DISENTANGLED SPEAKER EMBEDDING FOR ROBUST SPEAKER VERIFICATION

Lu YI and Man-Wai MAK

Department of Electronic and Information Engineering
The Hong Kong Polytechnic University, Hong Kong SAR

ABSTRACT

Entanglement of speaker features and redundant features may lead to poor performance when evaluating speaker verification systems on an unseen domain. To address this issue, we propose an InfoMax domain separation and adaptation network (InfoMax–DSAN) to disentangle the domain-specific features and domain-invariant speaker features based on domain adaptation techniques. A frame-based mutual information neural estimator is proposed to maximize the mutual information between frame-level features and input acoustic features, which can help retain more useful information. Furthermore, we propose adopting triplet loss based on the idea of self-supervised learning to overcome the label mismatch problem. Experimental results on VOiCES Challenge 2019 demonstrate that our proposed method can help learn more discriminative and robust speaker embeddings.

Index Terms— Speaker verification, domain adaptation, mutual information, self-supervised learning.

1. INTRODUCTION

Speaker verification (SV) has recently attracted increasing attention, and it has been widely used in real-world biometric applications. Traditional state-of-the-art SV systems use factor analysis to extract i-vectors as speaker embeddings [1]. With the development of deep learning, more and more researchers are committed to using deep neural models to improve speaker verification performance. Among them, the x-vector [2] proposed by Snyder *et al.* has achieved great success and become state-of-the-art. However, deep neural network (DNN) embeddings require a large amount of training data to achieve accurate predictions or classifications. Unfortunately, acquiring sufficient labelled training data is a great challenge because the data collection process is time consuming and labour intensive. The performance will also be degraded when deploying an existing system to an unseen environment because of the domain mismatch between the trained and deployed environments.

Domain mismatch occurs when the speech is collected from different acoustic environments. While data augmentation [2] can increase the number of possible acoustic environments a system may encounter, it can only partially alleviate the domain mismatch problem. A better solution is domain adaptation (DA). Domain adaptation has been applied to either adapt PLDA models to fit the target data [3] or find a shared embedding space in which the features contain more speaker information but less unrelated information. Finding a common embedding space is a more general approach in that it is independent of the backend. Previous methods tried to align the features by minimizing the cosine distance and mean squared error [4], or by minimizing the maximum mean discrepancy (MMD) [5] at the embedding layer. More recent research is based on adversarial learning [6, 7, 8]. The essence of these methods is the same: find a shared space where the feature distributions have low discrepancy across multiple domains. However, most of these studies treat DA as a transformation of speaker embedding vectors instead of using DA to extract domain-invariant embedding directly.

Although invariant representations can be obtained [5], transformation-based DA may not perform well on the target domain because of the difference in label distributions [9]. For speaker verification, the speaker IDs of different domains are different, which increases the difficulty of DA. Therefore, it is hard to train a speaker embedding network with domain-invariant property. In addition, insufficient data from the target domain would also limit the model to find a common embedding space for all possible domains.

To overcome these difficulties, we propose a novel framework to perform domain adaptation. The framework (1) incorporates domain adaptation directly into the training of speaker embedding extractor, (2) uses self-supervised learning to overcome the label mismatch problem without using labels from the target domain, and (3) introduces a frame-based mutual information neural estimator to maximize the mutual information between the frame-level features and input acoustic features. These actions are beneficial for the network to learn informative representations. Experimental results on the VOiCES 2019 dataset demonstrate the effectiveness of maximizing frame-based mutual information in speaker embedding networks.

This work was supported by the RGC of Hong Kong SAR, Grant No. PolyU 152137/17E and National Natural Science Foundation of China (NSFC), Grant No. 61971371.

