Interspeech 2024
1-5 September 2024, Kos, Greece

Contrastive Learning and Inter-Speaker Distribution Alignment Based Unsupervised Domain Adaptation for Robust Speaker Verification

Zuoliang Li[1], *Wu Guo*[1], *Bin Gu*[1], *Shengyu Peng*[1], *Jie Zhang*[1,*]

[1] NERC-SLIP, University of Science and Technology of China, Hefei, China

lzl2001@mail.ustc.edu.cn, guowu@ustc.edu.cn, jzhang6@ustc.edu.cn

Abstract

Unsupervised domain adaptation (UDA) can tackle the mismatch between the source and target domains for real-world speaker verification applications. In this paper, we propose an UDA method by leveraging the target-domain data through a self-supervised method. Firstly, we use momentum contrastive learning to effectively utilize the latent speaker labels in the target domain, enhancing intra-speaker compactness and inter-speaker separability simultaneously. Secondly, we improve the inter-speaker feature distribution alignment loss, ensuring the stability of the source-domain statistics and mitigating the impact of false negative pairs. These two methods are further combined with conventional supervised learning in the source domain. Using Voxceleb2 as the source domain and CN-Celeb1 as the target domain, experimental results demonstrate the effectiveness of our proposed method.

Index Terms: speaker verification, unsupervised domain adaptation, contrastive learning, distribution alignment

1. Introduction

Automatic Speaker Verification (ASV) aims to verify the identity of a speaker in an acoustic signal [1, 2]. In recent years, the rapid development of deep neural network (DNN) has significantly improved the ASV performance [3–6]. However, training ASV models requires a large amount of labeled data, which is often unavailable in practice, resulting in a performance drop of ASV models that are trained in the source domain but tested in a new target domain. This is often-times called domain mismatch. Compared to the labeled data, unlabeled target-domain data is relatively easy for collection. A feasible approach to solve the domain mismatch problem is by adapting the model trained in the well-labeled source domain to the target domain using only a small amount of unlabeled target-domain data via unsupervised domain adaptation (UDA) [7, 8].

In literature, many UDA methods were proposed to tackle the domain mismatch issue. For example, in the embedding space domain adversarial training is a straightforward alternative to extract domain-invariant features by minimizing the loss of the domain discriminator [9–12]. Discrepancy-based approaches [13–15] aims to reduce the difference between the distributions of the source and target domains. These approaches align the global distributions of the two domains. To our knowledge, speaker identity clues, channel and environment characteristics are coupled in speech signals, and the in-between relation (e.g., in terms of features) is non-linear. Neglecting the relationship among the speakers in the target domain will inevitably undermine the discriminative ability of features [16], especially when the domain difference is relatively large.

In addition, it is also feasible to apply clustering techniques by generating pseudo-labels for the unlabeled data in the target domain and then using the pseudo-labels as supervised training [17], while the number of speakers is usually unknown in prior to clustering and the generated pseudo-labels certainly contain noises, sometimes leading to a negative transfer effect on the ASV.

Recently, self-supervised speaker verification was shown to be effective in dealing with large amounts of unlabeled data [18–21], which aims to learn discriminative speaker representations from input data directly without any supervision. For instance, most contrastive learning (CL) based self-supervised methods attempts to pull the representations of the same speaker closer and push that of different speakers further. This trick is naturally appropriate for constructing discriminative representations in the unlabeled target-domain context of ASV, as corroborated by previous studies [17, 22]. Motivated by this, in this work we propose a new UDA method to fully utilize the target domain data and latent speaker labels as well as to maximize speaker separability in the unlabeled target domain, concurrently adapt the speaker discriminative capabilities from the source domain.

