INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Adversarial Domain Adaptation for Speaker Verification using Partially Shared Network

Zhengyang Chen, Shuai Wang, Yanmin Qian[†]

MoE Key Lab of Artificial Intelligence
SpeechLab, Department of Computer Science and Engineering
AI Institute, Shanghai Jiao Tong University, Shanghai
{zhengyang.chen, feixiang121976, yanminqian}@sjtu.edu.cn

Abstract

Speaker verification systems usually suffer from large performance degradation when applied to a new dataset from a different domain. In this work, we will study the domain adaption strategy between datasets with different languages using domain adversarial training. We introduce a partially shared network based domain adversarial training architecture to learn an asymmetric mapping for source and target domain embedding extractor. This architecture can help the embedding extractor learn domain invariant feature without sacrificing the ability on speaker discrimination. When doing the evaluation on cross-lingual domain adaption, the source domain data is in English from NIST SRE04-10 and Switchboard, and the target domain data is in Cantonese and Tagalog from NIST SRE16. Our results show that the usual adversarial training mode will indeed harm the speaker discrimination when the source and target domain embedding extractors are fully shared, and in contrast the newly proposed architecture solves this problem and achieves $\sim$25.0% relative average Equal Error Rate (EER) improvement on SRE16 Cantonese and Tagalog evaluation.

Index Terms: Adversarial Training, Domain Adaption, Partially Shared Weights, Speaker Verification

1. Introduction

The speaker verification task, which aims to verify a user's claimed identity given his or her speech segment, has gained significant improvement since the deep neural network (DNN) based speaker embedding was proposed. Researchers have investigated different DNN architectures [1, 2, 3, 4] and different loss functions [5, 6, 7, 8, 9, 10, 11] to enhance the discrimination of DNN based speaker embeddings.

Despite the success of DNN embeddings for speaker verification, DNN training usually requires a huge amount of well-annotated data with speaker labels. On the other hand, we know that the performance of a model trained from one domain will degrade dramatically when applied to a different domain where the data distribution is not the same. Training domain-specific models for each application scenario is a naive solution, however, collecting and labeling data for each domain is time-consuming and very expensive. So it is necessary to find an effective method to fast adapt an existing model trained on a well-labeled source domain dataset to a new target domain in which only the weakly-labeled or even unlabeled data is available.

Different approaches have been proposed to tackle the domain adaption problem for speaker verification, where the most commonly used one is utilizing the adversarial learning to make the representation domain-invariant and reduce the mismatch between the source and target domain data. The mismatch may be from different channels, noise types, and languages, etc. For instance, [12, 13, 14, 15] proposed to use channel adversarial training to make the speaker embeddings more channel-invariant. Similar ideas could also be found in [16, 17, 18]. However, in most of the current work, the data from the source and target domain share the same feature extractor, which might be sub-optimal. For example, in [12], it is non-trivial to make the adversarial trained network consistently outperform the baseline. More recently, some researchers from the computer vision community tried to use different feature extractors for the source and target data, while sharing parts of the parameters [19], and obtained consistent improvements on some image-related tasks. Accordingly, we are inspired to apply a similar idea to enhance speaker embedding with adversarial learning, which can be very useful for speaker verification.

In this paper, we show that the fully shared network indeed hurts the discrimination of the learned speaker embeddings, and a partially shared neural network architecture is designed and introduced to address this problem. The impact of different weight sharing strategies is thoroughly explored on NIST SRE 16 dataset. Domain mismatch problem is one main focus of recent NIST evaluations (NIST SRE16 and SRE18), and SRE16 [20] mainly focuses on the mismatch between different languages. In this setup, the source domain data is in English from NIST SRE04-10 and Switchboard, while the target domain data is Tagalog and Cantonese from NIST SRE16. Thus, in this paper, our proposed methods are evaluated on this cross-lingual speaker verification task, while they can also be easily extended to other domain mismatch scenarios. The main contributions of this paper are described as follows:

- Wasserstein GAN (WGAN) loss is used for adversarial training, aiming to learn domain invariant embeddings.
- Different from the fully shared feature extractor for both source and target domain, a partially shared network based domain adversarial training is designed and introduced to generate better representations for speaker verification task.
- The impact of different weight sharing strategies is fully explored for speaker verification, and it shows that sharing either lower of higher layers is better than the other positions. The best strategy gives a large relative $\sim$25.0% EER reduction on standard NIST SRE16 evaluation.

