Meta-Generalization for Domain-Invariant Speaker Verification

Hanyi Zhang, *Student Member, IEEE*, Longbiao Wang, Kong Aik Lee, *Senior Member, IEEE*, Meng Liu, Jianwu Dang, *Member, IEEE*, and Helen Meng, *Fellow, IEEE*

***Abstract*—Automatic speaker verification (ASV) exhibits unsatisfactory performance under domain mismatch conditions owing to intrinsic and extrinsic factors, such as variations in speaking styles and recording devices encountered in real-world applications. To ensure robust performance under unseen conditions, domain generalization has been explored. However, an inherent contradiction exists between model discrimination and domain generalization, in which the discrimination ability may be reduced while learning to generalize. In this paper, to extract discriminative yet domain-invariant representations, we propose the meta-generalized speaker verification (MGSV) via meta-learning. Specifically, we propose a metric-based distribution optimization and a gradient-based meta-optimization to simultaneously supervise the spatial relationship between embeddings and improve the generalization ability of the model on unseen domains. In addition, we design multiple-single (MS) and simulated speaker verification (SSV) sampling strategies based on single-domain (SD) and single-single (SS) strategies to simulate the train/test domain mismatch more relevantly, thereby mining transferable speaker-related knowledge. SSV is chosen as the most effective method, as it substantially improves the domain generalization by ensuring that the model has learned to discriminate efficiently. Additionally, to intuitively reflect the model performance on the unseen domains, the proposed method is validated on cross-genre, cross-device, and cross-dataset tasks. The experimental results demonstrate that our proposed method achieves remarkable performance in handling domain mismatch issues in speaker verification.**

***Index Terms*—Speaker verification, domain mismatch, meta-learning, meta-generalized speaker verification.**

I. Introduction

AUTOMATIC speaker verification (ASV) aims to validate the identity of a speaker using their voice. It has been widely applied in several fields, such as access control for biometric authentication. In traditional implementations [1]–[3], speaker information is extracted from utterances by using probabilistic models. With the development of deep neural networks (DNNs) in the past few years, deep-embedding-based methods [4]–[6] have improved the performance of ASV systems to a higher level.

However, the ASV systems deployed in real-world applications are adversely affected by domain mismatch caused by intrinsic and extrinsic variations, such as the recording device, speaking style, and background noise. Unlike channel mismatch (or session mismatch), which is due to the inconsistency of the statistical distribution between the enrollment and test utterances, domain mismatch is the result of inconsistency between the training and evaluation sets, that is, the evaluation set comprises unseen domains [7]. Obviously, in many applications, the numerous unknown domain variations make domain mismatch to be a more challenging problem. Additionally, because the enroll-test channel mismatch naturally exists in the unseen-seen and unseen-unseen verification pairs, domain mismatch is actually a compound problem encompassing channel mismatch.

Current strategies tackle domain mismatch from two aspects: 1) representation space optimization to narrow the cross-domain gaps between enrollment and test utterances and, 2) model generalization improvement for unseen domains. For the first point, recent studies [8]–[12] have investigated the use of scoring and learning strategies. In terms of scoring, probabilistic linear discriminant analysis (PLDA) and its variants [8]–[10] compensate for the speaker factor of utterances by spatial projection to reduce the negative impact of the enroll-test channel mismatch. In terms of model optimization, adversarial training [11] is adopted to remove variations across different channels to learn a channel-invariant subspace. For the second point, the typical approach [13]–[15] is to obtain knowledge that is universally applicable to different domains. Kang et al. [13] used a projection net based on model-agnostic meta-learning (MAML) [16] to fine-tune the output embedding of the pre-trained model and learn domain robust knowledge by choosing different domains for local and meta-updates.

Although recent studies have achieved impressive performance, some limitations still exist. First, there is no capability to simultaneously optimize the embedding space and domain generalization. Domain mismatch is a compound problem that involves channel mismatches and unseen domains. It is difficult for ASV systems with a single learning objective to provide stable performance for different types of cross-domain verification pairs to satisfy the demands of real-world applications. Second, there is no flexible trade-off between discrimination ability (i.e., model performance on the underlying distribution) and generalization ability (i.e.,

Manuscript created June, 2022. *(Corresponding author: Longbiao Wang and Kong Aik Lee.)*

Hanyi Zhang, Longbiao Wang, Meng Liu, and Jianwu Dang are with the Tianjin Key Laboratory of Cognitive Computing and Application, College of Intelligence and Computing, Tianjin University, Tianjin, China (email: hanyizhang@tju.edu.cn; longbiao_wang@tju.edu.cn; liumeng2017@tju.edu.cn; jdang@jaist.ac.jp).

Kong Aik Lee is with the Institute for Infocomm Research, A*STAR, Singapore (email: lee_kong_aik@i2r.a-star.edu.sg).

Jianwu Dang is also with the Japan Advanced Institute of Science and Technology, Ishikawa, Japan.

Helen Meng is with the Human-Computer Communications Laboratory (HCCL), Department of Systems Engineering and Engineering Management, Chinese University of Hong Kong, Hong Kong SAR, China (hmmeng@se.cuhk.edu.hk)).

robust performance on an unseen distribution). As reported in [17]–[19], domain generalization learning typically leads to a decrease in model discrimination. The root of this phenomenon is the difference in the feature representations. In [17], it was pointed out that neural network models tend to use the weakly correlated features of a specific distribution to achieve the desired performance. However, robust training results in models that cannot rely on these non-universal features, which in turn affects the discrimination of the underlying distribution.

In this paper, to reduce the adverse effects of domain mismatch, we propose meta-generalized speaker verification (MGSV) incorporating metric and gradient-based meta-learning. As a domain generalization method, MGSV can be directly deployed to process utterances from the seen and unseen domains without pre-training and fine-tuning. Specifically, we use episode-level sampling to build the meta-train and meta-test tasks. By so doing, the domain shift between the training and testing utterances is simulated, which facilitates the learning of domain-invariant knowledge by the model. Considering the trade-off between discrimination and generalization, and to improve the efficiency and quality of feature learning, we propose multiple-single (MS) and simulated speaker verification (SSV) sampling strategies based on single-domain (SD) and single-single (SS) strategies. Among these, the SSV method exhibits the most promising discrimination and generalization ability. Additionally, we propose distribution optimization and meta-optimization to train the speaker embedding space and learn the generalization ability to unseen domains. For distribution optimization, we adopt a metric loss to supervise the distance between embeddings to constrain the drift caused by domain mismatch. In addition, we use the classification loss to adjust the global embedding space. (The distribution optimization work was presented in part in [20].) Meta-optimization implemented based on high-order gradients is also proposed to learn domain-invariant and discriminative speaker embeddings. In the meta-train stage, the model first accumulates speaker-related knowledge. Subsequently, in the meta-test stage, we improve the generalization ability by feeding back the effect of the model with meta-trained parameters on the virtual test episode. Furthermore, an annealing strategy is adopted to balance the meta-train and meta-test losses to ensure the learning efficiency of the domain-invariant features.

We systematically evaluate the proposed MGSV from multiple perspectives. The optimization of MGSV in the embedding space is evaluated in a channel mismatch experiment. Further, to validate the effectiveness of MGSV on unseen domains, we specially design generalized cross-domain evaluations. The experimental results demonstrate that the proposed MGSV optimizes the embedding distribution and significantly improves the discrimination and generalization ability.

Our main contributions are summarized as follows:

1) We propose MGSV, which leverages the meta-learning technique. Compared to previous works, targeted optimization is performed from the perspective of embedding space and domain generalization.
2) A joint training strategy consisting of metric-based distribution optimization and gradient-based meta-optimization is also proposed. The distribution optimization supervises the spatial distribution of embeddings, whereas the meta-optimization motivates MGSV to improve the generalization ability to unseen domains.
3) We also propose the SSV episode-level sampling strategy. To our knowledge, this is the first work to explore simulated episode construction for speaker verification.
4) Generalized cross-domain evaluations are designed to validate the discrimination and generalization ability of MGSV. The results indicate that MGSV achieves remarkable performance on domain mismatch issues.

II. Related Work

A. Automatic Speaker Verification

The goal of automatic speaker verification (ASV) is to determine whether the speaker's voice matches the claimed identity. In the past few years, the popularity of DNNs has triggered the evolution of ASV systems from traditional probabilistic models [1]–[3] to deep embedding models [4]–[6]. The main components of a DNN-based ASV architecture include a neural network backbone, a pooling layer, a loss function, and a scoring strategy. The neural network backbone transforms acoustic features into representative feature maps. Current studies [6], [21] adopt the squeeze-and-excitation block [22], multi-scale [23] and other strategies to optimize the ability of the model to extract speaker information. The pooling layer followed by linear projection produces speaker embeddings. In [24]–[26], the representation ability of embeddings is improved by focusing on the important frames of utterances. During training, the neural network is optimized to reduce the loss function, for instance, the additive angular margin softmax (AM-Softmax) and its derivative work [27]–[29]. During deployment, the similarity between the enrollment and test utterances is measured using cosine similarity or PLDA [8]–[10] to provide verification results.

