CDMA: CROSS-DOMAIN DISTANCE METRIC ADAPTATION FOR SPEAKER VERIFICATION

Jianchen Li, Jiqing Han, Hongwei Song

School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

ABSTRACT

To solve the domain shift problem in speaker verification, one effective domain adaptation approach is to learn domain-invariant embeddings via aligning the source and target distributions in the embedding space. However, this approach could be problematic when the source and target domains are from the disjoint speaker label spaces as the embedding distributions of different speakers cannot be aligned. In this paper, we propose a Cross-domain Distance Metric Adaptation (CDMA) approach to alleviate the domain shift in the distance metric space, where the source and target domains share the same classes, i.e., within- and between-speaker. Specifically, the two target pairwise distance distributions are aligned with the source pairwise distance distributions and further separated to learn a domain-invariant metric, which is more suitable for speaker verification based on metric learning. Experiments indicate that CDMA significantly outperforms the approach proposed in the embedding space.

***Index Terms*—** Speaker verification, open-set domain adaptation, pairwise distance distributions

1. INTRODUCTION

Automatic Speaker Verification (ASV) is a task that determines whether a pair of utterances belong to the same speaker [1]. Recently, the deep speaker embedding models based on metric learning [2, 3] have become the state-of-the-art ASV systems due to the availability of large-scale annotated datasets [4]. However, the models will suffer from significant performance degradation when tested in a new target domain [5]. This is caused by the domain shift between the source training and the target test set. A typical solution is to collect and annotate enough training data for the target domain, but it is labor-intensive and expensive.

In the absence of sufficient data from the target domain, domain adaptation (DA) has emerged to alleviate the domain shift by transferring the knowledge from the resource-rich labeled source domain to the target domain [6]. For ASV models, traditional domain adaptation approaches are usually accomplished by adapting the PLDA model with the target domain data [7]. Recently, deep domain adaptation approaches have been proposed in the *input* and *embedding* spaces. For the input space, the CycleGAN- [8] and autoencoder-based [9] approaches learn a mapping between the source and target domains, and the transformed features are used to train ASV models. However, the performance of this approach completely depends on the quality of feature generation. For the embedding space, models are adapted by domain adversarial training [10, 11] and maximum mean discrepancy (MMD) minimization [12, 13], which align the cross-domain embedding distributions to learn domain-invariant embeddings. This approach has achieved the best performance and become the mainstream of deep domain adaptation in ASV [14].

However, directly applying the embedding space approach would be problematic for open-set domain adaptation in real-world ASV, where speakers differ in the source and target domains. The misalignment of the source and target embedding distributions not only comes from the domain shift but also the disjoint speaker label spaces of the source and target domains. Thus, directly aligning the embedding distributions of the different speakers from the source and target domains will impair the speaker discrimination ability. This open-set problem is ignored by previous works.

Given that learning a domain-invariant *metric* is the ultimate goal of metric learning-based ASV, we propose to alleviate the domain shift in the *distance metric* space instead of the embedding space. In the distance metric space, the source and target domains share the same classes, i.e., within- and between-speaker, so that the misalignment of the pairwise distance distributions only comes from the domain shift. This avoids the problem of the embedding space approach for open-set domain adaptation. Inspired by [15], the proposed CDMA optimizes the MMD loss to align the pairwise distance distributions in the target domain with the well-separated distance distributions in the source domain. Different from [15], CDMA also optimizes the MMD loss to further separate the within- and between-speaker distance distributions in the target domain. With the alignment and separation of the pairwise distance distributions, the overlap of the two pairwise distance distributions in the target domain can be largely decreased and the learned metric can effectively work on the target domain.

To the best of our knowledge, we are the first to highlight the open-set domain adaptation in ASV. We evaluate our proposed CDMA approach on a far-field dataset Voices Obscured

