LEARNING DOMAIN-INVARIANT TRANSFORMATION FOR SPEAKER VERIFICATION

Hanyi Zhang[1], *Longbiao Wang*[1,*], *Kong Aik Lee*[2,*], *Meng Liu*[1], *Jianwu Dang*[1,3], *Hui Chen*[1]

[1]Tianjin Key Laboratory of Cognitive Computing and Application,
College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]Institute for Infocomm Research, A*STAR, Singapore
[3]Japan Advanced Institute of Science and Technology, Ishikawa, Japan
{hanyizhang, longbiao_wang}@tju.edu.cn, lee_kong_aik@i2r.a-star.edu.sg

ABSTRACT

Automatic speaker verification (ASV) faces domain shift caused by the mismatch of intrinsic and extrinsic factors such as recording device and speaking style in real-world applications, which leads to unsatisfactory performance. To this end, we propose the meta generalized transformation via meta-learning to build a domain-invariant embedding space. Specifically, the transformation module is motivated to learn the domain generalization knowledge by executing meta-optimization on the meta-train and meta-test sets which are designed to simulate domain shift. Furthermore, distribution optimization is incorporated to supervise the metric structure of embeddings. In terms of the transformation module, we investigate various instantiations and observe the multilayer perceptron with gating (gMLP) is the most effective given its extrapolation capability. The experimental results on cross-genre and cross-dataset settings demonstrate that the meta generalized transformation dramatically improves the robustness of ASV systems to domain shift, while outperforms the state-of-the-art methods.

Index Terms— speaker verification, meta-learning, domain-invariant, meta generalized transformation

1. INTRODUCTION

Automatic speaker verification (ASV) aims to tackle the challenge of determining whether a speaker's voice matches the claimed identity. Compared to traditional methods [1, 2], most of the current works [3, 4] rely on deep neural network (DNN) based embeddings, which push the performance boundary of ASV to a higher level. However, in real-world applications, elements such as speaking style and recording device are complex and uncontrollable. The variations of speaker-independent statistical information in speech signals cause domain shifts and deteriorate the performance of ASV.

The approaches to address domain shift can be divided into two categories: domain adaptation and domain generalization. Domain adaptation [5, 6] aligns the source and target domains by applying feature-level methods such as generative adversarial nets (GAN) [7]. While domain generalization [8, 9, 10] learns a model which is robust to domain shift without the need of adaptation to the target domains. Li et al. [8] simulated the train/test domain shift during training and realized the meta-learning [11] based generalization method.

In ASV, domain shift is embodied as the mismatch of speaker-independent information. Probabilistic linear discriminant analysis (PLDA) and its derivative methods [12, 13] decouple the speaker and channel subspaces to reduce the interference of mismatch. Besides, the robustness of models is optimized through adversarial training [14, 15] and meta-learning [16, 17, 18]. Kang et al. [16] utilized the projection network based on meta-learning to fine-tune the embeddings of pre-trained model.

Nevertheless, these approaches still have some limitations. Firstly, the complexity of feature space caused by domain shift dramatically increases the difficulty of directly training the neural networks with domain generalization ability. Secondly, traditional methods tend to learn locally accurate embeddings for seen domains and lack the capability to adapt on unseen domains.

In this paper, we investigate the meta generalized transformation via meta-learning on virtual episodes to generate domain-invariant embeddings without pre-training and fine-tuning. Specifically, meta-train and meta-test sets are first constructed by episode-level sampling to simulate domain shift. Then, the backbone with representation ability is learned through meta-train. Next, in the meta-test phase, the domain generalization effect of the model on distinct ASV tasks is fed back to the transformation module. By doing so, the transformation module is trained to be robust to both seen and unseen domain shifts. Furthermore, metric and classification losses are fused to optimize cross-domain speaker verification by supervising the spatial distribution of embeddings. For the architecture of transformation module, we implement the multilayer perceptron with gating (gMLP) [19] and the multilayer convolution. Finally, the transformation module instantiated with gMLP is chosen.

