Adversarial Training Using Inter/Intra-Attention Architecture for Speech Enhancement Network

Yosuke SUGIURA* and Tetsuya SHIMAMURA*
* Faculty of Engineering, Saitama University, Saitama, Japan
E-mail: ysugiura, shima@mail.saitama-u.ac.jp Tel: +81-048-858-3776

***Abstract*—In this paper, we propose a new adversarial training for the end-to-end speech enhancement network. Taking the advantage of getting the paired training waveform, a new attention module is introduced into the proposed discriminator to incorporate the information of the desired waveform. Since this attention module has a role of the inter- and intra-attention mechanism, it helps the discriminator to distinctly distinguish the structural features underlying in the desired waveform and the waveform generated by the speech enhancement network. Unlike the other conditional generative adversarial networks, the proposed training architecture can simultaneously minimize the adversarial loss and the distortion loss. Through the simulation experiments for speech enhancement, we reveal that the proposed adversarial training can provide a significant performance.**

I. Introduction

Recently, the demands of speech communication and speech recognition have been increasing as the devices controlled with a voice user interface have been spread. Since such devices are mostly used in noisy environments, speech enhancement technique is growing in importance. In this paper, we focus on a single microphone speech enhancement technique.

Most of the state-of-the-art speech enhancement are operated in frequency domain or time-frequency domain [1], [2]. Although some of them produce excellent speech enhancement results, they usually require a little higher computational complexity because of the use of the Short Time Fourier Transform (STFT) or the wavelet analysis. The end-to-end speech enhancement, which is the time-domain speech enhancement, has the advantage of requiring low computational complexity. However, it is a challenging task since the waveform is more easily corrupted by noise than the spectral features.

Among the existing end-to-end models, Wave-U-Net [3] architecture significantly provides the outstanding performance. Wave-U-Net is composed of the stack of the downsampling fully-convolutional layers and the upsampling fully-convolutional layers. Although the several modifications of Wave-U-Net have been developed [4], [5], there remains a problem that the detailed structures are still degraded in a high noise-level situation.

There are two strategies for improving speech enhancement performance of the model: redesigning the network architecture [4]-[6], and developing a new training architecture to accelerate the learning efficiency. This paper investigates the latter strategy.

The mainstream of the training architecture to obtain high resolution is the adversarial training [7]-[9]. Speech Enhancement Generative Adversarial Network (SEGAN) [10] is one of the examples applying the adversarial training into speech enhancement. The SEGAN achieves the adversarial training using the discriminator which is designed based on DCGAN [11]. The Block diagram of the discriminator used in SEGAN is shown in Figure 1. This discriminator is composed of 11 dilated-convolutional layers. As seen in the Figure 1, the input of the discriminator is conditioned by the corresponding noisy waveform. The discriminator discriminates whether the input is the clean waveform or the waveform reconstructed by the generator, namely, real or fake. In the training scheme, the loss function composing of the distortion loss and the adversarial loss are minimized. The adversarial loss is related to the perceptual quality.

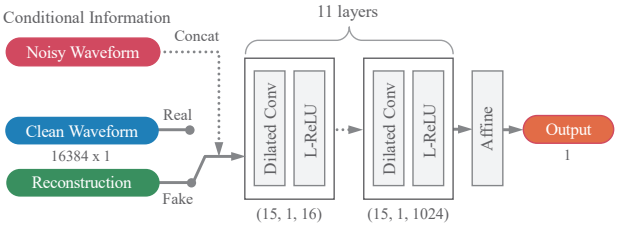

Fig. 1. Block diagram of the discriminator used in SEGAN.

Although SEGAN produced a very slight improvement in the speech enhancement performance, it may be difficult to make a furthermore improvement. Several papers [12], [13] reported that the benefit of the adversarial training is quite limited since there exists the incompatibility between the distortion loss and the adversarial loss in adversarial training. In fact, the adversarial training often induces an undesired distortion. The reference [14] also investigated the trade-off relationship existing between the degree of distortion and perceptual quality in more detail.

The resolution of the incompatibility is a major issue in the adversarial training. For speech enhancement, many researchers have been tackling this issue using multi-task learning. GSEGAN [15] introduces the acoustic feature loss as a multi-task cost function into FSEGAN [16], where the acoustic feature loss directly measures the distances of the log-power spectra and the Mel-frequency cepstral coefficients. HLGAN [17] adds a regularization of the latent vectors into the distortion loss to simultaneously minimize the variation

of the latent vector in the generator among the clean speech waveform input and the noisy waveform input. As the above methods, the additional various metrics contribute to avoiding convergence on a local solution induced by the incompatibility between the losses. Nevertheless, the additional metrics also induce another local solution.