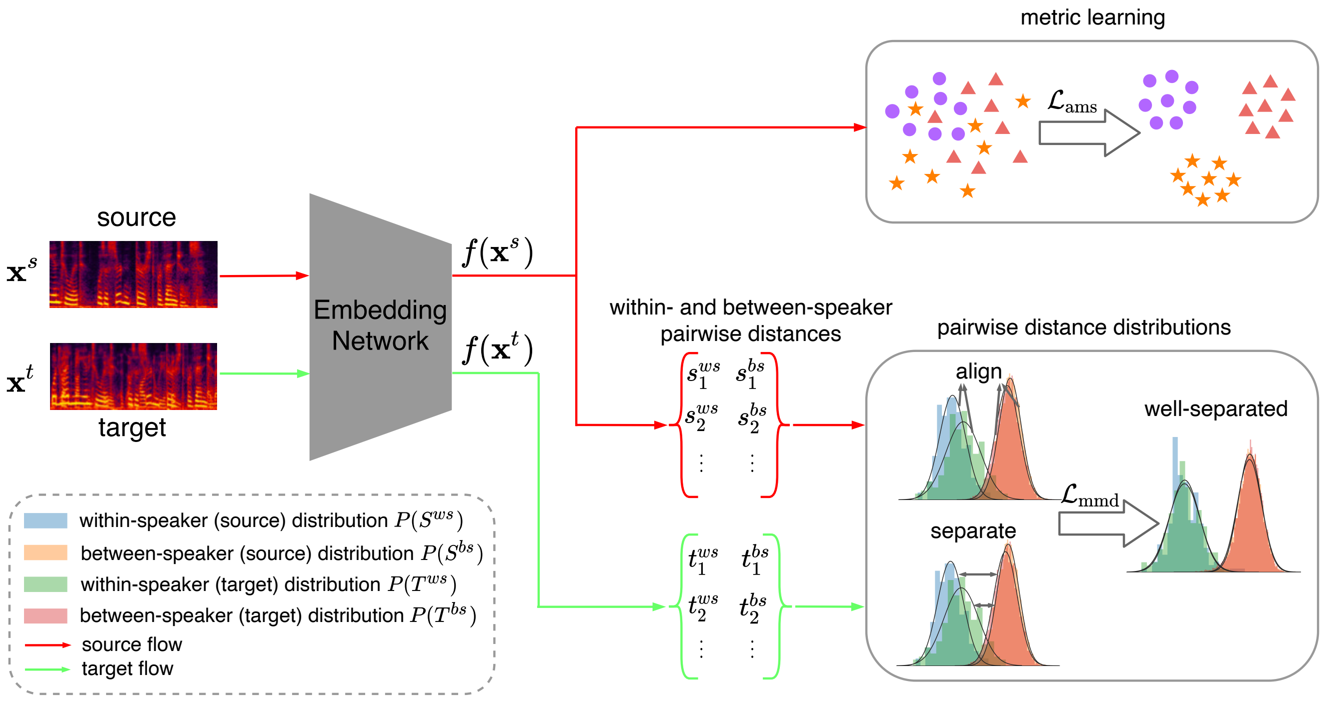

metric learning
$\mathcal{L}_{\text{ams}}$
source
$\mathbf{x}^s$
$\mathbf{x}^t$
target
Embedding Network
$f(\mathbf{x}^s)$
$f(\mathbf{x}^t)$
within- and between-speaker pairwise distances
s_1^{ws} s_1^{bs}
s_2^{ws} s_2^{bs}
t_1^{ws} t_1^{bs}
t_2^{ws} t_2^{bs}
pairwise distance distributions
align
well-separated
$\mathcal{L}_{\text{mmd}}$
separate
within-speaker (source) distribution $P(S^{ws})$
between-speaker (source) distribution $P(S^{bs})$
within-speaker (target) distribution $P(T^{ws})$
between-speaker (target) distribution $P(T^{bs})$
source flow
target flow

Fig. 1: Overview of the proposed CDMA approach. s^{ws} and s^{bs} are within- and between-speaker pairwise distances in the source domain, respectively. t^{ws} and t^{bs} are within- and between-speaker pairwise distances in the target domain, respectively.

in Complex Environmental Settings (VOiCES) [16]. Experimental results indicate that learning a domain-invariant metric is more suitable than learning domain-invariant embeddings for open-set domain adaptation in ASV. Our code is available at `https://github.com/matln/CDMA`.

2. METHODOLOGY

As shown in Fig. 1, the proposed CDMA approach involves two stages. First, the metric learning loss is used to pre-train the embedding network $f(\cdot;\theta)$ to obtain well-separated pairwise distance distributions in the labeled source domain $\{\mathbf{X}^s, \mathbf{Y}^s\}$. Then, the within- and between-speaker distance distributions from the target domain $\mathbf{X}^t$ are aligned with the well-separated distance distributions in the source domain, and further separated to reduce their overlap.

