Interspeech 2024
1-5 September 2024, Kos, Greece

Collaborative Contrastive Learning for Hypothesis Domain Adaptation

Jen-Tzung Chien[1], *I-Ping Yeh*[1], *Man-Wai Mak*[2]

[1]Inst. of Electrical and Computer Engr., National Yang Ming Chiao Tung University, Taiwan
[2]Dept. of Electrical and Electronic Engr., Hong Kong Polytechnic University, Hong Kong SAR
{jtchien,ping629.cs10}@nycu.edu.tw, man.wai.mak@polyu.edu.hk

Abstract

Achieving desirable performance for speaker recognition with severe domain mismatch is challenging. Such a challenge becomes even more harsh when the source data are missing. To enhance the low-resource speaker representation, this study deals with a practical scenario, called hypothesis domain adaptation, where a model trained on a source domain is adapted to a significantly different target domain as a hypothesis without access to source data. To pursue a domain-invariant representation, this paper proposes a novel collaborative hypothesis domain adaptation (CHDA) where the dual encoders are collaboratively trained to estimate the pseudo source data which are then utilized to maximize the domain confusion. Combined with the constrastive learning, this CHDA is further enhanced by increasing the domain matching as well as the speaker discrimination. The experiments on cross-language speaker recognition show the merit of the proposed method.

Index Terms: Domain adaptation, speaker verification, contrastive learning, collaborative learning

1. Introduction

Speaker verification aims to verify a person's identity based on his/her voice. In recent years, deep neural networks have played a pivotal role to achieve a remarkable success in speaker verification [1, 2, 3]. However, speaker models are susceptible to the robustness issue when faced with domain mismatches. Domain mismatches in speaker verification can arise from various variations in the collected speech data [4] due to background noise, recording equipment, language difference, etc. This paper deals with the practical applications where a pre-trained model is provided but the source data for training the model are absent. Domain adaptation (DA) methods are explored to mitigate the impact of domain mismatches in speaker verification.

To deal with the domain mismatch, many domain adaptation methods have been proposed for speaker verification. There are two common DA approaches to reducing the distribution shift between target and source domains. The first approach imposes the source domain information on the target data. This can be achieved by explicitly matching the feature distribution of the target domain with that of the source domain. For instance, a method called correlation alignment was proposed to minimize the domain shift by aligning the second-order statistics of source and target distributions [5]. In [6, 7], the maximum mean discrepancy was performed to reduce the domain mismatch by minimizing the mean-squared difference of the statistical features between two domains. These approaches leveraged the statistical measures to align the distributions between source and target domains, thereby reducing the mismatch and facilitating the domain adaptation. Other studies have been extended to use domain adversarial training [8, 9, 10], which involved training a domain classifier, also known as a domain critic, along with the primary task for speaker recognition. The speaker classifier and domain classifier were jointly trained in an adversarial manner to learn domain-invariant features that can generalize to the target domain. However, the data distributions in both domains were required. The overly matching in distribution might inadvertently reduce the capability of speaker discrimination. Besides, it is worth noting that the previous methods utilized the data distributions from source and target domains, which raise the concerns regarding data privacy and data portability. Thus, it is necessary to develop a practical method to efficiently adapt an existing model trained on source domain data to a new target domain where no source data are available.

This paper presents a collaborative contrastive learning method for hypothesis domain adaptation to tackle the challenges in source-free cross-language speaker recognition. Motivated by the deep Q-network (DQN) [11, 12] as a model-free, online, off-policy reinforcement learning (RL) method, a target network, copied from the Q-network, was used to approximate the Q-function and stabilize the training process of DQN by allowing the target network to be controlled while incorporating the most recent changes to the Q-network. Following the idea of collaborating two networks in DQN, this paper proposes a collaborative contrastive learning for domain adaptation and speaker recognition where the dual encoders are smoothly learned to enhance speaker embeddings in a target domain with unknown mismatches with the source domain.

