I-vector Domain Adaptation Using Cycle-Consistent Adversarial Networks for Speaker Recognition

Yi Liu[1,a)] Takahiro Shinozaki[1,b)]

Abstract: Speaker recognition systems often suffer from severe performance degradation due to the difference between training data 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 datasets 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.

Keywords: speaker recognition, i-vector, generate adversarial networks, CycleGAN, unpaired data

1. Introduction

Speaker recognition researches have obtained a significant improvement from the introduction of i-vector [1]. However, i-vector based speaker recognition systems suffer from severe performance degradation when the training data and the evaluation utterances come from different domains. This problem is called domain mismatch. In this paper, we apply cycle-consistent adversarial network based adaptation to i-vector features for speaker recognition. To reduce the domain mismatch, we investigate the use of the adversarial strategy and the cycle-consistent architecture to transform i-vectors from the source domain to the target domain.

2. I-vector and PLDA speaker recognition system

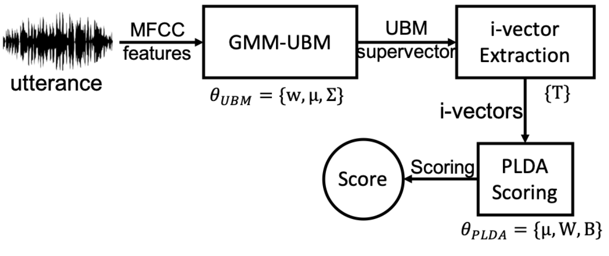

Fig. 1 Conventional speaker recognition system.

Figure 1 shows the basic structure of a conventional i-vector [1] and PLDA[2] based speaker recognition system. Firstly, Mel Frequency Cepstral Coefficients (MFCC) acoustic features are extracted from input utterances. The MFCC features serve as the input of a Gaussian mixture model based universal background model (GMM-UBM) [3] to compute high-dimensional supervectors. These supervectors are used for i-vector extraction through a total-variability matrix. Finally, a PLDA model uses i-vectors to calculate scores to recognize whether the input utterances belong to specific speakers.

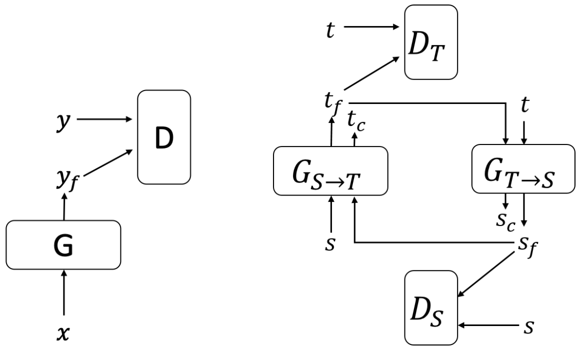

Fig. 2 GAN and CycleGAN

3. CycleGAN based domain adaptation

3.1 Generate Adversarial Networks (GAN)

The basic idea of GAN [4] is making a competition between two networks that have exactly opposite goals. These two networks are called generator and discriminator respectively. The generator aims at making fake data to cheat the discriminator. On the contrary, the discriminator aims to distinguish the generated fake data and real data.

[1] Tokyo Institute of Technology, Tokyo, Japan
[a)] liu.y.bc@m.titech.ac.jp
[b)] www.ts.ip.titech.ac.jp

The left part of Figure 2 shows the a GAN's structure. The optimization target is:

$$\begin{aligned}\min_{\boldsymbol{G}}\max_{\boldsymbol{D}} L_{GAN}(D,G) =&\mathbb{E}_{\boldsymbol{x}\sim p_{real}(\boldsymbol{x})}[logD(\boldsymbol{x})]\\ &+\mathbb{E}_{\boldsymbol{y}\sim p_{fake}(\boldsymbol{y})}[log(1-D(G(\boldsymbol{x})))], \quad (4.1)\end{aligned}$$

where $\boldsymbol{G}$ and $\boldsymbol{D}$ are the generator and discriminator in a GAN, respectively, and $\boldsymbol{x}$ and $\boldsymbol{y_f} = \boldsymbol{G(x)}$ are the real and fake data generated by $\boldsymbol{G}$.

By this way, there is an adversarial game between $\boldsymbol{G}$ and $\boldsymbol{D}$ in their training process. Finally, a powerful generator is trained to make fake data that is similar enough to the real data.

3.2 CycleGAN

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 [5] is used for this purpose. The right part of Figure 2 shows the structure of a CycleGAN. It consists of two GAN models and combines two transformations by the generator networks: $\boldsymbol{s_c} = \boldsymbol{G_{T\to S}(G_{S\to T}(s))}$, $\boldsymbol{t_c} = \boldsymbol{G_{S\to T}(G_{T\to S}(t))}$, $\boldsymbol{s_c}$ and $\boldsymbol{t_c}$ are called cycle data. This is basically an autoencoder-like structure.

The full objective function of the CycleGAN is :

$$\begin{aligned}L(G_{S\to T}, G_{T\to S}, D_S, D_T) =&L_{LSGAN}(D_T, G_{S\to T})\\ &+L_{LSGAN}(D_S, G_{T\to S})\\ &+\lambda L_{cyc}(G_{S\to T}, G_{T\to S}), \qquad (4.2)\end{aligned}$$

where $\boldsymbol{\lambda}$ is the coefficient of $\boldsymbol{L_{cyc}}$. Least square loss $\boldsymbol{L_{LSGAN}}$[17] is used to replace the log likelihood objective in $\boldsymbol{L_{GAN}}$ to stabilize the training of CycleGAN. $L_{cyc}(G_{S\to T}, G_{T\to S})$ is the cycle-consistent loss to ensure that the generated fake data can be highly recovered to the original data:

$$\begin{aligned}L_{cyc}(G_{S\to T}, G_{T\to S}) =&\mathbb{E}_{\boldsymbol{s}\sim p_{source}(\boldsymbol{s})}[||G_{T\to S}(G_{S\to T}(\boldsymbol{s})) - \boldsymbol{s}||_1]\\ &+\mathbb{E}_{\boldsymbol{t}\sim p_{target}(\boldsymbol{t})}[||G_{S\to T}(G_{T\to S}(\boldsymbol{t})) - \boldsymbol{t}||_1].\\ &\qquad (4.3)\end{aligned}$$

It guarantees that the generated fake data don't lose some speaker-relevant information during the domain conversion. It makes an additional constraint to keep the essential elements in transformed data unchanged. By this way, paired data are not required to guide GAN's training.

