# Cross-Domain Speaker Recognition using Cycle-Consistent Adversarial Networks

Yi Liu, Bairong Zhuang, Zhiyu Li, Takahiro Shinozaki
Tokyo Institute of Technology, Japan
www.ts.ip.titech.ac.jp

***Abstract*—Speaker recognition systems often suffer from severe performance degradation due to the difference between training and evaluation data, which is called domain mismatch problem. In this paper, we apply adversarial strategies in deep learning techniques and propose a method using cycle-consistent adversarial networks for i-vector domain adaptation. This method performs an i-vector domain transformation from the source domain to the target domain to reduce the domain mismatch. It uses a cycle structure that reduces the negative influence of losing speaker information in i-vector during the transformation and makes it possible to use unpaired dataset for training. The experimental results show that the proposed adaptation method improves recognition performance of a conventional i-vector and PLDA based speaker recognition system by reducing the domain mismatch between the training and the evaluation sets.**

## I. Introduction

Speaker recognition is widely used in electronic products and can be applied as an assistant tool for speech recognition systems to improve their performance by decreasing the negative effects of diverse speaker information in the speeches. For instance, smartphone voice assistants use speaker recognition technology to determine whether the voice command is from the actual owners.

In light of the utterance contents, speaker recognition systems can be classified into two categories: text-dependent and text-independent. As the name suggested, text-dependent means the system's input utterances should be the same as the predefined utterance template. On the contrary, text-independent means the utterance template and the input utterance can be different. According to the classification introduced above, our research in this paper will focus on the text-independent speaker verification task.

With the introduction of i-vector [1], which contains both speaker and channel information, researches on speaker recognition have acquired a significant improvement. However, in text-independent tasks, since the training data used to build the system and the evaluation utterances come from different domains, i-vector based speaker recognition systems would suffer from severe performance degradation. This problem is called domain mismatch. Language type, text content, speech duration and audio quality are the typical causes of such mismatch. The task of solving domain mismatch problem is called domain adaptation.

In this paper, we proposed a novel cross domain speaker recognition method that use CycleGAN to converse the i-vector from two different datasets, namely MIXER and SWB, then scored by the PLDA classifier. We designed 4 different CycleGAN based systems, and experimental results indicate that, among the 4 proposed CycleGAN based system, those with identity loss all obtain better performances than the original mismatch system, which means it has a noticeable effect on reducing the domain mismatch problem.

## II. I-vector and PLDA based speaker recognition

Figure 1 represents the basic structure of conventional text-independent speaker verification system. When using the pipeline shown in the Figure 1 for speaker recognition, we assume the i-vector used in training time and in testing time are from the same domain. Problem comes when we use i-vector from different domain in test time, which usually makes lower recognition accuracy. Formally, we denote the i-vector used in the training time to the the back-end PLDA classifier as target domain i-vector $\eta_T$, and the i-vector used in the testing time as source-domain i-vector $\eta_S$. Some domain adaptation techniques $f_{S\to T}(\cdot)$ could be applied to i-vector in the source domain to make it better adapt to the back-end classifier which was trained on the target domain i-vector.

### A. PLDA backends

Probabilistic Linear Discriminant Analysis (PLDA) [2][3] is a widely used back-end classifier for i-vectors channel compensation and scoring in the speaker recognition. Assuming that the training data of PLDA consists of $I$ speakers and each speaker provides $J$ utterances. $x_{ij}$ means the $j$-th utterance from the $i$-th speaker. PLDA models the data components by the following equation:

$$x_{ij} = \mu + Fh_i + Gw_{ij} + \epsilon_{ij} \tag{1}$$

This model comprises two parts, where the $\mu + Fh_i$ is the signal component that depends only on the identities of speakers. $Gw_{ij} + \epsilon_{ij}$ is the noise component that depends not only on the speaker identities but also on the channel information in different utterance. $\mu$ is the mean of the training data. The matrix $F$ contains the basis for the between-speaker subspace and $h_i$ represents the position. Similarly, the Matrix $G$ contains the basis for the within-speaker subspace, and $w_{ij}$ represents the position. other undescribed components are defined as $\epsilon_{ij}$, which is a Gaussian with diagonal covariance $\Sigma$. Thus, the parameter of a PLDA model can be represented as $\Theta = \{\mu, F, G, \sigma\}$. To training the PLDA model, EM algorithm usually employed to maximize the exceptation of