2. Advances in Domain Adaptation

2.1. Hypothesis domain adaptation

Traditional domain adaptation methods required the access to both labeled source data and unlabeled target data to implement domain adaptation. However, in a real scenario, it is often impractical to have the access to source data. It is emerging to develop the hypothesis domain adaptation (HDA) where such an access is waived [13, 14]. In [13], the source hypothesis transfer was proposed to employ a pseudo labeling strategy that combined the information maximization with the entropy minimization to adapt the trained classifier to fit the target features. In [14], the perturbations to model parameters through variational Bayesian inference were introduced to maintain the discriminative power of a model while performing model adaptation. A number of HDA methods have employed a strategy where the source classifier was frozen during the adaptation process to preserve the class information [15, 16, 17]. Pseudo-labels were then assigned to the target data based on the classifier's outputs. This was because, in a typical HDA, a closed set was assumed,

10.21437/Interspeech.2024-1800

i.e., the source and target label spaces are the same. However, the source and target label spaces are different in speaker verification due to the fact that speakers from different recordings or languages are inherently distinct. Therefore, the presence of domain mismatch causes a significant challenge in developing a robust speaker verification system.

2.2. Contrastive domain adaptation

Contrastive learning has shown remarkable advances in unsupervised or self-supervised learning representation [18, 19, 20] by effectively learning the instance discrimination. For visual learning representation, some studies have explored data augmentation methods to generate positive and negative paired samples for contrastive learning. Typically, the goal of contrastive learning is to maximize the similarity between positive pairs of samples while minimizing the similarity between negative pairs of samples. Some recent works reduced the domain discrepancy through minimization of a contrastive loss [21, 22, 23]. In [21], the class-wise contrastive learning was utilized to bridge the inter-domain gap to achieve contrastive alignment between the original input image and the strongly augmented target images. In [24], the generalization of contrastive learning was shown to be closely related to three key factors, including the alignment of positive samples, the divergence between class centers, and the concentration of augmented data. In this study, we leverage speaker representation learning and present a new data augmentation to enrich the learned embeddings to be robust to severe mismatch.

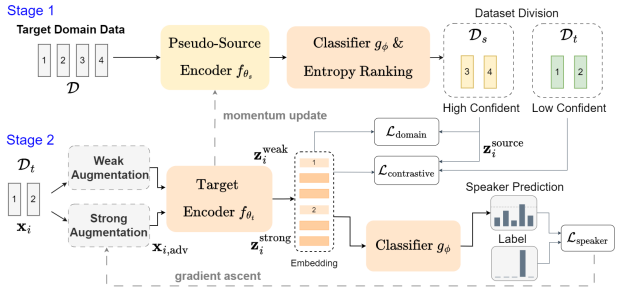

Figure 1: *Collaborative hypothesis domain adaptation for speaker recognition with two encoders and one classifier. The numbers in $\mathcal{D}_s$ and $\mathcal{D}_t$ indicate the speaker identities.*

3. Collaborative Hypothesis Adaptation

This study presents the collaborative hypothesis domain adaptation (CHDA) for source-free cross-language speaker verification. The overall process is depicted in Figure 1 with two stages. The first stage is to partition the target domain data into two parts with high and low confidence according to the prediction likelihood based on the source model. A pseudo-source encoder and a classifier are used. The second stage is to treat individual parts of samples with specific objectives for collaborative [25] and contrastive domain matching, which encourages to learn a discriminative target representation through both adaptation methods in instance level and distribution level where data augmentation with a target encoder is implemented. The process involves speaker representation learning, collaborative domain matching, and contrastive domain adaptation.

3.1. Speaker representation learning

Speaker verification is a task that involves accepting or rejecting an identity claim based on a speech sample. Numerous speaker verification systems have adopted the classification objective during model training. Our objective is to train the target encoder f_{θ_t} from labeled data in target domain. f_{θ_t} transforms the input $\mathbf{x}_i$ in $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ to an ℓ_2-normalized embedding vector $\mathbf{z}_i = f_{\theta_t}(\mathbf{x}_i)$. The classifier g_ϕ is used to output the predicted speaker. The training objective for speaker loss is calculated through the additive angular margin (AAM) loss [26], which is a margin-based softmax method. AAM-softmax optimizes the geodesic distance margin directly by leveraging the precise correspondence between angle and arc on the normalized hypersphere. Let $\mathbf{x}_i \in \mathbb{R}^d$ denote an original input speech sample with speaker label y_i, the target encoder θ_t and classifier ϕ are trained by minimizing the AAM-softmax loss as

$$\mathcal{L}_{\text{speaker}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{e^{s\cos(\phi_{y_i}+\alpha)}}{e^{s\cos(\phi_{y_i}+\alpha)}+\sum_{j=1,y_j\neq y_i}^{B}e^{s\cos(\phi_j)}} \tag{1}$$

where B is the batch size, s denotes the radius of a hypersphere of learned embeddings, $\cos\phi_{y_i}$ denotes the target logit, which is the dot product between the normalized class-weight vector and the normalized embedding vector $\mathbf{x}_i$, and α is an additive angular margin that increases intra-class compactness and inter-class disparity. A classification loss is minimized.