2. BASELINE ARCHITECTURE

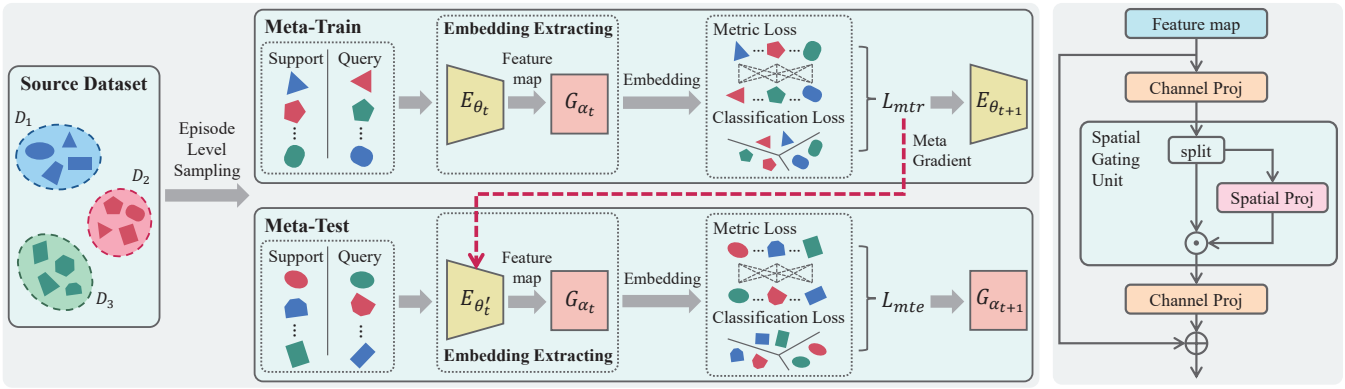
We use the r-vector [20, 21], which is an x-vector [3] variant implemented with ResNet [22], as our baseline. First, we intercept consecutive frames from each utterance and extract log Mel-filterbank features. Then, we adopt ResNet-18 [22] as the backbone. Next, global average pooling is placed to encode feature maps to utterance-level embeddings. Afterwards, the fully connected layer performs further processing. The softmax classifier is used to compute loss for training. And the verification results are calculated by the cosine similarity between the embeddings of enrolled and tested utterances.

3. META GENERALIZED TRANSFORMATION

3.1. Overview

In automatic speaker verification (ASV), the essence of domain mismatch is that data condition varies substantially from one ASV task to the next (e.g., wideboard interview speech, narrowboard

*Corresponding author



(a) The proposed meta generalized transformation. (b) The gMLP block.

Fig. 1. The architecture of meta generalized transformation and gMLP block. Three domains with different colors are presented in this figure. The same symbol indicates the utterances from the same speaker.

telephony speech). Inspired by this, we propose the meta generalized transformation via meta-learning to provide a domain-invariant embedding space. Different from the previous model-agnostic meta-learning (MAML) based methods [11, 16], we deliver the domain generalization functionality from the backbone to a transformation module, thereby reducing the difficulty of training the deep backbone. Simultaneously, we adopt the architecture with extrapolation ability to enhance the generalization of transformation module.

The architecture of our proposed meta generalized transformation is illustrated in Fig. 1(a). We first perform the episode-level sampling by randomly selecting utterances from the source dataset $\mathcal{D} = \{\mathcal{D}_1, ..., \mathcal{D}_\mathcal{B} | \mathcal{B} > 1\}$ which contains $\mathcal{B}$ seen domains to construct the task set $\mathcal{T} = \{\mathcal{T}_1, ..., \mathcal{T}_\mathcal{M} | \mathcal{M} > 1\}$. A task consists of a support set and a query set. Then, we use the backbone E_{θ_t} to extract feature maps for the meta-train set which is randomly sampled from the task set. The backbone E_{θ_t} corresponds to the deep network of baseline, which uses the structures such as ResNet-18. Next, we adopt the transformation module G_{α_t} to output utterance-level embeddings. We consider the metric structure of embeddings and the global classification situation to calculate the meta-train loss $\mathcal{L}_{mtr}$. Afterwards, the meta-gradient is obtained via meta-train loss to compute the updated parameters θ'_t of the backbone. In the meta-test stage, we sample a meta-test set which is different from the meta-train set. The backbone with updated parameters $E_{\theta_t'}$, transformation module G_{α_t}, and distribution optimization are used in turn to calculate the meta-test loss $\mathcal{L}_{mte}$. The purpose is to supervise the capability of transformation module to process different ASV tasks. Finally, we obtain the optimized backbone $E_{\theta_{t+1}}$ and transformation module $G_{\alpha_{t+1}}$ by meta-train loss and meta-test loss, respectively.