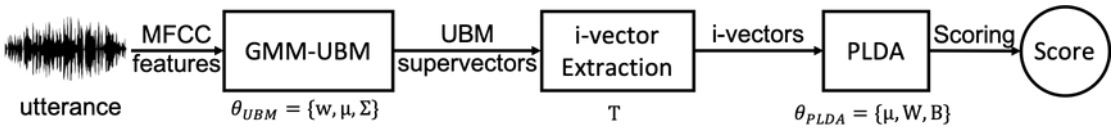


Fig. 1. Conventional speaker recognition system

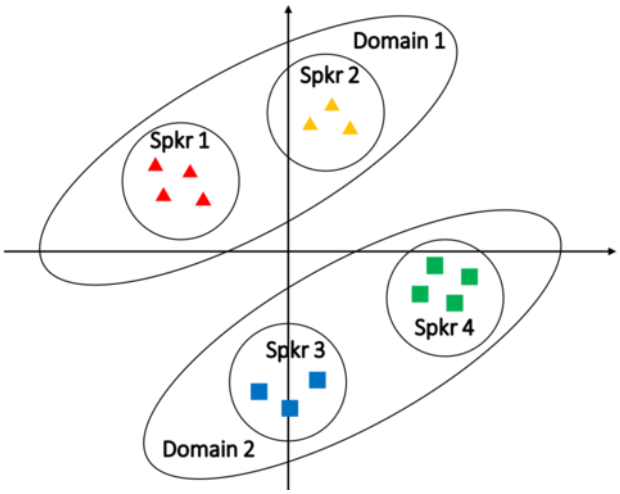


Fig. 2. i-vectors in the same domain and different domains

target speaker given its utterance. In the evaluation stage, assume they are two utterance $\eta_1$ and $\eta_2$ for comparison. The log likelihood score is computed by

$$score(\eta_1, \eta_2) = log \frac{p(\eta_1, \eta_2|\psi_s)}{p(\eta_1|\psi_d)(\eta_2|\psi_d)} \quad (2)$$

where $\psi_s$ represented the assumption that two utterance from a same speaker, and $\psi_d$ represented the assumption that two utterance from the different speaker. Higher score means higher probability that two utterance belonging to the same speaker.

### B. Domain mismatch problem

As described in previous section, training data and evaluating data usually come from different sources. This throws light on a fact that many speaker-irrelevant components in i-vectors affect the model performance. For instance, the training data and evaluating data usually have diverse audio properties (audio qualities, background noises, channel characteristics, etc.) due to the difference of recording devices, while other properties such as language or speech content may also degrade the speaker recognition performance.

In speaker recognition area, the word ”domain” is used to describe a general components similarity of features. As shown in Figure 2, circles serve as speakers, ovals represent domains, while triangle and square figures represent i-vectors. The i-vectors come from the same speaker share the same color, and they share the same shape when come from the same domain, vice versa. For i-vectors, only the speaker-relevant components are expected to take effect for speaker recognition, the difference of speaker-irrelevant components, nonetheless, vastly disturbs the recognition result. This problem is called domain mismatch.

### C. Related Work

To mitigate the effect of domain mismatch, domain adaptation [5] plays an essential role to alleviate the problem. During past years, most of domain adaptation researchers concentrated on the i-vector and PLDA based speaker recognition system as it still shows a dominant performance in speaker recognition field. In this context, some researchers try to modify the i-vector features. For instance, [6] proposes an inter dataset variability compensation (IDVC) method to remove a domain-related subspace from the total i-vector space; [7] suggests a domain mismatch modeling (DMM) method to discard a domain-related component from each i-vector. Differently, some researches aim at optimizing the PLDA back end in order to enhance the recognition ability of the system under mismatch situation. For example, [8] proposes an inter dataset variability modeling (IDVM) method to optimize the hyper-parameters in the PLDA model to obtain better domain robustness.

Recent years, benefited from the significant development in deep learning field, numerous pattern recognition works, including speech and speaker recognition, have attained tremendous progress with neural network based methods [9][10]. Among them, several neural network frameworks, such as an autoencoder based domain adaptation (AEDA) approach suggested in [11], which is based on autoencoder and denoising autoencoder networks, also take effect on domain adaptations tasks.