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

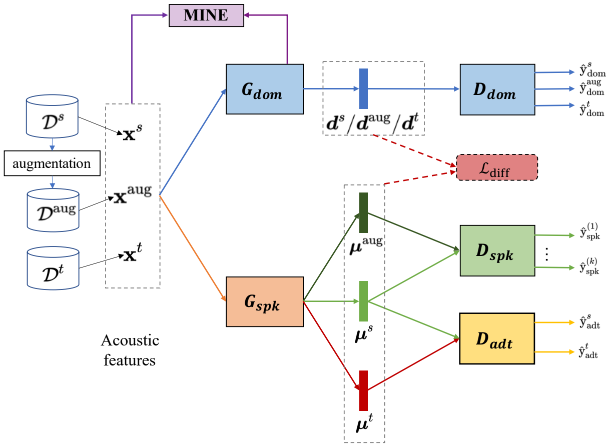

Fig. 1. Overview of the proposed InfoMax Domain Separation and Adaptation Network (InfoMax–DSAN).

2. INFOMAX DOMAIN SEPARATION AND ADAPTATION NETWORK

2.1. Overview

Traditional domain adaptation approaches utilize data from two domains: labelled source domain and unlabeled target domain. These methods generally require a large number of samples from the target domain; otherwise, the performance may degrade. To overcome this issue, our proposed framework, as Fig. 1 shows, considers three domains: the source domain $\mathcal{D}^s$, the augmentation of the source domain $\mathcal{D}^{\text{aug}}$, and the target domain $\mathcal{D}^t$.

Denote a labelled dataset from the source domain as $\mathbf{X}^s = \{(\mathbf{x}_i^s, \mathbf{y}_i)\}_{i=1}^{N_s}$, a labelled augmented set of the source domain as $\mathbf{X}^{\text{aug}} = \{(\mathbf{x}_i^{\text{aug}}, \mathbf{y}_i)\}_{i=1}^{N_{\text{aug}}}$, and an unlabelled dataset from the target domain as $\mathbf{X}^t = \{\mathbf{x}_i^t\}_{i=1}^{N_t}$. In addition to finding the domain-specific features from two totally different domains, we create a new domain $\mathcal{D}^{\text{aug}}$ that shares some properties with the source domain. The new domain consists of augmented samples that are created by introducing some background noises and reverbration to the labelled data. Therefore, background noise and reverberation are the factors causing the domain mismatch between $\mathcal{D}^s$ and $\mathcal{D}^{\text{aug}}$. In contrast, the domain mismatch between $\mathcal{D}^t$ and other two domains are caused by various factors, e.g., microphone variability, different types and levels of noises, and different room acoustics. Because the factors are unknown and speakers in the target domain are different from those in the source domain, constructing a domain with known factors can help training a feature extractor to extract domain-specific features.

As shown in Fig. 1, the desired speaker embeddings are extracted from the speaker-feature extractor G_{spk}. A speaker classifier D_{spk} is connected to G_{spk} to ensure that the embeddings $\boldsymbol{\mu}^s$ and $\boldsymbol{\mu}^{\text{aug}}$ are speaker discriminative. This is achieved by minimizing the speaker identification loss:

$$\begin{aligned}\mathcal{L}_{\text{spk}} &= \mathbb{E}_{\mathbf{x}^s\sim\mathcal{D}^s}\left[-\sum_{k=1}^{K} y_{\text{spk}}^{(k)} \log D_{\text{spk}}\left(G_{\text{spk}}\left(\mathbf{x}^s\right)\right)_k\right] \\ &\quad + \mathbb{E}_{\mathbf{x}^{\text{aug}}\sim\mathcal{D}^{\text{aug}}}\left[-\sum_{k=1}^{K} y_{\text{spk}}^{(k)} \log D_{\text{spk}}\left(G_{\text{spk}}\left(\mathbf{x}^{\text{aug}}\right)\right)_k\right],\end{aligned} \tag{1}$$

where k denotes the k-th output of the speaker discriminator and $y_{\text{spk}}^{(k)}$ is a one-hot encoded speaker label.