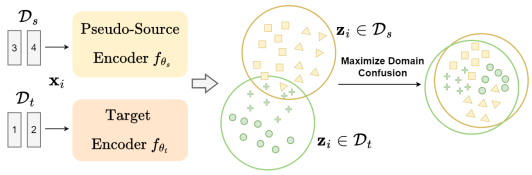

Figure 2: *Collaborative dual encoders for domain matching.*

3.2. Collaborative domain matching

Vanilla DA directly minimizes the domain discrepancy between source and target domains for domain matching. In contrast, HDA follows a different scenario where only a pre-trained source model is available, and there is no access to source data. HDA copes with the problem of unavailable source data by splitting the target domain data. Thus, CHDA tackles the task of transferring a source-trained encoder to fit a new target domain by employing two distinct encoders, including a target encoder f_{θ_t} and a pseudo-source encoder f_{θ_s}, which were both pre-trained on source domain data. The target and pseudo-source encoders have different updating methods. The pseudo-source encoder aims to generate embeddings that are closely related to the source domain, thereby maximizing the domain confusion through the data population generated by both encoders. It is important to point out that the pseudo-source encoder has been pre-trained on source domain data, so it is possible to simulate the source domain distribution by exploring the encoder's outputs. Figure 2 shows how collaborative domain matching works in the proposed CHDA framework.

First, following the prediction confidence or entropy score by using f_{θ_s}, the mini-batches of training samples are divided into a source-relevant set $\mathcal{D}_s$ and a source-irrelevant set $\mathcal{D}_t$

$$\mathcal{D}_t = \{\mathbf{x}_i \mid \text{top-}K(-p_i \log p_i), \forall \mathbf{x}_i \in \mathcal{D}\} \tag{2}$$

where $p_i = g_\phi(f_{\theta_s}(\mathbf{x}_i))$ is the prediction probability, K is the size of source-irrelevant set $\mathcal{D}_t$, and the source-relevant samples $\mathcal{D}_s$ are the remaining target data, where $\mathcal{D} = \{\mathcal{D}_s, \mathcal{D}_t\}$. Given the subsets $\mathcal{D}_s$ and $\mathcal{D}_t$, a kind of collaboration [27] between dual encoders $\{f_{\theta_s}, f_{\theta_t}\}$ is performed to encourage domain matching by minimizing the Kullback-Leibler (KL) divergence between source-relevant and source-irrelevant embeddings

$$\mathcal{L}_{\text{domain}} = \mathbb{E}_{\mathbf{x}_s \sim \mathcal{D}_s, \mathbf{x}_t \sim \mathcal{D}_t}[D_{\text{KL}}(f_{\theta_s}(\mathbf{x}_s) \| f_{\theta_t}(\mathbf{x}_t))]. \quad (3)$$

However, it was found that fixing the pseudo-source encoder during training resulted in poor performance. We hypothesize that this failure is caused by a rapidly changing target encoder, which reduces the consistency of the embeddings for the same instance. Therefore, this study adopts the momentum scheme, which was popular in optimization [28], into domain adaptation to stabilize the learning process through a smoothing function. Similar to the momentum in gradient updating, the parameter updating of pseudo-source encoder is done iteratively by

$$\theta_s \leftarrow m\theta_s + (1-m)\theta_t, \qquad m \in [0, 1) \quad (4)$$

where m is a momentum coefficient. Notably, only the parameter of target encoder θ_t is updated by back-propagation, and the momentum updating in Eq. (4) makes θ_s evolve more smoothly than θ_t. Such a smoothed domain adaptation addresses the potential generation of overly aggressive embeddings resulting from the excessively distribution matching. This strategy guides the embeddings to enhance the adaptation process through the collaborative dual encoders. Similar to the stabilization in DQN for RL, the pseudo-source encoder is a smoothed copy of target encoder to stabilize the collaborative learning.

