Channel**	**2D ResNet34 (C=32)**	Pre-train	9.913	0.832	12.995	0.951
			+FT-mix	**6.274**	**0.605**	**7.001**	**0.737**
			+FT-domain	7.285	0.735	8.003	0.812
5	**Multi-Channel**	**3D-ResNet34 (C=32)**	Pre-train	9.315	0.763	12.115	0.877
			+FT-mix	**5.485**	**0.532**	**6.042**	**0.649**
			+FT-domain	6.675	0.652	7.983	0.800
6	**Multi-Channel**	**C3DSE-ResNet34 (C=32)**	Pre-train	9.324	0.792	11.792	0.943
			+FT-mix	**5.433**	**0.503**	**5.085**	**0.604**
			+FT-domain	8.159	0.707	7.785	0.805
7	**Multi-Channel**	**3D2D-ResNet34 (C=32)**	Pre-train	9.983	0.808	12.192	0.914
			+FT-mix	**5.492**	**0.504**	**5.542**	**0.599**
			+FT-domain	6.883	0.624	8.207	0.805
8	**Multi-Channel**	**S3C2SE-ResNet34SE (C=32)**	Pre-train	9.358	0.770	15.200	0.998
			+FT-mix	**5.298**	**0.494**	**5.017**	**0.602**
			+FT-domain	7.275	0.666	7.476	0.759
9	**Multi-Channel**	**S3C2SE-ResNet34SE (C=64)**	Pre-train	9.127	0.736	12.010	0.887
			+FT-mix	**4.928**	**0.447**	**4.481**	**0.532**
			+FT-domain	6.957	0.634	7.064	0.721
Zhang *et al.* [82] (single model best)			5.78	0.57	-	-	
Gusev *et al.* [18] (single model best)			-	-	5.61	0.564	

TABLE VII
MODEL SIZE AND RUNNING TIME UNDER THE MULTI-CHANNEL TEST SCENARIO. $\times 4$ INDICATES THAT SINGLE-CHANNEL INPUT MODEL NEED COST 4 TIMES TO HANDLE THE 4-CHANNEL TEST DATA.

Model	Parameters (M)	Inference Time (ms)
2D ResNet34 (C=32)	5.45	18.93 $\times 4$
2D SE-ResNet34 (C=32)	5.53	19.00 $\times 4$
3D-ResNet34 (C=32)	16.00	51.54
CD3SE-ResNet34 (C=32)	16.08	56.19
3D2D-ResNet34 (C=32)	5.57	41.21
S3C2SE-ResNet34 (C=32)	5.66	45.59
S3C2SE-ResNet34 (C=64)	22.36	118.29

channel input data. The 2D ResNet model with multi-channel (SectionV-A) and single-channel input are the baseline systems here. The 2D ResNet model with MVDR indicates that the multi-channel input data apply MVDR and subsequently 2D ResNet model.

First, all experiment models conduct with FT-domain and FT-mix training strategies. The results present that FT-mix outperforms than FT-domain. With the stronger modeling ability of the model, the easier it is to overfit under the FT-domain strategy. Therefore, we only discuss the results of FT-mix in the following model performance discussion.

Second, we compare all single-channel input models. 2D ResNet with MVDR (Model ID 2) is slightly improved than 2D ResNet with multi-channel embedding fusion (Model ID 1). But the compute cost MVDR is much greater than the embedding fusion. The 2D SE-ResNet model with multi-channel embedding fusion (Model ID 3) outperforms other single-channel input models, which means that the benefits of the attention mechanism are larger than other methods.

Then, we take a comparison of the performances of single-channel 2D ResNet with MVDR (Model ID 2), multi-channel 2D ResNet34 (Model ID 4) and multi-channel 3D ResNet34 (Model ID 5). Although 2D ResNet with MVDR and multi-channel 3D ResNet34 explicitly or implicitly use spatial information, the performance of 3D ResNet34 is much better than 2D ResNet with MVDR. The input of 2D ResNet with MVDR is spatial filtering preprocessed, and the raw signal inevitably suffers. While the 3D ResNet input is the raw signal, the original information is preserved without damage. In addition, the convolution operation of the multi-channel 2D ResNet34 between channels is independent of multiple channels spatial location, the advantage of multi-channel is not reflected.

