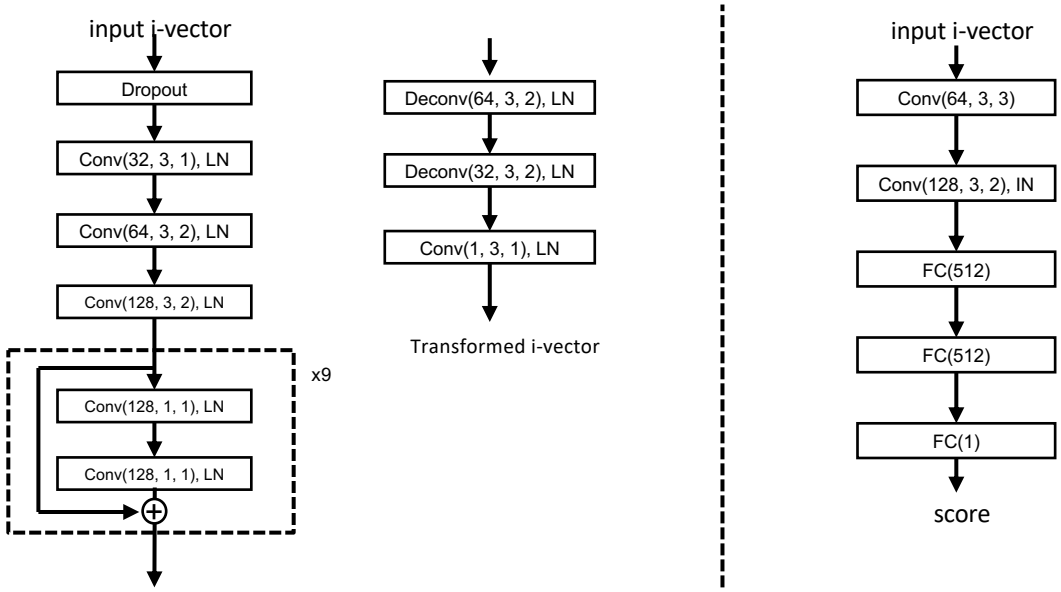

Fig. 5. Architecture of proposed generator (left hand side) and discriminator (right hand side) used in the experiments. The Conv means the convolutional operation and each parameter means # of kernel, stride and padding, respectively. LN means the Layer Normalization. Deconv means transposed convolutional 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.