3.3. Collaborative contrastive learning

To leverage the information from target domain data without access to source data, collaborative contrastive learning is proposed to enhance the cross-language speaker representation. Figure 3 shows how this learning strategy is performed. The idea is to pull together an anchor sample and a positive sample in the embedding space, and push apart this anchor from those negative samples. In the implementation, the weak and strong augmented data from source-irrelevant samples $\mathcal{D}_t$ are generated. For contrastive learning [29], the speaker characteristics in embedding vectors should remain consistent across different augmentations for the same instance, while the embeddings for different instances should exhibit distinct characteristics.

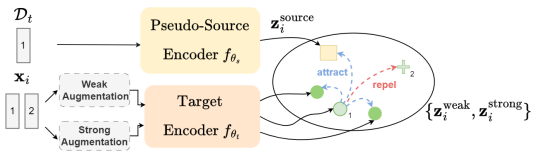

Figure 3: *Collaborative contrastive domain adaptation.*

Therefore, two different augmentation strategies are considered. For the original utterance in the target domain, the speaker embedding $\mathbf{z}_i$ via target encoder f_{θ_t} is viewed as an anchor. For weak and strong augmentation utterances, the embeddings $\mathbf{z}_i^{\text{weak}}$ and $\mathbf{z}_i^{\text{strong}}$ based on target encoder are obtained, respectively, and viewed as the positive samples. The embedding $\mathbf{z}_i^{\text{source}}$ via pseudo-source encoder f_{θ_s} can be also viewed as a positive sample. Thus, a set of positive samples $\mathcal{P}_i = \{\mathbf{z}_i^{\text{weak}}, \mathbf{z}_i^{\text{strong}}, \mathbf{z}_i^{\text{source}}\}$ corresponding to an anchor $\mathbf{z}_i$ are accessible. Then, the remaining utterances in the same mini-batch are treated as the negative samples. The contrastive loss [18] for domain adaptation is accordingly yielded by

$$\mathcal{L}_{\text{contrastive}} = -\frac{1}{K}\sum_{i=1}^{K}\sum_{\mathbf{z}^+ \in \mathcal{P}_i} \log \frac{e^{\text{sim}(\mathbf{z}_i, \mathbf{z}^+)/\tau}}{\sum_{j=1}^{K} e^{\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau}} \quad (5)$$

where $\text{sim}(\mathbf{z}_i, \mathbf{z}_j) = \mathbf{z}_i^\top \mathbf{z}_j / \|\mathbf{z}_i\| \|\mathbf{z}_j\|$, K is the size of source-irrelevant samples, and τ denotes a temperature parameter. By incorporating the embeddings from the pseudo-source encoder that are closely associated with source domain into the contrastive objective, the hybrid adaptation process encourages the matching not only at the distribution level as seen in Eq. (3) but also at the instance level as given in Eq. (5).

In this study, the weak augmentation was performed by randomly using various additive noises and reverberations based on MUSAN [30] and RIR [31] datasets, respectively. For strong augmentation, the adversarial augmentation method using the projected gradient descent (PGD) [32] was implemented. PGD is a well-known adversarial attack where the forward and backward processes are run to compute the perturbation δ by gradient ascent which is finally added to the input $\mathbf{x}$ as $\mathbf{x}_{\text{adv}} \leftarrow \mathbf{x} + \delta$. Given the logarithm of Mel filterbanks of original speech utterance $\mathbf{x}$ with speaker label y and the parameters of target encoder θ_t and classifier ϕ, the adversarial perturbation δ is estimated as a worst-case projection via a multi-step gradient ascent by maximizing the loss $\mathcal{L}_{\text{speaker}}(\mathbf{x}_{\text{adv}}, y)$. Therefore, the two-stage parameter learning for target encoder θ_t, pseudo-source encoder θ_s and speaker classifier ϕ is performed by minimizing $\mathcal{L}_{\text{chda}} = \mathcal{L}_{\text{speaker}} + \mathcal{L}_{\text{domain}} + \mathcal{L}_{\text{contrastive}}$ through Algorithm 1.

