A Loss With Mixed Penalty for Speech Enhancement Generative Adversarial Network

Jie Cao*, Yaofeng Zhou*, Hong Yu†, Xiaoxu Li*, Dan Wang‡ and Zhanyu Ma§
* Lanzhou University of Technology, Lanzhou, China
† Ludong University, Yantai, China
‡ South-to-North Water Diversion Middle Route Information Technology Co., Ltd., China
§ Beijing University of Posts and Telecommunications, Beijing, China
E-mail:{hy@ldu.edu.cn, xiaoxulilut@gmail.com, mazhanyu@bupt.edu.cn}

***Abstract*—Speech enhancement based on generative adversarial networks (GANs) can overcome the problems of many classical speech enhancement methods, such as relying on the first-order statistics of signals and ignoring the phase mismatch between the noisy and the clean signals. However, GANs are hard to train and have the vanishing gradients problem which may lead to generate poor samples. In this paper, we propose a relativistic average least squares loss function with a mixed penalty term for speech enhancement generative adversarial network. The mixed penalty term can minimize the distance between generated and clean samples more effectively. Experimental results on Valentini 2016 and Valentini 2017 dataset show that the proposed loss can make the training of GAN more stable, and achieves good performance in both objective and subjective evaluation.**

***Index Terms*—loss function, generative adversarial networks, speech enhancement, convolutional neural networks**

I. Introduction

Speech enhancement technology [1], [2] aims to improve speech quality and intelligibility by removing background noise from speech. It is an important research topic in the field of signal processing and has been widely used in mobile communication, smart home, human-computer interaction and other fields.

Classical speech enhancement methods, e.g. Wiener filtering [3], spectral subtraction [4], and statistical model-based methods [5] are all belongs to unsupervised methods. These methods are suitable for stationary noise environments, while in non-stationary or low signal-to-noise ratios (SNR) conditions, classical enhancement methods will produce much residual noise which affects the perceived quality and intelligibility of speech. Therefore, scholars develop many neural networks based enhancement methods to learn the nonlinear relationship between noisy and clean speeches in order to build a more adaptable speech enhancement method which are suitable for complicated noise environments. Lu et al. [6] proposed a stacked denoising autoencoder to generate enhanced speeches, which achieves better results on noise known conditions than classic unsupervised methods. Xu Yong et al. [7] proposed a deep neural network (DNN) speech enhancement method, which uses DNN to learn the non-linear mapping relationship between noisy and clean speeches. In paper [8], the author applied some different strategies to improve the performance of DNN based enhancement method in complex noise environments. Besides DNN, some other neural network architectures such as RNN and LSTM are all applied on speech enhancement tasks [9]. Recently, with the emergence of the generative adversarial network (GAN) [10], the method of speech enhancement has a new breakthrough. Daniel et al. [11] proposed a speech enhancement method based on conditional generative adversarial network (cGAN). The results show that the performance is comparable to classical methods and DNN. Due to the disadvantage of traditional Jensen-Shannon divergence (JS Div), the Wasserstein distance [12] is proposed to replace the JS distance, which improves the stability of the GAN. Shan Qin [13] combined Wasserstein distance with cGAN [14] to propose a new speech enhancement GAN. In addition, some time domain based speech enhancement methods have been proposed recently. Santiago et al. [15] proposed an end-to-end speech enhancement system based on least squares GAN. Deepak et al. [16] applied the relative discriminator to cGAN and added a gradient penalty term [17] to the discriminator, which improves the stability of the speech enhancement GAN. However, most of the current speech enhancement methods are still based on the short-time Fourier analysis framework [18], these methods ignore the effect of the short-time phase to speech enhancement. And other time domain methods based on GANs have the vanishing gradients problem and difficult to train, which lead to poor quality of generated samples.

In this work, we apply relativistic average discriminator to the least squares loss function [19] and add a mixed penalty term in the generator (G) loss. The proposed method is more stable for GAN training and can preserve a variety of feature in speech data. We explored the performance of the proposed method on noise unknown conditions and low SNR conditions. Experimental results show that the proposed method has better performance in both objective and subjective evaluation compared with other baseline methods.

