Interspeech 2022
18-22 September 2022, Incheon, Korea

Class-Aware Distribution Alignment based Unsupervised Domain Adaptation for Speaker Verification

Hang-Rui Hu[1], *Yan Song*[1], *Li-Rong Dai*[1], *Ian McLoughlin*[1,2], *Lin Liu*[3]

[1]National Engineering Research Center of Speech and Language Information Processing, University of Science and Technology of China, Hefei, China.
[2]ICT Cluster, Singapore Institute of Technology, Singapore.
[3]iFLYTEK Research, iFLYTEK CO. LTD., Hefei, China.

hhr@mail.ustc.edu.cn, {songy, ivm, lrdai}@ustc.edu.cn, liulin@iflytek.com

Abstract

Existing speaker verification (SV) systems usually suffer from significant performance degradation when applied to a new domain that lies outside the training distribution. Given the unlabeled target-domain dataset, most Unsupervised Domain Adaptation (UDA) methods aim to minimize the distribution divergence between different domains. However, global distribution alignment strategies fail to consider latent speaker label information and can hardly guarantee the feature discriminative capability in target domain. In this paper, we propose a novel UDA approach called WBDA (Within-class and Between-class Distribution Alignment), which aims to transfer the class-aware information (*i.e.*, within- and between-class distributions) learned from the well-labeled source-domain, to the unlabeled target-domain. Motivated by the recent progress of self-supervised contrastive learning, positive and negative pairs are constructed separately for source and target domains, from which the within- and between-class distribution can be estimated. And the SV system can then be learned by jointly optimizing the cross-domain class-aware distribution discrepancy loss and source-domain classification loss in an end-to-end manner. Evaluations on NIST SRE16 and SRE18 achieve a relative performance improvement of about 43.7% and 26.2% over the baseline in terms of Equal Error Rate (EER) separately, significantly outperforming the previous adaption methods based on global distribution alignment.

Index Terms: Speaker Verification, Unsupervised Domain Adaptation, End-to-End, Distribution Alignment

1. Introduction

Speaker verification (SV) aims to determine whether a speech utterance belongs to a given speaker or not. In recent years, a profusion of deep neural network (DNN) methods have achieved great success on SV tasks. To improve the compactness and discriminative capability of speaker embeddings, existing works mainly focus on designing different network architectures, pooling methods and optimizing objectives [1, 2, 3, 4, 5, 6, 7, 8, 9, 10].

Despite the success of SV using deep embedding learning, it is well known that such methods are generally sensitive to the *domain shift* issue, *i.e.*, performance degrades significantly when applied to a target-domain whose distribution lies outside the source-domain, (as shown in Fig. 1a). Since collecting and labeling target domain data is time-consuming and expensive, it is necessary to find an effective method to adapt an existing model trained on a well-labeled source-domain dataset to a target-domain where only weakly-labeled or even unlabeled data is available.

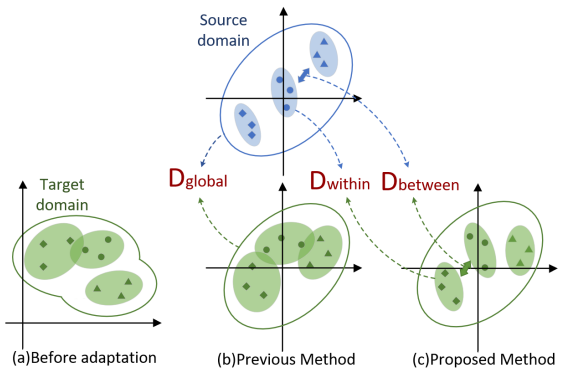

Figure 1: *Motivation for the proposed approach. (a) Domain distribution mismatch before adaptation. (b) Existing UDA methods only achieve global distribution alignment which cannot ensure discriminative improvement in target domain. (c) Using our within- and between-class distribution alignment (WBDA) methods, we further transfer the class-aware distribution information from well-labeled source-domain to target-domain.*