## III. CycleGAN based unsupervised adaptation

Within the area of deep learning, generated adversarial network (GAN) has become one of the hottest topics. The adversarial strategy has a wide range of applications in different types of tasks. A successful application of GAN is feature conversion, in which such strategy helps to generate images and speeches differ from the original ones to satisfy specific application situations. Inspired by previous works of GAN, this paper would like to propose a neural network based domain adaptation method for speaker recognition. This method uses both adversarial strategy and a cycle-consistent architecture to perform domain adaptation on i-vectors with unpaired training data.

The basic idea of Generative Adversarial Networks (GAN) [12] is making a competition between two networks that have exactly opposite goals. These two networks are called generator and discriminator respectively. The generator aims

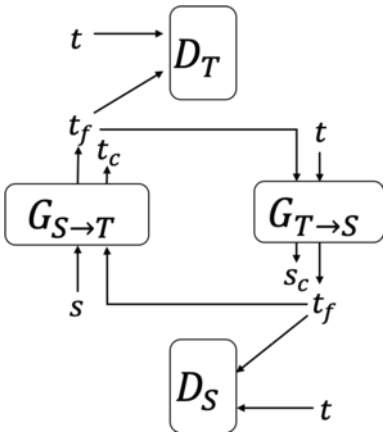


Fig. 3. Illustration of CycleGAN for domain adaptation

at making fake data to cheat the discriminator. On the contrary, the discriminator aims to distinguish the generated fake data and real data.

If an adaptation is performed using unpaired data, it is useful to reduce the effort to prepare the adaptation data. A variation of GAN with a cycle structure, called cycle-consistent adversarial network, or CycleGAN [13], is used for this purpose. As shown in Figure 3, it consists of two GAN models and combines two transformations by the generator networks: $s_c = G_{T\to S}(G_{S\to T}(s))$, $t_c = G_{S\to T}(G_{T\to S}(t))$, where the $s_c$ and $t_c$ are called cycle data.

The objective function of the CycleGAN is:

$$\begin{aligned} L(G_{S\to T}, G_{T\to S}, D_S, D_T) = & L_{LSGAN}(D_T, G_{S\to T}) \\ & + L_{LSGAN}(D_S, G_{T\to S}) \\ & + \lambda L_{cyc}(G_{S\to T}, G_{T\to S}) \end{aligned} \quad (3)$$

where the $\lambda$ is the coefficient of $L_{cyc}$. Mean square loss $L_{LSGAN}$ is used to replace the log likelihood objective in $L_{GAN}$ to stabilize the training of CycleGAN. $L_{cyc}(G_{S\to T}, G_{T\to S})$ is the cycle-consistent loss shown in the equation5 to ensure that the generated fake data can be highly recovered to the original data:

$$\begin{aligned} L_{cyc}(G_{S\to T}, G_{T\to S}) = & \mathbb{E}_{s\sim p_s(s)}[||G_{T\to S}(G_{S\to T}(s)) - s||_1] \\ & + \mathbb{E}_{t\sim p_t(t)}[||G_{S\to T}(G_{T\to S}(t)) - t||_1] \end{aligned} \quad (4)$$

Besides the cycle-consistent loss, an identity loss [14] is also introduced in the whole loss function to further strengthen the identity consistency:

$$\begin{aligned} L_{ide}(G_{S\to T}, G_{T\to S}) = & \mathbb{E}_{t\sim p_{target}(t)}[||G_{S\to T}(t) - t||_1] \\ & + \mathbb{E}_{s\sim p_{source}(s)}[||G_{T\to S}(s) - s||_1] \end{aligned} \quad (5)$$

In all, the extended objective function of CycleGan is:

$$\begin{aligned} L(G_{S\to T}, G_{T\to S}, D_S, D_T) = & L_{LSGAN}(D_T, G_{S\to T}) \\ & + L_{LSGAN}(D_S, G_{T\to S}) \\ & + \lambda L_{cyc}(G_{S\to T}, G_{T\to S}) \\ & + \gamma L_{ide}(G_{S\to T}, G_{T\to S}) \end{aligned} \quad (6)$$

Note that CycleGAN is proved to be successfully applied on image style translation, cross-domain speech recognition, etc. when applying CycleGAN on i-vector domain adaptation, the property of cycle-consistent guarantees that the generated fake i-vector don't lose some speaker-relevant information during the domain conversion by making an additional constraint to keep the essential elements in transformed data unchanged.