II. Speech Enhancement Based on RAGAN

A. *Least Squares Generative Adversarial Networks*

Comparing the standard GAN (SGAN) using cross entropy (CE) as loss function, the least squares generative adversarial

*Corresponding authors: Hong Yu and Xiaoxu Li.

network (LSGAN) is more stable for network training and can generate samples with better quality. In LSGAN, a novel least squares loss function which minimizes the euclidean distance between the distribution of generated samples and real samples is proposed to replace the classical CE loss function. The LSGAN loss functions are defined as follows:

$$\begin{aligned} L(D) = & \frac{1}{2} E_{x \sim r(x)} [(D(x) - 1)^2] \\ & + \frac{1}{2} E_{z \sim f(z)} [(D(G(z)))^2], \end{aligned} \tag{1}$$

$$L(G) = \frac{1}{2} E_{z \sim f(z)} [(D(G(z)) - 1)^2], \tag{2}$$

where D and G are discriminator and generator respectively, x is real sample, z is random noise vector and $G(z)$ is generated sample; 1 and 0 are the labels of the real and generated samples.

B. Relativistic Average Generative Adversarial Network

In the SGAN, the probability of real samples being real will not decrease as the probability of fake samples being real increases, which may affect the validity of the JS Div calculation and contradict the prior knowledge. To solve this problem, the relativistic average generative adversarial network (RaGAN) [20] is proposed. Instead of computing the distance between discriminator/generator outputs to ground truth labels, the objective function of RaGAN evaluate the probability of the given real samples is more realistic than fake samples, which significantly improve the stability of GANs and the quality of generated samples. With the objective functions are:

$$\begin{aligned} L(D) = & E_{x \sim r(x)} [f_1(D(x) - E_{z \sim f(z)} D(z)] + \\ & E_{z \sim f(z)} [f_2(D(z) - E_{x \sim r(x)} D(x))], \end{aligned} \tag{3}$$

$$\begin{aligned} L(G) = & E_{x \sim r(x)} [g_1(D(x) - E_{z \sim f(z)} D(z)] + \\ & E_{z \sim f(z)} [g_2(D(z) - E_{x \sim r(x)} D(x))], \end{aligned} \tag{4}$$

where f_1, f_2, g_1, g_2 are scalar-to-scalar functions.

C. The Proposed Loss

In this work, we design the speech enhancement GAN by using the convolutional neural network [21] which is an end to end system and can extract feature from time domain. In order to make the training of the model be more stable and improve the quality of generated samples, we apply the relativistic average discriminator to the least squares loss functions. Besides, some extra input $\tilde{x}$ are added to G and D to perform mapping and classification. The objection functions are:

$$\begin{aligned} L(D) = & \frac{1}{2} E_{x, \tilde{x} \sim r(x, \tilde{x})} [(D(x, \tilde{x}) - E_1 - 1)^2] + \\ & \frac{1}{2} E_{z \sim f(z), \tilde{x} \sim r(\tilde{x})} [(D(G(z, \tilde{x}), \tilde{x}) - E_2)^2], \end{aligned} \tag{5}$$

$$\begin{aligned} L(G) = & \frac{1}{2} E_{z \sim f(z), \tilde{x} \sim r(\tilde{x})} [(D(G(z, \tilde{x}), \tilde{x}) - E_2 - 1)^2] \\ & + \frac{1}{2} E_{x, \tilde{x} \sim r(x, \tilde{x})} [(D(x, \tilde{x}) - E_1)^2], \end{aligned} \tag{6}$$

with

$$E_1 = E_{z \sim f(z), \tilde{x} \sim r(x)} [D(G(z, \tilde{x}), \tilde{x})], \tag{7}$$

$$E_2 = E_{x, \tilde{x} \sim r(x, \tilde{x})} [D(x, \tilde{x})], \tag{8}$$

where E_1 and E_2 are the average of real samples and fake samples in a training batch, respectively.