Given the unlabeled target-domain dataset, most existing Unsupervised Domain Adaptation (UDA) methods rely on global distribution alignment including adversarial learning [11, 12, 13, 14, 15] or discrepancy based methods [16, 17, 18, 19, 20] to address the domain shift issue. Adversarial learning methods [11, 12, 13, 14, 15] encourage learning of embeddings that are domain-invariant by utilizing an additional adversarial domain discriminator. Discrepancy-based methods aim to minimize feature distributions discrepancy between different domains, which is usually based on maximum mean discrepancy (MMD) [21] or correlation alignment (CORAL) [22]. However, such global distribution based methods fail to take into account the latent speaker information of the target domain and fail to guarantee speaker discrimination of learned features (as shown in Fig. 1b). In [23], unsupervised clustering based domain adaptation was proposed to estimate pseudo-labels of target-domain data, and then perform self-supervised adaptation. More recently, the self-supervised learning based domain adaptation (SSDA) method leveraged potential label information from the target domain and adapted the discrimina-

Yan Song is the corresponding author. This work was supported by the Leading Plan of CAS (XDC08030200)

10.21437/Interspeech.2022-591

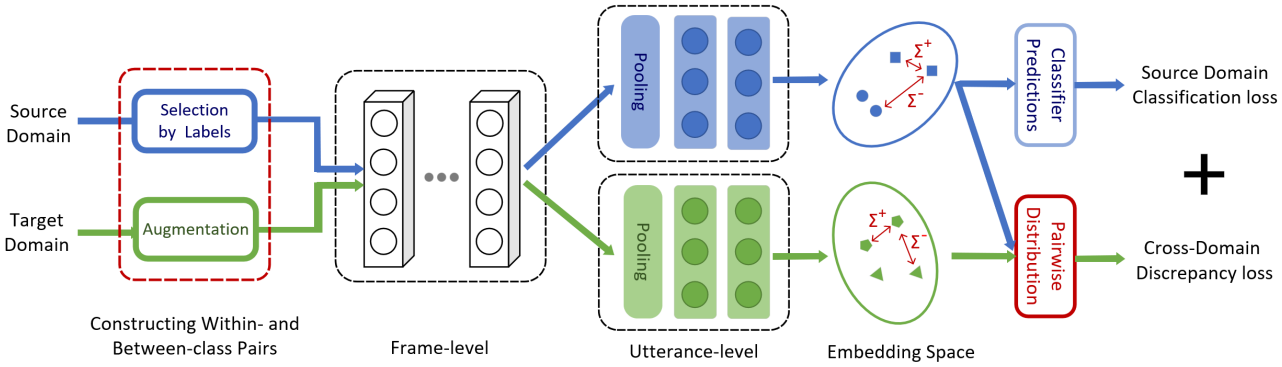

Figure 2: *Framework of the proposed Within-Class and Between-Class Distribution Alignment (WBDA) method. The network is jointly trained via optimizing the source domain classification loss and cross-domain discrepancy loss. Data is fed into the network in the form of positive and negative pairs to estimate class-aware statistics without target domain labels. Sample pairs from source domain are randomly selected according to the speaker ids, while those from target domain are constructed similar as unsupervised contrastive learning. Note that we don't share the parameters of utterance-level part to extract domain-specific knowledge and align distribution of target domain.*

tion ability from the source domain simultaneously [24]. However, noisy estimated label information in the target domain may hinder any performance improvement.

In this paper, we propose a novel unsupervised domain adaptation approach, called Within-class and Between-class Distribution Alignment (WBDA), as illustrated in Figs. 1c and as a block diagram in Fig. 2. WBDA aims to transfer class-aware information learned from the well-labeled source-domain to the unlabeled target-domain. Specifically, we construct within-class and between-class pairs for source and target domains separately. These positive and negative pairs are randomly selected according to the speaker identities. Then, for the unlabeled target domain, we construct positive pairs via data augmentation, motivated by the progress of recent contrastive learning. The main aim for WBDA is to focus on transferring the within-class and between-class distribution information, estimated from the source domain, to the target domain. Compared with existing global distribution alignment based UDA methods, WBDA is class-aware, which can effectively learn discriminative speaker embeddings for the unlabeled target domain. The network can be learned by jointly optimizing both cross-domain class-aware distribution discrepancy loss and the cross-entropy loss in an end-to-end manner. Evaluations on NIST SRE16 and SRE18 can achieve a relative performance improvement of 43.7% and 26.2% EER over the baseline, which is significantly better than the previous unsupervised adaption methods based on global distribution alignment.