Meanwhile, G_{spk} is trained adversarially to extract the common features $\boldsymbol{\mu}^s$ and $\boldsymbol{\mu}^t$ from the source and target domains. The adapataion network D_{adt} is trained to indicate whether the common features come from the source domain or the target domain. Therefore, D_{adt} and G_{spk} are trained with the following minimax loss:

$$\begin{aligned}\min_{G_{\text{spk}}}\max_{D_{\text{adt}}} \mathcal{L}_{\text{adt}} &= \mathbb{E}_{\mathbf{x}^s\sim\mathcal{D}^s}\left[\log D_{\text{adt}}\left(G_{\text{spk}}\left(\mathbf{x}^s\right)\right)\right] \\ &\quad + \mathbb{E}_{\mathbf{x}^t\sim\mathcal{D}^t}\left[\log\left[1 - D_{\text{adt}}\left(G_{\text{spk}}\left(\mathbf{x}^t\right)\right)\right]\right].\end{aligned} \tag{2}$$

The domain-feature extractor G_{dom} extracts domain-specific features $(\boldsymbol{d}^s, \boldsymbol{d}^{\text{aug}}, \boldsymbol{d}^t)$ from the acoustic features. A domain discriminator D_{dom} disentangles the domain-specific features by predicting to which of the three domains the domain-specific features should belong. To this end, the loss function for D_{dom} is set to

$$\begin{aligned}\mathcal{L}_{\text{dom}} &= \mathbb{E}_{\mathbf{x}^t\sim\mathcal{D}^t}\left[-\log D_{\text{dom}}\left(G_{\text{dom}}\left(\mathbf{x}^t\right)\right)_0\right] \\ &\quad + \mathbb{E}_{\mathbf{x}^s\sim\mathcal{D}^s}\left[-\log D_{\text{dom}}\left(G_{\text{dom}}\left(\mathbf{x}^s\right)\right)_1\right] \\ &\quad + \mathbb{E}_{\mathbf{x}^{\text{aug}}\sim\mathcal{D}^{\text{aug}}}\left[-\log D_{\text{dom}}\left(G_{\text{dom}}\left(\mathbf{x}^{\text{aug}}\right)\right)_2\right].\end{aligned} \tag{3}$$

where $()_k$ denotes the k-th output of D_{dom} and $k = 0, 1, 2$ correspond to the target, source, and augmented-source, respectively.

To further ensure that the domain features $\boldsymbol{d}$'s contain domain information from the inputs, a frame-based mutual information neural estimator (MINE) is applied. Details will be given in Section 2.2.

To disentangle the domain-specific features from the speaker features, we introduce a constraint ($\mathcal{L}_{\text{diff}}$ in Fig. 1) to make the speaker embeddings ($\boldsymbol{\mu}$) and the domain embeddings ($\boldsymbol{d}$) different. In this study, we computed the difference loss in two ways: (1) making $\boldsymbol{d}$ and $\boldsymbol{\mu}$ othorgonal as suggested in [10] and (2) minimizing their mutual information as proposed in [11].

2.2. Mutual Information Neural Estimation

Feature learning based on mutual information has attracted increased interest in recent years. In speaker verification, many researchers applied MINE to compute the mutual information between the embeddings and frame-level features [12, 13], or between multiple embeddings [14]. Studies seldom attempt to employ MINE on frame-level features. However, when computing the mutual information between the frame-level and segment-level features, the problem of dimensionality imbalance or information loss would occur. To address these issues

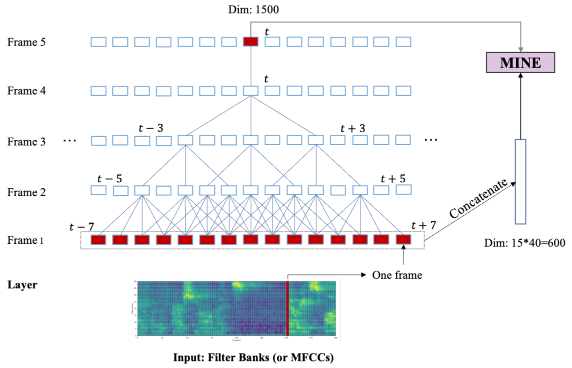

Fig. 2. Frame-based mutual information neural estimation.The TDNN is part of G_{dom} in Fig. 1.

and to make mutual information estimation more reliable, we propose to employ frame-based MINE on the time-delay neural network (TDNN) as shown in Fig. 2.