In the loss function of G, we add a mixed penalty term consisting of L_1 norm and mean square error (MSE). In the experiment, we found that the mixed penalty term can improve the quality of generated speeches and be more effective than only L_1 norm added. The distance between generated and real samples can be minimized by adjusting the hyper-parameters of L_1 norm and MSE.

$$L_1 = \| G(z, \tilde{x}) - x \|_1. \tag{9}$$

$$L_{MSE} = \frac{1}{N} \sum_{i=1}^{N} (G^{(i)}(z, \tilde{x}) - x^{(i)})^2. \tag{10}$$

Finally, the G loss becomes

$$L_{our}(G) = L(G) + aL_1 + bL_{MSE}, \tag{11}$$

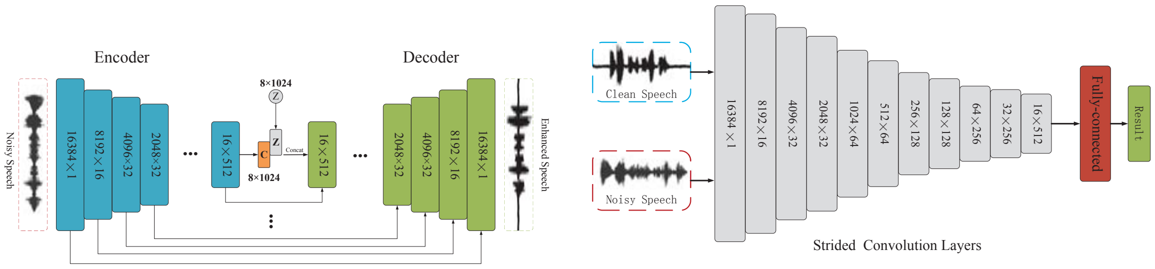
where *a* and *b* are hyper-parameters that control L_1 norm and L_{MSE}.

III. Experimental Setup

A. Datasets

To evaluate the performance of our method, we use the same Valentini 2016 dataset [22] as the SEGAN [15]. The dataset includes 30 speakers (11572 utterances) where 28 speakers are used as train set and the utterances of the other 2 speakers are used as testing. In order to generate the noisy training set, we consider 40 different noise conditions with 10 types of noise (two artificial noise and eight real noise collected from the DEMAND [23]) and four different SNRs (15, 10, 5, and 0dB). Every training speaker has approximately 10 different sentences in each condition. The noise conditions of the test set are different from the training set. We consider 20 different noise conditions with five types of noise from the DEMAND and four different SNRs (17.5, 12.5, 7.5, and 2.5dB). Every test speaker has approximately 20 different sentences in each condition.

To evaluate the performance of our method at low SNR conditions, we select two speakers from Valentini 2017 dataset to make another test set. There are 20 different noise conditions are considered where includes five types of noise from the DEMAND and four different SNRs (15, 10, 5, and 0dB). Each test speaker has approximately 20 different sentences in each condition.



(a) Network structure of the generator

(b) Network structure of the discriminator

Fig. 1. (a) The G network is an encoder-decoder architecture. C and Z are the thought vector and the latent vector respectively. The arrow lines denote skip connections. (b) The D network has two input channels and we use VBN and LeakyReLU in it.

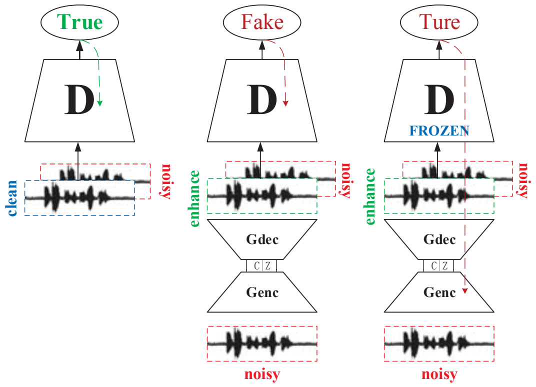


Fig. 2. Training process of GAN. The model trains the D network and the G network alternately until convergence and the parameters are updated by back propagation algorithm.