B. Meta-Learning

Meta-learning aims to accumulate meta-knowledge from multiple episodes to improve learning ability. In the context of speaker verification, meta-knowledge refers to any experience that can guide model learning in different speaker tasks, such as a generalized embedding space and initial parameters. Recently, meta-learning has been applied to compression, few-shot learning, face recognition, and other tasks. Meta-learning strategies can be grouped into three categories [30]: optimization-based, metric-based, and model-based. Optimization-based meta-learning [16], [31], [32] is achieved via optimization strategies, such as gradient or reinforcement learning. MAML [16] is foundational work that learns a set of initialization parameters by leveraging the gradients from an inner and an outer loop for fast adaptation to new tasks. This gradient-based architecture is applied to domain generalization by meta face recognition (MFR) [33] and meta-learning domain generalization (MLDG) [34]. Raghu et al. [35] reported that the effectiveness of MAML is due to the fact that meta-initialization itself produces high-quality

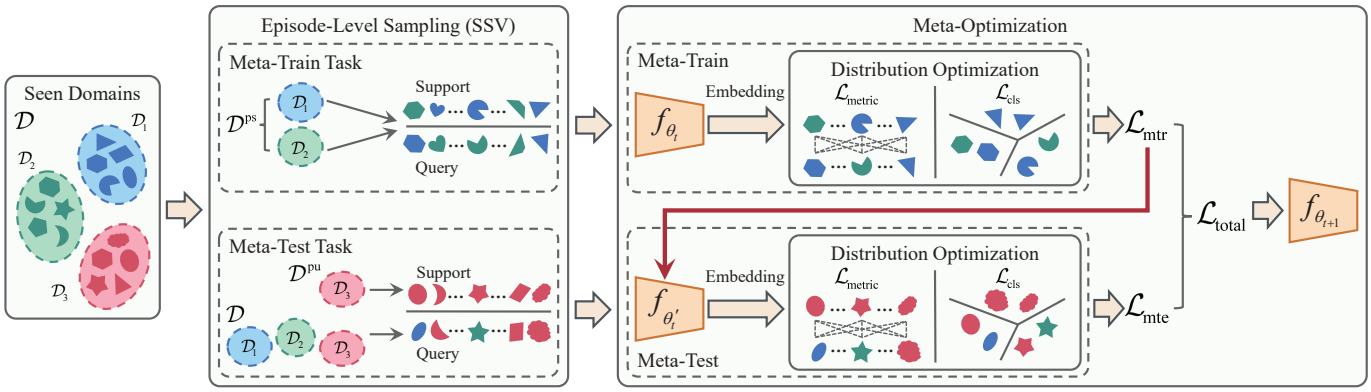


Fig. 1. Architecture of our proposed meta-generalized speaker verification (MGSV) method with simulated speaker verification (SSV) sampling. The same symbol denotes utterances from the same speaker. The training data $\mathcal{D}$ consist of three domains indicated by different colors. In the iteration shown above, we assume that $\{\mathcal{D}_1, \mathcal{D}_2\}$ forms the pseudo-seen domain $\mathcal{D}^{\text{ps}}$ and $\{\mathcal{D}_3\}$ forms the pseudo-unseen domain $\mathcal{D}^{\text{pu}}$.

representation instead of its ability for rapid learning. This explained the feasibility of MAML-based domain generalization methods from that perspective. Metric-based meta-learning [36]–[39] trains an embedding network. The network learns the similarity relationship between the support and query instances to generate a high-representational embedding space that is suitable for new tasks. Model-based meta-learning [40]–[42] is implemented by wrapping task-specific learning in the feed-forward process of a single model. In this study, our work mainly involved optimization and metric-based meta-learning.

C. Domain Generalization

The aim of domain generalization is to obtain a consistent performance in different domains without the need for adaptation. The existing domain generalization methods can be grouped into three main approaches. The first approach involves learning domain-invariant representations. This type of work [43]–[45] reduces the sensitivity of the learned representations to domain shift by shrinking the gap between the source domains caused by different statistical distributions, such that the model can generalize well on unseen domains. Adversarial training and meta-learning are typically used in these approaches. In [43], multiple-domain discriminators and gradient-reversal layers [46] were adopted for domain-invariant representation learning. Our proposed MGSV also belongs to the first approach. The second approach focuses on disentangled representations [47]–[49] whereby source data are decomposed into domain-dependent and domain-independent features. Techniques such as generative adversarial networks [50] have been used to purify domain-independent discriminative representations for cross-domain tasks. The third approach is based on data augmentation. In addition to the traditional hand-crafted transformations [51], model-based [52] and feature-level [53] augmentations have also been widely applied in many investigations.

III. Meta-Generalized Speaker Verification

A. Problem Definition

In many applications, the extrinsic and intrinsic variations are complex and unpredictable, which causes the domain mismatch problem and affects the speaker verification performance. The domain mismatch problem arises in unseen domains that are not included in the training set. Given that the unseen-seen and unseen-unseen verification pairs of domain mismatch naturally involve distribution inconsistencies between the enrollment and test utterances, domain mismatch is actually a compound problem encompassing enroll-test channel mismatch. Optimizing the embedding space perturbed by the channel mismatch is also necessary. Thus, the question of *how to motivate the model to optimize the embedding space and generalize well to unseen domains* is a key question in dealing with the domain mismatch problem. Additionally, the inherent contradiction between discrimination and domain generalization ability [17]–[19] may weaken the optimization of domain mismatch caused by robustness improvement. To learn domain-invariant knowledge, domain generalization methods (such as meta-learning) tend to mine moderately correlated features and filter out domain-specific features. However, for the underlying distribution (known domains), owing to the discarding of weakly correlated features, such operations reduce the information-extraction efficiency and affect the discrimination ability. Thus, the question of *how to deal with the trade-off between discrimination and generalization abilities* is also a key issue.

B. Proposed Method

To eliminate the adverse effects of domain mismatch, we propose the meta-generalized speaker verification (MGSV) method. In contrast to previous works [13], [16], [34], we adopt metric-based distribution optimization to improve the spatial distribution of embeddings and simultaneously use gradient-based meta-optimization to enhance the generalization ability. In addition, we design the episode-level sampling method called simulated speaker verification (SSV) sampling and introduce an annealing strategy to ensure the discrimination ability while learning to generalize. The proposed MGSV is model-agnostic and is therefore applicable to various ASV architectures, such as x-vector [5] and r-vector [54], [55].

The training process for our proposed MGSV is illustrated in Fig. 1 and summarized in Algorithm 1. We as-

sume that $\mathcal{M}$ domains are available in the training set $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_\mathcal{M}\}$. First, we perform episode-level sampling to facilitate the learning of moderately correlated knowledge. Using a leave-one-out strategy, we randomly split $\mathcal{M}$ domains into $\mathcal{M}-1$ pseudo-seen domains $\mathcal{D}^{\text{ps}}$ and one pseudo-unseen domain $\mathcal{D}^{\text{pu}}$ to simulate the domain mismatch encountered in real-world applications. Next, we construct meta-train and meta-test tasks using the pseudo-seen and pseudo-unseen domains. Each task contains a support set and a query set. In the meta-train phase, the DNN $f_{\theta_t}(\cdot)$ extracts utterance embeddings of the meta-train tasks. The meta-train loss $\mathcal{L}_{\text{mtr}}$, consisting of a metric loss and a classification loss, is then computed to supervise the distribution of the embedding space from a local and global perspective. The *meta-trained parameters* $\theta_t{}'$ of the DNN are obtained using the meta-train loss $\mathcal{L}_{\text{mtr}}$ and gradient backpropagation. By so doing, we accumulates the speaker-related knowledge from the meta-train task. In the meta-test phase, we use the DNN $f_{\theta_t{}'}(\cdot)$ updated with the *meta-trained parameters* to calculate the meta-test loss $\mathcal{L}_{\text{mte}}$ for utterances from the meta-test tasks. The meta-test loss $\mathcal{L}_{\text{mte}}$ feeds back the generalization effect of the meta-trained model on the unseen domains, which is beneficial for improving model generalization. Finally, the total loss $\mathcal{L}_{\text{total}}$ is used to train the *optimized parameters* θ_{t+1}. The details of the meta-optimization stages are further presented in Section V-B.

The proposed MGSV is model-agnostic and can be used with architectures such as x-vector. First, the trained DNN extracts the embeddings for the enrollment and test utterances. The verification result is then obtained by measuring the similarity between the enrollment and test embeddings using cosine similarity or PLDA.

IV. Episode-Level Sampling

To find a good trade-off between discrimination ability and domain generalization, we explore several episode-level sampling strategies for speaker verification.

A. Single-Domain Sampling

In [13], [16], the domain mismatch between the training and test sets is simulated to learn transferable knowledge. We refer to this basic approach as single-domain (SD) sampling, as shown in Fig. 2(a). Although the SD approach contains only one domain in each iteration, a domain mismatch exists between the iterations. In the i-th iteration, we randomly select a domain $\mathcal{D}^{\text{SD}}$ from $\mathcal{D}$ and choose $\mathcal{K}$ speakers from the selected domain $\mathcal{D}^{\text{SD}}$. Next, for each speaker, we sample $\mathcal{P}$ utterances as the support set and $\mathcal{Q}$ utterances as the query set to build a meta-train task. Thus, a batch of meta-train tasks containing $\mathcal{K}$ speakers and $\mathcal{N} = \mathcal{K} \cdot (\mathcal{P} + \mathcal{Q})$ utterances is established. The meta-test tasks are built in the same manner by randomly re-sampling another $\mathcal{K}$ speakers and $\mathcal{N}$ utterances from the selected domain $\mathcal{D}^{\text{SD}}$. For each iteration, an episode containing a batch of meta-train tasks and a batch of meta-test tasks is obtained through the above steps. For each episode, the speakers and utterances do not overlap between the meta-train and meta-test tasks.