2. Overview of the proposed framework

The structure of the proposed within-class and between-class distribution alignment (WBDA) adaption method is shown in Fig. 2. The entire network is trained in an end-to-end manner, aiming to learn discriminative deep representations in the well-labeled source domain, while transferring the compact class structure knowledge to the unlabeled target domain. The overall loss has two parts;

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda \mathcal{L}_{WBDA} \tag{1}$$

where $\mathcal{L}_{CE}$ denotes the standard cross-entropy loss for classification training in the source domain. $\mathcal{L}_{WBDA}$ denotes a proposed cross-domain discrepancy loss where λ is the hyper-parameter to weight $\mathcal{L}_{CE}$ and $\mathcal{L}_{WBDA}$. This aligns the second-order statistics of within-class and between-class distributions.

To calculate class-relevant statistics without touching the target domain labels, data is fed into the network in the form of positive and negative pairs. In the source domain, pairs can be directly constructed based on ground-truth labels. In the target domain, positive pairs are obtained by two data augmentations (such as RandomCrop, SpecAug [25], adding noise, etc.) from the same utterance, while samples from different utterances are treated as negative pairs, which is similar to the assumption of unsupervised contrastive learning. The difference is that these sample pairs are used for domain-knowledge transfer instead of direct metric learning. We will show that, when sample pairs are properly constructed, the within- and between-class statistics can be efficiently and accurately estimated in each mini-batch, allowing the entire network to be trained in an end-to-end manner.

The network structure is similar to previous works, with parameters of shallow layers being domain-shared since they can extract general frame-level local representations. However, sharing the statistics of Batch Normalization (BN) layers is inappropriate when the domain shift is significant, we thus replace all the BNs with DABNs [26] to separate the statistics of each domain, while sharing the affine transformation. Likewise, we do not share the parameters of deep layers, and incorporate two utterance-level branches with weight regularization in [15], to better seek domain-specific knowledge and adjust the distribution of the target domain.

3. Method

We utilize commonly used second-order statistics in domain adaptation such as covariance or correlation, to ascertain or measure cross-domain distributional discrepancy. In this section, we will first briefly introduce the covariance matrix, which can actually be divided into two parts: within-class and between-class. Then, we will derive the equivalent forms of these statistics based on positive and negative pairs. Finally, we will present the WBDA loss, which is designed to align the within- and between-class distributions of different domains within each batch.

3.1. Definition of Covariance Matrix

Consider a training data set X containing N examples, where each example is a column vector of length d and belongs to one of K classes. The $d \times d$ global covariance matrix is computed as follows:

$$\Sigma_{cov} = \frac{1}{N}\sum_{x}(x-\mu)(x-\mu)^T \tag{2}$$

where $\mu = \frac{1}{N}\sum_{x\in X} x$ is the global mean vector. The total covariance matrix can actually be divided into two parts: Within-class covariance Σ_{cov}^W indicates the degree of dispersion between samples and the corresponding class center, and between-class covariance Σ_{cov}^B represents the separability of different class centers:

$$\begin{aligned}
\Sigma_{cov} &= \Sigma_{cov}^W + \Sigma_{cov}^B \\
\Sigma_{cov}^W &= \frac{1}{N}\sum_{k}\sum_{x}(x^{(k)}-\mu^{(k)})(x^{(k)}-\mu^{(k)})^T \\
\Sigma_{cov}^B &= \frac{1}{N}\sum_{k} n_k(\mu^{(k)}-\mu)(\mu^{(k)}-\mu)^T
\end{aligned} \tag{3}$$

where $x^{(k)}$ are examples of class k, n_k is the number of samples in this class, and $\mu^{(k)} = \frac{1}{n_k}\sum_x x^{(k)}$ is the class center.