B. Network Structure

In this work, we use the similar structure mentioned in the SEGAN [15]. As shown in Fig. 1 (a), we use a fully convolutional auto-encoder as the G network, which performs the enhancement. The G network includes 22 one-dimensional strided convolutional layers with parametric rectified linear units (PReLU) as activation. The filter-width of each layer is set as 31 and the stride is set as two. For GAN training, strided convolution is more stable than other pooling methods. In encoding stage, the dimensions of each layer (samples×feature maps) are 16348×1, 8192×16, 4096×32, 2048×32, 1024×64, 512×64, 256×128, 128×128, 64×256, 32×256, 16×512 and 8×1024. The decoding stage is the mirroring process of the encoding stage with the same number of filters and network parameter configurations. Besides, skip connections are added in G network to pass fine-grained information of speech to the decoding stage.

The D network is typically a binary classifier which has the similar one-dimensional strided convolution structure as the G's encoding stage (Fig. 2 (b)). However, D network has two input channels and it uses Virtual Batch Normalization (VBN) [24] in convolutional layer before Leaky ReLUs. In the last activation layer there is an additional one-dimensional convolution layer with one filter of width one, which can reduce the amount of parameters required for the final classification of neurons.

C. Experimental Parameter Settings

The model is trained for 100 epoch with RMSprop [25] and the learning rate is set as 0.0002, using an effective batch size of 100. Fig. 2 shows how GAN is trained. In the experiment, all original speeches should be down-sampled to 16kHz, and we use a sliding windows with 500 ms length (50% overlap) to extract chunks of speech samples. Each chunk has about 16384 samples. We also apply a high frequency pre-emphasis filter with a coefficient of 0.95 to all input speech samples. In additions, we set a = 100 and b = 20 in equation (11).

IV. Experimental Results

A. Objective Evaluation

To evaluate the performance of our method we compute the following objective measures: Perceptual Evaluation of Speech Quality (PESQ) [26]; MOS prediction of the signal distortion attending only to the speech (CSIG) [27]; MOS Prediction of the intrusiveness of background noise (CBAK) [27]; MOS of prediction of the overall effect (COVL) [27]; Short-Time Objective Intelligibility (STOI) [28]. All metrics will test on the entire dataset and the higher values mean the better performance.

TABLE I
Objective evaluation results comparing the baseline methods and Our method on Valentini 2016.

Metric	Noisy	Wiener	SEGAN	Ralsgan-L_1	Ours
PESQ	1.97	2.22	2.16	2.2483	**2.2646**
CSIG	3.35	3.23	3.48	3.5361	**3.5707**
CBAK	2.44	2.68	2.94	**2.9821**	2.9475
COVL	2.63	2.67	2.80	**2.8721**	2.8313
STOI	0.9210	0.9144	0.9250	0.9270	**0.9316**

Table I shows the objective evaluation results of different speech enhancement methods on Valentini 2016 dataset. We can observe that our method has higher PESQ and STOI scores than other baseline methods. Comparing with SEGAN and Wiener, the PESQ of our method increases 4.8% and 2%

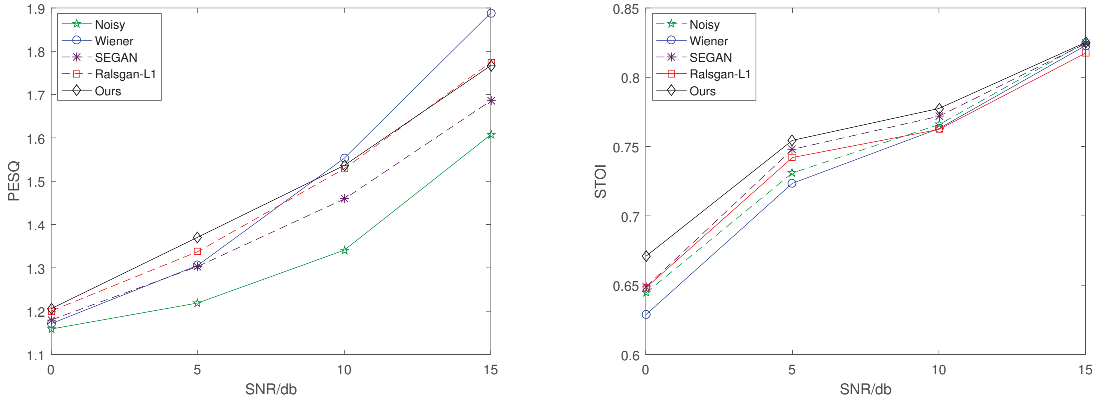


Fig. 3. PESQ and STOI results on Valentini 2017 at different SNR conditions.

