CycleGAN-based speech enhancement for the unpaired training data

Jing Yuan* and Changchun Bao†
*Beijing University of Technology, Beijing, China
E-mail: yuanjings@emails.bjut.edu.cn Tel: +86-18810673150
†Beijing University of Technology, Beijing, China
E-mail: baochch@bjut.edu.cn Tel: +86-13671188296

***Abstract*—Speech enhancement is an important task of improving speech quality in noise scenario. Many speech enhancement methods have achieved remarkable success based on the paired data. However, for many tasks, the paired training data is not available. In this paper, we present a speech enhancement method for the unpaired data based on cycle-consistent generative adversarial network (CycleGAN) that can minimize the reconstruction loss as much as possible. The proposed model employs two discriminators and two generators to preserve speech components and reduce noise so that the network could map features better for the unseen noise. In this method, the generators are used to generate the enhanced speech, and two discriminators are employed to discriminate real inputs and the outputs of the generators. The experimental results showed that the proposed method effectively improved the performance compared to traditional deep neural network (DNN) and the recent GAN-based speech enhancement methods.**

I. INTRODUCTION

Speech enhancement is used to improve speech quality and intelligibility of the degraded speech [1]. Speech enhancement covers a wide range of the applications, including teleconferencing, military eavesdropping, hearing aid devices and speech recognition devices. Moreover, it is a pre-processing module for the speech coding and recognition systems. Conventional single-channel speech enhancement methods, such as spectral subtraction [2], Wiener filtering [3], statistical model-based methods [4], and subspace algorithms [5, 6] often cause inaccurate spectral estimation of clean speech under non-stationarity noise environment. With the advance of the deep learning, deep neural network (DNN) has been applied in speech enhancement effectively. For example, the masking estimation method [7] was proposed based on the input features of noisy speech using a DNN. This method transformed the speech enhancement problem into a classification problem, in which the mapping function that is well trained could minimize the loss between the features of the enhanced speech and clean speech. In this case, the clean speech and the enhanced speech are paired for the training so that the supervised learning system is conducted. Generally, this kind of the paired data is impossible since the features and energy of noise varies with the time and scenario, that is, the varying noise could not match the speech signal. This unpaired data easily results in a mismatch of energy distribution of speech in frequency domain and makes the generalization skill of the DNN decreased.

The Generative Adversarial Networks (GANs) [8] have provided a possibility for the unpaired training data since it could generate the required output from the distribution of real data via adversarial training. At least, the GANs [9, 10] have provided better performance than the DNN in the paired data or supervised system. For example, Santiago Pascual et al. first applied GANs into the supervised speech enhancement based on the paired data [9]. Since obtaining the paired training data is a difficult and expensive task, the CycleGAN was considered for the unpaired training data in [11]. The basic idea of the CycleGAN is that the forward and backward mappings are simultaneously learned with the adversarial loss [12] and the cycle-consistency loss [13], where the cycle-consistency loss is used to constrain the parts of input information and the adversarial loss is used to identify the generated output or real input. These two losses are comprised of final cost function.

In this paper, we propose a new speech enhancement method that uses the CycleGAN to improve enhancement performance for the unpaired data. It is known that the CycleGAN has successful application in the fields of image processing. Due to the special structure of the CycleGAN, it's possible for it to enhance noisy speech recorded by different devices in real life. Specifically, our proposed model contains two generators, namely G and F and two discriminators, namely D_x and D_y. The function of generator G is to finish a mapping from x to y such that the outputs $\hat{y} = G(x)$ while the generator F is used to finish a mapping from y to x. Thus, the G and F could keep an inverses relationship. Two discriminators are used to discriminate real inputs and the generated outputs. The single optimization of adversarial cost function often

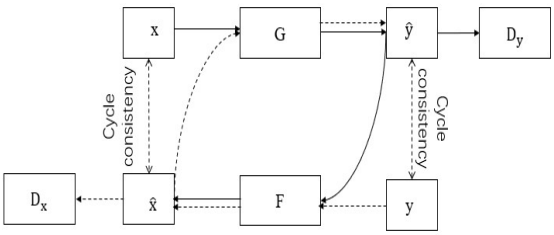

Fig. 1 Training procedure of the CycleGAN

leads to the mode collapse, that is, all inputs have the same output and the optimization process cannot be performed, so the cycle consistency loss [13] is added to meet $F(G(x)) \approx x$ and $G(F(y)) \approx y$. In this case, the speech is well preserved while the noise is effectively reduced for the unseen data.