In this paper we address to mediate the incompatibility between the distortion loss and the adversarial loss by developing a new discriminator. The proposed discriminator includes the attention mechanism to measure the similarity of the structural features between the clean speech waveform and the reconstruction. In the attention mechanism, the clean speech waveform and the input waveform corresponds to the target and the source, respectively. The effect of the attention mechanism is changed depending on whether the input waveform is the real or the fake, namely the clean waveform or the reconstruction. In the former case, the attention mechanism is equivalent to Self-Attention GAN (SAGAN) [18]. This architecture behaves as inter-attention and extracts the structural features characterizing the clean speech waveform. Meanwhile, the attention mechanism of the latter case obtains an effect of intra-attention. The discriminator in this case works to enlarge the distance of the underlying structural features between the clean speech waveform and the reconstruction.

Focusing only on the structure of the discriminator, the proposed architecture can be considered as using the discriminator which incorporates the feature map projection into the projection discriminator [20], or the discriminator which replaces the concatenation of the feature maps with the inner-product operator in the fusion discriminator [21]. The two conventional discriminators are designed based on cGAN [22] to capture the distribution of the underlying structural features. Unlike such the cGAN-based methods, the proposed method having the attention mechanism is specialized in supervised learning with paired training data to directly measure the similarity of the structures between the clean speech waveform and the reconstruction. The proposed method is therefore expected to be useful not only for the speech enhancement task but also the high-dimensional reconstruction task such as the super-resolution task.

Some experiments were conducted to evaluate the performance of the proposed method. Through the comparison with the results of the several conventional models, we reveal that the proposed method can solve the incompatibility between the distortion loss and the adversarial loss, and improve the speech enhancement performance.

II. Discriminator with Inter/intra Attention

In this section, we explain a new framework of the discriminator for speech enhancement, which discriminates a probability distribution of the structures refined by the attention map. The detail of the discriminator architecture is given in Figure 2. In this paper, both the clean speech waveform and the reconstruction are set to 16,384 samples. The discriminator is composed of a stack of the convolutional layer with two branches, the attention module coupling two branches, two convolutional layers, and an affine layer. The dimensions of the outputs from the successive layers of the network are: 16384x1 (input, conditional information), 16384x64, 16384x64, 8192x128, 4096x256, 1 for the output. The filter size of each convolutional layer is set to 15. The stride sizes of two convolutional layers after the attention module are set to 2x1 to implement the downsampling operation. The leaky ReLU with a slope of 0.1 for the negative part is used as the activation function in all layers except in the output layer.

The attention module behaves differently depending on whether the input of the discriminator is real or fake, namely, the clean speech waveform or the reconstruction. When the discriminator receives the clean speech waveform as the real input, the attention mechanism is equivalent to a Self-Attention (SA) architecture, which has an effect of intra-attention. In this case, the discriminator works to extract the structures characterizing the clean speech waveform under unsupervised learning. Meanwhile, the attention mechanism having the reconstruction as the fake input is so similar to the Source-Target Attention (STA) which has an effect of inter-attention. The mask generated by the inter-attention mechanism calculates the structural similarity between the reconstruction and the clean speech waveform, and so it helps to distinguish between the common and the different structures lying in them.

In this paper, $\boldsymbol{x}$ and $\boldsymbol{y}$ stand for the noisy speech waveform and the corresponding clean speech waveform, respectively. Using p to designate the true distributions of pair data $(\boldsymbol{x}, \boldsymbol{y})$, the discriminator cost function relaxed by the hinge function [18] is given by

$$\begin{aligned} L_D &= E_{\boldsymbol{y}\sim p}\left[\max\left\{0, 1 - D(\boldsymbol{y}, \boldsymbol{y})\right\}\right] \\ &\quad + E_{(\boldsymbol{x},\boldsymbol{y})\sim p}\left[\max\left\{0, 1 + D\left(G(\boldsymbol{x}), \boldsymbol{y}\right)\right\}\right]. \end{aligned} \quad (1)$$

The generator cost function is defined by

$$L_G = \left\|G(\boldsymbol{x}) - \boldsymbol{y}\right\|_1 - \lambda E_{(\boldsymbol{x},\boldsymbol{y})\sim p}\left[D\left(G(\boldsymbol{x}), \boldsymbol{y}\right)\right]. \quad (2)$$

The first term and the second term on the right side in (2) are called the distortion loss and the adversarial loss, respectively. In (2), λ is a regularization parameter with a positive value, which adjusts the regularization strength of the adversarial loss against the distortion loss. To guarantee that the discriminator satisfies 1-Lipshitz constraint, we use an additional regularization of spectral normalization [19] for training.

