Interspeech 2018
2-6 September 2018, Hyderabad

A Multi-Discriminator CycleGAN for Unsupervised Non-Parallel Speech Domain Adaptation

Ehsan Hosseini-Asl, Yingbo Zhou, Caiming Xiong, Richard Socher

Salesforce Research

{ehosseiniasl,yingbo.zhou,cxiong,rsocher}@salesforce.com

Abstract

Domain adaptation plays an important role for speech recognition models, in particular, for domains that have low resources. We propose a novel generative model based on cyclic-consistent generative adversarial network (CycleGAN) for unsupervised non-parallel speech domain adaptation. The proposed model employs multiple independent discriminators on the power spectrogram, each in charge of different frequency bands. As a result we have 1) better discriminators that focus on fine-grained details of the frequency features, and 2) a generator that is capable of generating more realistic domain-adapted spectrogram. We demonstrate the effectiveness of our method on speech recognition with gender adaptation, where the model only has access to supervised data from one gender during training, but is evaluated on the other at test time. Our model is able to achieve an average of 7.41% on phoneme error rate, and 11.10% word error rate relative performance improvement as compared to the baseline, on TIMIT and WSJ dataset, respectively. Qualitatively, our model also generates more natural sounding speech, when conditioned on data from the other domain.

Index Terms: generative models, speech domain adaptation, non-parallel data, unsupervised learning

1. Introduction

Neural-based acoustic models have shown promising improvements in building automatic speech recognition (ASR) systems [1, 2, 3, 4]. However, it tends to perform poorly when evaluated on out-of-domain data, because of mismatch between the training and testing distribution (Table 1).

Domain mismatch is mainly due to variation in non-linguistic features, such as different speaker identity, unseen environmental noise, large accent variations, etc. Therefore, training a robust ASR system is highly dependent on factorizing linguistic features (text) from non-related variations, or adapting the inter-domain variations of source and target.

Voice conversion (VC) has been widely used to adapt the non-linguistic variations, such as statistical methods [5, 6, 7], and Neural-based models [8, 9, 10, 11, 12, 13, 14]. However, traditional VC methods require parallel data of source and target that is difficult to obtain in practice. In addition, the requirement of parallel data prevent these methods from using more abundant unsupervised data. Therefore, an unsupervised domain adaptation is desirable for building a robust ASR system.

In this paper, we propose a new generative model based on CycleGAN [15] for unsupervised non-parallel domain adaptation. Since differences in magnitude of frequency is the main variation across domains for spectrogram representations, it is imperative that CycleGAN correctly catch the spectro-temporal variations between different frequency bands across domains during training. This will allow the generator to learn the mapping function which can convert spectrogram from source to target domain. In this paper, we show that the original CycleGAN model is failing to learn the correct mapping function between domains, and the generator collapses into learning an identity mapping function, which results in generating a noisy and unnatural-sounding audio.

Table 1: *ASR prediction mismatch when train/test on different genders, and when adapting using Multi-Discriminator CycleGAN, on WSJ (eval92) dataset*

		Train on Male
Test on Female	True	CIBA AGREED TO REMEDY THE OVERSIGHT
	Female	SEVEN AGREED TO REMEDY THE OVER SITE
	Female→Male	CIBA AGREED TO REMEDY THE OVER SITE
	True	A LITTLE ...NEWS COULD SOFTEN THE MARKET'S RESISTANCE
	Female	A LITTLE ...NEWS COULD SOUTH IN THE MARKETS RESISTANCE
	Female→Male	A LITTLE ...NEWS COULD SOFTEN THE MARKET'S RESISTANCE
		Train on Female
Test on Male	True	THEY EXPECT COMPANIES TO GROW OR DISAPPEAR
	Male	THE DEBUT COMPANIES TO GO ON DISAPPEAR
	Male→Female	THEY EXPECT COMPANIES TO GROW OR DISAPPEAR
	True	MR POLO ALSO OWNS THE FASHION COMPANY
	Male	MR PAYING ALSO LONG THE FASHION COMPANY
	Male→Female	MR POLO ALSO OWNS THE FASHION COMPANY

To accommodate generative adversarial network for training on non-parallel spectrogram domains, the generator should be back-propagated with multiple gradient signals (from different discriminators), that each represents the variations between source and target domains at a specific frequency band. To achieve this goal, we propose to use multiple and independent discriminators for each domain, similar to generative multi adversarial network (GMAN) [16]. We show that the proposed Multi-Discriminator CycleGAN, without pretraining the discriminators, outperforms CycleGAN [15] with pretrained discriminator, for spectrogram adaptation. Furthermore, we show that the multi discriminator architecture can overcome the checkerboard artifacts problem caused by deconvolution layer in generator [17], and generates natural clean audio. To evaluate the performance of the proposed model, gender-based domains are selected as domain adaptations.