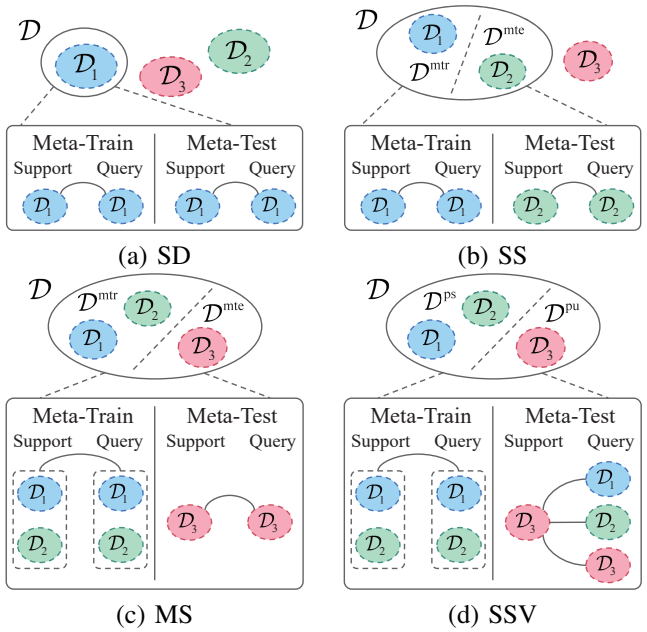


Fig. 2. Domain split details of four different episode-level sampling strategies in one iteration. We assume that the seen domains $\mathcal{D}$ in the training set contain three domains $\{\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3\}$. (a) Single-Domain (**SD**) Sampling. (b) Single-Single (**SS**) Sampling. (c) Multiple-Single (**MS**) Sampling. (d) Simulated Speaker Verification (**SSV**) Sampling.

B. Single-Single Sampling

In SD sampling, a domain mismatch does not exist between the meta-train and meta-test tasks within the same iteration. This limits the ability of the model to reflect on unseen domains. To address this issue, single-single (SS) sampling [13], as shown in Fig 2(b), is proposed. In this method, we select two distinct domains at each iteration to simulate cross-domain tasks. For the i-th iteration, we randomly sample two different domains, one for meta-train $\mathcal{D}^{\text{mtr}}$ and the other for meta-test $\mathcal{D}^{\text{mte}}$. The meta-train tasks consisting of the support and query sets are constructed by sampling utterances from $\mathcal{D}^{\text{mtr}}$, whereas the meta-test tasks consist of utterances from the selected meta-test domain $\mathcal{D}^{\text{mte}}$.

C. Multiple-Single Sampling

Recent results [56] show that SS outperforms SD in most settings, although the improvement is marginal. The inherent contradiction [17]–[19] between discrimination and generalization learning is the primary cause. We argue that a meta-train set with only a single domain tends to learn domain-specific features. However, these weakly correlated features are filtered out by the model because they are not applicable to distinct domains in the meta-test phase. This situation limits the efficiency of the model in extracting correlated knowledge and is detrimental to discrimination learning and model performance. To overcome this deficiency, we increase the number of domains contained in the meta-train phase to improve the probability of mining universally correlated knowledge in this phase, thereby optimizing the efficiency and quality of feature learning. We call this strategy multiple-single (MS) sampling, as shown in Fig. 2(c). For each iteration, MS

sampling first randomly samples $\mathcal{M}-1$ domains from the $\mathcal{M}$ source domains $\mathcal{D}$ as the meta-train domain $\mathcal{D}^{\mathrm{mtr}}$. The remaining domain is adopted as the meta-test domain $\mathcal{D}^{\mathrm{mte}}$. We then construct the meta-train and meta-test tasks in the same manner as in SD and SS.

D. Simulated Speaker Verification Sampling

In some applications, utterances from the unseen domains may form verification pairs with utterances from all domains (including seen and unseen domains), rather than the pairs involving only a single unseen domain, such as the simulated meta-test task of MS. Thus, on the basis of MS, we propose *simulated speaker verification* (SSV) sampling to deliberately optimize the generalization ability of the model on the unseen-seen and unseen-unseen pairs in the meta-test phase. As shown in Fig. 2(d), the support set of the meta-test task is drawn from the pseudo-unseen domain, whereas its query set is drawn from all domains. Specifically, we first select $\mathcal{M}-1$ domains from $\mathcal{M}$ domains $\mathcal{D}$ as pseudo-seen domains $\mathcal{D}^{\mathrm{ps}}$ and select the remaining domain as pseudo-unseen domain $\mathcal{D}^{\mathrm{pu}}$. Then, the meta-train tasks are constructed in the same manner as MS. Next, we randomly choose $\mathcal{K}$ speakers from the pseudo-unseen domain $\mathcal{D}^{\mathrm{pu}}$ and $\mathcal{P}$ utterances for each speaker as the support set of the meta-test task. Subsequently, we sample $\mathcal{Q}$ utterances for each of the chosen $\mathcal{K}$ speakers from all source domains $\mathcal{D}$ as the query set.

V. Optimization

In this section, we present the algorithm utilized for distribution optimization of the embeddings and meta-optimization for improving model generalization to unseen domains.

A. Distribution Optimization

Distribution optimization includes a metric loss and a classification loss, which are used in both the meta-train and meta-test iterations.

Metric Loss. Domain mismatch leads to inconsistencies in the statistical distribution, which causes data drift. The drift further impacts the speaker similarity between utterances and degrades the verification results. Thus, we introduce a metric loss to constrain the embedding-level drift by decreasing the intra-class distance and increasing the inter-class distance. The metric loss is realized by cosine similarity and the calculation pipeline includes the following steps. First, we obtain the centroid representations of $\mathcal{K}$ speakers by averaging $\mathcal{P}$ utterance embeddings of each speaker in the support set. The centroid representation $\boldsymbol{C}_k$ of the k-th selected speaker is computed as follows:

$$\boldsymbol{C}_k = \frac{1}{\mathcal{P}} \sum_{p=1}^{\mathcal{P}} f_\theta(\boldsymbol{U}_p^k) \tag{1}$$

where $\mathcal{P}$ denotes the number of utterances for each speaker in the support set, $f_\theta(\cdot)$ denotes the DNN model and $\boldsymbol{U}_p^k$ denotes the p-th support utterance of the k-th speaker. Then, we calculate the cosine similarities $\boldsymbol{d}_q = \{\boldsymbol{d}_{q,1}, \boldsymbol{d}_{q,2}, \cdots, \boldsymbol{d}_{q,\mathcal{K}}\}$ between the embeddings of the q-th query utterances in the task batch and the centroid representations $\boldsymbol{C} = \{\boldsymbol{C}_1, \boldsymbol{C}_2, \cdots, \boldsymbol{C}_\mathcal{K}\}$ of all speakers. By normalizing the query embeddings and centroid representations, the calculation of $\boldsymbol{d}_q$ can be simplified into the following form,

$$\boldsymbol{d}_q = f_\theta(\boldsymbol{U}_q^{\mathrm{query}}) \cdot \boldsymbol{C}^\top \tag{2}$$

where $\boldsymbol{U}_q^{\mathrm{query}}$ is the q-th utterance of all $\mathcal{N}_{\mathrm{query}} = \mathcal{K} \cdot \mathcal{Q}$ query utterances in the task batch, $\mathcal{K}$ is the number of selected speakers and $\mathcal{Q}$ is the number of utterances of each speaker in the query set. Finally, the metric loss $\mathcal{L}_{\mathrm{metric}}$ is defined as the cross-entropy,

$$\mathcal{L}_{\mathrm{metric}} = -\frac{1}{\mathcal{N}_{\mathrm{query}}} \sum_{q=1}^{\mathcal{N}_{\mathrm{query}}} \boldsymbol{S}_q \cdot \log(\boldsymbol{d}_q) \tag{3}$$

where $\boldsymbol{S}_q$ is the one-hot vector of the selected speaker corresponding to the q-th utterance in the batch of tasks.

Classification Loss. The metric loss optimizes the similarity relationship between cross-domain embeddings from a focused local perspective. To further stabilize the model performance, a softmax-based classification loss is introduced to supervise the distribution of speaker clusters from a global perspective. In particular, we use all speakers $y = 1, 2, \cdots, \mathcal{Y}$ as the ground truth to compute the classification loss $\mathcal{L}_{cls}$ for all utterances in the batch of tasks, including both the support and query sets. The calculation pipeline is as follows,

$$\mathcal{L}_{\mathrm{cls}} = -\frac{1}{\mathcal{N}_{\mathrm{task}}} \sum_{i=1}^{\mathcal{N}_{\mathrm{task}}} \boldsymbol{y}_i \cdot \log(f_\alpha(f_\theta(\boldsymbol{U}_i^{\mathrm{task}}))) \tag{4}$$

where $\mathcal{Y}$ is the total number of speakers in the source domain, $\mathcal{N}_{\mathrm{task}} = \mathcal{K} \cdot (\mathcal{P} + \mathcal{Q})$ is the total number of utterances in the batch of tasks, $\boldsymbol{y}_i$ is the one-hot vector of the speaker ID of the i-th utterance, $f_\alpha(\cdot)$ is the classification linear layer followed by softmax and $\boldsymbol{U}_i^{\mathrm{task}}$ is the i-th utterance in the task batch.