[†] corresponding author

http://dx.doi.org/10.21437/Interspeech.2020-2226

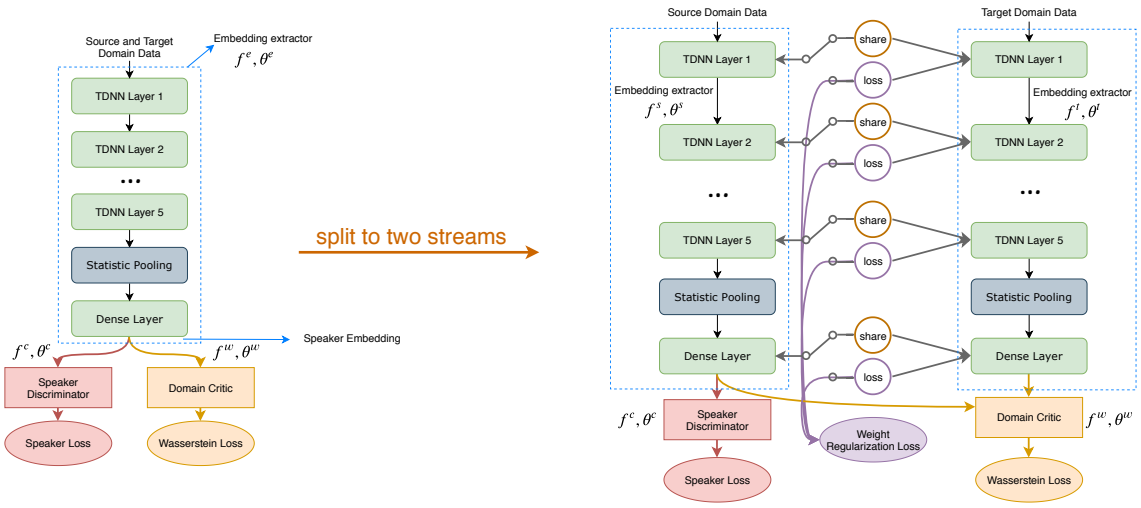

Figure 1: *Left:* ***Fully Shared Network (FSN)*** *with adversarial training criterion, and data from the source and target domains share the same embedding extractor; Right:* ***Partially Shared Network (PSN)*** *with adversarial training criterion. Two parallel embedding extractors are designed for data from the source and target domain, while the weight of the corresponding layers can be shared or constrained with a weight regularization loss.*

.

2. Partially Shared Network for Adversarial Learning

2.1. Fully Shared Network

For a typical domain adversarial architecture, a common feature extractor is used to learn domain-invariant features with the supervision of the adversarial training loss. Such a strategy is investigated for speaker embedding learning in [12]. As shown in Figure. 1 (left), we used f^e, f^c, f^w to denote embedding extractor, speaker discriminator and domain critic [21], which are parameterized by θ^e, θ^c and θ^w respectively. We assume a labeled source dataset $X^s = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$, and an unlabeled target domain dataset $X^t = \{(x_i^t)\}_{i=1}^{n_t}$, where x denotes utterances and y denotes speaker labels. And the total loss of Fully Shared Network (FSN) is defined below:

$$\mathcal{L}_{FSN} = \mathcal{L}_c + \lambda_w \mathcal{L}_w \tag{1}$$

where $\mathcal{L}_c$ is the normal cross entropy loss defined as $\mathcal{L}_c = \mathtt{CE}(f^c(f^e(x_i)), y_i)$, and $\mathcal{L}_w$ is WGAN loss [22] defined as:

$$\mathcal{L}_w = \mathcal{L}_{wd} + \gamma \mathcal{L}_{grad} \tag{2}$$

where $\mathcal{L}_{wd}$ is Wasserstein distance defined as:

$$\mathcal{L}_{wd} = f^w(f^e(x^s)) - f^w(f^e(x^t)) \tag{3}$$

$\mathcal{L}_{grad}$ represents the 1-Lipschitz constraint on the gradient of domain critic's parameters, which makes $\mathcal{L}_{wd}$ as an approximation of the Wasserstein distance,

$$\mathcal{L}_{grad}(\hat{h}) = \left(\left\| \nabla_{\hat{h}} f_w(\hat{h}) \right\|_2 - 1 \right)^2 \tag{4}$$

where $\hat{h}$ is the linear combination of a paired h^s ($h^s = f^e(x^s)$) and h^t ($h^t = f^e(x^t)$) [22].

2.2. Partially Shared Network

2.2.1. Model Architecture

Instead of fully sharing the embedding extractor, in this paper, we propose the partially shared network. As shown in Figure. 1 (right), two parallel embedding extractors are adopted for the data from the source domain and target domain, respectively. The parameters at the same layer position from two branches are either shared or not. Source and target domain data are fed to the two streams separately to generate the embeddings.

2.2.2. Loss Function

In the partially shared network (PSN), the common embedding extractor f^e defined in FSN will be split to parallel extractors f^s and f^t, which are parameterized by θ^s and θ^t respectively. We use θ_j^s and θ_j^t to denote the parameters of the j^{th} layer (not including the statistic pooling layer). Besides $\mathcal{L}_{wd}$ and $\mathcal{L}_{grad}$ in the loss of FSN, another weight regularization loss is integrated to constrain the weight distribution of θ^s and θ^t. The total loss of PSN is defined in equation 5.

$$\mathcal{L}_{PSN} = \mathcal{L}_c + \lambda_w \mathcal{L}_w + \lambda_r \mathcal{L}_r \tag{5}$$

where $\mathcal{L}_r$ is defined as

$$\mathcal{L}_r = \sum_{j \in \Omega} \left[exp\left(\left\| \theta_j^s - \theta_j^t \right\|^2 \right) - 1 \right] \tag{6}$$

The $\mathcal{L}_r$ loss constrains the θ^t to be similar to θ^s, which is used to avoid target extractor overfitting on domain-invariant features learning task and losing speaker-discriminative ability. The definition of $\mathcal{L}_r$ is modified from the exponential form weight regularization loss in [19], in which the exponential calculation can punish harder on the difference between θ_j^s and θ_j^t, and we removed the linear transformation in the original definition because it makes the training unstable in our experiments. Ω is the set of layers and defined as $\Omega = \{1 \cdots 6\}$ in the x-vector based architecture.

2.2.3. Training Algorithm

The whole training procedure is shown in Algorithm 1, which can be divided into two iterative steps. In the first step, the WGAN domain critic is trained for multiple iterations so that

the domain critic network can discriminate the embedding from different domains. Then the speaker classification loss and the well-trained domain critic network will guide the embedding extractor to learn speaker-discriminative and domain-invariant embeddings.

Algorithm 1: Partially Shared Network for Adversarial Training

Initialize source and target domain embedding extractors, speaker discriminator and domain critics parameterized by θ^s, θ^t, θ^c and θ^w.
repeat
  Sample minibatch $\{(x_i^s, y_i^s)\}, (x_i^t)\}$
  Step 1
  for $k = 1, ..., n$ **do**
    $h^s \leftarrow f^s\left(x^s\right), h^t \leftarrow f^t\left(x^t\right)$
    Sample $\hat{h}$ as the random points between h^s and h^t pairs.
    $\hat{h} = \eta h^s + (1-\eta)h^t, \eta \in (0,1)$
    $\theta^w \leftarrow$
    $\theta^w + \alpha \nabla_{\theta^w}\left[\mathcal{L}_{wd}\left(x^s, x^t\right) - \gamma \mathcal{L}_{grad}(\hat{h})\right]$
  Step 2
  $\theta^c \leftarrow \theta^c - \alpha \nabla_{\theta^c} \mathcal{L}_c\left(x^s, y^s\right)$
  $\theta^s \leftarrow \theta^s - \alpha \nabla_{\theta^s}\left[\mathcal{L}_c\left(x^s, y^s\right) + \lambda_r \mathcal{L}_r(\theta^s, \theta^t)\right]$
  $\theta^t \leftarrow \theta^t - \alpha \nabla_{\theta^t}\left[\lambda_r \mathcal{L}_r(\theta^s, \theta^t) + \lambda_w \mathcal{L}_{wd}(x^t)\right]$
until *Reaching max iteration*;