1.1. Related Work

Generative Adversarial Network (GAN) [18] is a family of non-parametric density estimation models which learn to model the data generating distribution using adversarial training. Conditional GANs (CGAN) [19] was first proposed for supervised (parallel) domain adaptation, where the goal is to convert source distribution to match the target. CGAN has been used in various data domains, especially image domains, both for parallel [20, 21] and non-parallel domain adaptation [22, 23, 15].

Sound demos can be found at https://einstein.ai/research/a-multi-discriminator-cyclegan-for-unsupervised-non-parallel-speech-domain-adaptation

10.21437/Interspeech.2018-1535

Recently, CGAN is used for speech enhancement on parallel datasets [19, 20]. Speech denoising is achieved by conditioning the generator on noisy speech to learn the de-noised version [24, 25]. Donahue et al. [26] proposed a GAN model on audio (WaveGAN) and spectrogram (SpecGAN), which is actually trained CGAN on parallel domains. Kaneko et al. [27] proposed a cycle-consistent adversarial network (CycleGAN) [15] with gated convolutional neural network (CNN) as the generator part, where the model is trained on Mel-cepstral coefficients (MCEPs) features. Hsu et al. [28] proposed a combination of variational inference network, using variational autoencoder (VAE) [29], and adversarial network, using Wasserstein GAN (WGAN) [30]. In [28], the goal is to disentangle the linguistic from nuisance latent variables via VAE using spectra (SP for short), aperiodicity (AP), and pitch contours (F0) features, followed by adversarial training to learn the target distribution from the inferred linguistic latent distribution. A recurrent VAE is also proposed [31, 32] to capture the temporal relationships in the disentangled representation of sequence data, using Mel-scale filter bank (FBank).

Contributions of the proposed generative model are, **(1)** It is a robust GAN model developed for non-parallel unsupervised domains, compared to parallel-based SpecGAN and WaveGAN [26], **(2)** The choice of multiple discriminator is adjustable to the spectro-temporal structure of the intended domains, compared to domain-specific model design of [27], **(3)** Proposed GAN model training is robust and invariant to the choice of adversarial objective, i. e. binary cross-entropy or least square (LS-GAN [33]), while the CycleGAN in [27] is only stable using least square loss, with additional using of identity mapping loss in generator, **(4)** Source and target domains in [27] are sampled from same speakers, both including male and female, only uttering different sentences, while our approach is more natural as source and targets distribution is strongly diverged due to different speaker, gender, and uttered sentences. **(5)** Compared to FHVAE [32], our models improves ASR performance on TIMIT female set by 2.067% PER (Table 3), when only trained on male.

2. Proposed Model

In this section, the proposed model is explained. We first describe the generative model based on adversarial network. Generative Model based on adversarial training (GAN) has been proposed by Goodfellow et al. [18] to model the data generating distribution. Training GAN is based on minimizing Jensen-Shannon divergence between data generating distribution $p_{data}(x)$ and model data distribution $p_z(z)$. Learning is through minimization of the adversarial loss between generator network $G(z)$, which learns a mapping function $G : Z \rightarrow X$, and discriminator network $D(x)$. The generator is learning to model the data distribution $p_{data}(x)$ by generating indistinguishable samples $\hat{x} = G(z)$ from x, using a source noise signal z to minimize (1), whereas discriminator is learning to discriminate between real data x and generated $\hat{x}$ by maximizing the adversarial loss,

$$\begin{aligned}\mathcal{L}_{GAN}(G, D) =& \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \\ & \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))]\end{aligned} \tag{1}$$