Let $P\left(S^{ws}\right)$ and $P(S^{bs})$ denote the within- and between-speaker pairwise distance distributions in the source domain, respectively. Let $P\left(T^{ws}\right)$ and $P(T^{bs})$ denote the within- and between-speaker pairwise distance distributions in the target domain, respectively. CDMA minimizes the discrepancy between $P\left(S^{ws}\right)$ and $P\left(T^{ws}\right)$ and the discrepancy between $P(S^{bs})$ and $P(T^{bs})$ to align with the source distance distributions. It also maximizes the discrepancy between $P\left(S^{ws}\right)$ and $P(T^{bs})$ and the discrepancy between $P(S^{bs})$ and $P(T^{ws})$ to separate the two pairwise distance distributions in the target domain.

2.1. Source-domain Pre-training

The metric learning loss is optimized on the source domain to minimize the distance between the within-speaker embedding pairs and maximize the distance between the between-speaker embedding pairs. Specifically, AM-softmax loss [17] is adopted in this paper,

$$\mathcal{L}_{\text{ams}} = -\frac{1}{n_s}\sum_{i=1}^{n_s}\log\frac{e^{d\cdot\left(\cos\delta_{y_i^s,i}-m\right)}}{e^{d\cdot\left(\cos\delta_{y_i^s,i}-m\right)}+\sum_{c=1,c\neq y_i^s}^{C}e^{d\cdot\cos\delta_{c,i}}}, \tag{1}$$

where d denotes the scaling factor, n_s denotes the batch size, m denotes the margin, C denotes the total number of speakers from the source domain, $y_i^s \in \{1, 2, \ldots, C\}$ is the ground-truth speaker label of $\mathbf{x}_i^s \in \mathbf{X}^s$, and $\delta_{c,i}$ is the angle between the weight vector $\mathbf{w}_c$ and the embeddings $f\left(\mathbf{x}_i^s\right)$. By optimizing $\mathcal{L}_{\text{ams}}$, the pairwise distance distributions in the source domain are well separated.

2.2. Cross-domain Distance Metric Adaptation

After pre-training the embedding network, the distances of embedding pairs are computed via cosine dissimilarity to estimate the source and target pairwise distance distributions. Specifically, the within-speaker source pairwise distance s^{ws} and target pairwise distance t^{ws} are computed as,

$$s^{ws}\left(\mathbf{x}_{c,i}^{s}, \mathbf{x}_{c,j}^{s}\right) = 1 - \cos\left(f\left(\mathbf{x}_{c,i}^{s}\right), f\left(\mathbf{x}_{c,j}^{s}\right)\right), \quad (2)$$

$$t^{ws}\left(\mathbf{x}_{c',i}^{t}, \mathbf{x}_{c',j}^{t}\right) = 1 - \cos\left(f\left(\mathbf{x}_{c',i}^{t}\right), f\left(\mathbf{x}_{c',j}^{t}\right)\right), \quad (3)$$

where $\mathbf{x}_{c,i}^{s}$ is the i-th sample of the speaker c from the source domain, $i \neq j$, $c \neq c'$. The between-speaker source pairwise distance s^{bs} and target pairwise distance t^{bs} can be computed in the same way.

MMD [18] is used to measure the discrepancy between two pairwise distance distributions, which is based on the fact that if two distributions are identical, all of the statistics should be the same. For example, the MMD loss between $P\left(S^{ws}\right)$ and $P\left(T^{ws}\right)$ is defined as,

$$\mathcal{L}_{\text{mmd}} \triangleq \sup_{\phi \in \Phi}\left(\mathbb{E}_{S^{ws}}\left[\phi\left(S^{ws}\right)\right] - \mathbb{E}_{T^{ws}}\left[\phi\left(T^{ws}\right)\right]\right), \quad (4)$$

where $\phi(\cdot)$ is a function mapped to $\mathbb{R}$, Φ is the unit ball in the Reproducing Kernel Hilbert Space (RKHS). In practice, the value of the MMD loss is estimated with the empirical kernel mean embeddings [18],