III. Comparison with other Adversarial Training

In the proposed discriminator, the clean speech waveform has a role as the conditional information, whereas other conventional methods based on cGAN normally treat the noisy speech waveform as the conditional information of the discriminator. From this standpoint, our design concept of the discriminator is essentially different with cGAN, but rather similar to the attention-based GAN.

The discriminator naively conditioned by the additional information has a problem that it is difficult to capture the

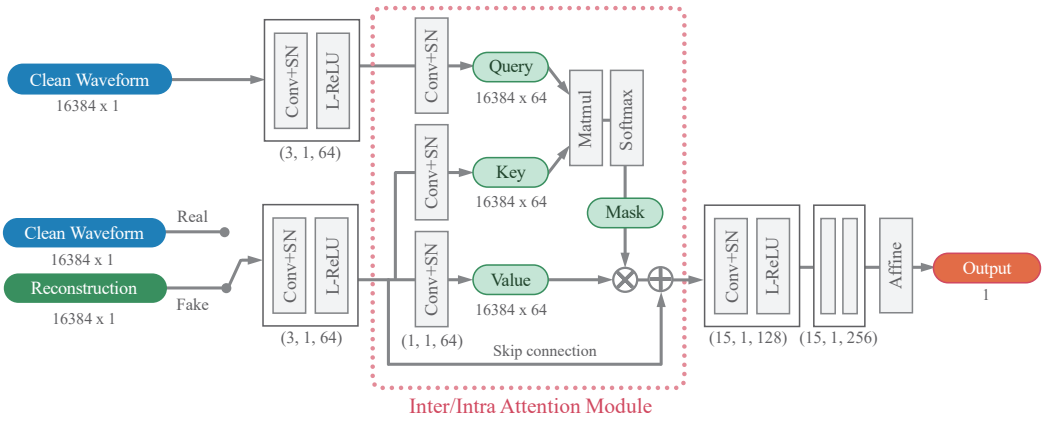

Fig. 2. Block diagram of the discriminator with Inter/Intra attention module for speech enhancement.

statistical features of the potential structures. The projection discriminator [20] and the fusion discriminator [21] were designed to solve that problem by incorporating the conditional information using the inner-product or the fusion of the feature map. Although they assess the discrepancy of the structural probability distribution between the unpaired data of the truth and the target, they can not directly assess the similarity of the structures between the paired data of the generated data and the desired data owing to the framework of cGAN.

Meanwhile, the attention based GAN, such as SAGAN [18] and spatial attention GAN [23], achieves either inter-attention or intra-attention but have no framework to simultaneously implement both features.

Thus, the proposed discriminator contributes to capturing the structural features to distinctly distinguish between the clean waveform and the reconstruction with the inter/intra-attention module.

IV. Archtecture of Generator

In this paper, the Wave-U-Net [3] is used as the generator. The Wave-U-Net is a state-of-the-art architecture for end-to-end speech enhancement. Figure 3 shows the architecture of the Wave-U-Net. The encoder and decoder half and double the resolution of the feature map in each downsampling block and upsampling block, respectively. The feature skip-connections from the encoder layers and the decoder layers help to restore the fine structures.

In the proposed method, the spectral normalization is used to stabilize the training of the generator, unlike the original training architecture. The spectral normalization guarantees that the generator satisfies the following relation:

$$\max_{x} \frac{\|G(\boldsymbol{x})\|_2}{\|\boldsymbol{x}\|_2} \leq 1 \;\Rightarrow\; \|\boldsymbol{x}\|_2 \geq \|G(\boldsymbol{x})\|_2 \,. \tag{3}$$

Since this relation plays a similar role as a bounded restriction on the generator discussed in [1], the spectral normalization brings another benefit that the solution space can be efficiently reduced in the speech enhancement task.