2.1. Domain Adaptation via GAN

For domain adaptation between parallel domains X and Y, Conditional GAN (CGAN)[19, 20] is proposed, using a generator that directly learns the mapping function $G : X \rightarrow Y$, by minimizing parallel conditional adversarial loss $\mathcal{L}_{P-CGAN}$,

$$\begin{aligned}\mathcal{L}_{P-CGAN}(G, D) = & \mathbb{E}_{(x,y)\sim p_{data}(x,y)}[\log D(x, y)] + \\ & \mathbb{E}_{x\sim p_{data}(x), z\sim p_z(z)}[\log(1 - D(x, G(z, x)))]\end{aligned} \tag{2}$$

where D is discriminating between pair of real parallel data (x, y) and generated pair $(x, G(z, x))$. To apply CGAN for adaptation between non-parallel domains X and Y, a conditional GAN using cycle consistent adversarial loss (CycleGAN) has been proposed [15, 22, 23]. In CycleGAN [15], there are two conditional generators, i. e., $G_X : X \rightarrow Y$ and $G_Y : Y \rightarrow X$, each trained in adversarial setting with D_Y and D_X, respectively. In other words, there are two pairs of Non-parallel conditional adversarial loss $\mathcal{L}_{NP-CGAN}(G_X, D_Y)$ and $\mathcal{L}_{NP-CGAN}(G_Y, D_X)$, where,

$$\begin{aligned}\mathcal{L}_{NP-CGAN}(G_X, D_Y) = & \mathbb{E}_{(y)\sim p_Y(y)}[\log D_Y(y)] + \\ & \mathbb{E}_{x\sim p_X(x), z\sim p_z(z)}[\log(1 - D_Y(G_X(z, x)))]\end{aligned} \tag{3}$$

In non-parallel situation, the goal is to find the correct pseudo pair (x, y) across X and Y domains in an unsupervised way. To ensure that G_X and G_Y will learn such mapping function, CycleGAN[15] proposed to minimize a cycle consistency loss using ℓ_1 norm,

$$\begin{aligned}\mathcal{L}_{cycle} =& \mathbb{E}_{x\sim p_X(x)}[\| G_Y(G_X(x)) - x \|_1] + \\ & \mathbb{E}_{y\sim p_Y(y)}[\| G_X(G_Y(y)) - y \|_1]\end{aligned} \tag{4}$$

Therefore, CycleGAN[15] learns unsupervised mapping functions between X and Y domains by combining (3) and (4), to maximize the adversarial loss $\mathcal{L}_{CycleGAN}$, where,

$$\begin{aligned}\mathcal{L}_{CycleGAN} =& \mathcal{L}_{NP-CGAN}(G_X, D_Y) + \\ & \mathcal{L}_{NP-CGAN}(G_Y, D_X) + \\ & - \lambda\mathcal{L}_{cycle}(G_X, G_Y)\end{aligned} \tag{5}$$

2.2. Multi-Discriminator CycleGAN (MD-CycleGAN)

In this section, we propose a multiple discriminator generative model based on cycle consistency loss (5). The model is based on generative multi adversarial network (GMAN) [16]. In this paper, X and Y represents spectrogram feature datasets of different speech domains. Spectrogram feature represents the frequency variation of audio data through time dimension. In order to allow CycleGAN to learn the mapping function of spectrogram between different speech domains, the generators $\{G_X, G_Y\}$ should be able to learn the variations in each frequency band for each aligned time window, across domains.

In order to learn the frequency-dependent mapping functions $\{G_X, G_Y\}$ that catch the variation per each frequency bands, we define multiple frequency-dependent discriminators $\left\{D_X^{f_{j\in n}}, D_Y^{f_{i\in m}}\right\}$, where $f_{j\in n}$ represents the i^{th} frequency band of domain X with n frequency bands, and $f_{i\in m}$ represents j-th frequency band od domain Y, respectively. The frequency band definition in each domain can share a portion of frequency spectrum, or be exclusive, based on the domain spectrogram distribution. We are also using the non-saturating version of GAN[18], NS-GAN, where the generator G is learned through maximizing the probability of predicting generated samples $\hat{x}$ as drawn from data generating distribution $p_{data}(x)$. Accordingly, the adversarial loss for each pair of generator and discriminator $\left\{\left(G_X, D_Y^{f_{i\in m}}\right), \left(G_Y, D_X^{f_{j\in n}}\right)\right\}$ in (3) and (5) is

$$\begin{aligned}\mathcal{L}_{MD-CGAN}\left(G_X, D_Y^{f_{i\in m}}\right) = & \mathbb{E}_{(y)\sim p_Y(y)}\left[\sum_{i=1}^{m}\log D_Y^{f_i}(y)\right] \\ + & \mathbb{E}_{x\sim p_X(x), z\sim p_z(z)}\left[\sum_{i=0}^{m}\log\left(D_Y^{f_i}(G_X(z, x))\right)\right]\end{aligned} \tag{6}$$