## IV. Proposed Method

In this paper, we proposed to use CycleGAN to converse the i-vector from source domain to the target domain. The conversed i-vector then scored by the PLDA model trained on the target domain data. In contrast to the baseline system which is indicated in Figure 1, the structure of our proposed system is shown in Figure 4. The difference lies on the CycleGAN part, which will be explained next.

Since the original generator and discriminator used in the CycleGAN mainly processing the image-like data, which the property is quite different from the speaking embedding, e.g. i-vector. We modified the architecture of generator and discriminator to make it better to fit our goal. The architecture of proposed CycleGAN model is shown in the Figure 5.

The left hand side of Figure 5 (divided by the dotted line) demonstrates the structure of generator. The generator contains down-sampling, residual block [15], and up-sampling networks. In the down-sampling part, we apply convolutional operation to down-sample the 600 dimensional inputted i-vectors to 150 dimension internal representation. There are several residual block follows the down-sampling network, where we apply 1x1 convolution operation to the internal i-vector representation with the expectation that help the model to converse the i-vector from source domain to target domain well. For the up-sampling network, we applied transpose convolutional operations. Detailed information can be found in the figure.

The architecture of discriminator is shown in the right hand side of Figure 5. We first apply convolutional operation to the input i-vector, followed by 3 fully-connected layers. Since we

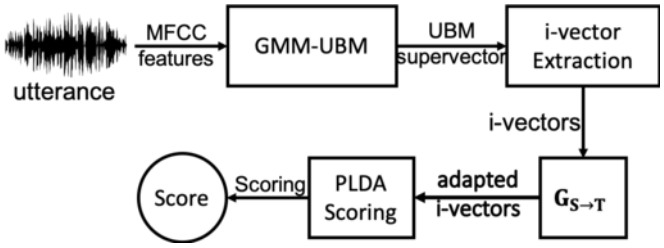


Fig. 4. Proposed GAN-based system for cross-domain speaker recognition

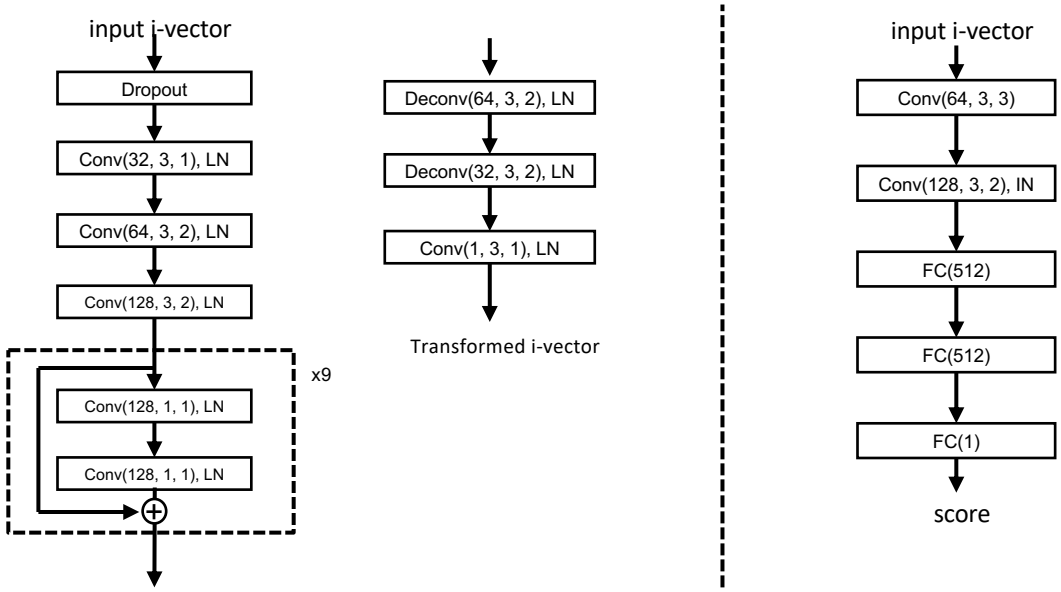


Fig. 5. Architecture of proposed generator (left hand side) and discriminator (right hand side) used in the experiments. The Conv means the convolutional operation and each parameter means # of kernel, stride and padding, respectively. LN means the Layer Normalization. Deconv means transposed convolutional operation. FC means the fully-connected layer. Except for the residual block and discriminator, we use Leaky RELU [4] with slope 0.2 as the activation function, we use RELU in the other parts.

treat the problem for discriminator as the binary classification problem, we then apply sigmoid function to the output of discriminator, obtain the probability that represent the inputted i-vector is truly from the target domain.