$$\begin{aligned}\mathcal{L}_{\text{mmd}}\left(S^{ws}, T^{ws}\right) = & \frac{1}{n_{ws}^{2}} \sum_{i=1}^{n_{ws}} \sum_{j=1}^{n_{ws}} k(s_i^{ws}, s_j^{ws}) \\ & + \frac{1}{n_{ws}^{2}} \sum_{i=1}^{n_{ws}} \sum_{j=1}^{n_{ws}} k(t_i^{ws}, t_j^{ws}) \\ & - \frac{2}{n_{ws}^{2}} \sum_{i=1}^{n_{ws}} \sum_{j=1}^{n_{ws}} k(s_i^{ws}, t_j^{ws}),\end{aligned} \quad (5)$$

where $k(\cdot, \cdot)$ is the RBF kernel. n_{ws} is the number of the within-speaker pairwise distances in the mini-batch. The MMD losses between $P(S^{bs})$ and $P(T^{bs})$, $P\left(S^{ws}\right)$ and $P(T^{bs})$, $P(S^{bs})$ and $P\left(T^{ws}\right)$ can be computed in the same way.

The overall objective function of CDMA is the weighed sum of the AM-softmax loss and the MMD losses,

$$\begin{aligned}\mathcal{L}_{\text{CDMA}} = & \mathcal{L}_{\text{ams}}\left(f\left(\mathbf{X}^{s}\right), \mathbf{Y}^{s}\right) \\ & + \lambda_1 \mathcal{L}_{\text{mmd}}\left(S^{ws}, T^{ws}\right) + \lambda_2 \mathcal{L}_{\text{mmd}}\left(S^{bs}, T^{bs}\right) \\ & - \lambda_3 \mathcal{L}_{\text{mmd}}\left(S^{ws}, T^{bs}\right) - \lambda_4 \mathcal{L}_{\text{mmd}}\left(S^{bs}, T^{ws}\right),\end{aligned} \quad (6)$$

where λ_1, λ_2, λ_3, and λ_4 are trade-off parameters. $\mathcal{L}_{\text{ams}}$ is optimized to maintain the well-separated pairwise distance distributions in the source domain. The second and third losses are to align the cross-domain pairwise distance distributions. The last two losses are to separate the within- and between-speaker pairwise distance distributions.

3. EXPERIMENTS

3.1. Datasets

The source domain is the combination of the development sets of VoxCeleb1 [19] and VoxCeleb2 [4]. The former contains 148,642 utterances from 1,211 speakers, and the latter contains 1,092,009 utterances from 5,994 speakers. We do not apply any data augmentation strategies to the source domain.

The target domain comprises utterances from the Voices Obscured in Complex Environmental Settings (VOiCES) Challenge 2019 corpus [16], which has a disjoint speaker label space from the source domain. This corpus consists of a development set with 15,904 utterances from 196 speakers, and an evaluation set with 11,392 utterances. We use the utterances in the development set to adapt the pre-trained model and test the performance on the evaluation set.

3.2. Implementation Details

For the input features, 64-dimensional log-mel filter banks are extracted within a 25ms sliding window for every 10ms. Cepstral mean normalization (CMN) is performed within a 3-second sliding window. We do not use voice activity detection (VAD) to remove silence frames. During training, each utterance is cut into 400-frame chunks to keep the same input length.

ResNet34 [20] is adopted as the speaker embedding network. For pre-training, the network is optimized by the stochastic gradient descent (SGD) optimizer with an initial learning rate of 0.02 and a weight decay of 5e-4. ReduceLROnPlateau scheduler is adopted to update the learning rate, and the batch size is 256. The scaling factor d and margin m of AM-softmax loss are 30 and 0.2, respectively.

For adaptation, the CDMA is evaluated in both supervised and unsupervised adaptation settings, based on the presence or absence of the speaker labels in the target domain. A speaker-balanced sampling strategy is adopted to construct the within- and between-speaker pairs, i.e., each speaker in the mini-batch contains the same number of chunks. We randomly select 4 chunks of each speaker within batches of size 128. For the supervised setting, the speaker-balanced mini-batch can be directly sampled. For the unlabeled target data in the unsupervised setting, we follow the strategy of self-supervised learning [21] that treats each unlabeled utterance as a distinct class, and samples within- and between-speaker pairs from the same and different utterances, respectively. The network is adapted with an initial learning rate of 0.001 decreasing by 10% every 10 epochs. The trade-off parameters λ_1 and λ_2 are 2 and 1, respectively. As for λ_3 and λ_4, the values are {0.05, 0.03} in the unsupervised setting, and {0.1, 0.05} in the supervised setting.