From Table V, we observe that all 3D ConvNet models outperform the 2D ConvNet. Although the 2D ResNet34 with multi-channel input has the same input size as 3D ConvNet,

the 2D convolution kernel only slides along the time and frequency axis. Thus the spatial information between channels can not be learned by 2D Conv. Moreover, although all multi-channel pre-train data are simulated, the 3D ConvNet model is also better than the 2D ConvNet model in the pre-train phase. Therefore, 3D ConvNet is better than 2D one in multi-channel training.

In addition, as shown in Table V, the 3D2D-ResNet model (Model ID 7) outperforms the 3D ResNet (Model ID 5) model in most case. However, the parameter size of the 3D2D-ResNet in Table V is smaller than the size of the 3D-ResNet model. This indicates that the first 3D convolution residual block of the 3D2D-ResNet provides enough spatial information as the fully 3D model 3D-ResNet. Therefore, the 3D2D-ResNet is more efficient, which achieves better performance while using less parameter size.

Regarding the 3D attention module, the ResNet with C3DSE module (Model ID 6) has about 6% relative performance improvement over the 3D-ResNet model (Model ID 5) in the term of mDCF. Compared with the 3D channel-wise SE module, the proposed S3C2SE-ResNet34 module (Model ID 8), which focus on more spatial information (e.g., speaker location, locations of noise sources), achieves the best performance with a similar parameters size. In order to further explore better performance, we expand the channel to [64,128,256,512], and the model has 10% relative improvement but with 3 times additional parameters.

Finally, we discuss the model performances based on the parameter size and the inference time. Since the memory subsystems, internal subsystems, compute cores, and caches will influence the inference time, we infer the network over a certain amount of examples and then average the results. We randomly choose 100 examples to test the inference time for each model under the `Intel(R) Xeon(R) Gold 5215 CPU @ 2.50GHz`. Results are shown in Table VII. The inference time of all models are close. If we consider the factors of model performance, model size, and inference time together, the S3C2SE-ResNet model performs the best in our far-field scenarios.

2) Scenario 1 and 3: : Table VI reports the Scenario 1 and 3 results. The proposed Meta-Learning method achieves about 15% related improvements to the baseline system in Scenario 1. For Scenario 3, we adopt the $K = 500$ cluster to perform semi-supervised learning. The results indicate that the model performances tend to be stable in Round 2. The final model obtains the 30% related improvement to the pre-trained model.

VIII. Conclusion

This paper proposes multiple training strategies and models for far-field speaker verification regarding the different in-domain data availability scenarios. First, given large-scale out-of-domain labeled data and few in-domain labeled data, transfer learning is adopted to fine-tune the in-domain data. In this case, the FT-Mix strategy achieves the best performance in far-field speaker verification. Second, given zero out-of-domain data and few in-domain labeled data, we use the proposed meta-learning training strategy to perform domain adaptation. Moreover, when given large-scale out-of-domain labeled data and few in-domain unlabeled data, the semi-supervised approach is adopted to generate pseudo labels for unlabeled data before fine-tuning. The final performances of the models trained by pseudo labels are close to the supervised learning method. For the multi-channel training models, we conduct experiments comparing the 3D ConvNet with the 2D ConvNet. Results show that the 3D ConvNet outperforms the 2D ConvNet. In addition, compared with the fully 3D ConvNet, the model incorporating 3D Conv with 2D Conv achieves 60% relative parameter reduction while moderately improving the performance. To further explore the channel dependencies, we extend the SE attention module and propose the channel-wise 3D SE module and spatial-wise 3D SE module. Finally, with the proposed spatial-wise 3D SE attention module, the model incorporating 3D Conv with 2D Conv obtains the best performance with a few additional parameters. In future works, we will further explore semi-supervised learning in cross-domain scenarios and enhance the robustness of ASV systems in the far-field multi-channel scenarios.

IX. Acknowledgment

This research is funded by the Kunshan Government Research Fund (R97030019S).

References

[1] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN Embeddings for Speaker Recognition," in *Proc. ICASSP*, 2018, pp. 5329–5333.

[2] 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.

[3] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks," in *Proc. Interspeech*, 2018, pp. 3743–3747.