Algorithm 1 Collaborative hypothesis domain adaptation

Input: target domain data $\mathcal{D}$, target encoder θ_t, pseudo-source encoder θ_s, classifier ϕ, epoch number N, augmentation strategy $\mathcal{A}$
Output: target encoder θ_t
initialize θ_t, θ_s by pre-trained source encoder
for epochs $n = 0, \ldots, N-1$ **do**
  sample a mini-batch $\{\mathbf{x}_i, y_i\}_{i=1}^{B} \sim \mathcal{D}$
  Stage 1 ($\mathcal{D}, \theta_s, \phi$)
  compute $\mathcal{L}_{\text{speaker}}(\mathbf{x}, y)$ by Eq. (1)
  update speaker classifier ϕ with $\mathcal{L}_{\text{speaker}}$
  divide $\mathcal{D}$ into $\mathcal{D}_s = \{\mathbf{x}_s, y_s\}$, $\mathcal{D}_t = \{\mathbf{x}_t, y_t\}$
  Stage 2 ($\mathcal{D}_s$,$\mathcal{D}_t$, θ_s,θ_t, ϕ)
  compute $\mathcal{L}_{\text{speaker}}(\mathbf{x}_t, y_t)$ by Eq. (1)
  compute $\mathcal{L}_{\text{domain}}(\mathbf{x}_s, \mathbf{x}_t)$ by Eq. (3)
  compute $\mathcal{L}_{\text{contrastive}}(\mathbf{x}_s, \mathbf{x}_t, \mathcal{A}(\mathbf{x}_t))$ by Eq. (5)
  compute $\mathcal{L}_{\text{chda}} = \mathcal{L}_{\text{speaker}} + \mathcal{L}_{\text{domain}} + \mathcal{L}_{\text{contrastive}}$
  update target encoder θ_t with $\mathcal{L}_{\text{chda}}$
  update pseudo-source encoder θ_s by Eq. (4)
end for
return θ_t

4. Experiments

4.1. Experimental settings

This study conducted the experiments on using VoxCeleb2 [33] and CNCeleb [34]. VoxCeleb2 comprised more than 1 million utterances from over 6,000 celebrities, extracted from YouTube videos, in both English and Spanish. CNCeleb was a mainland Mandarin accent speech dataset collected from bilibili videos. CNCeleb contains more than 130K utterances from 1,000 Chinese celebrities, covering 11 diverse domains, and each speaker might have speech samples in multiple domains. Compared with VoxCeleb (the source domain), the size of CNCeleb (the

target domain) is much smaller. The encoder was initially pre-trained on VoxCeleb and subsequently transferred to CNCeleb. The performance was evaluated on CNCeleb, which comprises 18,024 target pairs and 3,586,776 non-target pairs. The backbone speaker model is an ECAPA-TDNN [35] with 512 channels. The input features consist of 80-dimensional log Mel-filterbanks. The embedding with a dimension of 192 was extracted from the final layer before AAM-softmax layer. The Adam optimizer was used to update model parameters with the weight decay parameter 2e-5 and initial learning rate of 0.001. The number of training epochs was set to 15. The AAM's loss scale s and the temperature parameter τ in Eq. (5) were set to 32 and 0.07, respectively, and K was set such that $\mathcal{D}_t$ constitutes 80% of the target domain data $\mathcal{D}$.

For evaluation, the verification score was determined by calculating the cosine similarity between the pairs of test and enrollment segments and then averaging them. The metrics of equal error rate (EER) and minimum detection cost function (minDCF) were measured. Test data in CNCeleb were used for evaluation. Target-domain adaptation data were used for domain matching via CHDA without access to source data in VoxCeleb. Only the pre-trained backbone model was used for computing the speaker embedding vectors.

4.2. Experimental results

The proposed CHDA is evaluated as shown in Table 1. CHDA is compared with the baseline with EER=12.35%, which involves only AAM-softmax loss $\mathcal{L}_{\text{speaker}}$ but without backbone model from VoxCeleb2. When comparing the performance of the same architecture on the test set of VoxCeleb (EER=1.01%) [35] with that on CNCeleb, it is obvious that the limited and heterogeneous CNCeleb is seen as a challenging dataset for speaker verification. When the pre-trained encoder from VoxCeleb was evaluated on the CNCeleb data, an EER of 13.74% was obtained. This result highlights a considerable domain mismatch between the two datasets even though VoxCeleb is much larger than CNCeleb. The pre-trained encoder from VoxCeleb fine-tuned on CNCeleb using a classification loss $\mathcal{L}_{\text{speaker}}$ (11.64%) performs better than the standard training (13.74%). This is because the pre-trained encoder already contained the knowledge to distinguish speakers. When fine-tuning on the target domain, the encoder does not have to start from scratch.