3.2. Covariance in Pairwise Form

We can see from the definition in Eqn. (3) that Σ_{cov}^W and Σ_{cov}^B depend on the speaker label and class center, both of which are difficult to compute in the unlabeled target domain. Even though we could find pseudo-labels by clustering, the number of speakers is difficult to determine, and the clustering results may not be accurate enough in practice. We therefore look for solutions through easily-achievable pairwise difference computation – and will show that the Gram matrices of residuals for positive and negative pairs, denoted by Σ^+ and Σ^-, are equivalent to twice the within-class and between-class covariance.

3.2.1. Within-Class Covariance

Each element of within-class covariance Σ_{ij}^W represents the covariance between i-th and j-th dimensions of features within the same class:

$$\Sigma_{ij}^W = \underset{k,x}{E}\left[\left(x_i^{(k)}-m_i^{(k)}\right)\left(x_j^{(k)}-m_j^{(k)}\right)\right] \tag{4}$$

where $m^{(k)} = E_x[x^{(k)}]$ is the expectation for the k-th class. If we replace $m^{(k)}$ with another independent sampling $p^{(k)}$ from the same class, we can find it is equivalent to $2\Sigma_{ij}^W$ as follows

$$\begin{aligned}
\Sigma_{ij}^+ &= \underset{k,x,p}{E}\left[\left(x_i^{(k)}-p_i^{(k)}\right)\left(x_j^{(k)}-p_j^{(k)}\right)\right] \\
&= \underset{k,x,p}{E}\left[\left((x_i^{(k)}-m_i^{(k)})-(p_i^{(k)}-m_i^{(k)})\right)\right. \\
&\quad \left.\left((x_j^{(k)}-m_j^{(k)})-(p_j^{(k)}-m_j^{(k)})\right)\right] \\
&= \Sigma_{ij}^W - 0 - 0 + \Sigma_{ij}^W = 2\Sigma_{ij}^W
\end{aligned} \tag{5}$$

where

$$\underset{k,x,p}{E}\left[(x_i^{(k)}-m_i^{(k)})(p_j^{(k)}-m_j^{(k)})\right] = 0 \tag{6}$$

In this form, we do not explicitly use class centers. Only the residuals of positive pairs $x-p$ are needed to equivalently calculate the within-class covariance matrix.

3.2.2. Between-Class Covariance

Similarly, we can compute Σ_{cov}^B from the negative pairs. By definition, we can obtain:

$$\Sigma_{ij}^B = \underset{k}{E}\left[\left(m_i^{(k)}-m_i\right)\left(m_j^{(k)}-m_j\right)\right] \tag{7}$$

where $m = E_x[x]$ is the expectations of all the samples. Suppose $x^{(k)}$ and $n^{(l)}$ are two independent samples from different classes k and l, then we have

$$\Sigma_{ij}^- = \underset{k\neq l,x,n}{E}\left[\left(x_i^{(k)}-n_i^{(l)}\right)\left(x_j^{(k)}-n_j^{(l)}\right)\right] = 2\Sigma_{ij}^B \tag{8}$$

thus between-class covariance is obtainable from negative pairs.