3. Experimental Setup

3.1. Dataset

Audios from previous NIST-SRE evaluations (2004-2010) and Switchboard Cellular are used to train the baseline system. The same data augmentation strategy following [3] is applied. We randomly select $128,000$ augmented data and add them to the clean speech. After that, the silent parts are removed using an energy-based voice activity detector. Besides, we remove the utterances less than 0.5s and speakers with fewer than eight utterances. Finally, there are 4805 speakers consisting of 193551 utterances left.

For doing adversarial training, we also augment the SRE16 major data following the strategy in [3]. We combine all the augmented copies with the clean speech, ending up with 11360 recordings. These recordings will be considered as the target domain data when doing adversarial training and the data illustrated in the above paragraph will be considered as the source domain data.

3.2. System Configuration

23-dimensional MFCC features extracted using Kaldi [23] are used for the neural network training. The training utterances are cut into 2s-4s chunks, whereas the whole utterance will be used to extract embedding during the evaluation period. Our baseline system is a standard x-vector using the same configuration as in [3], and the whole training pipeline follows Kaldi SRE16 recipe.

The same x-vector architecture used in the baseline is adopted for the embedding extractor in both FSN and the two parallel extractors in the proposed PSN, containing five TDNN layers and a dense layer. The embedding extractors for both FSN and the proposed PSN are initialized with the well-trained baseline x-vector system. The domain critic network is a simple feed-forward network with the dimension 512 x 512 x 512 x 1, while ReLU [24] is used as the non-linearity function. The domain critic network is initialized randomly. We set the parameters of the adversarial training in Algorithm 1 to $\gamma = 10.0$, $\alpha = 0.0001$, $\lambda_w = 0.1$, $\lambda_r = 0.001$ and $n = 5$.

150-dimensional LDA is first used to reduce the embedding dimension, after which PLDA is used for scoring. Both the LDA and PLDA are trained on the NIST SRE04-10 dataset. Besides, the evaluations data is centered using the mean of the NISE SRE16 unlabeled development set.

4. Results and discussion

In our proposed partially shared network, the corresponding layers of the two parallel extractors could be either shared or not shared but constrained by a regularization loss. In the training phase, two modes are performed and compared: 1) jointly train both the source and target extractors; 2) fix the source extractor and only update the target extractor.

4.1. Mode#1: Jointly train the source and target extractors

With both extractors trainable, the results of different weight sharing strategies could be found in Figure 2.

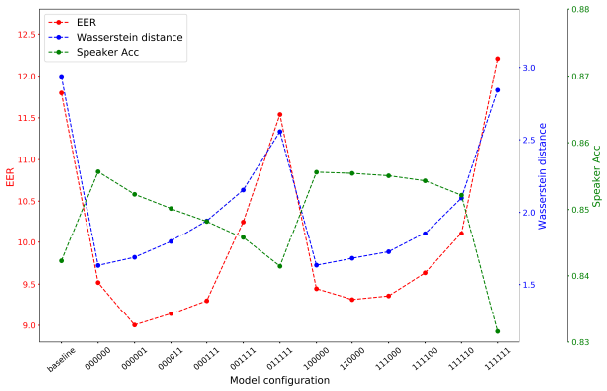

Figure 2: *The results of different weight sharing strategies with jointly training the source and target extractors, and EER (%) denotes the pooled results on SRE16. On the $x-$axis,* 1 *or* 0 *denotes whether or not to share the weights of the corresponding layer (from the lowest to the highest layer, low means close to the input layer), e.g.* 100000 *means only the parameters of the lowest layer is shared.*

The x-vector baseline only trained on the source domain data achieves 11.81% EER, when the parameters of the source and target extractors are fully shared, the domain adversarial trained network (correspond to configuration 111111 in Figure 2) obtains even worse EER at 12.21%. Similar performance degradation is also observed in [12] with the usual fully shared structure. Moreover, the speaker classification accuracy during training of this configuration is the lowest, too, which means imposing domain invariance may hurt speaker discrimination via simply shared the whole embedding extractor for different domain data.