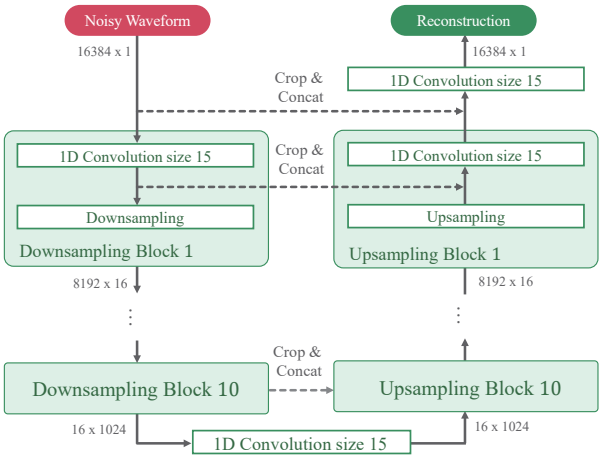

Fig. 3. The Wave-U-Net architecture

V. Experimental Setup

A. Dataset

To evaluate the performance of the proposed architecture on the speech enhancement task, we employed VCTK speech dataset [24] and DEMAND dataset [25], which are the same datasets used in [10]. For the training set, 10 types of noise and 10 different sentences with 4 signal-to-noise ratio (SNR) (15, 10, 5, and 0 dB) were used. For the test set, 5 types of noise and 4 different sentences with 4 SNR (17.5, 12.5, 7.5, and 2.5 dB) were used. Each speech data has the sampling frequency of 16kHz. During training, the speech waveform segments with length of 16,384 samples were extracted from the training data with 50% overlap. During testing, the length of the speech waveform segment was the same as that of training, but the ratio of the overlap was changed to 75%. As in [10], a high-frequency pre-emphasis filter of coefficient

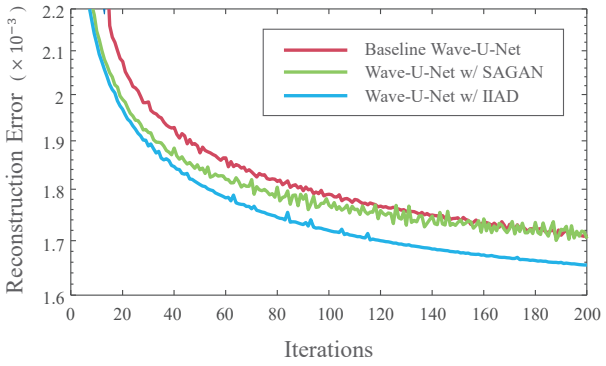

Fig. 4. Learning curves of three different models on the training set

0.95 was applied to each segment of the waveform. The segments of the waveform produced by the trained model were de-empasized and eventually concatenated to reconstruct the enhanced speech waveform.

B. Experiments

To confirm the effect of the proposed training architecture, we experimented with three models applied over the aforementioned data, where the three models are Wave-U-Net, Wave-U-Net with SAGAN, and Wave-U-Net with the proposed discriminator that we name Inter/Intra-Attention Discriminator (IIAD). For the training of all models, we used the Adam optimizer [26] with $\beta_1 = 0.9$ and $\beta_2 = 0.9$. The learning rates for the generator and the discriminator are respectively 0.0001 and 0.0005, which are designed based on two-timescale update rule (TTUR) [27]. All models were trained for 200 epochs with random minibatches of size 20. As mentioned above, the proposed training architecture used the spectral normalization in the discriminator as well as in the generator for stable training. For more stable training, the zero-centered gradient penalty [28] was used only in the discriminator, which is a modification of the gradient penalty used in Wasserstein GAN with gradient penalty (WGAN-GP) [29].

Figure 4 shows the learning curves of three models. In this figure, the vertical axis shows the distortion loss defined by $\|G(\boldsymbol{x}) - \boldsymbol{y}\|_1$ and the horizontal axis shows the number of epochs. The displayed values of the distortion loss are averaged over all the iterations in each epoch. As seen in this figure, the proposed training architecture provides the lowest distortion error, while the SAGAN induced a little worse convergence around 200 epochs than the baseline Wave-U-Net because of the incompatibility between the distortion loss and the adversarial loss. At 200 epochs, the proposed method reduced about 12.9% of the distortion error in comparison with the other models. As seen from this result, the proposed method can relax the trade-off relationship between the distortion loss and the adversarial loss, and thus improve the training efficiency.