3.2. Episode-Level Sampling

We introduce episode-level sampling to simulate training and testing in real-world scenarios. By doing so, the model is encouraged to learn transferable knowledge and achieve domain generalization.

Specifically, we first randomly select $\mathcal{S}$ identities from all speakers in the training set. Then, from the known domains, each identity randomly selects $\mathcal{P}$ utterances as support and $\mathcal{Q}$ utterances as query. A task containing a support set and a query set is created. Next, we randomly allocate distinct tasks from the task set to the meta-train and meta-test sets to simulate the real-world episodes.

3.3. Distribution Optimization

In both meta-train and meta-test stages, we adopt metric loss and classification loss to optimize cross-domain embedding distribution.

Metric Loss. The mismatch of speaker-independent information may affect the distance between embeddings, and thus interfere with the verification results. To this end, we adopt the metric loss to map the utterances of the same speaker into nearby embeddings and map the embeddings of different speakers apart from each other. In this work, we introduce the metric based on cosine similarity. We first obtain the representations $\boldsymbol{R}_s$ of selected speakers $s = 1, ..., \mathcal{S}$ in the support set by computing the average utterance embeddings.

$$\boldsymbol{R}_s = \frac{1}{\mathcal{P}} \sum_{p=1}^{\mathcal{P}} G_\alpha(E_\theta(\boldsymbol{U}_{s,p}^{\text{support}})) \tag{1}$$

where $G_\alpha(E_\theta(\boldsymbol{U}_{s,p}^{\text{support}}))$ is the embedding of p-th utterance of s-th speaker in support set. Then, we normalize the embeddings of query utterances and the speaker representations in the task. Through this operation, the calculation of the similarity set $\boldsymbol{d}_q = [d_{q,1}, ..., d_{q,S}]$ between the embedding of q-th query utterance and the whole speaker representations $\boldsymbol{R} = [\boldsymbol{R}_1, \boldsymbol{R}_2, ..., \boldsymbol{R}_{S-1}, \boldsymbol{R}_S]$ can be simplified into the following form.

$$\boldsymbol{d}_q = G_\alpha(E_\theta(\boldsymbol{U}_q^{\text{query}})) \cdot \boldsymbol{R}^\top \tag{2}$$

where $G_\alpha(E_\theta(\boldsymbol{U}_q^{\text{query}}))$ is the embedding of q-th query utterance in the task. The metric loss $\mathcal{L}_{metric}$ is defined using cross-entropy:

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

where $\boldsymbol{s}_q$ is the one-hot form of the selected speaker corresponding to the q-th utterance, and $\mathcal{N}_{\text{query}} = \mathcal{S} \times \mathcal{Q}$ is the total number of query utterances in the task.

Classification Loss. We utilize all speaker IDs $y = 1, ..., \mathcal{Y}$ as labels to calculate the classification loss $\mathcal{L}_{cls}$ of all utterances in the task, where y_i is the speaker label of the i-th utterance, F_β is the linear layer followed by softmax for classification, and $\mathcal{N}_{\text{task}} = \mathcal{S} \times \mathcal{P} + \mathcal{S} \times \mathcal{Q}$ is the number of all utterances in the task.

$$\mathcal{L}_{cls} = -\frac{1}{\mathcal{N}_{\text{task}}} \sum_{i=1}^{\mathcal{N}_{\text{task}}} y_i \cdot \log(F_\beta(G_\alpha(E_\theta(\boldsymbol{U}_i)))) \tag{4}$$

Table 1. EERs (%) of MGT and state-of-the-art methods on the Cross-Genre and Cross-Dataset trials. MGT (w/o MO) is MGT without meta-optimization (MO). MGT-Conv is MGT with multilayer convolution. MGT-gMLP is MGT with gMLP.