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

[5] W. Cai, J. Chen, and M. Li, "Exploring the Encoding Layer and Loss Function in End-to-End Speaker and Language Recognition System," in *Proc. Odyssey*, 2018, pp. 74–Sd81.

[6] T. Zhou, Y. Zhao, and J. Wu, "ResNeXt and Res2Net Structures for Speaker Verification," in *Proc. IEEE SLT*, 2021, pp. 301–307.

[7] M. K. Nandwana, J. van Hout, C. Richey, M. McLaren, M. A. Barrios, and A. Lawson, "The VOiCES from a Distance Challenge 2019," in *Proc. Interspeech*, 2019, pp. 2438–2442.

[8] 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.

[9] H. Zeinali, K. A. Lee, J. Alam, and L. Burget, "SdSV Challenge 2020: Large-Scale Evaluation of Short-Duration Speaker Verification," in *Proc. Interspeech*, 2020, pp. 731–735.

[10] H. Zeinali, K. A. Lee, J. Alam and L. Burget, "Short-duration Speaker Verification (SdSV) Challenge 2021: the Challenge Evaluation Plan," in *arXiv:1912.06311*.

[11] The VoxSRC21 website. [Online]. Available: https://www.robots.ox.ac.uk/~vgg/data/voxceleb/competition2021.html

[12] M. K. Nandwana, J. V. Hout, M. McLaren, C. Richey, A. Lawson, and M. A. Barrios, "The VOiCES from a Distance Challenge 2019 Evaluation Plan," *arXiv:1902.10828*, 2019.

[13] X. Qin, M. Li, H. Bu, R. K. Das, W. Rao, S. Narayanan, and H. Li, "The FFSVC 2020 Evaluation Plan," *arXiv:2002.00387*, 2020.

[14] M. Ladislav, P. Oldich, M. Pavel, N. Ondej, and C. Jan, "Dereverberation and Beamforming in Robust Far-Field Speaker Recognition," in *Proc. Interspeech*, 2018, pp. 1334–1338.

[15] L. Zhang, J. Wu, and L. Xie, "NPU Speaker Verification System for INTERSPEECH 2020 Far-Field Speaker Verification Challenge," in *Proc. Interspeech*, 2020, pp. 3471–3475.

[16] S. Dowerah, R. Serizel, D. Jouvet, M. Mohammadamini, and D. Matrouf, "Compensating noise and reverberation in far-field Multichannel Speaker Verification," 2022, working paper or preprint.

[17] L. Yi and M.-W. Mak, "Adversarial Separation and Adaptation Network for Far-Field Speaker Verification," in *Proc. Interspeech*, 2020, pp. 4298–4302.

[18] A. Gusev, V. Volokhov, A. Vinogradova, T. Andzhukaev, A. Shulipa, S. Novoselov, T. Pekhovsky, and A. Kozlov, "STC-Innovation Speaker Recognition Systems for Far-Field Speaker Verification Challenge 2020," in *Proc. Interspeech*, 2020, pp. 3466–3470.

[19] L. Li, D. Wang, J. Kang, R. Wang, J. Wu, Z. Gao, and X. Chen, "A principle solution for enroll-test mismatch in speaker recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2022.

[20] W. Xu, X. Wang, H. Wan, X. Guo, J. Zhao, F. Deng, and W. Kang, "Jointing multi-task learning and gradient reversal layer for far-field speaker verification," in *Chinese Conference on Biometric Recognition*, 2021, pp. 449–457.

[21] Y. Shi, Q. Huang, and T. Hain, "Robust Speaker Recognition Using Speech Enhancement And Attention Model," *arXiv:2001.05031*, 2020.

[22] L. Moner, O. Plchot, L. Burget, and J. H. ernock, "Multi-channel speaker verification with conv-tasnet based beamformer," in *Proc. ICASSP*, 2022, pp. 7982–7986.

[23] L. Moner, O. Plchot, J. Rohdin, and J. ernock, "Utilizing VOiCES Dataset for Multichannel Speaker Verification with Beamforming," in *Proc. Odyssey*, 2020, pp. 187–193.