B. Meta-Optimization

Meta-optimization combines the meta-train and meta-test stages and is designed to improve the generalization ability of the model to unseen domains. The overall meta-optimization procedure is illustrated in Fig. 1 and Algorithm 1.

Meta-Train. Owing to the trade-off between discrimination and generalization abilities, the improvement in domain generalization may affect model discrimination. Thus, in the meta-train phase, to ensure the quality of discrimination learning, we build the meta-train task involving multiple pseudo-seen domains using SSV and motivate the model to accumulate sufficient moderately correlated knowledge. Specifically, we use the DNN model $f_{\theta_t}(\cdot)$ with parameters θ_t and distribution optimization to calculate the meta-train loss $\mathcal{L}_{\mathrm{mtr}}$ for the meta-train tasks,

$$\mathcal{L}_{\mathrm{mtr}} = \mathcal{L}_{\mathrm{metric}}(\mathcal{X}_{\mathrm{mtr}}; \theta_t) + \mathcal{L}_{\mathrm{cls}}(\mathcal{X}_{\mathrm{mtr}}, \mathcal{Y}_{\mathrm{mtr}}; \theta_t), \tag{5}$$

where $\mathcal{X}_{\mathrm{mtr}}$ and $\mathcal{Y}_{\mathrm{mtr}}$ denote the utterances and speaker IDs of the meta-train tasks. Through gradient backpropagation, the *meta-trained parameters* ${\theta_t}'$ of the model are obtained,

$${\theta_t}' = \theta_t - \sigma \nabla_{\theta_t} \mathcal{L}_{\mathrm{mtr}}(\theta_t) = \theta_t - \sigma {\mathcal{L}'}_{\mathrm{mtr}}(\theta_t) \tag{6}$$

where σ is the learning rate and $\nabla_{\theta_t}\mathcal{L}_{\text{mtr}}(\theta_t) = \mathcal{L}'_{\text{mtr}}(\theta_t)$ is the meta-trained gradient of the model.

Meta-Test. The meta-test phase regularizes the model optimization to improve the generalization ability for unseen domains. Based on the SSV episode-level sampling strategy, we construct a batch of meta-test tasks to fully simulate the unseen-seen and unseen-unseen verification pairs in real-world scenarios. Accordingly, in the meta-test phase, the robust performance of the model with *meta-trained parameters* θ_t' on the pseudo-unseen domains is fed back to the optimized gradient to boost the generalization ability. Formally, we use the model $f_{\theta_t'}(\cdot)$ with *meta-trained parameters* and distribution optimization to calculate the meta-test loss $\mathcal{L}_{\text{mte}}$:

$$\mathcal{L}_{\text{mte}} = \mathcal{L}_{\text{metric}}(\mathcal{X}_{\text{mte}}; \theta_t') + \mathcal{L}_{\text{cls}}(\mathcal{X}_{\text{mte}}, \mathcal{Y}_{\text{mte}}; \theta_t') \quad (7)$$

where $\mathcal{X}_{\text{mte}}$ and $\mathcal{Y}_{\text{mte}}$ denote the utterances and labels of the meta-test tasks. Then, by weighting the meta-train loss $\mathcal{L}_{\text{mtr}}$ and meta-test loss $\mathcal{L}_{\text{mte}}$, the total loss $\mathcal{L}_{\text{total}}$ is computed,

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{mtr}}(\theta_t) + \gamma\mathcal{L}_{\text{mte}}(\theta_t - \sigma\mathcal{L}'_{\text{mtr}}(\theta_t)). \quad (8)$$

where γ denotes the weight with the annealing strategy to balance the meta-train loss and meta-test losses. It controls the model to accumulate speaker-related features efficiently in the early stage of training and learns universal knowledge with high-quality for the unseen domains in the late stage. Thus, the effectiveness of the learned discrimination knowledge is ensured while learning the domain generalization. Through total loss and high-order gradients, the final *optimized parameters* θ_{t+1} of the model are obtained,

$$\theta_{t+1} = \theta_t - \mu\frac{\partial(\mathcal{L}_{\text{mtr}}(\theta_t) + \gamma\mathcal{L}_{\text{mte}}(\theta_t - \sigma\mathcal{L}'_{\text{mtr}}(\theta_t)))}{\partial\theta_t} \quad (9)$$

where μ is the learning rate.

Algorithm 1: Meta-Generalized Speaker Verification

Input: Source domains $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_{\mathcal{M}}\}$.
Init: Model parameters θ. Hyperparameters σ, γ, μ.
Output: Optimized parameters θ of the model.

for *ite* ***in*** *iterations* **do**
- Sample $\mathcal{M}-1$ domains from $\mathcal{D}$ as $\mathcal{D}^{\text{ps}}$;
- The remaining one domain of $\mathcal{D}$ is $\mathcal{D}^{\text{pu}}$;
- `// Meta-Optimization`
- **Meta-Train:**
- Sample a batch of meta-train tasks $(\mathcal{X}_{\text{mtr}}, \mathcal{Y}_{\text{mtr}})$ from the pseudo-seen domains $\mathcal{D}^{\text{ps}}$ using SSV;
- $\mathcal{L}_{\text{mtr}} = \mathcal{L}_{\text{metric}}(\mathcal{X}_{\text{mtr}}; \theta) + \mathcal{L}_{\text{cls}}(\mathcal{X}_{\text{mtr}}, \mathcal{Y}_{\text{mtr}}; \theta)$;
- Update meta-trained parameters by: $\theta' = \theta - \sigma\nabla_\theta\mathcal{L}_{\text{mtr}}(\theta) = \theta - \sigma\mathcal{L}'_{\text{mtr}}(\theta)$;
- **Meta-Test:**
- Sample a batch of meta-test tasks $(\mathcal{X}_{\text{mte}}, \mathcal{Y}_{\text{mte}})$ from the pseudo-unseen domain $\mathcal{D}^{\text{pu}}$ and the source domains $\mathcal{D}$ using SSV;
- $\mathcal{L}_{\text{mte}} = \mathcal{L}_{\text{metric}}(\mathcal{X}_{\text{mte}}; \theta') + \mathcal{L}_{\text{cls}}(\mathcal{X}_{\text{mte}}, \mathcal{Y}_{\text{mte}}; \theta')$;
- **Optimization:** Optimize θ $\theta \leftarrow \theta - \mu\frac{\partial(\mathcal{L}_{\text{mtr}}(\theta) + \gamma\mathcal{L}_{\text{mte}}(\theta - \sigma\mathcal{L}'_{\text{mtr}}(\theta)))}{\partial\theta}$;

return optimized parameters θ of the model;

VI. Experimental Setups

The dataset settings and implementation details of these methods are described in this section.

A. Datasets

To validate the universal effectiveness of MGSV in cross-domain challenges, we evaluate on the cross-genre, cross-device, and cross-dataset tasks.

Cross-Genre. CN-Celeb [56], a dedicated cross-genre corpus, is used to test model performance. CN-Celeb contains in-the-wild utterances of 3,000 speakers from 11 different genres, such as vlogs, interviews, and dramas. Considering data quality, we discard speakers with fewer than 5 utterances and the "advertisement" genre with fewer than 100 speakers. Finally, we construct a training set with 2,768 speakers and a test set with 200 speakers. For trials, utterances belonging to a certain genre are paired with random utterances from all genres to form the trial of that genre, and the cross-genre trial is a union of all genre trials. The above settings ensure that the trials more accurately and precisely reflect the model ability to process different genre mismatch issues.

Cross-Device. HI-MIA [57] is used to evaluate the cross-device model performance. In a real home environment, Devices 2, 5, 6, and 7 are placed at different locations to record utterances. Device 7 is a close-talking microphone recorded at a sampling frequency of 44.1 kHz. Devices 2, 5, and 6 are circular microphone arrays recorded at 16 kHz, which are placed in different positions. In our experiments, we use recordings from channel 06-13 of the circular microphone arrays. The training and test sets contain 254 and 44 speakers, respectively. For trials, similar to cross-genre, we pair the utterances of a certain device with the utterances of all devices to construct the trial of that device. The cross-device trial includes verification pairs of all single-device trials.

Cross-Dataset. In real-world applications, speaker verification systems are required to obviate domain mismatch issues with multiple uncontrollable variables. To evaluate the model performance in complex situations more relevantly, we adopt a cross-dataset setting that includes CN-Celeb [56], HI-MIA [57], FFSVC [58] and Voxceleb [59]. Genre mismatches exist in the CN-Celeb corpus. In the HI-MIA corpus, the recording devices and locations are different. The utterances in the FFSVC are time-varying and noisy. Data marked "I0.25M" are used in our experiments. Voxceleb, a popular clean corpus, is also used. We randomly select 1,307 speakers for training and 175 for testing. The design of the cross-dataset trial is similar to that of the cross-genre and cross-device trials.