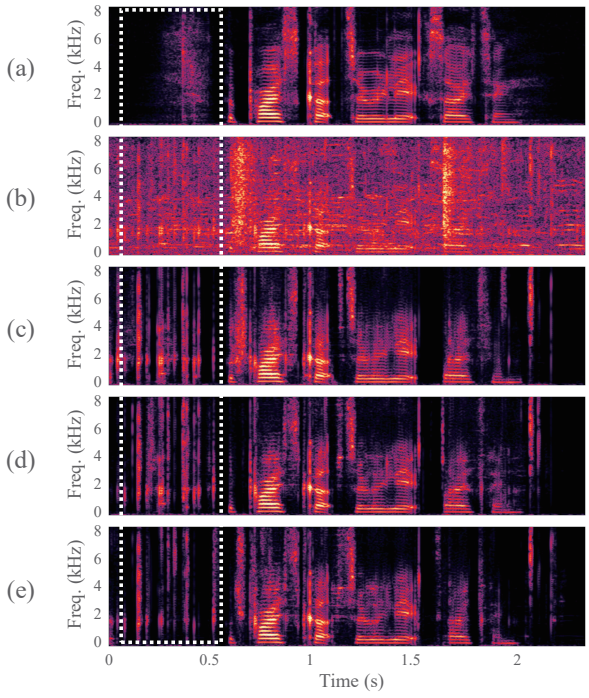

Fig. 5. Spectrograms of speech speech p257_070.wav in the test: (a) clean speech, (b) noisy speech, (c) enhanced speech by Wave-U-Net, (d) enhanced speech by Wave-U-Net with SAGAN, (e) enhanced speech by Wave-U-Net with IIAD.

C. Objective Evaluation

To assess the quality of the reconstruction signals, we used six objective metrics including PESQ [30], CSIG, CBAK, COVL [31], Segmental SNR (SSNR), and STOI [32]. PESQ measures speech quality, which returns a score from 4.5 to -0.5, with higher scores indicating better quality. CSIG is a MOS predictor of speech distortion (from 1 to 5), CBAK is a MOS predictor of intrusiveness of background noise (from 1 to 5), and COVL is a MOS predictor of overall processed speech quality (from 1 to 5). STOI whose score ranges from 0 to 1 is a measure used to predict the intelligibility of speech.

For evaluation, we compared other five end-to-end models in addition to the above three models: SEGAN, HLGAN, WGAN-GP, SERGAN [33], DSEGAN. Table I shows the experimental results of the different models. This table summarizes that the proposed training architecture outperforms the other conventional methods at all assessment methods. Only Wave-U-Net with SAGAN provides a similar performance to Wave-U-Net with IIAD.

Figures 5 demonstrates the resulting spectrograms of p257_070.wav which is a noisy female speech signal with low SNR included in the test set. Focusing on the silenced region enclosed in the white dashed box, both Wave-U-Net and Wave-U-Net with SAGAN have an insufficient performance in removing the ambient noise. Meanwhile, the Wave-U-Net with IIAD successes in suppressing the noise compared with

TABLE I
OBJECTIVE EVALUATION RESULTS OF EIGHT DIFFERENT MODELS ON THE TEST SET OF VCTK DATASET

Model	PESQ	CSIG	CBAK	COVL	SSNR	STOI
Noisy	1.97	3.35	2.44	2.63	1.68	0.921
SEGAN [10]	2.16	3.43	2.94	2.80	7.73	-
HLGAN [17]	2.48	3.65	3.19	3.05	9.21	-
WGAN-GP [33]	2.54	-	-	-	-	0.937
SERGAN [33]	2.62	-	-	-	-	0.940
DSEGAN [6]	2.39	3.46	3.11	2.90	8.72	0.933
Wave-U-Net [3]	2.40	3.52	3.24	2.96	9.97	-
Wave-U-Net with SAGAN	2.76	4.09	**3.38**	3.42	**10.3**	**0.948**
Wave-U-Net with IIAD	**2.80**	**4.11**	3.37	**3.45**	10.0	0.944

the other methods.

VI. CONCLUSION

In this paper, we proposed a new training architecture for the end-to-end speech enhancement network. The proposed discriminator adopts the attention module fusing inter-attention and intra-attention. This discriminator architecture can break the strong trade-off relationship between the distortion loss and the adversarial loss existing in the common adversarial training. From the experimental results, we reveal the effectiveness of the proposed training architecture.

REFERENCES

[1] H. Choi, J. H. Kim, J. Huh, A. Kim, J. W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex U-Net," in *Proc. of ICLR 2019*, New Orleans, USA, April 2019.

[2] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le, R. R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Proc. of LVA/ICA 2015*, Liberec, Czech Republic, Aug. 2015.

[3] C. Macartney and T. Weyde, "Improved speech enhancement with the Wave-U-Net," in *Proc. of NIPS 2018*, Montreal, Canada, Nov. 2018.