[24] L. Mosner, P. Matejka, O. Novotny, and J. H. Cernocky, "Dereverberation and Beamforming in Far-Field Speaker Recognition," in *Proc. ICASSP*, 2018, pp. 5254–5258.

[25] 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. Interpseech*, 2019, pp. 4045–4049.

[26] S. Chen, A. S. Subramanian, H. Xu, and S. Watanabe, "Building state-of-the-art distant speech recognition using the CHiME-4 challenge with a setup of speech enhancement baseline," *arXiv:1803.10109*, 2018.

[27] D. Cai, X. Qin, and M. Li, "Multi-Channel Training for End-to-End Speaker Recognition under Reverberant and Noisy Environment," in *Proc. Interspeech*, 2019, pp. 4365–4369.

[28] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in *Proc. CVPR*, 2018, pp. 7132–7141.

[29] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating smallroom acoustics," *JASA*, vol. 65, no. 4, pp. 943–950, 1979.

[30] X. Qin, H. Bu, and M. Li, "HI-MIA: A Far-Field Text-Dependent Speaker Verification Database and the Baselines," in *Proc. ICASSP*, 2020, pp. 7609–7613.

[31] T. H. Falk and W.-Y. Chan, "Modulation spectral features for robust far-field speaker identification," *TASLP*, vol. 18, no. 1, pp. 90–100, 2010.

[32] S. O. Sadjadi and J. H. Hansen, "Hilbert envelope based features for robust speaker identification under reverberant mismatched conditions," in *Proc. ICASSP*, 2011, pp. 5448–5451.

[33] S. Ganapathy, J. Pelecanos, and M. K. Omar, "Feature normalization for speaker verification in room reverberation," in *Proc. ICASSP*, 2011, pp. 4836–4839.

[34] Q. Jin, R. Li, Q. Yang, K. Laskowski, and T. Schultz, "Speaker identification with distant microphone speech," in *Proc. ICASSP*, 2010, pp. 4518–4521.

[35] S. O. Sadjadi and J. H. L. Hansen, "Blind spectral weighting for robust speaker identification under reverberation mismatch," *TASLP*, vol. 22, no. 5, pp. 937–945, 2014.

[36] C. Kim and R. M. Stern, "Power-Normalized Cepstral Coefficients (PNCC) for Robust Speech Recognition," *TASLP*, vol. 24, no. 7, pp. 1315–1329, 2016.

[37] D. Cai, X. Qin, W. Cai, and M. Li, "The DKU System for the Speaker Recognition Task of the 2019 VOiCES from a Distance Challenge," in *Proc. Interspeech 2019*, 2019, pp. 2493–2497.

[38] T. Yamada, L. Wang, and A. Kai, "Improvement of distant-talking speaker identification using bottleneck features of DNN," in *Proc. Interspeech*, 2013, pp. 3661–3664.

[39] X. Liu, M. Sahidullah, and T. Kinnunen, "Parameterized Channel Normalization for Far-field Deep Speaker Verification," in *Proc. ASRU*, 2021.

[40] X. Zhao, Y. Wang, and D. Wang, "Robust speaker identification in noisy and reverberant conditions," *TASLP*, vol. 22, no. 4, pp. 836–845, 2014.

[41] M. Kolbk, Z.-H. Tan, and J. Jensen, "Speech enhancement using long short-term memory based recurrent neural networks for noise robust speaker verification," in *Proc. IEEE SLT*, 2016, pp. 305–311.

[42] Z. Oo, Y. Kawakami, L. Wang, S. Nakagawa, X. Xiao, and M. Iwahashi, "DNN-Based Amplitude and Phase Feature Enhancement for Noise Robust Speaker Identification," in *Proc. Interspeech*, 2016, pp. 2204–2208.

[43] S. E. Eskimez, P. Soufleris, Z. Duan, and W. Heinzelman, "Front-end speech enhancement for commercial speaker verification systems," *Speech Communication*, vol. 99, pp. 101–113, 2018.

[44] H. Taherian, Z.-Q. Wang, and D. Wang, "Deep Learning Based Multi-Channel Speaker Recognition in Noisy and Reverberant Environments," in *Proc. Interspeech 2019*, 2019, pp. 4070–4074.