As it was shown in [22] that joint training with labeled source-domain data and unlabeled target-domain data is beneficial for performance, the conventional source-domain supervised learning remains a pivotal component of the proposed method. For the self-supervised learning in the target domain, similarly to typical contrastive learning [23–25], we use the momentum contrast (MoCo) [24] framework to construct positive and negative pairs in the target domain, maximize the similarity between positive pairs (from the same speaker) and minimize the similarity between negative pairs (from different speakers), i.e., enhancing intra-speaker compactness and inter-speaker discrepancy in the embedding space. In order to transfer the discriminative ability from the source-domain supervised training to the target domain, we propose an improved inter-speaker feature distribution alignment between the two domains to filter false negative pairs in the target domain and transfer the stable distribution from the source domain. The remainder of this paper is structured as follows. Section 2 details the proposed method. Section 3 describes the experimental setup, followed by results in Section 4. Finally, Section 5 concludes this work.

2. Methodology

Figure 1 illustrates the structure of the proposed UDA method, which consists of three modules: supervised learning in the source domain, contrastive learning in the target domain and inter-speaker feature distribution alignment. For the supervised

*Correspondence: jzhang6@ustc.edu.cn

10.21437/Interspeech.2024-973

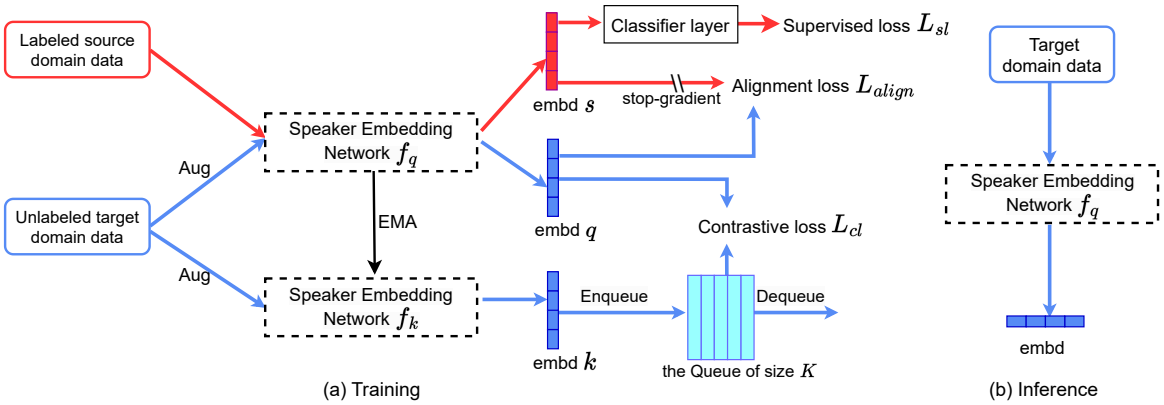

Figure 1: *An illustration of the proposed Contrastive Learning and Inter-speaker Distribution Alignment based UDA method: (a) training; (b) inference. For a source domain segment with a ground-truth label and a target domain segment without label, we employ the speaker embedding network f_q to extract features. Simultaneously, another target domain segment is fed into the network f_k.*

learning, we consider the methods in [6, 26] and use AAM-softmax as the loss function [27] in this work.

2.1. Momentum Contrast (MoCo) in the target domain

Both SimCLR [23] and MoCo [24] are successful representatives in the context of self-supervised speaker verification by optimizing the distance between positive (and/or negative) pairs. As there exist no speaker labels in the target domain, we treat two sub-segments from the same utterance as positive pairs as they are from the same speaker, two segments from different utterances are thus regarded as negative pairs. To improve the model robustness and avoid encoding undesired channel information into the representations [20], we perform different data augmentation on two sub-segments of the same utterance by adding noise or reverberation.

The SimCLR is optimized using the InfoNCE loss, where for each mini-batch of size N, $2N$ samples can be constructed via truncation and data augmentation with one positive and $2(N-1)$ negative samples. Due to the GPU memory, the SimCLR constructs negative samples exclusively within the current mini-batch, leading to a scarcity of negative samples and thus a poor capacity of separating negative instances in the embedding space. To introduce more negative samples for training, we therefore exploit the MoCo in this work.