In this comparison, the proposed CHDA obtains the lowest EER (9.68%). Language mismatch between VoxCeleb and CNCeleb is compensated by domain adaptation via fine-tuning. Furthermore, comparison with CHDA without (w/o) different loss terms was conducted to demonstrate the benefit of the proposed learning strategy. Results show that removing domain discrepancy loss results in the most significant drop of EER.

Table 1: *Performances on the test set of CNCeleb.*

Pre-training	Adapt loss function	EER(%)	minDCF
CNCeleb	None	12.35	0.6038
VoxCeleb	None	13.74	0.5971
VoxCeleb	$\mathcal{L}_{\text{speaker}}$	11.64	0.6379
VoxCeleb	CHDA w/o $\mathcal{L}_{\text{domain}}$	10.42	0.5282
VoxCeleb	CHDA w/o $\mathcal{L}_{\text{contrastive}}$	10.19	0.5506
VoxCeleb	CHDA	**9.68**	**0.5091**

In Table 2, the proposed CHDA is further compared with the hypothesis domain adaptation (HDA) in [36]. The latter focused on self-supervised contrastive domain adaptation, where the encoder uses ResNet34. However, simply using self-supervised learning may lose the speaker discrimination on the target domain, so the method in [36] further adopted the labeled data in the source domain to conduct joint training using speaker classification loss $\mathcal{L}_{\text{speaker}}$. In the self-supervised setting, the encoder was trained on CNCeleb by using the generalised end-to-end (GE2E) objective (13.34%), which is slightly worse than standard supervised training (12.35%). This result indicates that supervised training outperforms self-supervised training in speaker verification. In supervised setting, both VoxCeleb and CNCeleb data were simultaneously fed into the encoder during training, where VoxCeleb2 was used to calculate the classification loss $\mathcal{L}_{\text{speaker}}$, while CNCeleb was used in the GE2E loss. It is found that the proposed CHDA is superior to the other HDA variants by using self-supervised training on the target domain (CNCeleb) even without access to source-domain data.

Table 2: *Comparison with the related method in [36]. 'SF' denotes source-free or adaptation without access to source data.*

Source loss function	Adapt loss function	SF	EER(%)
$\mathcal{L}_{\text{speaker}}$	GE2E	✗	10.24
None	GE2E	✓	13.34
None	CHDA	✓	**9.68**

Table 3: *Performances of applying CHDA target encoder θ_t by using different m for updating the pseudo-source encoder θ_s.*

Parameter update for θ_s	EER(%)	minDCF
No update	11.36	0.6092
Momentum update ($m = 0.2$)	10.00	0.4996
Momentum update ($m = 0.3$)	9.77	**0.4995**
Momentum update ($m = 0.4$)	**9.68**	0.5091
Momentum update ($m = 0.5$)	10.08	0.5539

This paper proposes using the momentum update to address the severe inconsistency between the embeddings of the same speaker produced by two different encoders, which interferes the adaptation process. The performance of different momentum values m is examined as shown in Table 3. When the pseudo-source encoder is used as the source-trained encoder and then fixed during training (the first row), the results deteriorate. The best result is obtained when m=0.4. The embeddings produced by the pseudo-source encoder using this value of m contain valuable source information while also maintaining a manageable gap between source and target domains. A lower momentum value results in a weaker influence of the source-trained encoder on the target domain, leading to a diminished performance in capturing the characteristics of source data.

5. Conclusions

This paper presented a novel collaborative hypothesis domain adaptation method for speaker verification that leveraged dual encoders with different updating strategies to address cross-language source-free domain adaptation. The key idea is to exploit the concept of momentum updating and incorporate it into domain adaptation, which reduced the inconsistencies of embeddings from two different encoders while preserving the source information. The target domain data were partitioned according to the prediction confidence. This method implements the distribution-level and instance-level domain adaptation and matching. The experiments on speaker recognition show that the proposed method outperforms the related methods and mitigates the language domain mismatch without the access to source domain data.

6. References

[1] M.-W. Mak and J.-T. Chien, *Machine learning for speaker recognition*. Cambridge University Press, 2020.

[2] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 5329–5333.

[3] J.-T. Chien and K.-T. Peng, "Neural adversarial learning for speaker recognition," *Computer Speech & Language*, vol. 58, pp. 422–440, 2019.