The rest of this paper is organized as follows. Brief introduction of the CycleGAN and the detailed description of the proposed method are presented in Section II. The experimental results compared to the reference methods are given in Section III. Finally, we draw a conclusion in Section IV.

II. PROPOSED METHOD

A. Cycle-consistency Adversarial Network

The GANs work well for the paired data. For the unpaired data, it has a problem, that is, the output $\hat{y} = G(x)$ of the generator G cannot be distinguished from output domain of y by an adversarial network. In principle, the cost function can make the output y close to the input x through optimizing G [8]. Such mapping produced by G cannot guarantee one-to-one relationship between input x and output y. This often causes mode collapse of the networks. Based on this view, an inverse generator F that maps output y to input x is considered for optimizing the cost function in this paper. Except for two mappings generated by G and F, two adversarial discriminators D_x and D_y are used in this paper, where D_x is used to distinguish x and $F(y)$, D_y is used to distinguish y and $G(x)$.

In addition, the proposed method employs two cost functions, namely adversarial cost function and cycle consistency cost function, where the adversarial cost function can make the generated output similar to the input and the cycle consistency cost function can prevent G and F from contradiction each other. The training procedure is illustrated in Fig.1. As shown in Fig. 1, the training process is divided into the forward cycle consistency represented by the solid line and the backward cycle consistency represented by the dashed line. The purpose of the forward cycle consistency is to bring each input from domain x back to the output $\hat{x}$ after passing through the generator G and the generator F. The backward is in the similar way. In addition, the cycle-consistency cost function is introduced between input x (or y) and output $\hat{x}$ (or $\hat{y}$) to optimize two generators.

For the adversarial cost function, the discriminator D_y is used to discriminate the estimated data $\hat{y}$ when the training data of domain x passing through the generator G is false and the training data of domain y is true, we can describe data distribution of domain *x* and *y* as $x \sim p_{data}(x)$ and $y \sim p_{data}(y)$, respectively. The training is done through minimization of adversarial loss between the generator $G(x)$ which learns a mapping from *x* to *y* and the discriminator $D_y(y)$. Thus, the adversarial cost function can be defined as follows according to [12]:

$$\begin{aligned} L_{GAN}(G, D_y, x, y) = & E_{y \sim p_{data}(y)}\left[\log D_y(y)\right] \\ & + E_{x \sim p_{data}(x)}\left[\log(1 - D_y(G(x)))\right] \end{aligned} \quad (1)$$

where symbol $E_*(*)$ denotes the expectation about all the inputs of data domain x or y. $D_y(y)$ represents the probability that y came from the real data rather than generator's distribution. This adversarial cost function implies that G attempts to generate output G(x) that are close to the value of domain y and the discriminator D_y aims to distinguish output $\hat{y}$ generated by G and input x. $D_y(y)$ is the average prediction result that y passes through the discriminator network. G aims to minimize this cost function against the adversary D_y that tries to maximize cost function, that is, the generator G is obtained as follows:

$$G^* = \min_G \max_{D_y} L_{GAN}(G, D_y, x, y) \quad (2)$$

Given generator F and discriminator D_x, similar to (1), the second adversarial cost function $L_{GAN}(F, D_x, y, x)$ can be defined as well and the generator F is obtained as follows:

$$F^* = \min_F \max_{D_x} L_{GAN}(F, D_x, y, x) \quad (3)$$

The optimal adversarial cost function could make the outputs of the generators G and F have same distribution as the target domains of y and x, respectively. However, due to the infinitely great of domains y and x, the network may map an input value to a random value of target domain. In order to reduce dynamic range of the mapping operation, the best way is to make the mapping operation cyclically consistent. We use F to translate $\hat{y}$ back to the domain x, and constrain $F(\hat{y} = G(x))$ to be close to the input x. The similar processing is for generator G. Thus, the cycle-consistent cost function can be defined as follows [13]:

$$\begin{aligned} L_{cyc}(G, F) = & E_{x \sim p_{data}(x)}\left[\left\|F(G(x)) - x\right\|_1\right] \\ & + E_{y \sim p_{data}(y)}\left[\left\|G(F(y)) - y\right\|_1\right] \end{aligned} \quad (4)$$

where $\|\cdot\|_1$ means the L1 norm [14].

Combing three cost functions, the overall cost function can be defined as follows:

$$L(G,F,D_x,D_y) = L_{GAN}(G,D_y,x,y) + L_{GAN}(F,D_x,y,x) + \lambda L_{cyc}(G,F) \quad (5)$$

where constant λ is used to control relative importance of cycle-consistent cost function. Finally, the estimated two generators are solved as follows:

$$G^*,F^* = \arg\min_{G,F}\max_{D_x,D_y} L(G,F,D_x,D_y) \quad (6)$$

The above CycleGAN can be implemented by training two auto-encoders that have special internal structure, that is, the input can be mapped to itself through intermediate presentation layer. This setup can also be seen as a special case of adversarial auto-encoder which uses the adversarial loss to train the bottleneck layer to match any target distribution.

B. Modification of the CycleGAN structure

In order to apply the CycleGAN in the speech enhancement system under the unpaired training data, two generators G and F are combined with the modified DNN structure and the identity-mapping loss given in [15] is adopted in the CycleGAN.

Modified-DNN: Here, two generators both employ the modified-DNN structure in our method, in which the generator G transforms noisy speech into clean speech and the generator F transforms clean speech into noisy speech. Fig. 2 shows the architecture of the generator G. In the Fig. 2, $w_l(l \in \{1,\cdots,L\})$ is the weights between two adjacent layers, L is the number of hidden layers. The original output layer, $h_l(l \in \{1,\cdots,L\})$ is the output of the hidden layer. Given the input feature, namely logarithmic power spectrum (LPS) of noisy speech, the generator G could predict both speech power spectrum $\hat{P}_s(f)$ and noise power spectrum $\hat{P}_n(f)$. Thus, the Wiener filter can be embedded in the network to obtain magnitude spectrum S(f) of the enhanced speech, that is, S(f)=Y(f)H(f) where Y(f) is the magnitude spectrum of noisy speech and H(f) is the transfer function of Wiener filter.

Identity-mapping loss: The two generators are used to data generation from one domain to another. If we ask an input of a domain to pass through any generator, the output of generator should fall in the same domain. Thus, we can calculate the loss between input and output using the identity mapping that could preserve speech components without relying on extra modules. In addition, introducing additional identity-mapping loss encourages mapping to preserve a combination between the input x (or y) and output F(x) (or G(y)). When a sample of the target domain is provided as input, the identity loss function is defined as:

$$L_{id}(G,F) = E_{y\sim p_{data}(y)}[\|G(y)-y\|_1] + E_{x\sim p_{data}(x)}[\|F(x)-x\|_1] \quad (7)$$

The effectiveness of (7) has been proven in [11].