[45] J.-Y. Yang and J.-H. Chang, "Joint Optimization of Neural Acoustic Beamforming and Dereverberation with x-Vectors for Robust Speaker Verification," in *Proc. Interspeech 2019*, 2019, pp. 4075–4079.

[46] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech Dereverberation Based on Variance-Normalized Delayed Linear Prediction," *TASLP*, vol. 18, no. 7, pp. 1717–1731, 2010.

[47] K. Kinoshita, M. Delcroix, H. Kwon, T. Mori, and T. Nakatani, "Neural Network-Based Spectrum Estimation for Online WPE Dereverberation," in *Proc. Interspeech*, 2017, pp. 2017–733.

[48] S. Novoselov, A. Gusev, A. Ivanov, T. Pekhovsky, A. Shulipa, G. Lavrentyeva, V. Volokhov, and A. Kozlov, "STC Speaker Recognition Systems for the VOiCES from a Distance Challenge," in *Proc. Interspeech*, 2019, pp. 2443–2447.

[49] A. Jati, R. Peri, M. Pal, T. J. Park, N. Kumar, R. Travadi, P. Georgiou, and S. Narayanan, "Multi-Task Discriminative Training of Hybrid DNN-TVM Model for Speaker Verification with Noisy and Far-Field Speech," in *Proc. Interspeech*, 2019, pp. 2463–2467.

[50] L.-j. Zhao and M.-W. Mak, "Channel interdependence enhanced speaker embeddings for far-field speaker verification," in *Proc. ISCSLP*, 2021, pp. 1–5.

[51] Y. Tu, M. Mak, and J. Chien, "Variational Domain Adversarial Learning With Mutual Information Maximization for Speaker Verification," *TASLP*, vol. 28, pp. 2013–2024, 2020.

[52] 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.

[53] D. Cai and M. Li, "Embedding Aggregation for Far-Field Speaker Verification with Distributed Microphone Arrays," in *Proc. IEEE SLT*, 2021, pp. 308–315.

[54] C. Liang, J. Chen, S. Guan, and X. Zhang, "Attention-based Multi-channel Speaker Verification with ad-hoc Microphone Arrays," *arXiv:2107.00178*, 2021.

[55] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," in *Artificial Neural Networks and Machine Learning – ICANN 2018*. Springer International Publishing, 2018, pp. 270–279.

[56] 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, pp. 2317–2321.

[57] S. M. Kye, Y. Jung, H. B. Lee, S. J. Hwang, and H. Kim, "Meta-Learning for Short Utterance Speaker Recognition with Imbalance Length Pairs," in *Proc. Interspeech*, 2020, pp. 2982–2986.

[58] 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, pp. 2977–2981.

[59] J. Wang, K.-C. Wang, M. T. Law, F. Rudzicz, and M. Brudno, "Centroid-based Deep Metric Learning for Speaker Recognition," in *Proc. ICASSP*, 2019, pp. 3652–3656.

[60] J. Snell, K. Swersky, and R. S. Zemel, "Prototypical Networks for Few-shot Learning," *Proc. NIPS*, p. 40804090, 2017.

[61] D. Cai, W. Cai, and M. Li, "Within-Sample Variability-Invariant Loss for Robust Speaker Recognition Under Noisy Environments," in *Proc. ICASSP*, 2020, pp. 6469–6473.

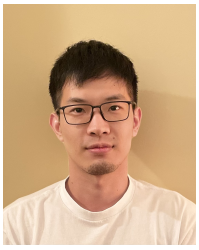
[62] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," in *Proc. CVPR*, 2019, pp. 4685–4694.

[63] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," *arXiv:1610.02242*, 2016.

[64] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in Neural Information Processing Systems*, 2017.