## V. Experimental Setup

In our experiments, we use Domain Adaptation Challenge 2013 (DAC13) [16] data standard for our experiments. The training data consists of two datasets: source domain data MIXER [17][18][19] and target domain data SWB [20]. The details of these two datasets are shown in Table I.

TABLE I
DATASET STATISTICS

| | ***SWB*** | ***MIXER*** |
|---|---|---|
| # of speakers | 3114 | 3790 |
| Males | 1461 | 1115 |
| Females | 1653 | 2675 |
| Files | 33039 | 36470 |
| Avg. files/spkr | 10.6 | 9.6 |
| Avg. phone num/spkr | 3.8 | 2.8 |

Two baselines, match and mismatch systems, are built with the system structure (except the domain adaptation part) shows in Figure 4. For the training of the systems GMM-UBM [21], i-vector extractor and PLDA parts, the match system uses source domain data MIXER, while the mismatch systems uses target domain data SWB. We evaluate the systems on SRE2010 C5 extended task [22]. The evaluation criteria are equal error rate (EER) and minimum detection cost function (minDCF).

The proposed CycleGAN based system uses MIXER as the source domain data and SWB as the target domain data for training. We use the trained $G_{S\to T}$ to obtain domain-adapted SRE10 evaluation i-vectors. Other parts are the same as the mismatch baseline system.

We design 4 different CycleGAN based systems for comparison:

**Cyc-basic** is the basic CycleGAN model described in above section.

**Cyc-ide** appends an identity loss to the full loss of CycleGAN

**Cyc-WGAN-ide** uses Wasserstein GAN (WGAN) [23], which is a modified GAN structure, to stabilize the training and avoid inherent problems of GANs training such as model collapse.

**Cyc-ide-GRL** adds another network to the CycleGAN model, which is called domain predictor. This domain predictor is trained to be domain-discriminative, but its loss is reversely combined to the full loss of CycleGAN through a gradient reversal layer (GRL) between generator and domain predictor [24][25]. As a result, the generated i-vectors tend to be more domain-confusing so that this strategy has a positive effect on the training objective of GAN.

During the training, we use Adabound [26] as the optimizer, and each model was trained for 40 epochs.

## VI. Result

Results are shown in Table II. Compared to the match system, speaker recognition performance of the Mismatch system was significantly worse. This fact shows the noticeable performance degradation caused by domain mismatch. The Cyc-basic system didn't outperform the mismatch baseline system in all evaluation criteria. Other adapted systems outperformed

TABLE II
EXPERIMENT RESULT ON MIXER-SWB DATASET

| | EER | $DCF10^{-2}$ | $DCF10^{-3}$ |
|---|---|---|---|
| Match | 4.46 | 0.3918 | 0.5940 |
| Mismatch | 12.25 | 0.6450 | **0.7706** |
| Cyc-basic | 14.44 | 0.7781 | 0.9102 |
| Cyc-ide | **10.05** | 0.6493 | 0.8069 |
| Cyc-WGAN-ide | 11.44 | **0.6376** | 0.7760 |
| Cyc-ide-GRL | 11.06 | 0.6549 | 0.7951 |

the baseline system in EER. The Cyc-ide system performed best in EER (17.9% better than mismatch baseline), while the Cyc-WGAN-ide system performed best in $\mathrm{DCF10}^{-2}$. However, Under $\mathrm{DCF10}^{-3}$ metric, the mismatch system even performed better than all CycleGAN based systems. This result indicates that the adapted i-vectors may have the disadvantage of increasing the false-alarm rate in the evaluation.

## VII. CONCLUSION

This paper proposed a CycleGAN based i-vector domain adaptation method for text-independent speaker recognition system. It reduces the domain mismatch components in i-vector and has the advantage of utilizing unpair datasets for adaptation. Experimental results indicate that the proposed method improves the performance in EER of an i-vector and PLDA based speaker recognition system.