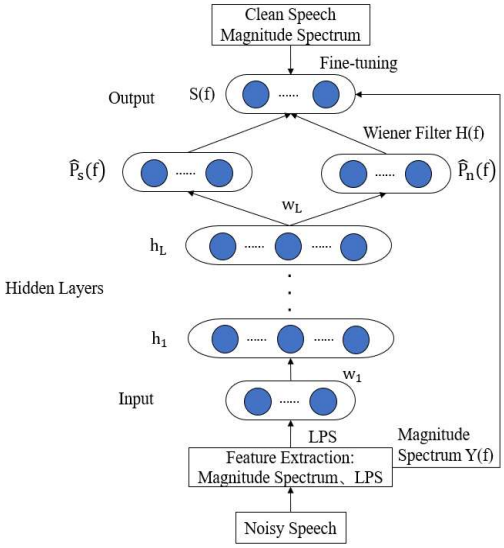

Fig. 2 The architecture of the generator G.

Because of the instability of the training process and vanishing gradient problem, the least-squares approach [17] is used instead of the cross-entropy loss in (1). Thus, (1) is changed to the following form:

$$L_{GAN}(G,D_Y,X,Y) = E_{y\sim p_{data}(y)}[(D_y(y)-1)^2] + E_{x\sim p_{data}(x)}[(D_y(G(x)))^2] \quad (8)$$

C. The proposed speech enhancement system

A block diagram of the proposed speech enhancement system is illustrated in Fig.3. The generators and discriminators are trained through adversarial way. In the training process, the goal of the generator is to generate real data to fool the discriminator, and the goal of the discriminator is to separate the generated data from the real data. In this way, the generator and the discriminator form a dynamic adversarial process. The proposed method contains two stages: training stage and enhancement stage.

In the training stage, for the forward cycle process, the LPS and magnitude spectrum of noisy speech are inputted into the generator G to estimate magnitude spectrum of clean speech. The LPS and magnitude spectrum of clean speech are inputted into the generator F to estimate magnitude spectrum of noisy speech. Two generators are optimized by comparing the cycle-consistency loss between the estimated magnitude spectrum of noisy speech and the original magnitude spectrum of noisy speech. For the backward cycle process, the estimated LPS and magnitude spectrum of clean speech are then used as the input of the generator F to obtain magnitude spectrum of the estimated noisy speech, this magnitude spectrum of the estimated noisy speech is inputted into the generator G again to obtain the magnitude spectrum of the estimated clean speech. Two generators are optimized again by comparing the cycle-consistency loss between the magnitude spectra of the estimated clean speech and original clean speech.

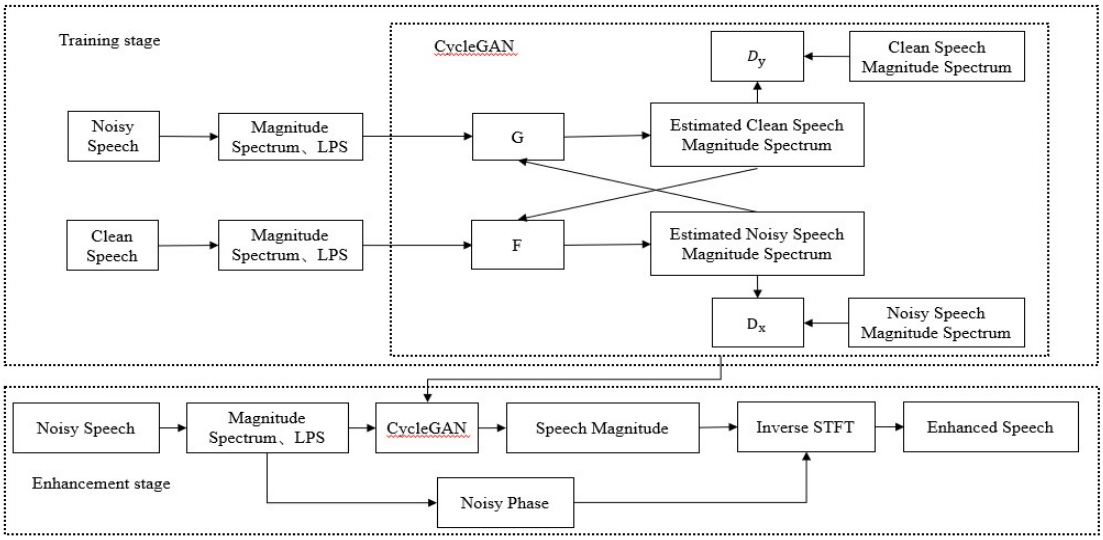

Fig. 3 The block diagram of proposed speech enhancement system.