Method	Cross-Genre	Cross-Dataset
r-vector	23.80	11.47
x-vector	22.57	10.93
MAML	26.10	13.40
MGT (w/o MO)	20.52	7.95
MGT-Conv	19.23	7.38
MGT-gMLP	**18.56**	**6.83**

3.4. Meta-Optimization

To improve the generalization, we motivate the model to build a domain-invariant embedding space via meta-train and meta-test.

Meta-Train. Over-reliance on the deep backbone to achieve rigorous domain generalization may increase the difficulty of training, and affect the representation ability of embeddings. Thus, we incorporate the generalization task to transformation module and focus on backbone in meta-train. Concretely, we still use transformation module G_{α_t} to generate embeddings for feature maps which are extracted by backbone E_{θ_t}, and calculate meta-train loss $\mathcal{L}_{mtr}$ by distribution optimization. However, we only compute the meta-gradient of backbone $\nabla_{\theta_t}\mathcal{L}_{mtr}(\theta_t, \alpha_t)$ and obtain its updated parameters θ'_t.

$$\mathcal{L}_{mtr} = \mathcal{L}_{metric}(\mathcal{X}_{mtr}; \theta_t, \alpha_t) + \mathcal{L}_{cls}(\mathcal{X}_{mtr}; \theta_t, \alpha_t) \quad (5)$$

$$\theta'_t = \theta_t - \lambda\nabla_{\theta_t}\mathcal{L}_{mtr}(\theta_t, \alpha_t) \quad (6)$$

where $\mathcal{X}_{mtr}$ is the utterances in meta-train, θ_t is the backbone parameters, α_t is the transformation module parameters, λ is learning rate. By doing so, the model performance on meta-train task is fully reflected to backbone, so that it can learn deep features more purely.

Meta-Test. By episode-level sampling, we simulate the distinct ASV tasks in real-world applications to motivate the model to learn transferable knowledge. During the meta-test phase, we mainly focus on the transformation module which is adopted to improve generalization. First, the backbone $E_{\theta_t'}$ with updated parameters θ'_t and the transformation module G_{α_t} are used to extract the embeddings of meta-test set. Then, we evaluate the robustness of the model with updated backbone on distinct ASV tasks by meta-test loss $\mathcal{L}_{mte}$.

$$\mathcal{L}_{mte} = \mathcal{L}_{metric}(\mathcal{X}_{mte}; \theta'_t, \alpha_t) + \mathcal{L}_{cls}(\mathcal{X}_{mte}; \theta'_t, \alpha_t) \quad (7)$$

where $\mathcal{X}_{mte}$ denotes the meta-test utterances. Afterwards, we truncate the gradient backpropagation at the tail of backbone to completely feed back the generalization results to the transformation module. Hence, we train a transformation module that performs well on different ASV tasks via learning to learn.

Summary. Finally, we learn the optimized parameters θ_{t+1} of backbone $E_{\theta_{t+1}}$ through meta-train loss $\mathcal{L}_{mtr}$, while the meta-test loss $\mathcal{L}_{mte}$ is adopted to obtain the optimized parameters α_{t+1} of transformation module $G_{\alpha_{t+1}}$, as follows:

$$\alpha_{t+1} = \alpha_t - \mu\nabla_{\alpha_t}\mathcal{L}_{mte}(\theta'_t, \alpha_t) \quad (8)$$

where μ is the learning rate. For the next iteration, θ_{t+1} is optimized to the value of θ'_t.

3.5. Transformation Module

We implement the architecture of multilayer perceptron with gating (gMLP) [19] and multilayer convolution for transformation module.

Table 2. The grouping results of ten genres in CN-Celeb.

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

Table 3. Generalized cross-genre evaluation (GCG) benchmark. The seen domains are for both training and testing, the unseen domain is only for testing and is unknowable during training.