3.2.3. WBDA loss

Assuming that each batch contains N_p positive pairs and N_n negative pairs in each domain, we first compute the residuals between sample pairs, denoted as R_p and R_n respectively. Then, the actual within- and between-class covariance matrices for each domain are calculated from Σ^+ and Σ^-, the Gram matrices of positive and negative residuals:

$$\begin{aligned}
\Sigma_{cov}^W &= \frac{1}{2}\Sigma^+ = \frac{1}{2N_p}R_pR_p^T \\
\Sigma_{cov}^B &= \frac{1}{2}\Sigma^- = \frac{1}{2N_n}R_nR_n^T
\end{aligned} \tag{9}$$

Note that if we only focus on the *direction* of the distribution, we can normalize the covariance matrix of both domains to compute the correlation matrix. In this work, we use the correlation matrix for the within-class part because it performs better in practice:

$$\begin{aligned}
\Sigma_{corr}^W &= \Sigma_{cov}^W \Big/ \sqrt{Diag\left(\Sigma_{cov}^W\right) Diag\left(\Sigma_{cov}^W\right)^T} \\
\Sigma_{corr}^B &= \Sigma_{cov}^B \Big/ \sqrt{Diag\left(\Sigma_{cov}^B\right) Diag\left(\Sigma_{cov}^B\right)^T}
\end{aligned} \tag{10}$$

where $Diag(\cdot)$ extracts the diagonal elements of the input matrix as a column vector, and $\sqrt{\cdot}$ gets the square root of the matrix elements.

The proposed WBDA loss aims to minimize the discrepancy in second-order statistics for within-class and between-class distributions respectively between different domains.

$$\mathcal{L}_{WBDA} = \lambda_W\|\Sigma_S^W - \Sigma_T^W\|_F^2 + \lambda_B\|\Sigma_S^B - \Sigma_T^B\|_F^2 \tag{11}$$

where $\|\cdot\|_F^2$ denotes the squared matrix Frobenius norm. Σ_S and Σ_T denote the statistics (*i.e.* correlation or covariance matrices computed in eqns (9),(10)) of the source and target domain, respectively, while λ_W and λ_B are the corresponding loss weight hyper-parameters.

4. Experiments

4.1. Experimental Settings

Datasets: Experiments are conducted on the NIST SRE16 and SRE18 CMN2. Training data primarily consists of telephone speech from past issues of NIST-SRE (2004-2010) plus Switchboard. The SRE16 task incorporates Tagalog and Cantonese telephone speech, and the SRE18 CMN2 task contains speech in Tunisian Arabic. In addition, a small development dataset

Table 1: *Cosine EER (%) results of the comparison systems on NIST-SRE 2016 and 2018 evaluation.*

System	SRE16			SRE18
	Pooled	Tagalog	Cantonese	CMN2
Baseline	12.74	17.75	7.85	12.16
CORAL [22]	9.77	13.92	5.81	10.44
MK-MMD [21]	10.01	13.68	6.06	10.59
WBDA	7.16	10.28	4.11	8.92

from the unlabeled target domain, with roughly 2k samples, is used to adapt systems. The Kaldi toolkit [27] is used to extract 41-dimensional FBank from 25ms windows with 10ms shift between frames. We apply mean-normalization over a 3s sliding window, and use voice activity detection (VAD) to remove silent segments. Training set features are randomly truncated into short slices ranging in length from 2 to 4s.

Model configuration: The baseline model uses the ResNet-34 backbone as in [6]. The number of heads in the attentive bilinear pooling (ABP) layer is set to 8 and the scaling factor after L2-norm is set to 30. The weight regularization [15] of domain-specific parameters is set to 0.01. The batchsize is 512, with each batch consisting of 128 positive pairs from the source and target domains. The networks are optimized using stochastic gradient descent (SGD), with momentum of 0.9 and weight decay of 5e-4. An initial learning rate of 0.1 is used to train the first 20 epochs, gradually declining to 0.0001 for the remaining 40 epochs. The loss weights λ_W and λ_B are set to 0 for the first 30 epochs, and their final value are selected by grid search, empirically making each loss comparable.

4.2. Main Results

Since the proposed WBDA is a DNN-based adaption method, we examine its effectiveness using a cosine distance measure, in terms of Equal Error Rate (EER). The main results are reported in Table 1, which compares against previously published domain discrepancy loss methods Deep CORAL [22] and Multi-Kernel MMD [21] applied to our baseline system. From the results, it can be seen that the WBDA loss performs significantly better than either Deep CORAL or MK-MMD loss for the same conditions. WBDA achieves a large relative EER reduction of 43.7% and 26.2% on SRE16 and SRE18, respectively. We believe the primary reason is that WBDA loss is able to achieve a finer class-level distribution alignment rather than just domain-level. This can transfer compact and discriminative class structure knowledge learned from a well-labeled source domain to the target domain, thus benefitting the adaptation.