The speaker accuracy represents the speaker discrimination ability and the Wasserstein distance denotes the mismatch extent of data from the source and target domain. As expected,

the system performance in terms of EER is clearly positively correlated to the speaker accuracy and negatively correlated to the Wasserstein distance.

It's interesting to see that only sharing the lowest layers (100000 or 110000) or the highest layers (000001 or 000011) can significantly boost the system performance and more detailed results are shown in Table 1.

Table 1: *Results with different partially shared configurations.*

System	Config	EER(%)		
		Pooled	Cantonese	Tagalog
baseline	-	11.81	8.36	15.38
FSN	111111	12.21	7.30	17.20
PSN	100000	9.45	5.73	13.25
	110000	9.32	5.55	13.16
	000001	9.00	5.43	12.67
	000011	9.14	5.38	12.98

These results are not very consistent with the findings in [19], in which the best configuration occurs in the condition that when the first or last few layers are unshared. And the possible explanation is that these good configurations in Figure. 2 all achieve great speaker accuracy improvement, which may play an important part in final EER promotion in target domain. And another experiment avoiding the influence of speaker classification accuracy change is analysed in the next section.

4.2. Mode#2: Fix the source extractor

Since the main task of this paper is to compensate the domain mismatch, we decide to keep the speaker discriminative ability of the source extractor by freezing its parameters and focusing on optimizing the Wasserstein distance. The results are illustrated in Figure. 3.

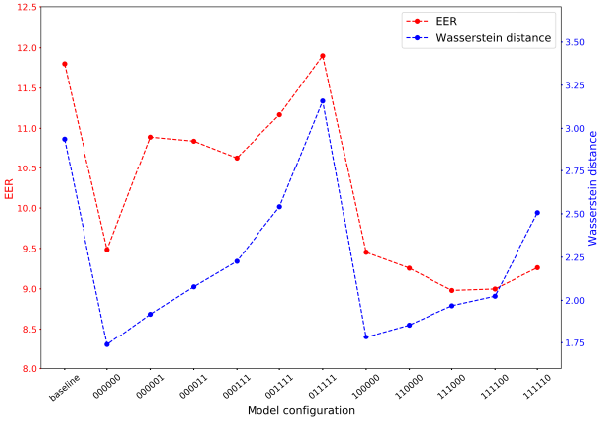

Figure 3: *The results comparison of different model configurations. In this experimental setup, the parameters of the source domain embedding extractor are fixed.*

Results show that the less layers are shared, the more similar distribution (smaller Wasserstein distance) can be achieved between source and target domain speaker embeddings. This observation means the carefully selected parameters for the source domain data is not suitable for the target domain data, demanding a different set of parameters to learn the difference. Better results are obtained when the top layers of the embedding extractor are not shared. Besides, the results in Figure. 2 and Figure. 3 both show that only untying the embedding extractor weights at the higher layers, i.e. the last layer or the last two layers can obtain good performance. A possible explanation the high-level information such as language is mainly abstracted in the higher layers, so it's helpful to keep different parameters at high layers for the two extractors.

The best system with partially shared network based adversarial training proposed in this paper and normally fully shared model are compared in Table 2. The best configuration of partially shared weights architecture outperforms the baseline by a large margin, ∼25.0% relative improvement on the pooled EER compared to the baseline system.

Table 2: *Results comparison using different weight sharing strategies.*

System	EER(%)		
	Pooled	Cantonese	Tagalog
baseline	11.81	8.36	15.38
FSN	12.21	7.30	17.20
PSN	**8.98**	**5.18**	**12.90**

4.3. The impact of the regularization loss

Finally, we explore the effectiveness of the weight regularization loss. The results are shown in Table 3. We can find that when λ_w is small, e.g. $\lambda_w = 0.1$, the weight regularization loss contributes very little to the final improvement. But when λ_w is larger, e.g. $\lambda_w = 1.0$, the model almost loses the discriminative ability on the speaker embedding without the weight regularization. So, the weight regularization loss makes the model more robust to the other hyper-parameters and plays an important role when keeping the speaker discrimination of the target domain embeddings.