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

Multilayer Convolution. The multilayer convolution based transformation module uses the residual structure and three two-dimensional convolution layers. It takes the output of the backbone directly as input and combines pooling layer to extract embeddings.

gMLP. The architecture of gMLP block is shown in Fig. 1(b). This block consists of channel projections and spatial gating units, which captures the hidden knowledge from the perspective of spatial interactions. In our work, we first average the temporal dimension of the extracted feature maps. Then, the residual module with three gMLP blocks and the pooling module are adopted to transform the raw feature maps into a domain-invariant embedding space.

4. EXPERIMENTS AND ANALYSIS

4.1. Experimental Setup

Our proposed meta generalized transformation (MGT) is evaluated on cross-genre and cross-dataset settings.

Cross-Genre. We use CN-Celeb [23] to test the performance of models on cross-genre. This corpus contains 11 genres including drama, singing, and so on. We discard the speakers with fewer than 5 utterances and the "advertisement" genre with fewer than 100 speakers in the training set. Finally, we adopt 2,768 speakers for training and 200 speakers for testing. In addition, we obtain a more difficult trial by constructing cross-genre verification pairs.

Cross-Dataset. We use CN-Celeb [23], HI-MIA [24], FFSVC [25], and VoxCeleb [26] to achieve cross-dataset evaluation. HI-MIA consists of different recording devices and locations. FFSVC has time-varying and noise, we only use the data marked "I0.25M". VoxCeleb is the clean dataset that is widely applied. Through random sampling, we construct the train set with 1,307 speakers and the test set with 175 speakers. Similarly, we adopt a dedicated trial with cross-dataset verification pairs for testing.

Implementation Details. For MGT, the backbone uses ResNet-18, which is consistent with the r-vector (baseline). Each task randomly selects 80 speakers. Each speaker randomly selects one utterance as support and two utterances as query. The code and trials will be available at https://github.com/MiukkaZh/MGT.

4.2. Comparative Experiments

To measure the performance of MGT, we introduce r-vector (baseline) [20, 21], x-vector [3], and MAML [11] for comparison.

Table 4. EERs (%) of MGT and state-of-the-art methods on the GCG benchmark. The unseen domain of each protocol is marked with *.

Protocol	Method	Cross-Genre	Genre I			Genre II			Genre III		Genre IV	
			lb	vl	sp	en	in	pl	dr	mo	si	re
GCG I (* Genre IV)	r-vector	24.26	21.49	21.24	18.38	24.13	25.21	27.38	26.07	33.42	28.61	14.91
	x-vector	22.36	17.97	20.81	15.61	22.69	23.45	28.15	23.13	30.95	27.50	**11.98**
	MGT-gMLP	**19.16**	**16.89**	**17.30**	**14.57**	**19.38**	**19.71**	**20.23**	**20.39**	**24.71**	**23.55**	12.75
GCG II (* Genre III)	r-vector	24.07	21.91	22.19	17.62	24.03	25.30	25.85	26.20	30.87	27.16	14.18
	x-vector	22.73	18.98	21.02	16.53	23.19	23.81	27.38	23.72	30.70	26.05	**11.94**
	MGT-gMLP	**19.25**	**18.58**	**17.66**	**14.31**	**19.84**	**19.65**	**20.23**	**20.24**	**25.73**	**21.93**	13.15
GCG III (* Genre II)	r-vector	26.68	22.56	23.31	20.10	27.06	28.92	26.69	27.66	34.51	29.06	15.86
	x-vector	25.03	21.57	22.63	**17.71**	25.77	26.69	30.46	24.86	30.90	27.49	**12.34**
	MGT-gMLP	**22.84**	**21.24**	**20.06**	17.87	**23.50**	**24.30**	**24.00**	**23.53**	**28.00**	**24.42**	14.14
GCG IV (* Genre I)	r-vector	23.95	22.10	22.31	17.11	23.95	25.43	28.31	24.99	31.56	26.02	14.84
	x-vector	22.68	19.11	21.74	16.18	22.93	23.82	28.15	23.75	29.59	25.92	**13.22**
	MGT-gMLP	**19.42**	**17.67**	**18.01**	**14.65**	**19.93**	**20.05**	**21.46**	**20.98**	**24.73**	**22.47**	13.99