The Multi-Discriminator CycleGAN (MD-CycleGAN) is training by maximizing $\mathcal{L}_{MD-CycleGAN}$, where,

$$\begin{aligned}\mathcal{L}_{MD-CycleGAN} =& \mathcal{L}_{MD-CGAN}(G_X, D_Y^{f_{i\in m}}) + \\ & \mathcal{L}_{MD-CGAN}(G_Y, D_X^{f_{j\in n}}) + \\ & -\lambda\mathcal{L}_{cycle}(G_X, G_Y)\end{aligned} \tag{7}$$

A natural extension to the proposed MD-CycleGAN is to use multiple frequency-dependent generators [34] jointly with discriminators as well. This can follow in two configurations. In one-one setting, each generator is trained on a specific frequency band with the corresponding discriminator, i. e., set of $\left\{\left(G_X^{f_i}, D_Y^{f_i}\right) : i \in m\right\}$. Additionally, in one-many setting, each frequency-dependent generator is trained with all frequency-dependent discriminators, i. e., set of $\left\{\left(G_X^{f_j}, D_Y^{f_{i\in m}}\right) : j \in n\right\}$ trained in adversarial setting.

3. Experiment

We used TIMIT [35] and Wall Street Journal (WSJ) corporas to evaluate the performance of proposed model on domain adaptation. TIMIT dataset contains broadband 16kHz recordings of phonetically-balanced read speech of 6300 utterances (5.4 hours). Male/Female ratio of speakers across train/validation/test sets are approximately 70% to 30%. WSJ contains $\approx$ 80 hours of standard *si284/dev93/eval92* for train/validation/test sets, with equally distributed genders.

The spectrogram representation of audio is used for training the CycleGAN and ASR models, which is computed with a 20ms window and 10ms step size. Each spectrogram is normalized to have zero mean and unit variance. To implement MD-CycleGAN, three non-overlapping frequency bands are defined, i. e. $m = n = 3$ with $[53, 53, 55]$ bandwidth, for male and female domains. We denote the size of the convolution layer by the tuple (C, F, T, SF, ST), where C, F, T, SF, and ST denote number of channels, filter size in frequency dimension, filter size in time dimension, stride in frequency dimension and stride in time dimension respectively. CycleGAN model architecture is based on [15] with some modifications. The generator is based on U-net [36] architecture with 4 convolutional layers of sizes (8,3,3,1,1), (16,3,3,1,1), (32,3,3,2,2), (64,3,3,2,2) with corresponding deconvolution layers. We noticed that the discriminator in [15] outputs a vector with dimension equal to the channel size of final convolution layer, instead of outputting a scalar [18]. It was observed that this causes instability in a balanced training between generator and discriminator. We modified this by adding a fully connected layer as final layer, to match the discriminator in [18]. Discriminator has 4 convolution layers of sizes (8,4,4,2,2), (16,4,4,2,2), (32,4,4,2,2), (64,4,4,2,2), as default kernel and stride sizes in [15]. We used Griffin-lim algorithm [37] for audio reconstruction, to assess its quality. ASR model is based on [38], trained with maximum likelihood, and no policy gradient. The model has one convolutional layer of size (32,41,11,2,2), and five residual convolution blocks of size (32,7,3,1,1), (32,5,3,1,1), (32,3,3,1,1), (64,3,3,2,1), (64,3,3,1,1) respectively. Following the convolutional layers are 4 layers of bidirectional GRU RNNs with 1024 hidden units per direction per layer, one fully-connected hidden layer of size 1024 and final output layer.

3.1. Quantitative Evaluation

In this section, ASR model is employed to evaluate the performance of proposed model, where domains are different genders.