The MoCo employs a dynamic dictionary to store negative samples in the training process. This dictionary is maintained as a queue of data samples, where the encoded embeddings of the current mini-batch are enqueued, and the oldest are dequeued to keep the queue size consistent. As depicted in Fig. 1(a), K is the size of the queue, which can be set to a relatively large number. The embeddings $\mathbf{q}_i$ and $\mathbf{k}_i$ are obtained using the two sub-segments from the i-th utterance within a mini-batch in the target domain as inputs for speaker embedding network f_q and f_k, respectively. To maintain the consistency of negative samples, the parameters θ_k of f_k is slowly updated by the Exponential Moving Average (EMA) of θ_q of f_q, given by

$$\theta_{\mathrm{k}} \leftarrow m\theta_{\mathrm{k}} + (1-m)\theta_{\mathrm{q}}, \tag{1}$$

where the momentum coefficient m is set to 0.999 in experiments. The InfoNCE loss is used for training the speaker embedding network, which is defined as

$$\mathcal{L}_{\mathrm{MoCo}} = -\frac{1}{N} \times \sum_{i=1}^{N} \log \frac{\exp\left(\mathbf{q}_i' \cdot \mathbf{k}_i^+ / \tau\right)}{\exp\left(\mathbf{q}_i' \cdot \mathbf{k}_i^+ / \tau\right) + \sum_{j=1}^{K} \exp\left(\mathbf{q}_i' \cdot \mathbf{k}_j' / \tau\right)}, \tag{2}$$

where $\mathbf{q}_i' = \mathbf{q}_i / \|\mathbf{q}_i\|_2$ and $\mathbf{k}_i^+ = \mathbf{k}_i / \|\mathbf{k}_i\|_2$ are two normalized embeddings of positive pairs, $\mathbf{k}_j'$ the negative samples of $\mathbf{q}_i'$ from the queue, '$\cdot$' donates the dot product. As depicted in Fig. 1(b), we only use the speaker embedding network f_q to extract speaker embeddings at the inference stage.

2.2. Inter-speaker feature distribution alignment

We use second-order statistics of negative pairs to measure the difference of feature distributions between the two domains, enhance inter-speaker discrepancy in the target domain and compute the final loss function.

2.2.1. *Equivalent form of inter-speaker covariance*

It was shown in [28] that the inter-speaker covariance can be computed using negative pairs. For a speaker embedding $\mathbf{x} = (x_1, ..., x_D)^T$ of dimension D, letting $\mathbf{x}^{(k)}$ and $\mathbf{n}^{(l)}$ denote the embeddings from speaker k and l, respectively, the inter-speaker covariance $\mathbf{\Sigma}$ can be computed equivalently using the residuals of the negative pairs $\mathbf{x}^{(k)}$ and $\mathbf{n}^{(l)}$ as

$$\Sigma_{ij} = \frac{1}{2}\Sigma_{ij}^{-} = \frac{1}{2} \underset{k \neq l, \mathbf{x}, \mathbf{n}}{E} \left[\left(x_i^{(k)} - n_i^{(l)} \right) \left(x_j^{(k)} - n_j^{(l)} \right) \right], \tag{3}$$

where Σ_{ij} is the (i,j)-th entry of $\mathbf{\Sigma}$, and $x_i^{(k)}$ and $x_j^{(k)}$ are the i-th and j-th element of $\mathbf{x}^{(k)}$, similarly for $n_i^{(k)}$ and $n_j^{(k)}$, and $E(\cdot)$ denotes mathematical expectation.

Given N_n negative pairs in the mini-batch from two domains, (3) can be reformulated as

$$\mathbf{\Sigma} = \frac{1}{2N_n} \mathbf{R}_n \mathbf{R}_n^T \tag{4}$$

where $\mathbf{R}_n \in \mathbb{R}^{D \times N_n}$ is the matrix that consists of the residuals $\mathbf{x}^{(k)} - \mathbf{n}^{(l)}$ of all the negative pairs.