Mutual information (MI) is a measure of information shared between random variables. The value of mutual information will be high if two random variables are highly dependent. To address the difficulties in estimating mutual information for unknown probability distributions, Belghazi *et al.* [15] proposed a neural network based MI approximator called MINE. The approximated MI is given by

$$\begin{aligned} I_{\Theta}^{\text{KL}}(X, Z) &= D_{\text{KL}}\left(\mathbb{P}_{XZ} \| \mathbb{P}_X \otimes \mathbb{P}_Z\right) \\ &= \sup_{\theta \in \Theta} \mathbb{E}_{\mathbb{P}_{XZ}}\left[T_\theta\right] - \log\left(\mathbb{E}_{\mathbb{P}_X \otimes \mathbb{P}_Z}\left[e^{T_\theta}\right]\right), \end{aligned} \tag{4}$$

where $T_\theta : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ is a function parametrized by a deep neural network, $\mathbb{P}_{XZ}$ is the joint distribution, and $\mathbb{P}_X \otimes \mathbb{P}_Z$ is the product of the marginal distributions.

Since the precise value of mutual information is not necessary, we may use a MINE that is based on Jensen-Shannon divergence [16] instead of KL-divergence. Then, the mutual information estimation becomes

$$I_{\Theta}^{\text{JS}}(X, Z) = \mathbb{E}_{\mathbb{P}_{XZ}}\left[-\text{sp}\left(-T_\theta\right)\right] - \mathbb{E}_{\mathbb{P}_X \otimes \mathbb{P}_Z}\left[\text{sp}\left(T_\theta\right)\right], \tag{5}$$

where $\text{sp}(z) = \log\left(1 + e^z\right)$ is the softplus function.

Let $\mathbf{x}_{i,t}$ represents the t-th frame of utterance i and $\mathbf{f}_{i,t}$ represents the frame-level feature extracted from the last frame-level layer (before the statistics pooling layer) at frame t. According to the TDNN architecture described in [2], each frame-level feature $\mathbf{f}_{i,t}$ aggregates information from 15 frames of the input features, i.e., $[\mathbf{x}_{i,t-7} \colon \mathbf{x}_{i,t+7}]$. Therefore, we train a MINE to maximize the average MI between $\mathbf{f}_{i,t}$ and $\mathbf{x}_{i,t-7\colon t+7}$, where $\mathbf{x}_{i,t-7\colon t+7}$ is the concatenation of $[\mathbf{x}_{i,t-7} \colon \mathbf{x}_{i,t+7}]$. The negative samples from $\mathbb{P}_X \otimes \mathbb{P}_Z$ (as indicated in Eq. 5) can be obtained by presenting $\mathbf{f}_{i,t}$ and $\mathbf{x}_{j,t-7\colon t+7}$ $(i \neq j)$ to the MINE. Denote T as the number of frames in each utterance and N as the total number of utterances. Then, the set that aggregates $\mathbf{x}_{i,t-7\colon t+7}$ and $\mathbf{f}_{i,t}$ can be defined as $X = \{\mathbf{x}_{i,t-7\colon t+7}, i = 1, \ldots, N, t = 1, \ldots, T\}$ and $Z = \{\mathbf{f}_{i,t}, i = 1, \ldots, N, t = 1, \ldots, T\}$, respectively. We used zero padding for those non-existing frames, i.e., $\mathbf{x}_{i,t<1}$. Therefore, the mutual information loss becomes $\mathcal{L}_{\text{MINE}} = -I_{\Theta}^{\text{JS}}(X, Z)$.