The discriminator D_y is used to discriminate magnitude spectrum of clean speech and magnitude spectrum of the estimated clean speech generated by the generator G. The discriminator D_x is used to discriminate real magnitude spectrum of noisy speech and the magnitude spectrum of the estimated noisy speech generated by the generator F.

In the enhancement stage, the magnitude spectrum and LPS of noisy speech are used as the input of the well-trained CycleGAN to produce enhanced magnitude spectrum of noisy speech. Combined with noisy speech phase [18] and an inverse short time Fourier transform (ISTFT), the enhanced speech is obtained.

III. EXPERIMENTS

A. *Datasets used for experiments*

The proposed framework is evaluated on the TIMIT [19] corpus. In the experiment, the noisy speech and the clean speech are trained without paired, that is, they are not the paired one-one. We chose a quarter of the 4620 sentences from different speakers as the clean speech of the training set. The 102 noise types including 100 kinds of environmental noise, F16 and Babble noise are used as the noise of the training set. In addition, the other three-quarters of the 4620 sentences of clean speech and 102 noises are artificially mixed at four different signal-to-noise ratio (SNR) levels from -5 to 10dB spaced by 5dB, and then an 8-hour noisy training set is built. All signals are down sampled to 8 kHz.

The test set contains around 200 sentences from the TIMIT [19] test set. The noisy speech has 4 types of noises in which the two noises (Office, Babble) are in the training set and others two noises (Street, Factory) are outside the training set with 4 different SNR ranging from -5 to 10 dB by step of 5dB. The length of the testing set is about 10 minutes.

B. *Experimental Setups*

The LPS of noisy speech and the magnitude spectrum of noisy speech are extracted on a 32ms hamming widow with a half window overlap. The extracted LPS are normalized to have zero mean and unit variance. The modified-DNN model with 3 hidden layers including 2048 neurons are used for two generators. The feedforward multilayer perceptions (MLPs) is used for two discriminators to directly maps noisy speech LPS into magnitude spectrum of clean speech as in [20]. All the CycleGAN models used are implemented with PyTorch [21]. The networks are trained with the Adaptive Moment Estimation (Adam) algorithm [22] and a learning rate of 0.0002. The Rectified linear unit (ReLU) [23] is used as the activation function, and the mini-batch size is set to 128 for both G and D. The total epoch is 100 by step of 10 to update the generators and the discriminators.

Under the unpaired training data, our proposed method denoted as CycleGAN for comparison with reference method denoted as GAN without cycle consistency cost function. In addition, under the paired data, our proposed method is compared with the GAN network without the cycle consistency cost function and the GAN network denoted as GAN+fc that only has the forward cycle.

C. *Experimental Results*

We evaluate the enhance performance in terms of Perceptual Evaluation of Speech Quality (PESQ) [24] and Short-Time Objective Intelligibility (STOI) [25] in which the STOI is able to accurately predict the intelligibility of speech by the verification of experiments [25].

Fig.4 shows the average PESQ scores for the proposed method and the reference methods with the unpaired data at four different SNR levels. In the case

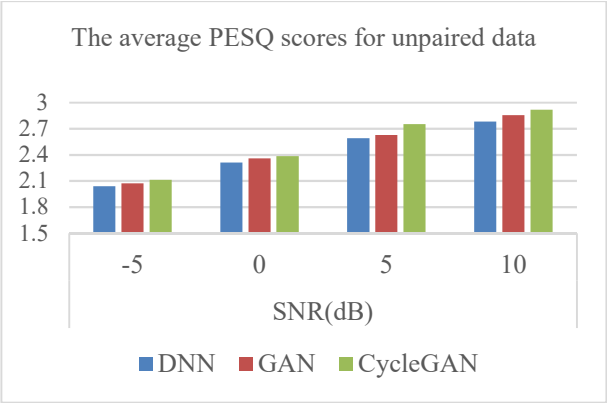

Fig. 4 The average of PESQ scores of DNN, GAN and CycleGAN.