B. Implementation Details

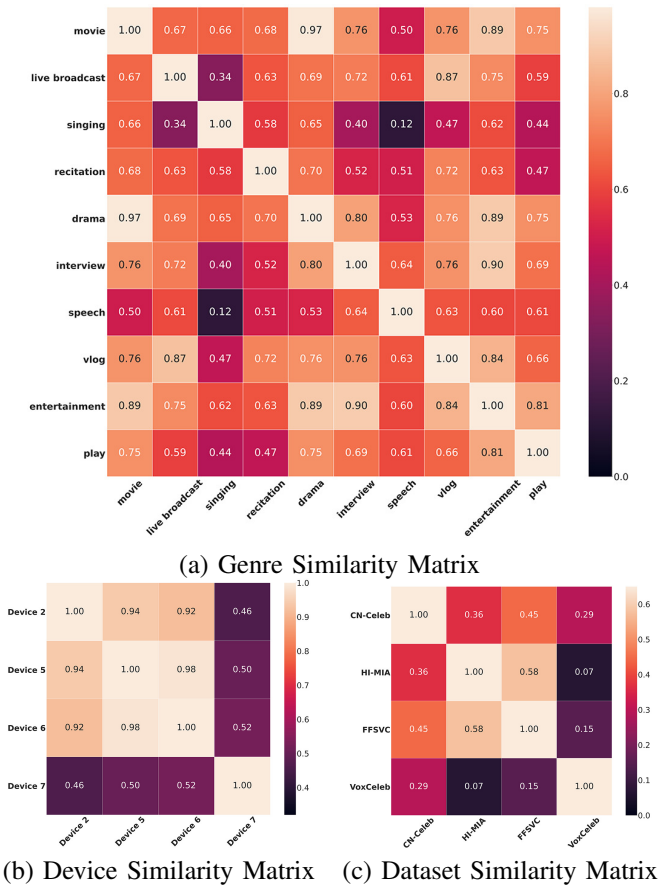
In our experiments, we introduce other methods for comparison purpose. Note that all the methods uniformly use log Mel-filterbank features and cosine similarity.

Acoustic Feature Extraction. We randomly intercept 200 consecutive frame segments from each utterance. The 40-dimensional log Mel-filterbank features are extracted with a 25 ms window and 10 ms window step.

X-Vector and R-Vector. The model architectures adopted by x-vector [5] and r-vector [54], [55] are the same, their difference lies in the neural network backbone used. For r-vector, ResNet-18 [60] referring to the configuration in [54], is used as the backbone. Specifically, ResNet-18 first transforms the acoustic features into deep features. Then, the global average pooling (GAP) maps the deep features into the utterance-level embeddings. After the pooling layer, a linear layer performs further processing and outputs 256-dimensional embeddings. In the training stage, softmax is used to calculate the loss. For x-vector, a time-delay neural network (TDNN) with the configuration in [5] is used as the neural network backbone, and the remaining details are consistent with r-vector.

MAML. MAML [16] uses the r-vector architecture and softmax loss. This implementation used is based on [34]. The meta-train and meta-test sets of each iteration use the same domain. Further, each iteration uses a random domain, similar to the single-single strategy.

MGSV. Because MGSV is model-agnostic, it can be easily applied to different neural network backbones without any additional parameters. In our experiments, we implement MGSV-rvec and MGSV-xvec based on the r-vector and x-vector, respectively. Specifically, 100 speakers ($\mathcal{K} = 100$) are randomly selected for a batch of tasks. Each speaker randomly samples one utterance as a support ($\mathcal{P} = 1$) and two utterances as a query ($\mathcal{Q} = 2$). The weight γ balancing the meta-train and meta-test losses, which adopts annealing scheduling, is initialized to 0.001 and updated with increasing epochs at $\gamma_{ep} = 0.001 \times 1.2^{ep}$, where ep denotes the training epoch. Both learning rates σ and μ are initialized to 0.05.



(a) Genre Similarity Matrix

(b) Device Similarity Matrix (c) Dataset Similarity Matrix

Fig. 3. Domain similarity matrices of genres, devices, and datasets. Each value represents the cosine similarity between the domains corresponding to the x-axis and the y-axis. The lighter the color, the higher is the similarity.

VII. Experiments and Analysis

In this section, we present the design and analysis of the domain similarity experiments, comparative experiments on channel mismatch, episode-level sampling comparisons, and several generalized cross-domain benchmarks.

A. Domain Similarity Experiments

To visualize the domain mismatch in speaker verification, we compute the domain similarity matrix. Specifically, we first train an r-vector model on Voxceleb. Then, we randomly sample 1,000 utterances for each domain and use the trained r-vector to extract the speaker embeddings. Next, the speaker embedding centroid of each domain, namely domain embedding, is obtained by averaging 1,000 utterance embeddings of each domain. Cosine similarity is used to reflect the relationship between the domains.

We consider the genre, recording device, and dataset containing multiple factors as domains to explore different types of cross-domain issues. The domain similarity matrices of the genres, recording devices, and datasets are presented in Fig. 3. Note that for the r-vector trained on Voxceleb, the domains are unseen. It can be observed that the domain embeddings are not coincident and the similarities between domains differ significantly. For example, the similarity between the "drama" and "movie" genres is 0.97, whereas for the "singing" and "speech" genres, it is only 0.12. This implies that the domain shift is ubiquitous and may interfere with the distribution of the embedding space. Additionally, the unpredictable similarity between domains increases the difficulty of dealing with unseen domains. Thus, optimizing the speaker embedding space and learning domain generalization are desired to develop a robust speaker verification system.

B. Comparative Experiments on Channel Mismatch

The optimization effect of the methods on the embedding space is reflected in the channel mismatch experiment, where the enrollment and test utterances are from different domains. Furthermore, as mentioned in Section III-A, domain mismatch is a compound challenge that involves channel mismatch. To this end, we design comparative experiments on the channel mismatch.

The results of the cross-genre, cross-device and cross-dataset experiments are shown in Table I. It can be observed that the proposed MGSV is significantly superior to other methods in terms of optimizing the embedding distribution. Compared to r-vector, MGSV-rvec achieves 18.26%, 11.09%, and 29.28% relative reductions in terms of equal error rate (EER) on cross-genre, cross-device, and cross-dataset, respectively. In addition, on average, MGSV-xvec obtains a 12.61% relative reduction in EER compared to x-vector. The main

TABLE I
EQUAL ERROR RATES (EERS) (%) OF MGSV AND OTHER METHODS ON THE CROSS-GENRE, CROSS-DEVICE, AND CROSS-DATASET TRIALS. **MGSV-RVEC** DENOTES MGSV WITH R-VECTOR ARCHITECTURE. **MGSV-XVEC** DENOTES MGSV WITH X-VECTOR ARCHITECTURE.

Method	Cross-Genre	Cross-Device	Cross-Dataset
r-vector	24.09	13.25	12.16
x-vector	22.55	9.08	10.82
MAML	24.23	12.25	13.34
MGSV-rvec	**19.69**	11.78	**8.60**
MGSV-xvec	20.39	**8.21**	8.80

reason for this is that MGSV adopts distribution optimization to supervise the metric distances between embeddings, thereby reducing the offsets caused by channel mismatch. Additionally, meta-optimization promotes the learning of domain-invariant knowledge, which also contributes to the adjustment of speaker embeddings. Furthermore, the results of MGSV-rvec and MGSV-xvec verify that MGSV universally optimizes the distribution of the embedding space on different neural network backbones.

We also introduce the classic MAML meta-learning method for comparison. The results indicate that MAML is not substantially superior to r-vector, which is similar to the results in [56]. This phenomenon is mainly attributed to the lack of optimization of the similarity between embeddings in the training stage. This limitation makes MAML indirectly applicable to speaker verification that emphasizes the similarity of verification pairs. Our proposed MGSV ameliorates these issues and uses distribution optimization to achieve remarkable improvement.

C. Generalized Cross-Genre Evaluation

Because of the complex and uncontrollable factors in real-world applications, it is essential that the speaker verification system be able to process cross-domain issues in which both the enrollment and test utterances may originate from seen and unseen domains. Thus, to intuitively reflect the generalization ability of the methods in unseen domains, we design three types of generalized cross-domain evaluations, as presented in Sections VII-C, VII-D and VII-E.

For the generalized cross-genre evaluation (GCG) settings, we first group the more similar genres into a group according to the results in Fig. 3(a). By so doing, we minimize the correlation between genre groups to ensure that GCG can fully verify the domain generalization ability. As shown in Table II, we divide the ten genres of CN-Celeb into four groups. We choose three of the four groups as the seen domains for training. The remaining group, as the unseen domain, is only used for testing. Hence, we design four protocols using different genre groups as the unseen domains, as shown in Table III. Note that the trial construction for each genre is presented in Section VI-A.

Table IV shows the results obtained on the GCG benchmark. It can be seen that the proposed MGSV outperforms the other algorithms in terms of both generalization and discrimination. In Table IV, the unseen domain trials (e.g., "pl" in GCG III),

TABLE II
THE GROUPING RESULTS OF THE TEN GENRES IN CN-CELEB.

Group	Genre Types
Genre I	speech (sp), live broadcast (lb), vlog (vl)
Genre II	interview (in), play (pl), entertainment (en)
Genre III	drama (dr), movie (mo)
Genre IV	recitation (re), singing (si)

TABLE III
THE SETTINGS OF GENERALIZED CROSS-GENRE EVALUATION (GCG). THE SEEN DOMAINS ARE AVAILABLE DURING THE TRAINING STAGE. THE UNSEEN DOMAIN IS USED SOLELY FOR TESTING.