Table 1 shows the comparative results on cross-genre and cross-dataset. It is observed that the performance of our proposed MGT-gMLP is superior to other state-of-the-art algorithms significantly. In terms of cross-genre, MGT-gMLP achieves 22.02% and 17.77% relative reduction in the equal error rate (EER) compared to r-vector (baseline) and x-vector, respectively. For cross-dataset, we adopt multiple datasets which involve different types of domain shifts to more relevantly simulate the complex real-world scenarios. The results indicate that the EER of MGT-gMLP is relatively reduced by 40.45% compared to that of r-vector in the cross-dataset trial where the enrolled and tested utterances come from the random datasets. This is because the meta-learning performed on the virtual episodes and the transformation module with extrapolation capability promote MGT-gMLP to capture the transferable speaker knowledge from utterances, thereby improving the model robustness to distinct ASV tasks. Additionally, the metric structure of embedding space is also supervised by distribution optimization. Furthermore, the comparative results between MGT-gMLP and MGT-Conv validate that the gMLP based architecture is more effective.

As the classic meta-learning method, MAML [11] is also adopted for comparison. It can be seen that the EER of MAML is obviously higher than that of MGT-gMLP and r-vector. The main reason is that the ASV backbone itself is difficult to train due to its deep structure, while it is also required to learn generalization in the domains with distinct speaker-independent statistical information. Overly complex tasks cause instability in the training, which reduces the representation ability of the embeddings. Thus, in our work, we specially design the transformation module to take over the generalization task from the backbone so as to ensure the representation ability and robustness of the embedding space.

To quantitatively evaluate the effect of meta-optimization (MO), we carry out an ablation study. For MGT (w/o MO), we only reserve the distribution optimization and use the three linear layers without MO instead of the transformation module. It can be seen that MGT-gMLP achieves 11.82% relative reduction in EER compared to the MGT (w/o MO). This result indicates that meta-optimization further transforms the features into domain-invariant embeddings.

4.3. Generalized Cross-Genre Evaluation

Due to the complexity and variation of the acoustic environment in real-world applications, ASV is desired to process utterances from distinct and unseen domains. Hence, to intuitively reflect the performance and robustness of the model on the seen and unseen domains, we introduce the generalized cross-genre evaluation (GCG).

We first divide the ten genres of CN-Celeb into four groups according to the degree of association, as shown in Table 2. Then, we take the genre groups as units to design GCG benchmark. Table 3 shows the detailed settings. We split the whole genre groups into the seen and unseen domains to facilitate evaluation. In each protocol, we choose three out of the four groups as the seen domains which can be used during training. The remaining one group as the unseen domain is only for testing. Additionally, in the Cross-Genre trial, all pairs containing a certain genre are selected as the trial of this genre.

Table 4 shows the results on the GCG benchmark. It is observed that MGT-gMLP significantly outperforms other algorithms. More concretely, for the seen and unseen domains, MGT-gMLP obtains 17.64% and 16.45% relative reduction in terms of EER on average compared to the r-vector, respectively. This benefits from the capture of domain-invariant speaker knowledge via meta-optimization and the powerful extrapolation capability of gMLP. Compared to x-vector, MGT-gMLP has better performance in the majority of genres, but is inferior to this method in some genres. The reason may be that MGT-gMLP and x-vector use different deep backbone networks. Since our algorithm is model-agnostic, MGT can be extended to other backbones in future work to further improve the verification effect. On the whole, the comparative results indicate that our proposed MGT can dramatically improve the performance and generalization of ASV on both seen and unseen domains.