[4] R. Giri, U. Isik, and A. Krishnaswamy, "Attention Wave-U-Net for speech enhancement," in *Proc. of WASPAA 2019*, New York, USA, Dec. 2019.

[5] A. Pandey and D. Wang, "A new framework for CNN-based speech enhancement in the time domain," *IEEE/ACM Trans. Audio, Speech, and Language Process.*, vol. 27, no.7, pp.1179–1188, July 2019.

[6] H. Phan, I. V. McLoughlin, L. Pham, O. Y. Chen, P. Koch, M. De Vos, and A. Mertins, "Improving GANs for speech enhancement," in *arXiv preprint arXiv:2001.05532*, 2020.

[7] I. Goodfellow, J. P. Abadie, M. Mirza, B. Xu, D. W. Farley, et al., "Generative adversarial nets," in *Proc. of NIPS 2014*, Montreal, Canada, pp. 2672–2680, Dec. 2014.

[8] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proc. of ICLR 2018*, Vancouver, Canada, Nov. 2018.

[9] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *Proc. of ICLR 2019*, New Orleans, USA, May 2019.

[10] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech enhancement generative adversarial network," in *Proc. of Interspeech 2017*, Stockholm, Sweden, Aug. 2017.

[11] A. Radford, M. Luke, and C. Soumith, "Unsupervised representation learning with deep convolutional generative adversarial networks," arXiv preprint arXiv:1511.06434, 2015.

[12] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. of CVPR 2017*, Hawaii, USA, July 2017.

[13] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, et al., "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proc. of ECCV 2018*, Munch, Germany, Sep. 2018.

[14] Y. Blau and T. Michaeli, "The Perception-Distortion Tradeoff," in *Proc. of CVPR 2018*, Utah, USA, June 2018.

[15] S. Pascual, J. Serrà, and A. Bonafonte, "Towards generalized speech enhancement with generative adversarial networks," in *Proc. of INTERSPEECH 2019*, Graz, Austria, Sep. 2019.

[16] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," in *Proc. of ICCASP 2018*, Alberta, Canada, April 2018.

[17] F. Yang, Z. Wang, J. Li, R. Xia, and Y. Yan, "Improving generative adversarial networks for speech enhancement through regularization of latent representations," *Speech Communication*, vol. 118, pp. 1–9, April 2020.

[18] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proc. of ICML 2019*, California, USA, June 2019.

[19] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *Proc. of ICLR 2018*, Vancouver, Canada, April 2018.

[20] T. Miyato and M. Koyama, "cGANs with projection discriminator," in *Proc. of ICLR 2018*, Vancouver, Canada, April 2018.

[21] F. Mahmood, W. Xu, N. J. Durr, J. W. Johnson, and A. Yuille, "Structured prediction using cGANs with fusion discriminator," in *Proc. of ICLR 2018*, Louisiana, USA, April 2018.

[22] M. Mirza and S. Osindero, "Conditional generative adversarial nets," in *arXiv preprint arXiv:1411.1784*, Nov. 2014.

[23] G. Zhang, M. Kan, S. Shan, and X. Chen, "Generative adversarial network with spatial attention for face attribute editing," in *Proc. of ECCV 2018*, Munich, Germany, Sept. 2018.

[24] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *Proc. of Oriental COCOSDA*, Sep. 2013.

[25] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multichannel acoustic noise database: A database of multichannel environmental noise recordings," *J. Acoust. Soc. Am.*, vol.133, no.5, pp. 3591—3591, Sep. 2013.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc, of ICLR 2015*, Vancouver, Canada, May 2015.

[27] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," in *Proc, of NIPS 2017*, California, USA, Dec. 2017.

[28] K. Roth, A. Lucchi, S. Nowozin, and T. Hofmann, "Stabilizing training of generative adversarial networks through regularization," in *Proc, of NIPS 2017*, California, USA, Dec. 2017.

[29] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein GANs," in *Proc, of NIPS 2017*, California, USA, Dec. 2017.

[30] ITU-T Std., *P.862.2: Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs*, ITU-T, 2007.

[31] Y. Hu, and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol.16, no.1, pp.229–238, Dec. 2007.

[32] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. of ICASSP 2010*, Texas, USA, Mar. 2010.

[33] D. Baby and S. Verhulst, "SERGAN: Speech enhancement using relativistic generative adversarial networks with gradient penalty," in *Proc. of ICCASP 2019*, Brighton, UK, 2019.