Protocol		Seen Domains	Unseen Domain
GCG	I	Genre I, Genre II, Genre III	Genre IV
	II	Genre I, Genre II, Genre IV	Genre III
	III	Genre I, Genre III, Genre IV	Genre II
	IV	Genre II, Genre III, Genre IV	Genre I

including unseen-seen and unseen-unseen verification pairs, can illustrate the generalization ability of the *unknown* domains. It can also be seen that, compared to r-vector, MGSV-rvec achieves a 12.07% relative EER reduction on the unseen domain trials on average, which validates the improvements in the domain generalization ability. Additionally, the results of the seen domain trials (e.g., "sp" in GCG II) containing the seen-seen verification pairs indicate that MGSV has a more promising discrimination ability on the *known* domains. Furthermore, the cross-genre trial that combines the seen and unseen domain trials comprehensively reflects the discrimination and generalization abilities of the methods. The results show that MGSV-rvec and MGSV-xvec achieve 15.02% and 7.18% relative EER reductions in the cross-genre trial on average compared to r-vector and x-vector, respectively.

The reason for the above results is that we employ meta-optimization based on a high-order gradient and balance the relationship between discrimination learning and generalization learning. First, as stated in Section V-B, by simulating the train/set domain mismatch and regularizing model gradients during training, MGSV learns transferable knowledge from different domains, which is conducive to model generalization in unseen domains. Second, as described in [17]–[19], the trade-off between model discrimination ability and its robustness to drift data exits universally. The root of this phenomenon is the difference in the feature representations. To discover more universal knowledge, robust training (e.g., domain generalization) filters out some weakly correlated features on which discrimination ability depends, resulting in a decrease in learning efficiency. In response to this, we adopt the refined episode-level sampling and annealing strategies described in Sections IV-D and V-B to specifically optimize both the quality and efficiency of feature learning. By so doing, the performance improvement of MGSV resulting from generalization learning is not diluted owing to a decrease in discrimination ability, as in MAML. The results of the episode design are analyzed in detail in Section VII-F.

Distinctively, the results of the "recitation" ("re" in Table IV) genre are different from the other genres. This is because the "recitation" genre only contains the "id00877" speaker in the test set, which causes the "recitation" trial to lack positive

TABLE IV
EERs (%) OF MGSV AND OTHER METHODS ON THE GCG BENCHMARK. THE UNSEEN DOMAIN OF EACH PROTOCOL IS MARKED WITH *. ABBREVIATIONS SUCH AS "SP" DENOTE DIFFERENT TYPES OF GENRES, AND THEIR FULL NAMES ARE PRESENTED IN TABLE II.

Protocol	Method	Cross-Genre	Genre I			Genre II			Genre III		Genre IV	
			sp	lb	vl	in	pl	en	dr	mo	re	si
GCG I (*Genre IV)	r-vector	24.25	17.47	22.30	20.98	25.35	27.08	24.28	25.08	31.26	15.64	28.71
	x-vector	22.55	15.97	19.20	20.84	23.53	31.69	22.81	23.08	29.20	**12.42**	27.35
	MAML	24.50	16.90	22.32	22.21	25.62	29.92	24.16	26.32	32.00	13.63	29.86
	MGSV-rvec	**20.09**	16.01	**17.91**	**18.15**	**20.48**	**20.85**	**20.43**	**21.47**	25.12	13.70	**24.35**
	MGSV-xvec	20.99	**15.40**	18.35	20.52	21.14	24.00	21.71	22.16	**25.07**	15.13	25.89
GCG II (*Genre III)	r-vector	24.29	17.67	21.97	21.98	25.66	24.92	24.33	25.48	31.39	12.71	27.49
	x-vector	22.66	16.07	19.70	21.48	23.64	28.38	22.99	23.01	30.45	**11.79**	26.16
	MAML	25.02	18.36	21.39	23.08	26.46	28.31	25.52	26.09	33.60	16.96	27.33
	MGSV-rvec	**20.00**	**14.84**	19.36	**17.92**	**20.41**	**23.31**	**20.54**	**21.16**	**25.37**	13.52	**22.70**
	MGSV-xvec	20.44	15.22	**18.02**	20.43	20.55	24.77	21.49	21.88	26.03	13.96	23.87
GCG III (*Genre II)	r-vector	26.64	19.24	23.57	22.53	29.12	29.54	27.06	26.94	33.54	15.05	28.46
	x-vector	25.03	**17.33**	21.45	22.53	26.80	31.00	25.72	25.29	31.43	**12.23**	27.68
	MAML	26.92	19.23	23.02	23.84	29.38	29.62	27.33	27.67	34.90	15.46	30.49
	MGSV-rvec	23.74	18.75	23.87	**19.28**	24.99	28.77	**24.30**	**24.90**	28.81	14.87	**24.76**
	MGSV-xvec	**23.61**	17.59	**20.87**	22.73	**24.59**	**28.15**	24.90	25.22	**28.41**	14.54	26.01
GCG IV (*Genre I)	r-vector	24.09	17.68	21.01	21.87	25.52	26.23	24.12	25.46	33.77	**13.08**	27.05
	x-vector	22.79	16.38	19.14	21.99	23.93	28.85	23.28	23.78	30.36	13.70	25.51
	MAML	24.93	17.52	22.02	21.62	26.66	27.85	24.95	26.54	32.29	15.57	27.82
	MGSV-rvec	**20.63**	**15.24**	19.55	**19.84**	**20.96**	**21.23**	**21.10**	22.80	**25.61**	13.30	**23.44**
	MGSV-xvec	21.35	16.04	**18.39**	22.15	21.65	25.92	22.26	**22.68**	26.32	15.38	24.32

TABLE V
THE SETTINGS OF GENERALIZED CROSS-DEVICE EVALUATION (GCDV). THE SEEN DOMAINS ARE AVAILABLE DURING THE TRAINING STAGE. THE UNSEEN DOMAIN IS USED SOLELY FOR TESTING.

Protocol		Seen Domains	Unseen Domain
GCDV	I	Devices 2, 5, 7	Device 6
	II	Devices 2, 6, 7	Device 5
	III	Devices 5, 6, 7	Device 2

TABLE VI
EERs (%) OF MGSV AND OTHER METHODS ON THE GCDV BENCHMARK. THE UNSEEN DOMAIN OF EACH PROTOCOL IS MARKED WITH *.

Protocol	Method	EER (%)				
		Cross-Device	2	5	6	7
GCDV I (*Device 6)	r-vector	13.24	12.27	13.47	13.45	17.13
	x-vector	10.36	9.32	10.73	10.52	14.99
	MAML	15.46	13.87	15.30	15.28	20.71
	MGSV-rvec	12.43	11.62	12.67	12.81	15.69
	MGSV-xvec	**9.25**	**8.34**	**9.56**	**9.56**	**13.36**
GCDV II (*Device 5)	r-vector	13.65	12.79	14.01	13.94	15.71
	x-vector	10.47	9.30	10.88	10.73	14.66
	MAML	13.88	12.33	13.92	13.76	18.32
	MGSV-rvec	12.89	12.03	13.12	13.23	15.16
	MGSV-xvec	**9.65**	**8.66**	**10.01**	**10.07**	**13.10**
GCDV III (*Device 2)	r-vector	13.26	12.90	13.43	13.24	16.95
	x-vector	11.17	10.56	11.37	11.21	15.83
	MAML	14.17	12.74	14.17	14.08	20.58
	MGSV-rvec	13.16	12.62	13.36	13.50	15.57
	MGSV-xvec	**10.26**	**9.50**	**10.59**	**10.58**	**13.39**

pairs and be extremely biased toward methods of rigorously recognizing verification pairs as negative pairs.

D. Generalized Cross-Device Evaluation

The design of the generalized cross-device evaluation (GCDV) is similar to that of the GCG. Table V shows the details of the GCDV settings. In each protocol, we choose three devices from Devices 2, 5, 6, and 7 as the seen domains for training. The remaining device, as the unseen domain, is only used for testing, which can simulate unknown devices. Note that because Device 7 denotes the clean utterances recorded by the close-talking microphone, it will always be used as the seen domain.

The results for the GCDV benchmark are presented in Table VI. It can be seen that in the cross-device trial, MGSV-rvec and MGSV-xvec significantly surpass the baselines (r-vector and x-vector) with the same network architecture. The above results illustrate that the proposed MGSV is still effective on the cross-device problem, which validates that MGSV is flexible and can improve the model performance for different types of cross-domain issues. Additionally, promising improvements are achieved without increasing the size of the model, which means that MGSV with scalability can be easily applied to different neural network backbones.

TABLE VII
THE SETTINGS OF GENERALIZED CROSS-DATASET EVALUATION (GCDS). THE SEEN DOMAINS ARE AVAILABLE DURING THE TRAINING STAGE. THE UNSEEN DOMAIN IS USED SOLELY FOR TESTING.