Table 2: *TIMIT, Train set Female→Male domain adaptation.* Note: *Female&→Male means Female+Female→Male*

Model	Train	Male (PER) Val	Test
	Female	40.704	42.788
One-D CycleGAN	Female→Male	40.095	42.379
	Female&→Male	39.200	42.211
Three-D CycleGAN	Female→Male	**29.838**	33.463
	Female&→Male	30.009	**33.273**
	Male (baseline)	20.061	22.516

Table 3: *TIMIT, Train set Male→Female domain adaptation.* Note: *Male&→Female means Male+Male→Female*

Model	Train	Female (PER) Val	Test
	Male	35.702	30.688
One-D CycleGAN	Male→Female	32.943	30.069
	Male&→Female	31.289	29.038
Three-D CycleGAN	Male→Female	28.80	25.448
	Male&→Female	**25.982**	**24.133**
FHVAE [32]	Male + $\mathbf{z}_1$		26.20
	Female (baseline)	24.51	23.215

Table 4: *WSJ, Train set Female↔Male domain adaptation, using Three-D CycleGAN trained on TIMIT train set.*

	Test -*eval92*			
	Male		Female	
Train	CER	WER	CER	WER
Female (baseline)	14.31	27.66	2.80	6.71
Female&→Male	**5.20**	**12.39**		
Male (baseline)	3.19	8.16	7.57	16.38
Male&→Female			**4.22**	**9.46**

First, gender generators $\{G_{M\to F}, G_{F\to M}\}$ [1] are trained on gender-separated train set. These generators are then evaluated for train→test and test→train adaptation using ASR model. In former, ASR model is retrained on the adapted train set, while in latter, a more applicable case, ASR model is fixed and evaluated on the new adapted test sets.

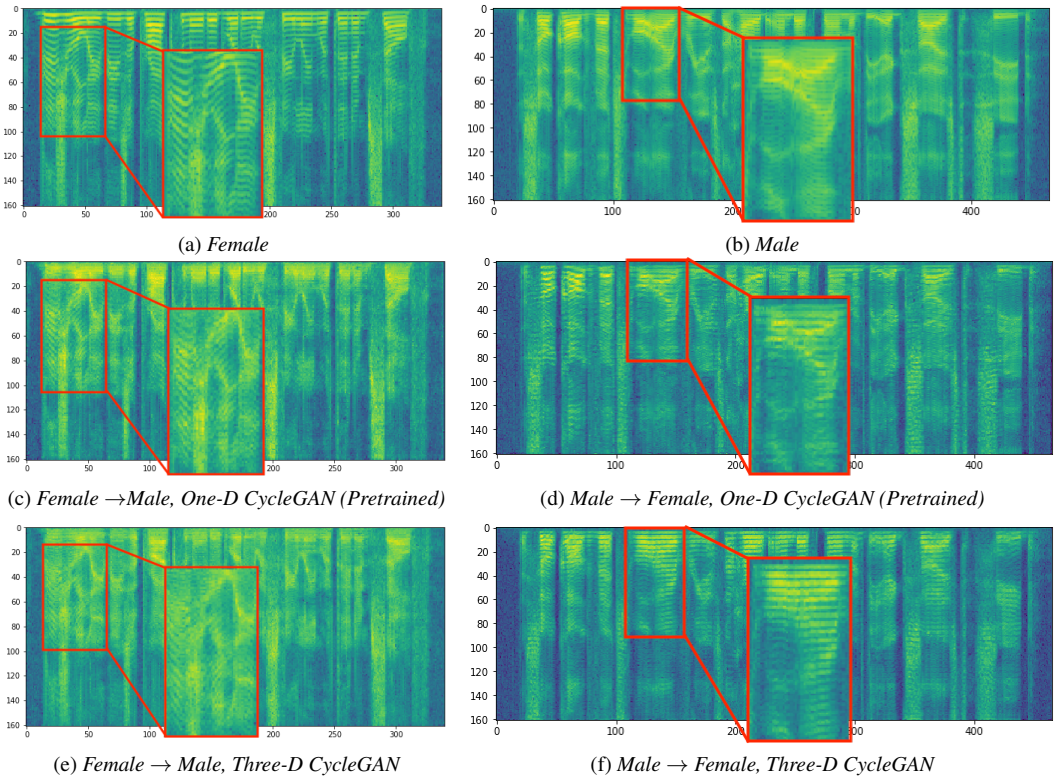
3.1.1. Train→Test Adaptation

Results on adapting TIMIT train set are shown in Table 2 and 3. As ablation study to CycleGAN-VC [27], performance is significantly improved with three discriminator compared to single one. Compared to FHVAE [32], phoneme error rate is improved by 2.067% in Table 3. To evaluate the generalization of the generators, we used them on WSJ dataset without retraining. As shown in Table 4, ASR performance is significantly improved by reducing the gap to the corresponding male and female baselines. For a fair comparison, ASR performance trained on WSJ train set is 5.55% WER. It is worth mentioning that relatively lower performance on TIMIT is due to smaller size of dataset.