2.3. Self-supervised Learning

Because the source and target domains typically have different label sets, forcing a low discrepancy between domain-dependent speaker embeddings is insufficient to reduce the domain mismatches. Ignoring the label mismatch issue will cause failure in domain adaptation [9], which is confirmed by our experiments where a negative transfer was observed. To tackle this problem, we used the triplet loss [17] to achieve self-supervised learning on the speech from the target domain. The primary purpose of triplet loss is to minimize the distance between an anchor and a positive sample ($d(a, p)$) and maximize the distance between an anchor and a negative sample ($d(a, n)$). To this end, the triplet loss can be defined as,

$$\mathcal{L}_{\text{triplet}} = \max(d(a, p) - d(a, n) + \text{margin}, 0). \tag{6}$$

In this work, the segments from the same utterance form the positive pair (a, p), while the negative pair contains segments from different utterances.

2.4. Overall Optimization

The overall objective of the proposed model is

$$\begin{aligned} \min_{G_{\text{dom}}, G_{\text{spk}}, D_{\text{dom}}, D_{\text{spk}}, T_\theta} \max_{D_{\text{adt}}} \mathcal{L}_{\text{spk}} + \mathcal{L}_{\text{dom}} + \alpha \mathcal{L}_{\text{MINE}} + \beta \mathcal{L}_{\text{diff}} \\ + \gamma \mathcal{L}_{\text{triplet}} + \lambda \mathcal{L}_{\text{adt}}, \end{aligned} \tag{7}$$

where α, β, λ are hyperparameters that control the contributions of the sub-losses to the whole loss. In our experiments, they were set to be $\alpha = 1.0, \beta = 1.0, \gamma = 0.5, \lambda = 0.5$.

3. EXPERIMENTAL SETTING

3.1. Data Preparation

The source domain comprises utterances from the development sets of VoxCeleb1 [18] and VoxCeleb2 [19]. We followed the data augmentation procedure in Kaldi to create the augmented set by adding noise, babble, and music from MUSAN [20] and reverberation from the RIR dataset [21]. The target domain comprises utterances from the VOiCES Challenge 2019 [22], but we only used the development set for training. We followed the Kaldi recipe to extract 40-dimensional filter bank features with a frame length of 25ms at 10ms shift and used Kaldi energy-based voice activity detection (VAD) to remove silence frames.

3.2. Implementation Details

In this study, both feature extractors G_{dom} and G_{spk} follow the x-vector framework [2], except that they only have one segment-level layer after the pooling layer. The speaker classifier D_{spk} comprises a linear layer and a softmax output layer.

Table 1. Performance on the original VOiCES 2019 development set and evaluation set. $\mathcal{D}^s$ comprises the development sets of VoxCeleb1&2, $\mathcal{D}^{\text{aug}}$ corresponds to the augmented set of VoxCeleb1&2, and $\mathcal{D}^t$ comprises utterances from the development set of VOiCES 2019. Two metrics were used for computing $\mathcal{L}_{\text{diff}}$: mutual information (MI) and orthogonality (ort).

Row	System	Source domain	Target domain	MINE	$\mathcal{L}_{\text{triplet}}$	D_{adt}	$\mathcal{L}_{\text{diff}}$	VOiCES dev.		VOiCES eval.	
								EER(%)	minDCF	EER(%)	minDCF
1	TDNN [4]	$\mathcal{D}^s$	$\mathcal{D}^{\text{aug}}$	×	×	✓	×	3.18	-	7.15	-
2	TDNN (G_{spk})	$\{\mathcal{D}^s, \mathcal{D}^{\text{aug}}\}$	×	×	×	×	×	2.35	0.2596	6.42	0.4398
3		$\{\mathcal{D}^s, \mathcal{D}^{\text{aug}}\}$	$\mathcal{D}^t$	×	✓	×	×	2.16	0.2627	6.32	0.4305
4		$\{\mathcal{D}^s, \mathcal{D}^{\text{aug}}\}$	×	✓	×	×	×	2.29	0.2453	5.98	0.4351
5		$\{\mathcal{D}^s, \mathcal{D}^{\text{aug}}\}$	$\mathcal{D}^t$	×	✓	✓	×	2.31	0.2645	6.29	0.4399
6		$\{\mathcal{D}^s, \mathcal{D}^{\text{aug}}\}$	$\mathcal{D}^t$	✓	✓	✓	×	2.54	0.2671	6.23	0.4379
7	InfoMax–DSAN (G_{dom}&G_{spk})	$\mathcal{D}^s$	$\mathcal{D}^{\text{aug}}$	✓	×	✓	MI	2.17	0.2418	5.93	0.4265
8		$\mathcal{D}^s$	$\mathcal{D}^{\text{aug}}$	✓	×	✓	ort	2.33	0.2577	5.98	0.4131
9		$\{\mathcal{D}^s, \mathcal{D}^{\text{aug}}\}$	$\mathcal{D}^t$	✓	×	✓	MI	3.29	0.3278	6.75	0.4614
10		$\{\mathcal{D}^s, \mathcal{D}^{\text{aug}}\}$	$\mathcal{D}^t$	×	✓	✓	MI	2.31	0.2490	6.02	0.4161
11		$\{\mathcal{D}^s, \mathcal{D}^{\text{aug}}\}$	$\mathcal{D}^t$	✓	✓	✓	MI	**2.06**	**0.2375**	**5.69**	**0.4127**