[65] W. Wang, D. Cai, X. Qin, and M. Li, "The DKU-DukeECE Systems for VoxCeleb Speaker Recognition Challenge 2020," *arXiv.2010.12731*.
[66] D. Cai and M. Li, "The DKU-DukeECE System for the Self-Supervision Speaker Verification Task of the 2021 VoxCeleb Speaker Recognition Challenge," *arXiv:2010.14751*.
[67] S. Ganapathy and V. Peddinti, "3-D CNN Models for Far-Field Multi-Channel Speech Recognition," in *Proc. ICASSP*, 2018, pp. 5499–5503.
[68] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks," in *Proc. ICCV*, 2015, pp. 4489–4497.
[69] J. Thienpondt, B. Desplanques, and K. Demuynck, "Integrating Frequency Translational Invariance in TDNNs and Frequency Positional Information in 2D ResNets to Enhance Speaker Verification," in *Proc. Interspeech*, 2021, pp. 2302–2306.
[70] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A Large-Scale Speaker Identification Dataset," in *Proc. Interspeech*, 2017, pp. 2616–2620.
[71] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep Speaker Recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.
[72] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "AISHELL-1: An open-source Mandarin Speech Corpus and A Speech Recognition Baseline," in *Proc. O-COCOSDA*, 2017, pp. 1–5.
[73] "Free st chinese mandarin corpus." [Online]. Available: www.surfing.ai
[74] Primewords Information Technology Co., Ltd., "Primewords chinese corpus set 1," 2018. [Online]. Available: https://www.primewords.cn
[75] Beijing DataTang Technology Co., Ltd, "A free Chinese Mandarin speech corpus." [Online]. Available: www.datatang.com
[76] Magic Data Technology Co., Ltd., "Magicdata mandarin chinese read speech corpus." [Online]. Available: http://www.imagicdatatech.com/index.php/home/dataopensource/data_info/id/101
[77] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," *arXiv:1510.08484*.
[78] R. Scheibler, E. Bezzam, and I. Dokmani, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. ICASSP*, 2018, pp. 351–355.
[79] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc.CVPR*, 2016, pp. 770–778.
[80] M. Souden, J. Benesty, and S. Affes, "On Optimal Frequency-Domain Multichannel Linear Filtering for Noise Reduction," *TASLP*, vol. 18, no. 2, pp. 260–276, 2010.
[81] E. Warsitz and R. Haeb-Umbach, "Blind Acoustic Beamforming Based on Generalized Eigenvalue Decomposition," *TASLP*, vol. 15, no. 5, pp. 1529–1539, 2007.
[82] P. Zhang, P. Hu, and X. Zhang, "Deep Embedding Learning for Text-Dependent Speaker Verification," in *Proc. Interspeech 2020*, 2020, pp. 3461–3465.

Xiaoyi Qin (Student Member, IEEE) received the M.S, degree in Electronics and Communication Engineering from Sun Yet-Sen University, Guangzhou, China. He is currently working toward the Ph.D. degree in Computer Science and Technology with Wuhan University, Wuhan, China. His research interests include speaker verification, anti-spoofing detection and speech signal processing.



Danwei Cai (Student Member, IEEE) is pursuing his Ph.D. degree in Electrical and Computer Engineering at Duke University. He received his bachelor's degree in Software Engineering and master's degree in Electronics and Communication Engineering from Sun Yet-Sen University in China. His primary research interests are in the area of speech processing, including speech recognition, speaker recognition, speaker diarization, and computational linguistics.

Ming Li received his Ph.D. in Electrical Engineering from University of Southern California in 2013. He is currently an Associate Professor of Electrical and Computer Engineering at Duke Kunshan University. He is also an Adjunct Professor at School of Computer Science, Wuhan University. His research interests are in the areas of audio, speech and language processing as well as multimodal behavior signal analysis and interpretation. He has published more than 140 papers and served as the member of IEEE speech and language technical committee, CCF speech dialogue and auditory processing technical committee, CAAI affective intelligence technical committee, APSIPA speech and language processing technical committee. He is the area chair of speaker and language recognition at Interspeech 2016, 2018 and 2020, as well as the technical program co-chair of Odyssey 2022 and ASRU 2023. Works co-authored with his colleagues have won first prize awards at Interspeech Computational Paralinguistic Challenge 2011, 2012 and 2019, ASRU 2019 MGB-5 ADI Challenge, Interspeech 2020 and 2021 Fearless Steps Challenge, VoxSRC 2021 and 2022 Challenge, ICASSP 2022 M2MeT Challenge. He received the IBM faculty award in 2016, the ISCA Computer Speech and Language 5-years best journal paper award in 2018 and the youth achievement award of outstanding scientific research achievements of Chinese higher education in 2020. He is a senior member of IEEE.