3.1.2. Test→Train Adaptation

Test set adaptation of TIMIT and WSJ are shown in Table 5 and 6. It is clear that using the proposed model, ASR perfor-

[1] $G_{M\to F} : Male \to Female, G_{F\to M} : Female \to Male$



(a) *Female* (b) *Male*

(c) *Female →Male, One-D CycleGAN (Pretrained)* (d) *Male → Female, One-D CycleGAN (Pretrained)*

(e) *Female → Male, Three-D CycleGAN* (f) *Male → Female, Three-D CycleGAN*

Figure 1: *Spectrogram conversion for (a,c,e) female→male, and (b,d,f) male→female, using One-D CycleGAN and Three-D CycleGAN on TIMIT test set.* ***Note:*** *The One-D CycleGAN generator converges only by pretraining the discriminator first, unless the generator will learn identity mapping function. However, the Three-D CycleGAN results are achieved without pretraining.*

mance is significantly improved by adapting test→train , compared to original CycleGAN. Qualitative assessment of ASR predictions are shown in Tables 1 and Appendix A.

Table 5: *TIMIT, Test set Male↔Female domain adaptation*

		Train	
Test (PER)	Model	Male	Female
Male (baseline)	–	22.516	42.788
Male→Female	One-D CycleGAN		43.427
	Three-D CycleGAN		**37.000**
Female (baseline)	–	32.085	23.215
Female→Male	One-D CycleGAN	32.606	
	Three-D CycleGAN	**25.758**	

Table 6: *WSJ, Test set Male↔Female domain adaptation*

	Train	
Test (CER / WER)	Male	Female
Male (baseline)	3.19 / 8.16	14.31 / 27.66
Male→Female		**6.82 / 15.68**
Female (baseline)	7.57 / 16.38	2.80 / 6.71
Female→Male	**5.93 / 13.18**	

3.2. Qualitative Evaluation

In this section, the quality of generated spectrogram for male↔female adaptation is assessed. The characteristic difference between male and female spectrograms is the variation rate of frequency for a fixed time window. As shown in Figure 1, top row depicts the original spectrograms, where male is characterized by smooth frequency variation, opposed to peaky and high-rate variations of female. Well-trained generators should catch these inter-domain variations. As ablation study, we are also showing the generated spectrogram by CycleGAN [15] (One-D CycleGAN), in middle row, comparing with Three-D CycleGAN in bottom row. One-D CycleGAN learns to convert the spectrogram only using a pretrained discriminator. It is noticeable that the converted spectrogram in One-D CycleGAN fails to match the target domain characteristics, at some frequency bands, and simply copied the source spectrogram. However, with no pretraining of Three-D CycleGAN, it learns a better mapping function, by either suitably smoothing the spectrogram (female→male), or generating peaky variations (male→female). The checkerboard artifacts [17] is a common problem in deconvolution-based generators. This problem is visible in One-D CycleGAN, with discontinuous artifacts through time and frequency dimensions, which results in a noisy and unnatural-sounding audio. This problem is mitigated in Three-D CycleGAN, by learning the target domain characteristics using multiple independent discriminators.

4. Conclusion and Future Directions

In this paper, a new cyclic consistent generative adversarial network based on multiple discriminators is proposed (MD-CycleGAN) for unsupervised non-parallel speech domain adaptation. Based on the frequency variation of spectrogram between domains, the multiple discriminators enabled MD-CycleGAN to learn an appropriate mapping functions that catch the frequency variations between domains. The performance of MD-CycleGAN is measured by ASR prediction, when train and test set are sampled from different genders. It was shown that MD-CycleGAN can improve the ASR performance on unseen domains. As future extension, this model will be evaluated on datasets adaptation, e. g. TIMIT↔WSJ, and accent, e. g. American↔Indian adaptations.

5. References

[1] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *INTERSPEECH*, 2014.

[2] T. N. Sainath, B. Kingsbury, G. Saon, H. Soltau, A. rahman Mohamed, G. E. Dahl, and B. Ramabhadran, "Deep convolutional neural networks for large-scale speech tasks," *Neural networks : the official journal of the International Neural Network Society*, vol. 64, pp. 39–48, 2015.