The domain discriminator D_{dom} has two fully connected layers with 128 nodes in each layer, while D_{adt} is a two-layer fully connected network with 512 nodes in each layer. The mutual information neural estimator (MINE) is a three-layer fully connected network with 512 nodes in each layer. Also, G_{dom} uses statistics pooling while G_{spk} uses attentive statistics pooling [23]. All models were trained using small chunks of acoustic sequences with a chunk length of 200 frames.

The proposed infoMax domain separation and adaptation network (InfoMax–DSAN) was trained using the RAdam optimizer [24] with an initial learning rate of 0.001. The dimension of domain-specific embeddings $\boldsymbol{d}$'s is 64, while that of the speaker embeddings $\boldsymbol{\mu}$'s is 256. Probabilistic linear discriminant analysis (PLDA) [25] was used as the backend for scoring. Before PLDA training, the extracted speaker embeddings were projected onto a 150-dimensional space by LDA, followed by whitening and length normalization. We also performed the symmetric score normalization (S-norm) [26] to normalize the scores of the VOiCES evaluation set, using the speech from the VOiCES development set as the cohort.

4. RESULTS AND DISCUSSIONS

Table 1 shows the performance of different systems on the VOiCES 2019 dataset. A "×" in the fourth column means that we used the development sets of VoxCeleb1&2 and their augmented sets to train the speaker embedding networks G_{spk} without domain adaptation. Rows 2–6 present the performance of G_{spk}-based systems (Fig. 1 without G_{dom} and D_{dom}). It can be observed that the performance can be improved when applying the MINE to maximize the mutual information between the speaker features and the input acoustic features (Row 4). This indicates that the proposed frame-based MINE can effectively help to extract informative features.

The lower part of Table 1 (Rows 7–11) shows the performance of InfoMax–DSAN based systems. The first two systems were trained by considering that the target domain can be perfectly simulated by the augmented data, i.e., we can replace $\mathbf{x}^t$ in Fig. 1 by $\mathbf{x}^{\text{aug}}$. Two metrics were used to evaluate $\mathcal{L}_{\text{diff}}$: mutual information (MI) and orthogonality (ort). Overall, minimizing the mutual information between domain features and speaker features performs better. The improvement of these two systems suggests that the extracted speaker embeddings are more robust. We also compared the performance of systems with and without the triplet loss (Rows 9 and 11). Without $\mathcal{L}_{\text{triplet}}$, the performance of the domain adaptation network performs even worse than the baseline. On the contrary, the best performance can be obtained when training the model with the triplet loss. This result demonstrates the importance of considering the label distribution of the target domain when applying domain adaptation. Ignoring the label mismatch problem would degrade performance. Rows 10 and 11 suggest that the MINE can help disentangle redundant features from speaker features. Another observation is that the TDNN system could not benefit from learning domain-invariant features only (Row 6). This implies that disentangling the domain-specific features is more effective.