Protocol		Seen Domains	Unseen Domain
GCDS	I	CN-Celeb, HI-MIA, Voxceleb	FFSVC
	II	CN-Celeb, FFSVC, Voxceleb	HI-MIA
	III	HI-MIA, FFSVC, Voxceleb	CN-Celeb

E. Generalized Cross-Dataset Evaluation

The ASV systems applied to real-world scenes typically face compound challenges involving multiple variations such as recording devices, speaking styles, and noise. Unlike the single-variable problem, a multivariable challenge is difficult

TABLE VIII
EERs (%) OF MGSV AND OTHER METHODS ON THE GCDS BENCHMARK. THE UNSEEN DOMAIN OF EACH PROTOCOL IS MARKED WITH *.

Protocol	Method	EER (%)				
		Cross-Dataset	CN-Celeb	HI-MIA	FFSVC	Voxceleb
GCDS I (*FFSVC)	r-vector	13.30	18.20	7.88	**6.21**	7.19
	x-vector	11.86	16.73	6.25	6.39	4.63
	MAML	14.27	20.04	9.40	6.59	8.02
	MGSV-rvec	10.73	15.38	7.05	7.47	4.82
	MGSV-xvec	**10.57**	**14.80**	**6.12**	7.84	**3.57**
GCDS II (*HI-MIA)	r-vector	13.75	17.56	**10.40**	5.43	6.80
	x-vector	12.17	15.64	11.30	5.15	4.16
	MAML	15.04	20.74	11.00	5.56	8.28
	MGSV-rvec	10.89	14.37	11.11	6.37	4.89
	MGSV-xvec	**10.73**	**13.56**	12.86	**5.10**	**3.44**
GCDS III (*CN-Celeb)	r-vector	12.75	20.08	7.92	5.12	7.16
	x-vector	11.95	20.39	**6.08**	**4.81**	4.90
	MAML	13.96	21.37	8.47	5.88	8.05
	MGSV-rvec	12.04	**16.56**	8.27	8.25	6.61
	MGSV-xvec	**11.46**	17.35	6.84	6.17	**4.27**

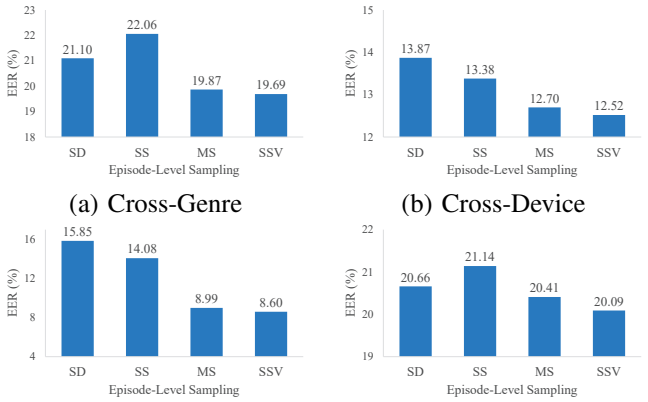
to solve by directly decoupling single objects, owing to the diversity of the involved factors. Thus, the multi-variable challenge is more complicated. In our experiment, we adopt a generalized cross-dataset evaluation (GCDS) to simulate the multivariable challenge. To the best of our knowledge, this is the first study to explore multivariable domain mismatch.

The GCDS settings are shown in Table VII. In each protocol, we choose two datasets from CN-Celeb, HI-MIA, and FFSVC together with the Voxceleb corpus as the seen domains for training. The remaining dataset is used as the unseen domain for the testing.

The comparative results on the GCDS benchmark are shown in Table VIII. From the results, it can be observed that MGSV-rvec and MGSV-xvec achieve remarkable improvement in the cross-dataset compared to the other methods, which indicates that MGSV can indeed tackle the multivariable challenge to a certain extent. In GCDS III, the results of the unseen CN-Celeb demonstrate that MGSV can also generalize well to difficult unseen domains. However, we find that MGSV obtains unsatisfactory and worse performance on HI-MIA and FFSVC. A possible reason for this is that the cross-dataset involves too many trade-off factors, resulting in the embeddings of a single space that may not be available to provide the best performance for different ASV tasks. This point is also reflected in other results. For example, r-vector shows good performance on HI-MIA and FFSVC, but it does not work well on Voxceleb. To solve this problem, in our future work, we plan to generate separate embedding spaces for specific tasks by adaptively fine-tuning the output embeddings, thereby enhancing the representation ability of embeddings for the multivariable cross-domain challenges.

F. Episode-Level Sampling Comparisons

To improve both the discrimination and generalization abilities, SD, SS, MS, and SSV sampling strategies are described in Section IV. In this section, we evaluate these four episode designs on cross-genre, cross-device, cross-dataset, and cross-genre (GCG I). Cross-genre (GCG I) denotes the cross-genre using the GCG I protocol, that is, the Genre IV group is not included in the training set.



(a) Cross-Genre (b) Cross-Device (c) Cross-Dataset (d) Cross-Genre (GCG I)

Fig. 4. Results comparison of episode-level sampling strategies on the cross-genre, cross-device, cross-dataset, and cross-genre (GCG I). The **cross-genre (GCG I)** represents the cross-genre obtained using the GCG I protocol. **S-D**: Single-Domain Sampling. **S-S**: Single-Single Sampling. **M-S**: Multiple-Single Sampling. **SSV**: Simulated Speaker Verification Sampling.

The results of the episode design are shown in Fig. 4. It can be seen that 1) the optimization of the model performance brought about by generalization improvement may be diluted owing to the reduction in discrimination ability; 2) improving the extraction efficiency of universal features can indeed alleviate the trade-off problem and boost the model performance; and 3) the proposed SSV is the most effective. More specifically, compared to the original SD, in SS, we design the domain mismatch between the meta-train and meta-test tasks to increase the generalization ability of the model. However, we find that after emphasizing generalization learning, the performance of SS is not consistently superior to that of SD, as expected. This is because the meta-test with simulated unseen domains motivates the model to filter out domain-specific information to refine domain-invariant knowledge, which leads to a decrease in the feature learning efficiency.

This phenomenon limits the discrimination ability of models with SS and further dilutes the performance optimization resulting from the improvement in the generalization ability. Consequently, we propose MS by introducing more available domains to increase the probability of learning universal knowledge in the meta-train stage. From the results, it can be seen that the optimized MS significantly outperforms SD and SS in terms of discrimination and generalization, which indicates that MS ensures discrimination learning and improves model performance after accumulating speaker-related knowledge more efficiently. Furthermore, on the basis of MS, we propose SSV sampling by simulating the application scenarios of the speaker verification systems. It is observed that SSV achieves a stable performance improvement compared to MS on the cross-genre, cross-device, cross-dataset, and cross-genre (GCG I) tasks. This is attributed to the supervision of SSV on the relationship between the simulated unseen domain and all the domains, which further optimizes the cross-domain embedding space.

VIII. Conclusions and Future Work

In this paper, we proposed the meta-generalized speaker verification (MGSV) method to reduce the adverse impact of compound domain mismatch. In particular, we constructed meta-train and meta-test tasks using episode-level sampling to simulate the domain mismatch between the training and testing sets mimicking real-world scenarios. To perform targeted optimization on the embedding space and unseen domains, we systematically designed metric-based distribution optimization and gradient-based meta-optimization. Additionally, to tackle the trade-off between model discrimination and domain generalization, we designed a sampling strategy called simulated speaker verification (SSV) and introduced an annealing mechanism to efficiently learn domain-invariant features. The experimental results obtained on multiple generalized cross-domain evaluation benchmarks indicate that our proposed method can substantially improve the discrimination and generalization abilities of cross-domain speaker verification. In future work, we plan to incorporate additional networks to adaptively generate separate embedding spaces for distinct tasks to tackle multivariable cross-domain challenges.

References

[1] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digit. Signal Process.*, vol. 10, no. 1-3, pp. 19–41, 2000.

[2] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1435–1447, 2007.

[3] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 788–798, 2010.

[4] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 4052–4056.

[5] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5329–5333.

[6] 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. 1–5.

[7] K. A. Lee, Q. Wang, and T. Koshinaka, "The coral+ algorithm for unsupervised domain adaptation of plda," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 5821–5825.

[8] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

[9] A. McCree, G. Sell, and D. Garcia-Romero, "Extended variability modeling and unsupervised adaptation for plda speaker recognition," *Proc. Interspeech*, pp. 1552–1556, 2017.

[10] L. Li, Y. Zhang, J. Kang, T. F. Zheng, and D. Wang, "Squeezing value of cross-domain labels: a decoupled scoring approach for speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 5829–5833.

[11] X. Fang, L. Zou, J. Li, L. Sun, and Z.-H. Ling, "Channel adversarial training for cross-channel text-independent speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6221–6225.

[12] C. Luu, P. Bell, and S. Renals, "Channel adversarial training for speaker verification and diarization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 7094–7098.

[13] J. Kang, R. Liu, L. Li, Y. Cai, D. Wang, and T. F. Zheng, "Domain-invariant speaker vector projection by model-agnostic meta-learning," *Proc. Interspeech*, pp. 3825–3829, 2020.

[14] G. Bhattacharya, J. Monteiro, J. Alam, and P. Kenny, "Generative adversarial speaker embedding networks for domain robust end-to-end speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6226–6230.

[15] A. Kanagasundaram, D. Dean, and S. Sridharan, "Improving out-domain plda speaker verification using unsupervised inter-dataset variability compensation approach," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 4654–4658.

[16] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1126–1135.

[17] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, "Robustness may be at odds with accuracy," in *Proc. Int. Conf. Learn. Represent.*, no. 2019, 2019.