[3] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *INTERSPEECH*, 2015.

[4] A. W. Senior, H. Sak, F. de Chaumont Quitry, T. N. Sainath, and K. Rao, "Acoustic modelling with cd-ctc-smbr lstm rnns," *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 604–609, 2015.

[5] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech and Audio Processing*, vol. 6, pp. 131–142, 1998.

[6] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 2222–2235, 2007.

[7] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, "Voice conversion using partial least squares regression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 912–921, 2010.

[8] L.-H. Chen, Z.-H. Ling, L.-J. Liu, and L.-R. Dai, "Voice conversion using deep neural networks with layer-wise generative training," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 1859–1872, 2014.

[9] T. Nakashika, T. Takiguchi, and Y. Ariki, "Voice conversion based on speaker-dependent restricted boltzmann machines," *IEICE Transactions*, vol. 97-D, pp. 1403–1410, 2014.

[10] S. Desai, A. W. Black, B. Yegnanarayana, and K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 954–964, 2010.

[11] S. H. Mohammadi and A. Kain, "Voice conversion using deep neural networks with speaker-independent pre-training," *2014 IEEE Spoken Language Technology Workshop (SLT)*, pp. 19–23, 2014.

[12] T. Nakashika, T. Takiguchi, and Y. Ariki, "High-order sequence modeling using speaker-dependent recurrent temporal restricted boltzmann machines for voice conversion," in *INTERSPEECH*, 2014.

[13] L. Sun, S. Kang, K. Li, and H. M. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4869–4873, 2015.

[14] T. Kaneko, H. Kameoka, K. Hiramatsu, and K. Kashino, "Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks," in *INTERSPEECH*, 2017.

[15] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

[16] I. P. Durugkar, I. Gemp, and S. Mahadevan, "Generative multi-adversarial networks," *CoRR*, vol. abs/1611.01673, 2016.

[17] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, 2016. [Online]. Available: http://distill.pub/2016/deconv-checkerboard/

[18] 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 27*, 2014, pp. 2672–2680.

[19] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *CoRR*, vol. abs/1411.1784, 2014. [Online]. Available: http://arxiv.org/abs/1411.1784

[20] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5967–5976, 2017.

[21] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," *CoRR*, vol. abs/1711.11585, 2017.

[22] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, 2017, pp. 1857–1865.

[23] Z. Yi, H. Zhang, P. Tan, and M. Gong, "DualGAN: Unsupervised dual learning for image-to-image translation," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2868–2876, 2017.

[24] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech enhancement generative adversarial network," in *INTERSPEECH*, 2017.

[25] D. Michelsanti and Z.-H. Tan, "Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification," in *INTERSPEECH*, 2017.

[26] C. Donahue, J. McAuley, and M. Puckette, "Synthesizing audio with generative adversarial networks," *CoRR*, vol. abs/1802.04208, 2018.

[27] T. Kaneko and H. Kameoka, "Parallel-Data-Free Voice Conversion Using Cycle-Consistent Adversarial Networks," *arXiv:1711.11293*, 2017.

[28] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," *arXiv:1704.00849*, 2017.

[29] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *CoRR*, vol. abs/1312.6114, 2013.

[30] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *ICML*, 2017.

[31] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised domain adaptation for robust speech recognition via variational autoencoder-based data augmentation," *arXiv:1707.06265*, 2017.

[32] W.-N. Hsu, Y. Zhang, and J. R. Glass, "Unsupervised learning of disentangled and interpretable representations from sequential data," in *NIPS*, 2017.

[33] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2813–2821, 2017.

[34] A. Ghosh, V. Kulharia, V. P. Namboodiri, P. H. S. Torr, and P. K. Dokania, "Multi-agent diverse generative adversarial networks," *CoRR*, vol. abs/1704.02906, 2017.

[35] J. S. Garofolo *et al.*, "TIMIT acoustic-phonetic continuous speech corpus LDC93S1," *Philadelphia: Linguistic Data Consortium*, 1993.

[36] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015.

[37] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time fourier transform," in *ICASSP*, 1983.

[38] Y. Zhou, C. Xiong, and R. Socher, "Improving end-to-end speech recognition with policy learning," *arXiv preprint arXiv:1712.07101*, 2017.