2.2.2. Improved inter-speaker distribution alignment

In the source domain, constructing negative pairs is straightforward using the oracle labels. However, in the absence of labels in the target domain, false negative pairs are inevitable, as different utterances may be erroneously considered as negative pairs. Moreover, the inter-speaker distribution of the source domain is susceptible to being influenced by the target-domain counterpart. This might degrade the model performance and lead the stability of the source-domain distribution to fluctuate during training.

To mitigate the impact of false negative pairs, we design a thresholding strategy for sample filtering, where the threshold η is obtained from the cosine similarity between the positive pairs within the current mini-batch in (2) as

$$\eta = 0.8 \times \frac{1}{N} \sum_{i=1}^{N} \mathbf{q}_i' \cdot \mathbf{k}_i^+. \tag{5}$$

Taking the threshold into account, (3) can be rewritten as

$$\begin{aligned} \Sigma_{ij} = \frac{1}{2} \underset{k \neq l, \mathbf{x}, \mathbf{n}}{E} \left[\left(x_i^{(k)} - n_i^{(l)} \right) \left(x_j^{(k)} - n_j^{(l)} \right) \right], \\ \text{s.t.} \quad \mathbf{x} \cdot \mathbf{n} < \eta. \end{aligned} \tag{6}$$

where $\mathbf{x}$ and $\mathbf{n}$ are two normalized embeddings.

To prevent the inter-speaker distribution in the source domain from being altered by the target domain, we employ a stop-gradient operation on the source domain branch. To do this, we define an improved loss for feature distribution alignment as

$$\mathcal{L}_{\text{align}} = \lambda \| \text{sg}[\mathbf{\Sigma}_S] - \mathbf{\Sigma}_T \|_F^2, \tag{7}$$

where $\text{sg}[\cdot]$ indicates the stop-gradient, λ is a weighting parameter, $\| \cdot \|_F^2$ denotes the Frobenius norm, $\mathbf{\Sigma}_S$ and $\mathbf{\Sigma}_T$ represent the inter-speaker covariance matrices in the source and target domains, respectively.

In addition, in order to ensure the stability of the source-domain statistics, $\mathbf{\Sigma}_S$ is updated during the training process via moving average as

$$\mathbf{\Sigma}_{S_t} \leftarrow 0.5 \times \mathbf{\Sigma}_{S_{t-1}} + 0.5 \times \frac{1}{2N_n} \mathbf{R}_n \mathbf{R}_n^T, \tag{8}$$

where the subscript t is the iteration index.

The final loss function is then formulated as

$$\mathcal{L} = \mathcal{L}_{\text{sl}} + \mathcal{L}_{\text{MoCo}} + \mathcal{L}_{\text{align}}, \tag{9}$$

where $\mathcal{L}_{\text{sl}}$ is the AAM-softmax loss for supervised training in the source domain.

3. Experimental setup

3.1. Datasets

In experiments, we use the development set of VoxCeleb2 [29] (denoted as Vox2 for brevity) as the source-domain data and CN-Celeb1 [30] (CN1) as the target domain. All speech files are sampled at a rate of 16 kHz. VoxCeleb2 comprises over 1 million utterances from 5,994 speakers, mostly spoken in English. CN-Celeb1 is a Mandarin corpus, where 797 and 200 speakers are used for training and evaluation, respectively. As many utterances in CN-Celeb1 are short, we consider combinations in the training set to create utterances longer than 5 seconds. This results in 58,276 utterances in total. We do not incorporate any speaker label or speaker number information of CN-Celeb1 during training. To simulate a variety of environments and reduce the effects of non-speaker components, we add noise and reverberation in both the source and target domains using the MUSAN corpus [31] and room impulse response (RIR) [32]. Additionally, speed perturbation is applied in the source domain.

Table 1: *The EER(%) of supervised training, where 'SL' means supervised learning and 'FT' fine-tuning on CN1.*

Method	Training data	Model	
		ECAPA	ResNet
SL	Vox2	13.02	11.19
SL	CN1	11.34	10.66
SL+FT	Vox2+CN1	8.28	8.11

The official trial list of the CN-Celeb1 test set contains 3,484,292 test pairs, which is scored using the cosine similarity. The ASV performance on CN-Celeb1 is evaluated in terms of equal error rate (EER).