Ablation Results: We conduct ablation experiments in which only one component of WBDA loss is involved, either within-class (or "W-") or between-class (or "B-"). We also study the effect of matching different second-order statics, *i.e.*, Covariance matrix (or "Cov"), and Correlation matrix (or "Corr"), in eqns (9),(10). The results are shown in Table 2.

First, we can see that EER on SRE16 improves from 12.73% to 7.7% with W-Corr. This indicates the important role of within-class distribution alignment which may improve the compactness of features in the target domain. It's interesting that the improvement achieved by W-Cov is much smaller than W-Corr, maybe because the magnitude of within-class perturbation can vary with the domain, and of course the positive pairs we construct in the target domain may not fully represent the true within-class distribution. Therefore, we find it more appropriate to focus on the *direction* of within-class distribution.

Table 2: *Cosine EER (%) results of ablation experiments with within- and between-class distribution loss*

System	SRE16			SRE18
	Pooled	Tagalog	Cantonese	CMN2
Baseline	12.74	17.75	7.85	12.16
W-Cov	11.76	16.43	7.17	11.91
W-Corr	7.70	10.86	4.33	9.24
B-Cov	9.81	13.90	5.77	10.51
B-Corr	11.57	15.56	7.26	11.48
WBDA	7.16	10.28	4.11	8.92

When matching between-class Covariance only, the performance of B-Cov is also better than the baseline, illustrating that between-class alignment may help to learn more discriminative features in the target domain. As expected, when combining both W-Corr and B-Cov, the performance of **WBDA** further improves, as it enables us to achieve a more comprehensive class-level distribution alignment.

Comparison with existing systems: In addition to the above evaluations, we compare the EER results with previous end-to-end adaptation systems using the same dataset in Table 3, where DANSE adopted an adversarial training strategy with a gradient reverse layer (GRL), LSGAN and FuseGan were two GAN-based systems, PSN performed adversarial training with partially shared network parameters, Mul-MMD minimized domain-wise MK-MMD loss on multiple layers, and APLDA refers to Kaldi's adaptive PLDA [27]. Thanks to the class-level alignment strategy in WBDA, our system achieves the best front-end performance, further demonstrating the effectiveness of our proposed method.

Table 3: *Comparison with other end-to-end UDA methods*

System	Metric	SRE16 Pooled	SRE18 CMN2
DANSE [12]	cosine	13.29	N/A
LSGAN [13]	cosine	11.74	N/A
FuseGAN [13]	cosine	10.88	N/A
PSN [15]	PLDA	8.98	N/A
WGAN [11]	PLDA	13.25	10.35
	APLDA	9.42	9.7
Mul-MMD[19]	PLDA	9.03	8.33
	APLDA	8.29	8.09
Baseline	cosine	12.74	12.16
WBDA	cosine	7.16	8.92

5. Conclusion

This paper has proposed a novel method to transfer class-aware information (*i.e.*, within- and between-class distributions), learned from a well-labeled source domain, to the unlabeled target domain. By separately constructing positive and negative pairs for source and target domain respectively, WBDA is able to effectively estimate the within- and between-class distributions. A deep domain-invariant embedding architecture can be learned in an end-to-end manner by jointly optimizing the cross-domain class-aware distribution discrepancy loss besides source-domain classification loss. Experimental results have demonstrated the superiority of the proposed WBDA method.

6. References

[1] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proc. Interspeech*, 2015, pp. 3214–3218.

[2] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey *et al.*, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. ICASSP*, 2018, pp. 5329–5333.

[3] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.

[4] Y. Liu, Y. Song, I. McLoughlin, L. Liu, and L.-r. Dai, "An effective deep embedding learning method based on dense-residual networks for speaker verification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6683–6687.