5. CONCLUSIONS

In this paper, we propose a domain adaptation framework to disentangle the task-related and task-irrelevant features in the front-end of speaker verification systems. A frame-based mutual information neural estimation is proposed to learn informative features by maximizing the mutual information between the input acoustic features and frame-level features from a TDNN. We also introduce the triplet loss to address the label mismatch problem in domain adaptation. The experiments on the VOiCES 2019 corpus demonstrate that our proposed approaches can effectively help extract more robust speaker features.

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

[3] D. Garcia-Romero and A. McCree, "Supervised domain adaptation for i-vector based speaker recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 4047–4051.

[4] Jonathan Huang and Tobias Bocklet, "Intel far-field speaker recognition system for VOiCES challenge 2019," in *Proc. Interspeech*, 2019, pp. 2473–2477.

[5] Wei-wei Lin, Man-Wai Mak, and Jen-Tzung Chien, "Multi-source i-vectors domain adaptation using maximum mean discrepancy based autoencoders," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 26, no. 12, pp. 2412–2422, 2018.

[6] 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 *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 4889–4893.

[7] Youzhi Tu, Man-Wai Mak, and Jen-Tzung Chien, "Variational Domain Adversarial Learning for Speaker Verification," in *Proc. Interspeech*, 2019, pp. 4315–4319.

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

[9] Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon, "On learning invariant representations for domain adaptation," in *Proc. International Conference on Machine Learning*, 2019, pp. 7523–7532.

[10] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan, "Domain separation networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 343–351.

[11] Mufan Sang, Wei Xia, and John HL Hansen, "DEAAN: Disentangled embedding and adversarial adaptation network for robust speaker representation learning," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 6169–6173.

[12] Min Hyun Han, Woo Hyun Kang, Sung Hwan Mun, and Nam Soo Kim, "Information preservation pooling for speaker embedding," in *Proc. The Speaker and Language Recognition Workshop (Speaker Odyssey)*, 2020.

[13] Youzhi Tu and Man-Wai Mak, "Mutual information enhanced training for speaker embedding," *Proc. Interspeech*, pp. 91–95, 2021.

[14] Yoohwan Kwon, Soo Whan Chung, and Hong Goo Kang, "Intra-class variation reduction of speaker representation in disentanglement framework," in *Proc. Interspeech*, 2020, vol. 2020, pp. 3231–3235.

[15] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Devon Hjelm, and Aaron Courville, "Mutual information neural estimation," in *International Conference on Machine Learning*, 2018, pp. 530–539.

[16] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio, "Learning deep representations by mutual information estimation and maximization," *arXiv preprint arXiv:1808.06670*, 2018.

[17] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[18] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Proc. Interspeech*, 2017, pp. 2616–2620.

[19] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018.

[20] David Snyder, Guoguo Chen, and Daniel Povey, "MUSAN: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[21] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 5220–5224.

[22] Colleen Richey, Maria Barrios, Zeb Armstrong, Chris Bartels, Horacio Franco, Martin Graciarena, Aaron Lawson, Mahesh Nandwana, Allen Stauffer, Julien Hout, Paul Gamble, Jeffrey Hetherly, Cory Stephenson, and Karl Ni, "Voices obscured in complex environmental settings (VOiCES) corpus," in *Proc. Interspeech*, 09 2018, pp. 1566–1570.

[23] Koji Okabe, Takafumi Koshinaka, and Koichi Shinoda, "Attentive statistics pooling for deep speaker embedding," in *Proc. Interspeech*, 2018, pp. 2252–2256.

[24] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han, "On the variance of the adaptive learning rate and beyond," in *Proc. International Conference on Learning Representations*, 2019.

[25] Sergey Ioffe, "Probabilistic linear discriminant analysis," in *European Conference on Computer Vision*, 2006, pp. 531–542.

[26] Pavel Matejka, Ondrej Novotnỳ, Oldrich Plchot, Lukas Burget, Mireia Diez Sánchez, and Jan Cernockỳ, "Analysis of score normalization in multilingual speaker recognition.," in *Proc. Interspeech*, 2017, pp. 1567–1571.