3.2. Implementation

For all the training data, we randomly truncate speech files into 2-seconds segments. Note that we use two sub-segments from the same utterance from the target-domain data as the input to the model. The 80-dimensional log-mel filter banks are extracted as speech features with a frame shift of 10 ms and a window length of 25 ms.

The well-known ECAPA-TDNN [6] (ECAPA) and ResNet34 [26] (ResNet) are considered as the speaker embedding networks. The channel size of ECAPA-TDNN is set to 1,024 and the embedding dimension is 192. The embedding dimension of ResNet34 embedding is set to 256. The AAM-Softmax loss for supervised training is defined with a margin of 0.2, scale of 30 for ECAPA-TDNN and 32 for ResNet34. The temperature parameter τ in (2) is set to 0.07, the weight hyper-parameter λ in (7) is set to 0 for the beginning 30 epochs and then increases to 5 for the rest epochs. The batch size for both the source and target domains is 128. The SGD optimizer is utilized to update the model parameters to minimize the total loss, with a momentum of 0.9 and weight decay of 0.0001. The implementation is based on the Wespeaker toolkit [33] using 2 NVIDIA RTX 3090 GPUs.

4. Results

4.1. Results of supervised training

Table 1 shows the results of supervised training in the source and target domains. The top row shows that in case ECAPA-TDNN and ResNet are trained on VoxCeleb2 and evaluated on the CN-Celeb1 test set, the EERs are 13.02% and 11.19%, respectively. This indicates a high mismatch between the two domains. In the following experiments, these two supervised systems are used as the baseline. The middle row lists the performance of supervised training on the CN-Celeb1 training set, which is better than the baseline but still not good enough due to the insufficient data even without domain mismatch. In the case of supervised training on both VoxCeleb2 and CN-Celeb1 training sets, the best EER (8.28% and 8.11% on ECAPA-TDNN and ResNet, respectively) is obtained. This result shows the up limit of performance for the ASV system when trained on these

Table 2: *Results of contrastive learning in target domain and supervised training in source domain on ECAPA-TDNN.*

Training loss	Queue size	EER (%)
$\mathcal{L}_{\text{sl}} + \mathcal{L}_{\text{MoCo}}$	K =16384	10.31
	K =32768	10.17
	K =65536	10.16
	K =131072	**10.14**

Table 3: *Performance comparison with existing methods independent on the target-domain speaker labels and number.*

Method	Model	EER (%)
EDITnet [34]	ECAPA	12.06
EDITnet [34]	SE-ResNet	9.60
CORAL in [34]	ECAPA	12.63
CORAL in [34]	SE-ResNet	10.45
SSDA [22]	ResNet	10.20
$\mathcal{L}_{\text{sl}} + \mathcal{L}_{\text{MoCo}} + \mathcal{L}_{\text{align}}$ (Proposed UDA)	ECAPA	**9.65**
	ResNet	**9.23**

two datasets, revealing the importance of both the data amount and domain consistency.

4.2. Impact of the size of queue in MoCo

As in Section 2.1, we used a queue to store the data samples in the MoCo, we then compare the performance with different K to select the optimal queue size. For simplicity but without loss of generality, we show the impact on using the ECAPA-TDNN model and using the VoxCeleb2 and unlabeled CN-Celeb1 training data as the training set.

The obtained results are presented in Table 2. It is clear that for the MoCo, with an increase in the queue size, the ASV model performance in terms of EER improves, this observation suggests that constructing more negative pairs in the target domain enhances the model's capability to distinguish negative instances in the embedding space, thereby enhancing the model's discriminative ability in the target domain. However, it's worth noting that the performance is not very sensitive to the number of negative samples. Note that a larger queue size would also consume a higher computational complexity. We thus set K to 65,536 for MoCo in the sequel.