Table 3: *Results with or without weight regularization. The model configuration corresponds to 111000 in Figure. 3.*

λ_w	λ_r	EER(%)		
		Pooled	Cantonese	Tagalog
1.0	0.0	26.72	26.60	26.84
	0.01	9.35	5.42	13.36
0.1	0.0	9.08	5.29	13.03
	0.01	**8.98**	**5.18**	**12.90**

5. Conclusion

This paper introduces the partially shared network based adversarial training architecture to do cross-lingual domain adaptation. Compared to the fully shared network, except for learning domain invariant embeddings, the partially shared network can learn more speaker-discriminative embeddings. And the proposed method outperforms the x-vector baseline with a large gain of ∼25.0% relative improvement on EER.

6. Acknowledgements

This work was supported by the China NSFC project No. U1736202. Experiments have been carried out on the PI supercomputers at Shanghai Jiao Tong University. The author Zhengyang Chen is supported by Wu Wen Jun Honorary Doctoral Scholarship, AI Institute, Shanghai Jiao Tong University.

7. References

[1] Y. Liu, Y. Qian, N. Chen, T. Fu, Y. Zhang, and K. Yu, "Deep feature for text-dependent speaker verification," *Speech Communication*, vol. 73, pp. 1–13, 2015.

[2] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification." in *Interspeech*, 2017, pp. 999–1003.

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

[4] Y. Zhu, T. Ko, D. Snyder, B. Mak, and D. Povey, "Self-attentive speaker embeddings for text-independent speaker verification." in *Interspeech*, 2018, pp. 3573–3577.

[5] X. Xiang, S. Wang, H. Huang, Y. Qian, and K. Yu, "Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition," *arXiv preprint arXiv:1906.07317*, 2019.

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

[7] 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.

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

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

[10] S. Wang, Z. Huang, Y. Qian, and K. Yu, "Discriminative neural embedding learning for short-duration text-independent speaker verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1686–1696, 2019.

[11] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4879–4883.

[12] J. Rohdin, T. Stafylakis, A. Silnova, H. Zeinali, L. Burget, and O. Plchot, "Speaker verification using end-to-end adversarial language adaptation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6006–6010.

[13] G. Bhattacharya, J. Monteiro, J. Alam, and P. Kenny, "Generative adversarial speaker embedding networks for domain robust end-to-end speaker verification," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6226–6230.

[14] W. Xia, J. Huang, and J. H. Hansen, "Cross-lingual text-independent speaker verification using unsupervised adversarial discriminative domain adaptation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5816–5820.

[15] Z. Chen, S. Wang, Y. Qian, and K. Yu, "Channel invariant speaker embedding learning with joint multi-task and adversarial training," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6574–6578.

[16] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7167–7176.

[17] J. Shen, Y. Qu, W. Zhang, and Y. Yu, "Wasserstein distance guided representation learning for domain adaptation," *arXiv preprint arXiv:1707.01217*, 2017.

[18] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[19] A. Rozantsev, M. Salzmann, and P. Fua, "Beyond sharing weights for deep domain adaptation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 4, pp. 801–814, 2018.

[20] S. O. Sadjadi, T. Kheyrkhah, A. Tong, C. S. Greenberg, D. A. Reynolds, E. Singer, L. P. Mason, and J. Hernandez-Cordero, "The 2016 nist speaker recognition evaluation." in *Interspeech*, 2017, pp. 1353–1357.

[21] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," *arXiv preprint arXiv:1701.07875*, 2017.

[22] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Advances in neural information processing systems*, 2017, pp. 5767–5777.

[23] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[24] A. F. Agarap, "Deep learning using rectified linear units (relu)," *arXiv preprint arXiv:1803.08375*, 2018.