[18] D. Su, H. Zhang, H. Chen, J. Yi, P.-Y. Chen, and Y. Gao, "Is robustness the cost of accuracy?–a comprehensive study on the robustness of 18 deep image classification models," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 631–648.

[19] H.-Y. Chen, J.-H. Liang, S.-C. Chang, J.-Y. Pan, Y.-T. Chen, W. Wei, and D.-C. Juan, "Improving adversarial robustness via guided complement entropy," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 4881–4889.

[20] H. Zhang, L. Wang, K. A. Lee, M. Liu, J. Dang, and H. Chen, "Meta-learning for cross-channel speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 5839–5843.

[21] J. Zhou, T. Jiang, Z. Li, L. Li, and Q. Hong, "Deep speaker embedding extraction with channel-wise feature responses and additive supervision softmax loss function." in *Proc. Interspeech*, 2019, pp. 2883–2887.

[22] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[23] S. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. H. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019.

[24] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," *Proc. Interspeech*, pp. 2252–2256, 2018.

[25] Y. Wu, C. Guo, H. Gao, X. Hou, and J. Xu, "Vector-based attentive pooling for text-independent speaker verification," *Proc. Interspeech*, pp. 936–940, 2020.

[26] S. M. Kye, Y. Kwon, and J. S. Chung, "Cross attentive pooling for speaker verification," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2021, pp. 294–300.

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

[28] M. Hajibabaei and D. Dai, "Unified hypersphere embedding for speaker recognition," *arXiv preprint arXiv:1807.08312*, 2018.

[29] X. Xiang, S. Wang, H. Huang, Y. Qian, and K. Yu, "Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2019, pp. 1652–1656.

[30] M. Huisman, J. N. van Rijn, and A. Plaat, "A survey of deep meta-learning," *Artif. Intell. Rev.*, pp. 1–59, 2021.

[31] H.-Y. Tseng, H.-Y. Lee, J.-B. Huang, and M.-H. Yang, "Cross-domain few-shot classification via learned feature-wise transformation," in *Proc. Int. Conf. Learn. Represent.*, 2020.

[32] K. Li and J. Malik, "Learning to optimize neural nets," *arXiv preprint arXiv:1703.00441*, 2017.

[33] J. Guo, X. Zhu, C. Zhao, D. Cao, Z. Lei, and S. Z. Li, "Learning meta face recognition in unseen domains," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6163–6172.

[34] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Learning to generalize: Meta-learning for domain generalization," in *Proc. AAAI Conf. Artif. Intell.*, 2018.

[35] A. Raghu, M. Raghu, S. Bengio, and O. Vinyals, "Rapid learning or feature reuse? towards understanding the effectiveness of maml," in *Proc. Int. Conf. Learn. Represent.*, 2019.

[36] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4080–4090.

[37] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, "Matching networks for one shot learning," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, pp. 3630–3638, 2016.

[38] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1199–1208.

[39] S. M. Kye, Y. Jung, H. B. Lee, S. J. Hwang, and H.-R. Kim, "Meta-learning for short utterance speaker recognition with imbalance length pairs," in *Proc. Interspeech*, 2020, pp. 2982–2986.

[40] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1842–1850.

[41] T. Munkhdalai and H. Yu, "Meta networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 2554–2563.

[42] N. Mishra, M. Rohaninejad, X. Chen, and P. Abbeel, "A simple neural attentive meta-learner," in *Proc. Int. Conf. Learn. Represent.*, 2018.

[43] Y. Li, X. Tian, M. Gong, Y. Liu, T. Liu, K. Zhang, and D. Tao, "Deep domain generalization via conditional invariant adversarial networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 624–639.

[44] Y. Balaji, S. Sankaranarayanan, and R. Chellappa, "Metareg: Towards domain generalization using meta-regularization," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, pp. 998–1008, 2018.

[45] S. Zhao, M. Gong, T. Liu, H. Fu, and D. Tao, "Domain generalization via entropy regularization," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020.

[46] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.

[47] G. Wang, H. Han, S. Shan, and X. Chen, "Cross-domain face presentation attack detection via multi-domain disentangled representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6678–6687.

[48] V. Piratla, P. Netrapalli, and S. Sarawagi, "Efficient domain generalization via common-specific low-rank decomposition," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 7728–7738.

[49] M. Ilse, J. M. Tomczak, C. Louizos, and M. Welling, "Diva: Domain invariant variational autoencoders," in *Proc. Int. Conf. Medical Imaging Deep Learn.*, 2020, pp. 322–348.

[50] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014.

[51] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *arXiv preprint arXiv:2103.02503*, 2021.

[52] K. Zhou, Y. Yang, T. Hospedales, and T. Xiang, "Learning to generate novel domains for domain generalization," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 561–578.

[53] M. Mancini, Z. Akata, E. Ricci, and B. Caputo, "Towards recognizing unseen categories in unseen domains," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 466–483.

[54] X. Qin, D. Cai, and M. Li, "Far-field end-to-end text-dependent speaker verification based on mixed training data with transfer learning and enrollment data augmentation." in *Proc. Interspeech*, 2019, pp. 4045–4049.

[55] S. Wang, Y. Yang, Z. Wu, Y. Qian, and K. Yu, "Data augmentation using deep generative models for embedding based speaker recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2598–2609, 2020.

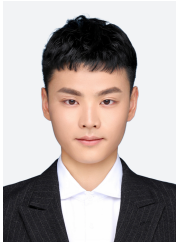
[56] L. Li, R. Liu, J. Kang, Y. Fan, H. Cui, Y. Cai, R. Vipperla, T. F. Zheng, and D. Wang, "Cn-celeb: multi-genre speaker recognition," *Speech Commun.*, 2022.

[57] X. Qin, H. Bu, and M. Li, "Hi-mia: A far-field text-dependent speaker verification database and the baselines," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 7609–7613.

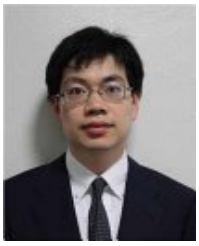
[58] X. Qin, M. Li, H. Bu, W. Rao, R. K. Das, S. Narayanan, and H. Li, "The interspeech 2020 far-field speaker verification challenge," *Proc. Interspeech*, pp. 3456–3460, 2020.

[59] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *Proc. Interspeech*, pp. 1086–1090, 2018.

[60] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.



Hanyi Zhang received the B.Eng. degree from Yunnan University, Yunnan, China, in 2020 and is currently working toward the M.Eng. degree in Tianjin University, Tianjin, China, under the supervision of Longbiao Wang. His research interests include speaker verification and meta-learning.



Longbiao Wang received his Dr. Eng. degree from Toyohashi University of Technology, Japan, in 2008. He was an assistant professor in the faculty of engineering at Shizuoka University, Japan from April 2008 to September 2012. From October 2012 to August 2016 he was an associate professor at Nagaoka University of Technology, Japan. Currently he is a Professor at the Tianjin University, China. His research interests include robust speech recognition and speaker recognition.

Kong Aik Lee (Senior Member, IEEE) received the Ph.D. degree from Nanyang Technological University, Singapore, in 2006. He is currently a Senior Scientist with Institute for Infocomm Research, A*STAR, Singapore. He was a recipient of the Singapore IES Prestigious Engineering Achievement Award 2013, for his contribution to voice biometrics technology, and the Outstanding Service Award by IEEE ICME 2020. He has been serving as an Editorial Board Member for Computer Speech and Language (Elsevier) since 2016, and an Associate Editor for IEEE/ACM Transactions on Audio, Speech and Language Processing since 2017. He is an Elected Member of IEEE Speech and Language Technical Committee.

Meng Liu received his B. Eng. degree from Changchun University of Science and Technology, China, in 2017. He received the M. Sc. degree in Tianjin University, China, in 2019. Currentyly he is working towards to the Dr. Eng. degree in Tianjin University under the supervision of Prof. Longbiao Wang and Dr. Kong Aik Lee. His research interests include speaker verification, multi-modal biometrics and anti-spoofing.

Jianwu Dang graduated from Tsinghua Univ. , China, in 1982, and got his M.S. at the same university in 1984. He worked for Tianjin Univ. as a lecture from 1984 to 1988. He was awarded the PhD from Shizuoka Univ. , Japan in 1992. He worked for ATR Human Information Processing Labs., Japan, as a senior researcher from 1992 to 2001. He joined the University of Waterloo, Canada, as a visiting scholar for one year from 1998. Since 2001, he has moved to Japan Advanced Institute of Science and Technology (JAIST). He joined the Institute of Communication Parlee (ICP), Center of National Research Scientific, France, as a research scientist the first class from 2002 to 2003. His research interests are in all the fields of speech production, speech synthesis, and speech cognition. He built a 3D physiological model for speech and swallowing, and endeavors to apply the model on clinics.

Helen Meng (Fellow, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge, MA, USA. In 1998, she joined the Chinese University of Hong Kong, Hong Kong, where she is currently the Chair Professor with the Department of Systems Engineering & Engineering Management. Her research interests include human–computer interaction via multimodal and multilingual spoken language systems, spoken dialog systems, speech processing. She was the Editor-in-Chief of the IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING between 2009 and 2011. She is a Fellow of ISCA, HKCS, and HKIE.