3.3 Proposed CycleGAN based i-vector adaptation method

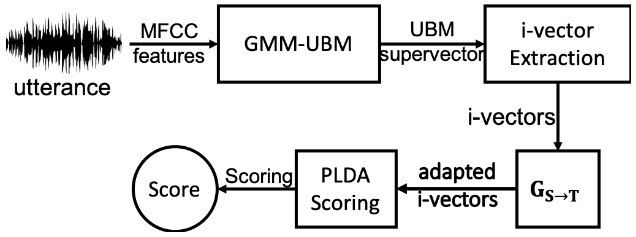

Fig. 3 System using proposed method

We propose a CycleGAN based i-vector domain adaptation method as shown in Figure 3. Among the two generators, one generator is used to transform the source-domain i-vector to the target domain.

4. Experiments

4.1 Datasets

We use Domain Adaptation Challenge 2013 (DAC13) [*1] data standard for our experiments. The training data consists of two datasets: source domain data MIXER and target domain data SWB. The details of these two datasets are shown in Table 4.1. We evaluate the systems on SRE2010 C5 extended task [*2] The evaluation criteria are equal error rate (EER) and minimum detection cost function (minDCF) .

Table 1 DAC13 datasets

	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

4.2 Systems design

Two baselines, match and mismatch systems, are built with the system structure in Figure 1. For the training of the systems GMM-UBM, i-vector extractor and PLDA parts, the match system uses source domain data MIXER, while the mismatch systems uses target domain data SWB.

The proposed CycleGAN based system uses MIXER as the source domain data and SWB as the target domain data for training. The training follows the data flow in right part of Figure 2. Then, we use the trained $\boldsymbol{G_{S\to T}}$ 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, as shown in Table 2. **Cyc-basic** is the basic CycleGAN model described in section 4.3. **Cyc-ide** appends an identity loss [6] to the full loss of CycleGAN:

$$\begin{aligned}L(G_{S\to T}, G_{T\to S}, D_S, D_T) =&L_{LSGAN}(D_T, G_{S\to T})\\ &+L_{LSGAN}(D_S, G_{T\to S})\\ &+\lambda L_{cyc}(G_{S\to T}, G_{T\to S})\\ &+\gamma L_{ide}(G_{S\to T}, G_{T\to S}), \qquad (5.1)\end{aligned}$$

where $\boldsymbol{\gamma}$ is the coefficient of $\boldsymbol{L_{ide}}$. **Cyc-WGAN-ide** uses Wasserstein GAN (WGAN) [7], 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

[*1] http://www.clsp.jhu.edu/user_uploads/workshops/ws13/DAC_description_v2.pdf

[*2] https://www.nist.gov/sites/default/files/documents/itl/iad/mig/NIST_SRE10_evalplan-r6.pdf

Table 2 CycleGAN based Systems

	System description
Cyc-basic	System using the basic CycleGAN network
Cyc-ide	System using a CycleGAN with an identity loss
Cyc-WGAN-ide	System using a CycleGAN with identity loss and W-GAN design
Cyc-ide-GRL	System using a CycleGAN with identity loss and a GRL [8] based domain predictor

a gradient reversal layer (GRL) [8] between generator and domain predictor. 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.

4.3 Experimental results

Table 3 Speaker recognition results

	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	**10.96**	0.6531	0.8022
Cyc-WGAN-ide	11.44	**0.6376**	0.7760
Cyc-ide-GRL	11.06	0.6549	0.7951

Results are shown in Table 3. 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 the baseline system in EER. The **Cyc-ide** system performed best in EER (10.3% better than mismatch baseline), while the **Cyc-WGAN-ide** system performed best in DCF10^{-2}. However, no system outperformed the baseline system with the DCF10^{-3} measure. We are currently investigating the reason for this.

5. Conclusion and Future work

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-vectors and has the advantage of utilizing unpaired 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.

Future work includes evaluating the proposed i-vector adaptation method in other conditions, and make the adaptation robust to the difference of the settings. In fact, we observed a decrease in performance when we applied a normalization step to the i-vector features. Comparison with other adaptation methods is also needed.

References

[1] Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P. and Ouellet, P.: Front-end factor analysis for speaker verification, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 19, No. 4, pp. 788–798 (2011).

[2] Prince, S. J. and Elder, J. H.: Probabilistic linear discriminant analysis for inferences about identity, *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, IEEE, pp. 1–8 (2007).

[3] Reynolds, D. A., Quatieri, T. F. and Dunn, R. B.: Speaker verification using adapted Gaussian mixture models, *Digital signal processing*, Vol. 10, No. 1-3, pp. 19–41 (2000).

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

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

[6] He, K., Zhang, X., Ren, S. and Sun, J.: Identity mappings in deep residual networks, *European conference on computer vision*, Springer, pp. 630–645 (2016).

[7] Arjovsky, M., Chintala, S. and Bottou, L.: Wasserstein gan, *arXiv preprint arXiv:1701.07875* (2017).

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