[5] P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet *et al.*, "Deep neural networks for extracting Baum-Welch statistics for speaker recognition," in *The Speaker and Language Recognition Workshop*, 2014, pp. 293–298.

[6] Y. Liu, Y. Song, Y. Jiang, I. McLoughlin, L. Liu, and L. Dai, "An effective speaker recognition method based on joint identification and verification supervisions," in *INTERSPEECH*, 2020, pp. 3007–3011.

[7] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.

[8] Z. Huang, S. Wang, and K. Yu, "Angular softmax for short-duration text-independent speaker verification." in *Interspeech*, 2018, pp. 3623–3627.

[9] J. Wang, K.-C. Wang, M. T. Law, F. Rudzicz, and M. Brudno, "Centroid-based deep metric learning for speaker recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3652–3656.

[10] C. Zhang and K. Koishida, "End-to-end text-independent speaker verification with triplet loss on short utterances." in *Interspeech*, 2017, pp. 1487–1491.

[11] J. Rohdin, T. Stafylakis, A. Silnova, H. Zeinali, and O. Plchot, "Speaker verification using end-to-end adversarial language adaptation," in *ICASSP 2019*, 2019.

[12] G. Bhattacharya, J. Alam, and P. Kenny, "Adapting end-to-end neural speaker verification to new languages and recording conditions with adversarial training," in *ICASSP 2019*. IEEE, 2019, pp. 6041–6045.

[13] G. Bhattacharya, J. Monteiro *et al.*, "Generative adversarial speaker embedding networks for domain robust end-to-end speaker verification," in *ICASSP 2019*. IEEE, 2019, pp. 6226–6230.

[14] S. Kataria, J. Villalba *et al.*, "Deep feature cyclegans: Speaker identity preserving non-parallel microphone-telephone domain adaptation for speaker verification," in *INTERSPEECH*, 2021, pp. 1079–1083.

[15] Z. Chen, S. Wang, and Y. Qian, "Adversarial domain adaptation for speaker verification using partially shared network." in *INTERSPEECH*, 2020, pp. 3017–3021.

[16] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, 2016.

[17] K. A. Lee, Q. Wang, and T. Koshinaka, "The coral+ algorithm for unsupervised domain adaptation of PLDA," in *ICASSP 2019*. IEEE, 2019, pp. 5821–5825.

[18] W.-W. Lin, M.-W. Mak *et al.*, "Reducing domain mismatch by maximum mean discrepancy based autoencoders." in *Odyssey*, 2018, pp. 162–167.

[19] W. Lin, M.-M. Mak, N. Li, D. Su, and D. Yu, "Multi-level deep neural network adaptation for speaker verification using MMD and consistency regularization," in *ICASSP 2020*. IEEE, 2020, pp. 6839–6843.

[20] W.-w. Lin, M.-W. Mak, and J.-T. Chien, "Multisource 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.

[21] A. Gretton, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, K. Fukumizu, and B. K. Sriperumbudur, "Optimal kernel choice for large-scale two-sample tests," *Advances in neural information processing systems*, vol. 25, 2012.

[22] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *European conference on computer vision*. Springer, 2016, pp. 443–450.

[23] S. H. Shum, D. A. Reynolds, D. GarciaRomero, and A. McCree, "Unsupervised clustering approaches for domain adaptation in speaker recognition systems," in *Proc. of Odyssey*, 2014, pp. 265–272.

[24] Z. Chen, S. Wang, and Y. Qian, "Self-supervised learning based domain adaptation for robust speaker verification," in *ICASSP*. IEEE, 2021, pp. 5834–5838.

[25] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[26] H.-R. Hu, Y. Song, Y. Liu, L.-R. Dai, I. McLoughlin, and L. Liu, "Domain robust deep embedding learning for speaker recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7182–7186.

[27] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, and N. G. et al, "The kaldi speech recognition toolkit," in *ASRU*, 2011.