4.3. Results of UDA

Further, we compare the proposed UDA method with existing methods, including CORAL [34, 35], EDITnet [34] and self-supervised learning-based domain adaptation (SSDA) [22]. Note that the results of comparison methods are directly drawn from the published papers. It is clear in Table 3 that the proposed UDA method can achieve obvious improvements over different baselines. Our proposed method maximizes the utilization of unlabeled target domain data and latent speaker labels, taking into account the internal structure of features in the target domain, ensuring the feature discrimination ability in the target domain. Simultaneously, it fully leverages the source domain data, leading to the best performance on ResNet.

Furthermore, the proposed method outperforms the contrastive approaches. In [22], Chen et al. proposed SSDA to use both supervised loss in the source domain and self-supervised loss in the target domain, leveraging the potential label information from unlabeled target domain data. However, SSDA constructs negative pairs within the current mini-batch, lacking a sufficient number of negative pairs, which results in the model's limited ability to separate negative instances in the embedding space. In [34], Li et al. proposed EDITnet, a method that transfers embeddings from the target domain to the source domain using a conditional variational auto-encoder, but EDITnet does not sufficiently take into account the internal structure of features, leading to the generation of lower quality and less controllable samples. Additionally, in the experiments detailed in [34], CORAL utilizes covariance to transfer the target distribution to the source distribution, which only aligns the global distribution between the two domains and does not ensure that the learned features are discriminative in the target domain. As we can see in Table 3, the proposed UDA method can achieve 9.23% EER in ResNet model, outperforming all the contrastive methods from 3.8% to 27% EER reduction.

4.4. Ablation study

Table 4: *The EER(%) of results of ablation experiment results. 'CN1.U' means CN-Celeb1 without data labels.*

Training loss	Training data	Model	
		ECAPA	ResNet
$\mathcal{L}_{\text{sl}} + \mathcal{L}_{\text{MoCo}} + \mathcal{L}_{\text{align}}$	Vox2+CN1.U	**9.65**	**9.23**
$\mathcal{L}_{\text{sl}} + \mathcal{L}_{\text{MoCo}}$	Vox2+CN1.U	10.16	9.57
$\mathcal{L}_{\text{sl}} + \mathcal{L}_{\text{align}}$	Vox2+CN1.U	17.61	15.58
$\mathcal{L}_{\text{sl}}$	Vox2	13.02	11.19
$\mathcal{L}_{\text{MoCo}}$	CN1.U	14.10	14.57

As the proposed method incorporates the MoCo in the target domain and feature distribution alignment in addition to the source-domain supervised training, we finally conduct ablation experiments to assess the impact of each module in Table 4. We can see from Table 4 that the removal of any loss component leads to an obvious performance drop, implying the necessity of these modules. And in the experiments, excluding the MoCo results in significantly degraded model performance. Feature distribution alignment aims to minimize discrepancies between source and target domain features but neglects to learn the discriminative abilities for speakers, it must be integrated with the MoCo loss for better performance, since MoCo facilitates the learning of discriminative features of speakers.

We also conducted an experiment using only unsupervised training with the MoCo module in the target domain, and the results are listed in the last row. While the performance is worse than supervised training on CN-Celeb1, it remains a viable approach for learning in the absence of labels.

5. Conclusions

In this paper, we propose leveraging the MoCo framework to effectively utilize unlabeled target-domain data and latent speaker labels in the target domain, ensuring the discriminative capability of learned features. Additionally, we introduce an improved inter-speaker featuer distribution alignment method, adapting the speaker discrimination ability learned from the source domain. Experimental results demonstrate the superiority and reliability of our proposed method, outperforming existing strategies.

6. References

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

[2] Z. Bai and X.-L. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Networks*, vol. 140, pp. 65–99, 2021.

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

[4] B. Gu, J. Zhang, and W. Guo, "A dynamic convolution framework for session-independent speaker embedding learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[5] B. Gu, W. Guo, and J. Zhang, "Memory storable network based feature aggregation for speaker representation learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 643–655, 2023.

[6] 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.

[7] P.-M. Bousquet and M. Rouvier, "On robustness of unsupervised domain adaptation for speaker recognition," in *Proc. Interspeech*, 2019.