5. CONCLUSIONS AND FUTURE WORK

In this study, to alleviate the negative impact of domain shift on the automatic speaker verification (ASV), we proposed the meta generalized transformation via meta-learning. Specifically, the transformation module learned to generate domain-invariant embeddings through a series of simulated ASV tasks. We also fused metric loss and classification loss to optimize the distribution of embeddings. Furthermore, the multilayer perceptron with gating (gMLP) was implemented as the ideal architecture for transformation module. The results indicated that our proposed meta generalized transformation could improve the performance of cross-domain speaker verification and was superior to other methods. In future work, we will extend the meta generalized transformation to other backbone networks and investigate virtual episode construction strategies.

6. ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (Grant No. 62176182).

7. REFERENCES

[1] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[2] Patrick Kenny, Pierre Ouellet, Najim Dehak, Vishwa Gupta, and Pierre Dumouchel, "A study of interspeaker variability in speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 980–988, 2008.

[3] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[4] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *Interspeech2020*, 2020, pp. 1–5.

[5] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[6] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in *International conference on machine learning*. PMLR, 2018, pp. 1989–1998.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[8] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales, "Learning to generalize: Meta-learning for domain generalization," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[9] Hung-Yu Tseng, Hsin-Ying Lee, Jia-Bin Huang, and Ming-Hsuan Yang, "Cross-domain few-shot classification via learned feature-wise transformation," in *International Conference on Learning Representations*, 2020.

[10] Jianzhu Guo, Xiangyu Zhu, Chenxu Zhao, Dong Cao, Zhen Lei, and Stan Z Li, "Learning meta face recognition in unseen domains," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6163–6172.

[11] Chelsea Finn, Pieter Abbeel, and Sergey Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1126–1135.

[12] Simon JD Prince and James H Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.

[13] Lantian Li, Yang Zhang, Jiawen Kang, Thomas Fang Zheng, and Dong Wang, "Squeezing value of cross-domain labels: a decoupled scoring approach for speaker verification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5829–5833.

[14] Qing Wang, Wei Rao, Sining Sun, Leib Xie, Eng Siong Chng, and Haizhou Li, "Unsupervised domain adaptation via domain adversarial training for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4889–4893.

[15] Xin Fang, Liang Zou, Jin Li, Lei Sun, and Zhen-Hua Ling, "Channel adversarial training for cross-channel text-independent speaker recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6221–6225.

[16] Jiawen Kang, Ruiqi Liu, Lantian Li, Yunqi Cai, Dong Wang, and Thomas Fang Zheng, "Domain-invariant speaker vector projection by model-agnostic meta-learning," *arXiv preprint arXiv:2005.11900*, 2020.

[17] Hanyi Zhang, Longbiao Wang, Kong Aik Lee, Meng Liu, Jianwu Dang, and Hui Chen, "Meta-learning for cross-channel speaker verification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5839–5843.

[18] Seong Min Kye, Youngmoon Jung, Hae Beom Lee, Sung Ju Hwang, and Hoi-Rin Kim, "Meta-learning for short utterance speaker recognition with imbalance length pairs," in *Interspeech 2020*. ISCA, 2020, pp. 2982–2986.

[19] Hanxiao Liu, Zihang Dai, David R So, and Quoc V Le, "Pay attention to mlps," *arXiv preprint arXiv:2105.08050*, 2021.

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

[21] Shuai Wang, Yexin Yang, Zhanghao Wu, Yanmin Qian, and Kai Yu, "Data augmentation using deep generative models for embedding based speaker recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2598–2609, 2020.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[23] Lantian Li, Ruiqi Liu, Jiawen Kang, Yue Fan, Hao Cui, Yunqi Cai, Ravichander Vipperla, Thomas Fang Zheng, and Dong Wang, "Cn-celeb: multi-genre speaker recognition," *arXiv preprint arXiv:2012.12468*, 2020.

[24] Xiaoyi Qin, Hui Bu, and Ming Li, "Hi-mia: A far-field text-dependent speaker verification database and the baselines," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7609–7613.

[25] Xiaoyi Qin, Ming Li, Hui Bu, Wei Rao, Rohan Kumar Das, Shrikanth Narayanan, and Haizhou Li, "The interspeech 2020 far-field speaker verification challenge," *arXiv preprint arXiv:2005.08046*, 2020.

[26] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.