## REFERENCES


[1] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Transcations on Audio, Speech, and Language Processing* vol. 19, pp. 788-798, August 2010.

[2] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Computer Vision, 2007. ICCV 2007. IEEE 11$^{\mathrm{th}}$ International Conference on. IEEE*, 2007, pp. 1-8

[3] F. Bahmaninezhad, and J. H. L. Hansen, "i-Vector/PLDA speaker recognition using support vectors with discriminant analysis," in *IEEE International Conference on Acoustic, Speech and Signal Processing(ICASSP)*, March 2017.

[4] V. nair and G. E. Hinton, "Rectified linear units imporve restricted boltzmann machines," in *Proceedings of the 27$^{\mathrm{th}}$ international conference on machine learning(ICML-10)*, 2010, pp. 807-814.

[5] S. J. Pan, and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345-1359, 2010.

[6] H. Aronowitz, "Inter dataset variability compensation for speaker recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), Florence, Italy*, July 2014

[7] M. H. Rahman, I. Himawan, D. Dean, and S. Sridharan, "Domain mismatch modeling of out-domain i-vectors for plda speaker verification," in *Proceedings of the 18$^{\mathrm{th}}$ Annual Conference of the International Speech Communication Association(INTERSPEECH 2017), International Speech Communication Association(ISCA), Stockholm, Sweden*, 2017

[8] H. Aronowitz, "Inter dataset variability modeling for speaker recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), New Orleans, LA, USA*, June 2017

[9] F. Richardson, D. Reynolds, and N. Dehak, "Deep neural network approaches to speaker and language recognition," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1671-1675, 2015.

[10] D. Snyder, D. Garcia-Romero, D. Povey, S. Khudanpur, "Deep Neural Network Embeddings for Text-Independent Speaker Verification," in *Proceedings of the 18$^{\mathrm{th}}$ Annual Conference of the International Speech Communication Association(INTERSPEECH 2017), International Speech Communication Association(ISCA), Stockholm, Sweden*, 2017

[11] S. Shon, S. Mun, W. Kim, and H. Ko, "Autoencoder based domain adaptation for speaker recognition under insufficient channel information," *arXiv preprint arXiv:1708.01227*, 2017.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672-2680.

[13] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *arXiv preprint*, 2017.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "identity mappings in deep residual networks," in *European conference on computer vision.* Springer, 2016, pp. 630-645.

[15] 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.

[16] "Jhu 2013 speaker recognition workshop," http://www.clsp.jhu.edu/user_uploads/workshops/ws13/DAC_description_v2.pdf, 2013.

[17] C. Cieri, W. Andrews, J. P. Campbell, G. Doddington, J. Godfrey, S. Huang, M. Liberman, A. Martin, H. Nakasone, M. Przybocki, and K. Walker, "The Mixer and transcript reading corpora: Resources for mulitlingual, cross-channel speaker recognition research," in *Proc. LREC*, 2006.

[18] L. Brandschain, C. Cieri, D. Graff, C. Caruso, A. Neely, and K. Walker, "Speaker recognition: Building the Mixer 4 and 5 corpora," in *Proc. LREC*, 2008.

[19] L. Brandschain, D. Graff, C. Cieri, K. Walker, C. Caruso, and A. Neely, "The Mixer 6 corpus: Resources for cross-channel and text independent speaker recognition," in *Proc. LREC*, 2010.

[20] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: telephone speech corpus for research and development," in *1992 IEEE International Conference on Acoustic, Speech, and Signal Processing(ICASSP)*, San Francisco, CA, USA, March 1992.

[21] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19-41, 2000.

[22] "The nist year 2010 speaker recognition evaluation plan," https://www.nist.gov/sites/default/files/documents/itl/iad/mig/NIST_SRE10_evalplan-r6.pdf, 2010.

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

[24] 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.

[25] 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 *2018 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). IEEE*, 2018, pp. 4889-4893.

[26] L. Luo, Y. Xiong, Y. Liu, and X. Sun, "Adaptive Gradient Methods with Dynamic Bound of Learning Rate," *arXiv preprint arXiv:1902.09843*, 2019.