[4] S. O. Sadjadi, T. Kheyrkhah, A. Tong, C. Greenberg, D. Reynolds, E. Singer, L. Mason, and J. Hernandez-Cordero, "The 2016 NIST speaker recognition evaluation," in *Proc. of Annual Conference of International Speech Communication Association*, 2017.

[5] M. J. Alam, G. Bhattacharya, and P. Kenny, "Speaker verification in mismatched conditions with frustratingly easy domain adaptation." in *Proc. of Speaker and Language Recognition Workshop*, 2018, pp. 176–180.

[6] W.-W. Lin, M. W. Mak, L. Li, and J.-T. Chien, "Reducing domain mismatch by maximum mean discrepancy based autoencoders," in *Proc. of Speaker and Language Recognition Workshop*, 2018, pp. 162–167.

[7] 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.

[8] 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*, vol. 33, no. 5, pp. 2236–2245, 2022.

[9] Y. Tu, M. W. Mak, and J. T. Chien, "Variational domain adversarial learning for speaker verification," in *Proc. of Annual Conference of International Speech Communication Association*, 2019, pp. 4315–4319.

[10] Y. Tu, M.-W. Mak, and J.-T. Chien, "Variational domain adversarial learning with mutual information maximization for speaker verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2013–2024, 2020.

[11] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing Atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[12] N. Vieillard, B. Scherrer, O. Pietquin, and M. Geist, "Momentum in reinforcement learning," in *Proc. of International Conference on Artificial Intelligence and Statistics*, 2020, pp. 2529–2538.

[13] J. Liang, D. Hu, and J. Feng, "Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation," in *Proc. of International Conference on Machine Learning*, 2020, pp. 6028–6039.

[14] M. Jing, X. Zhen, J. Li, and C. Snoek, "Variational model perturbation for source-free domain adaptation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 17 173–17 187, 2022.

[15] J. N. Kundu, N. Venkat, R. V. Babu *et al.*, "Universal source-free domain adaptation," in *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4544–4553.

[16] Y. Liu, W. Zhang, and J. Wang, "Source-free domain adaptation for semantic segmentation," in *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1215–1224.

[17] Z. Zhang, W. Chen, H. Cheng, Z. Li, S. Li, L. Lin, and G. Li, "Divide and contrast: Source-free domain adaptation via adaptive contrastive learning," in *Advances in Neural Information Processing Systems*, 2022.

[18] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[19] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. of International Conference on Machine Learning*, 2020, pp. 1597–1607.

[20] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[21] A. Singh, "CLDA: Contrastive learning for semi-supervised domain adaptation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 5089–5101, 2021.

[22] G. Kang, L. Jiang, Y. Yang, and A. G. Hauptmann, "Contrastive adaptation network for unsupervised domain adaptation," in *Proc. of IEEE/CVF Conference on Computer Vision and Pattern recognition*, 2019, pp. 4893–4902.

[23] S. Tang, P. Su, D. Chen, and W. Ouyang, "Gradient regularized contrastive learning for continual domain adaptation," in *Proc. of AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 2665–2673.

[24] W. Huang, M. Yi, X. Zhao, and Z. Jiang, "Towards the generalization of contrastive self-supervised learning," in *Proc. of International Conference on Learning Representations*, 2023.

[25] J.-T. Chien and W.-H. Chang, "Collaborative regularization for bidirectional domain mapping," in *Proc. of International Joint Conference on Neural Networks*, 2021, pp. 1–8.

[26] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[27] J.-T. Chien and C.-C. Chen, "Collaborative pseudo labeling for prompt-based learning," in *Proc. of Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, 2023, pp. 51–56.

[28] N. Qian, "On the momentum term in gradient descent learning algorithms," *Neural Networks*, vol. 12, no. 1, pp. 145–151, 1999.

[29] Y. Tu, M.-W. Mak, and J.-T. Chien, "Contrastive self-supervised speaker embedding with sequential disentanglement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2704–2715, 2024.

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

[31] 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 *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 5220–5224.

[32] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. of International Conference on Learning Representations*, 2018.

[33] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," *Proc. of Annual Conference of International Speech Communication Association*, pp. 1086–1090, 2018.

[34] 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 *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 7604–7608.

[35] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *Proc. of Annual Conference of International Speech Communication Association*, pp. 3830–3834, 2020.

[36] Z. Chen, S. Wang, and Y. Qian, "Self-supervised learning based domain adaptation for robust speaker verification," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 5834–5838.