For the back-end, the extracted embeddings are processed sequentially by LDA, centering, length normalization, and PLDA. For the unsupervised setting, the LDA and PLDA models are trained on the VoxCeleb1 set. For the supervised setting, the LDA and PLDA models are trained on the development set of VOiCES. All experiments are conducted using ASV-subtools [22].

3.3. Results

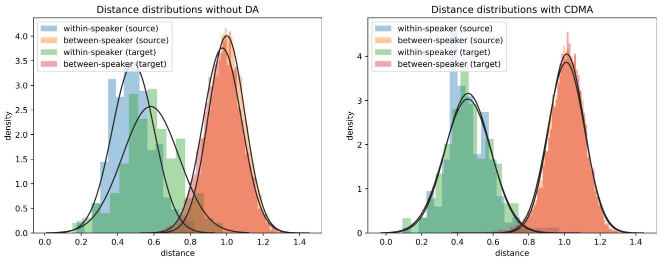
Table 1 and Table 2 show the results of various models in the unsupervised and supervised settings, respectively. For the supervised setting, the pre-trained ResNet34 model is fine-tuned (FT) on the development set of VOiCES, and CDMA is also combined with FT to train the model. MMD-E is the model that minimizes MMD loss in the embedding space. CDMA-align is the model that only aligns the pairwise distance distributions. Referring to Eq. 6, CDMA-align is implemented by setting λ_3 and λ_4 to zero. We compare the performance of CDMA with the pre-trained model and MMD-E. Meanwhile, CDMA-align and MMD-E are also compared to demonstrate the effectiveness of aligning in the distance metric space. The equal error rate (EER) and the minimum detection cost function (min-DCF) with $P_{\text{target}} = 0.01$ are adopted as the performance metrics.

Table 1: Results in the unsupervised setting.

Model	EER(%)	min-DCF
Pre-trained	8.14	0.566
MMD-E	6.36	0.525
Peri [23]	9.07	N/A
ADSAN + MINE [10]	6.76	0.599
CDMA-align	5.75	0.475
CDMA	**5.62**	**0.468**

Table 1 illustrates that CDMA can achieve the best results compared with the pre-trained model and MMD-E in the unsupervised setting. The relative reductions in EER are 31.0% and 11.6% respectively, and the relative reductions in min-DCF are 17.3% and 10.9% respectively. Compared with MMD-E, CDMA-align relatively reduces the EER and min-DCF by 9.6% and 9.5%, respectively. It indicates that aligning the pairwise distance distributions is more suitable for open-set domain adaptation than aligning the embedding distributions. There were recently proposed unsupervised domain adaptation approaches in the embedding space [23, 10], while the CDMA outperforms them by a large margin.

Table 2 illustrates that CDMA and CDMA-align outperform MMD-E in the supervised setting, and even outperform the fine-tuned model, which indicates that learning a domain-invariant metric can effectively boost the speaker verification performance. Compared with CDMA-align, CDMA relatively reduces the EER by 5.2%. It indicates that separating the two target distance distribution can boost the performance. When combined with fine-tuning, the EER and min-DCF of CDMA-align and CDMA can be further reduced, which indicates that they are complementary approaches of fine-tuning in the supervised setting. The CDMA also outperforms the recently proposed DEAAN model [11], which is a supervised domain adaptation approach in the embedding space.



(a) Before adaptation

(b) After adaptation

Fig. 2: Visualizations of the pairwise distance distributions.

Table 2: Results in the supervised setting. FT: fine-tuning.

Model	EER(%)	min-DCF
Pre-trained	6.12	0.467
FT	4.72	0.410
MMD-E	5.16	0.418
DEAAN [11]	5.21	0.394
CDMA-align	4.79	0.385
CDMA	**4.54**	**0.382**
CDMA-align + FT	4.54	**0.368**
CDMA + FT	**4.36**	0.369

To intuitively illustrate the effects of CDMA, we plot the within- and between-speaker pairwise distance distributions in Fig. 2. As can be seen, by applying CDMA, the target pairwise distance distributions are aligned with the source distributions, and the within- and between-speaker target distance distributions are effectively separated.