[8] A. Misra and J. H. Hansen, "Maximum-likelihood linear transformation for unsupervised domain adaptation in speaker verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1549–1558, 2018.

[9] J. Rohdin, T. Stafylakis, A. Silnova, H. Zeinali, L. Burget, and O. Plchot, "Speaker verification using end-to-end adversarial language adaptation," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*. IEEE, 2019, pp. 6006–6010.

[10] Q. Wang, W. Rao, S. Sun, L. Xie, E. S. Chng, and H. Li, "Unsupervised domain adaptation via domain adversarial training for speaker recognition," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*. IEEE, 2018, pp. 4889–4893.

[11] L. Li, M.-W. Mak, and J.-T. Chien, "Contrastive adversarial domain adaptation networks for speaker recognition," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[12] W. Xia, J. Huang, and J. H. Hansen, "Cross-lingual text-independent speaker verification using unsupervised adversarial discriminative domain adaptation," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*. IEEE, 2019, pp. 5816–5820.

[13] W.-W. Lin, M.-W. Mak, L. Li, and J.-T. Chien, "Reducing domain mismatch by maximum mean discrepancy based autoencoders." in *Odyssey*, 2018, pp. 162–167.

[14] 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.

[15] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*. Springer, 2016, pp. 443–450.

[16] S. Yao, Q. Kang, M. Zhou, M. J. Rawa, and A. Albeshri, "Discriminative manifold distribution alignment for domain adaptation," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 53, no. 2, pp. 1183–1197, 2022.

[17] H. Mao, F. Hong, and M.-w. Mak, "Cluster-guided unsupervised domain adaptation for deep speaker embedding," *IEEE Signal Processing Letters*, 2023.

[18] T. Lepage and R. Dehak, "Label-efficient self-supervised speaker verification with information maximization and contrastive learning," *arXiv preprint arXiv:2207.05506*, 2022.

[19] C. Zhang and D. Yu, "C3-dino: Joint contrastive and non-contrastive self-supervised learning for speaker verification," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1273–1283, 2022.

[20] H. Zhang, Y. Zou, and H. Wang, "Contrastive self-supervised learning for text-independent speaker verification," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*. IEEE, 2021, pp. 6713–6717.

[21] Y. Chen, S. Zheng, H. Wang, L. Cheng, and Q. Chen, "Pushing the limits of self-supervised speaker verification using regularized distillation framework," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*. IEEE, 2023, pp. 1–5.

[22] Z. Chen, S. Wang, and Y. Qian, "Self-supervised learning based domain adaptation for robust speaker verification," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*. IEEE, 2021, pp. 5834–5838.

[23] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[24] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[25] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *International Conference on Machine Learning*. PMLR, 2021, pp. 12 310–12 320.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[27] 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.

[28] H.-R. Hu, Y. Song, L.-R. Dai, I. McLoughlin, and L. Liu, "Class-aware distribution alignment based unsupervised domain adaptation for speaker verification." in *Proc. Interspeech*, 2022, pp. 3689–3693.

[29] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, p. 101027, 2020.

[30] Y. Fan, J. Kang, L. Li, K. Li, H. Chen, S. Cheng, P. Zhang, Z. Zhou, Y. Cai, and D. Wang, "Cn-celeb: a challenging chinese speaker recognition dataset," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*. IEEE, 2020, pp. 7604–7608.

[31] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[32] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*. IEEE, 2017, pp. 5220–5224.

[33] H. Wang, C. Liang, S. Wang, Z. Chen, B. Zhang, X. Xiang, Y. Deng, and Y. Qian, "Wespeaker: A research and production oriented speaker embedding learning toolkit," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*. IEEE, 2023, pp. 1–5.

[34] J. Li, W. Liu, and T. Lee, "EDITnet: A Lightweight Network for Unsupervised Domain Adaptation in Speaker Verification," in *Proc. Interspeech*, 2022, pp. 3694–3698.

[35] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, no. 1, 2016.