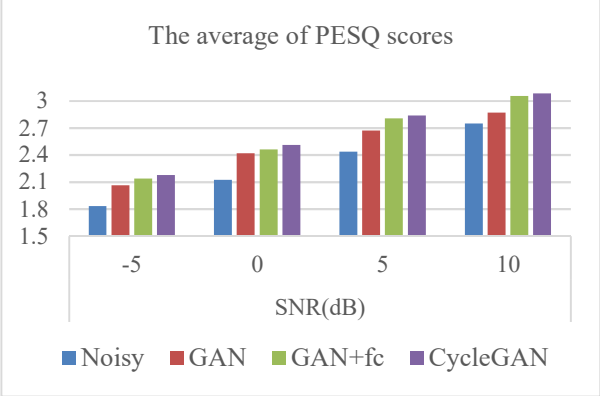

Fig. 5 The average of STOI scores of noisy speech, GAN, GAN+fc and CycleGAN at four different SNR levels.

that there is no one-to-one correspondence between noisy speech and clean speech, the proposed method and the reference methods are trained with unpaired training data. We find that the proposed method can effectively improve the speech quality by using the GAN network with cycle consistency loss function. The quality of the GAN network without the cycle consistency cost function is not much different from the traditional DNN method. This implies that for the unpaired data, the introduced cycle-consistency cost function does not lead a significant loss while transfer the input from one domain to another domain. Here, we do not take a comparison with the GAN+fc-based method for the unpaired data, because it was found through experiments that when the forward or backward cycle consistent occurs only, the instability and the mode collapse of the training are prone to be happened.

Fig.5 reports a comparison of the average PESQ scores under four noise types and four SNR levels in the paired data. We find that the proposed method is better than the reference methods under various SNR levels. In addition, the GAN network with only the forward cycle-consistency cost function is better than the GAN-only method. The GAN network with both the forward cycle loss and the backward cycle is better than the GAN+fc approach. This shows that even for paired data, the introduction of forward and backward cycle-consistency can improve network performance.

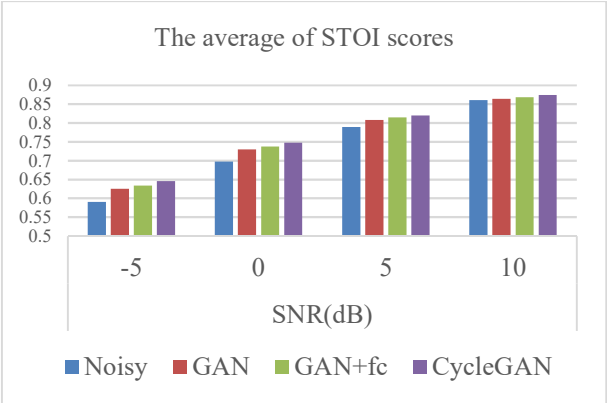

Fig. 6 The average of STOI scores of noisy speech, GAN, GAN+fc and CycleGAN at four different SNR levels.

In order to further reflect the effectiveness of the proposed method, the average intelligibility scores of the proposed method and the reference methods are compared under four different signal-to-noise ratios. It can be found from Fig. 6 that the proposed method is also improved compared with the reference methods, and the effect is obvious at the lower SNR levels. Under the condition of high SNR, no serious distortion of the speech is caused, and the influence of noise on the intelligibility is small, so the results are similar.