4. CONCLUSION

In this work, we highlighted the problem of directly aligning embedding distributions for open-set domain adaptation in ASV, and proposed to alleviate the domain shift in the distance metric space. The pairwise distance distributions were aligned and separated to learn a domain-invariant metric. Experimental results indicate that learning a domain-invariant metric is more suitable than learning domain-invariant embeddings for open-set domain adaptation in ASV.

5. ACKNOWLEDGEMENTS

This research is supported by the National Natural Science Foundation of China under grant No. U1736210.

6. REFERENCES

[1] John HL Hansen and Taufiq Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal processing magazine*, vol. 32, no. 6, pp. 74–99, 2015.

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

[3] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.

[4] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.

[5] Seyed Omid Sadjadi, Craig Greenberg, Elliot Singer, Douglas Reynolds, Lisa Mason, and Jaime Hernandez-Cordero, "The 2018 NIST Speaker Recognition Evaluation," in *Proc. Interspeech*, 2019, pp. 1483–1487.

[6] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira, "Analysis of representations for domain adaptation," in *Proc. NeurIPS*, 2006, pp. 137–144.

[7] Daniel Garcia-Romero and Alan McCree, "Supervised domain adaptation for i-vector based speaker recognition," in *Proc. ICASSP*, 2014, pp. 4047–4051.

[8] Saurabh Kataria, Jesús Villalba, Piotr Żelasko, Laureano Moro-Velázquez, and Najim Dehak, "Deep Feature CycleGANs: Speaker Identity Preserving Non-Parallel Microphone-Telephone Domain Adaptation for Speaker Verification," in *Proc. Interspeech*, 2021, pp. 1079–1083.

[9] Suwon Shon, Seongkyu Mun, Wooil Kim, and Hanseok Ko, "Autoencoder based domain adaptation for speaker recognition under insufficient channel information," in *Proc. Interspeech*, 2017, pp. 1014–1018.

[10] Lu Yi and Man-Wai Mak, "Adversarial Separation and Adaptation Network for Far-Field Speaker Verification," in *Proc. Interspeech*, 2020, pp. 4298–4302.

[11] Mufan Sang, Wei Xia, and John H.L. Hansen, "Deaan: Disentangled embedding and adversarial adaptation network for robust speaker representation learning," in *Proc. ICASSP*, 2021, pp. 6169–6173.

[12] Weiwei Lin, Man-Mai Mak, Na Li, Dan Su, and Dong Yu, "Multi-level deep neural network adaptation for speaker verification using mmd and consistency regularization," in *Proc. ICASSP*, 2020, pp. 6839–6843.

[13] Zhenyu Wang, Wei Xia, and John H.L. Hansen, "Cross-Domain Adaptation with Discrepancy Minimization for Text-Independent Forensic Speaker Verification," in *Proc. Interspeech*, 2020, pp. 2257–2261.

[14] Zhongxin Bai and Xiao-Lei Zhang, "Speaker recognition based on deep learning: An overview," *Neural Networks*, vol. 140, pp. 65–99, 2021.

[15] Djebril Mekhazni, Amran Bhuiyan, George Ekladious, and Eric Granger, "Unsupervised domain adaptation in the dissimilarity space for person re-identification," in *Proc. ECCV*, 2020, pp. 159–174.

[16] C. Richey, M.A. Barrios, Z. Armstrong, C. Bartels, H. Franco, M. Graciarena, Lawson A., M.K. Nandwana, A. Stauffer, J. van Hout, P. Gamble, Stephenson C. Hetherly, J., and K. Ni, "Voices Obscured in Complex Environmental Settings (VOiCES) Corpus," in *Proc. Interspeech*, 2018, pp. 1566–1570.

[17] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[18] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola, "A kernel method for the two-sample problem," *Journal of Machine Learning Research*, vol. 1, pp. 1–10, 2008.

[19] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," in *Proc. Interspeech*, 2017, pp. 2616–2620.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

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

[22] Fuchuan Tong, Miao Zhao, Jianfeng Zhou, Hao Lu, Zheng Li, Lin Li, and Qingyang Hong, "Asv-subtools: Open source toolkit for automatic speaker verification," in *Proc. ICASSP*, 2021, pp. 6184–6188.

[23] Raghuveer Peri, Monisankha Pal, Arindam Jati, Krishna Somandepalli, and Shrikanth Narayanan, "Robust speaker recognition using unsupervised adversarial invariance," in *Proc. ICASSP*, 2020, pp. 6614–6618.