IV. Conclusions

In this paper, we proposed a speech enhancement method based on cycle-consistent adversarial networks with unpaired training data. Because the paired speech data is difficult to obtain or expensive in real complex scenarios, we found that when the noisy speech and the clean speech do not correspond to each other, the speech is well preserved while the noise is effectively suppressed. By comparing with the reference methods, the proposed method can better improve the speech quality and intelligibility. In the future, we will integrate spatial and temporal information into the network, because spatial and temporal information is an integral sensory component of human hearing.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 61831019, No. 61471014 and No. 61231015).

References

[1] P. C. Loizou, “Speech Enhancement: Theory and Practice,” 2nd ed. Boca Raton, FL, USA: CRC Press, Inc., 2013.

[2] M. Berouti, R. Schwartz, and J. Makhoul, “Enhancement of speech corrupted by acoustic noise,” in Proc. of the Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), vol. 4, Apr 1979, pp. 208–211.

[3] J. Lim and A. Oppenheim, “All-pole modeling of degraded speech,” IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. 26, no. 3, pp. 197–210, Jun 1978.

[4] Y. Ephraim, “Statistical-model-based speech enhancement systems,” Proceedings of the IEEE, vol. 80, no. 10, pp. 1526–1555, Oct 1992.
[5] M. Dendrinos, S. Bakamidis, and G. Carayannis, “Speech enhancement from noise: A regenerative approach,” Speech Communication, vol. 10, no. 1, pp. 45–57, 1991.
[6] Y. Ephraim and H. L. Van Trees, “A signal subspace approach for speech enhancement,” IEEE Trans. on speech and audio processing, vol. 3, no. 4, pp. 251–266, 1995.
[7] Y. Wang, A. Narayanan, and D. Wang, " On training targets for supervised speech separation," IEEE Trans. on Audio, Speech, and Language Processing, vol. 22, no. 12, pp. 1849-1858,2014.
[8] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, “Generative adversarial nets,” in International Conference on Neural Information Processing Systems, 2014, pp. 2672–2680.
[9] Santiago Pascual, Antonio Bonafonte, and Joan Serrà, “SEGAN: speech enhancement generative adversarial network,” CoRR, 2017.
[10] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alexander Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, “Wavenet: A generative model for raw audio,” in Arxiv, 2016.
[11] J. Zhu, T. Park, P. Isola and A. A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 2242-2251.
[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In NIPS, 2014.
[13] T. Zhou, P. Krähenbühl, M. Aubry, Q. Huang and A. A. Efros, "Learning Dense Correspondence via 3D-Guided Cycle Consistency," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 117-126.
[14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, “Image-to-image translation with conditional adversarial networks,” arxiv, 2016.
[15] Y. Taigman, A. Polyak, and L. Wolf, “Unsupervised cross-domain image generation,” in ICLR, 2017.
[16] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, “A regression approach to speech enhancement based on deep neural networks,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 1, pp. 7–19, 2015.
[17] Xudong Mao, Qing Li, and et al., “Least squares generative adversarial networks,” arXiv:1611.04076, 2016.
[18] D. Wang and J. Lim, “The unimportance of phase in speech enhancement,” IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. 30, no. 4, pp. 679–681, Aug 1982.
[19] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, “DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1,” NASA STI/Recon technical report n, vol. 93, 1993.
[20] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, “A regression approach to speech enhacement based on deep neural networks,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 1, pp. 7–19, 2015.
[21] Adam Paszke, Sam Gross, and Soumith Chintala, “Pytorch,” 2017.
[22] Kingma, Diederik P, and J. Ba. "Adam: A Method for Stochastic Optimization.", Computer Science, 2014.
[23] Vinod Nair and Geoffrey E. Hinton, “Rectified linear units improve restricted boltzmann machines,” in International Conference on International Conference on Machine Learning, 2010, pp. 807–814.
[24] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs,” in Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on, 2001, pp. 749-752.
[25] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time–frequency weighted noisy speech,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 7, pp. 2125-2136, 2011.