TABLE II
Objective evaluation results comparing the baseline methods and Our method on Valentini 2017.

Metric	Noisy	Wiener	SEGAN	Ralsgan-L_1	Ours
PESQ	1.3316	**1.4797**	1.4067	1.4603	1.4701
CSIG	2.1596	2.0579	**2.5572**	2.5396	2.5570
CBAK	1.8189	1.8604	2.1325	2.1711	**2.2599**
COVL	1.6650	1.6693	1.9253	**1.9323**	1.8955
STOI	0.7416	0.7345	0.7480	0.7427	**0.7570**

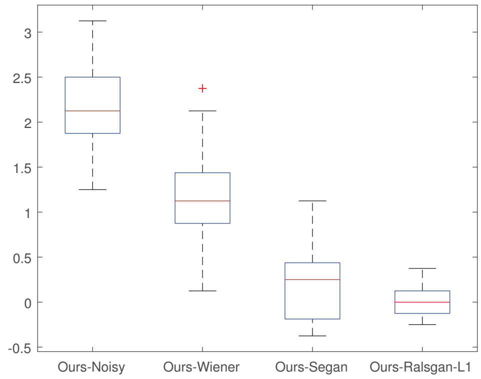


Fig. 4. CMOS box plot. The red central marks are the median, and the edges of the box are the 25th and 27th percentiles respectively. Positive values mean that our method is preferred

respectively, and it also slightly higher than Ralsgan-L_1 (Our method without MSE term). Table II shows the performances of different enhancement methods at low SNR conditions on Valentini 2017 dataset. It can be observed that our method performs slightly worse on PESQ than Wiener, but it has the highest STOI scores. Comparing with SEGAN and Ralsgan-L_1, our method can get better PESQ and STOI scores. Overall, our method performs better than other baseline methods at low SNR conditions.

In order to evaluate the speech enhancement performances of our method in different noise level, we select PESQ and STOI as assessment criteria. The experimental results on Valentini 2017 are shown in Fig.3. It can be observed that our method has higher score on PESQ compared with other three baseline methods when the SNR below 10dB. Especially, our method has a 2.9% and 2.2% improvement at 0 dB compared with Wiener and SEGAN on PESQ respectively, and it also performs slightly better than Ralsgan-L_1. In addition, our method always has better performance on STOI at different SNR conditions compared with other three baseline methods. These two experiments show that our method can obtain good performance at low SNR conditions.

TABLE III
Subjective evaluation results comparing the baseline methods and Our method on Valentini 2017.

Metric	Noisy	Wiener	SEGAN	Ralsgan-L_1	Ours
MOS	1.73	2.84	3.74	3.91	**3.96**

B. *Subjective Evaluation*

Subjective evaluation is a perceptual test which has a total of 16 listeners with 20 different sentences from Valentini 2017 dataset. We will give five forms (Noisy signal, Wiener-enhanced signal, SEGAN-enhanced signal, Ralsgan-L_1-enhanced signal and Our method enhanced signal) of speech files for each sentence randomly. The listeners use a scale from 1 to 5 to evaluate the overall quality and we calculate the average score as the final score for each method. In Table III, we can observe that our method has better perceived quality than other baseline methods.

Moreover, we calculate comparative MOS (CMOS) by subtracting the MOS of the two compared methods, describing which speech enhancement method the listener prefers to. Fig. 4 shows how the speech enhanced by our method are preferred. More specifically, compared with SEGAN system, 65% of the cases prefer our method, and 27.5% prefer SEGAN. Compared with Ralsgan-L_1, 42.5% of the cases prefer our method and 37.5% prefer Ralsgan-L_1 (no preference in 20% of the case).

V. CONCLUSIONS

In this work, we proposed a novel GAN based speech enhancement model which uses relativistic average discriminator on least squares loss function and add a mixed penalty term in the G loss. The experimental results on Valentini 2016 and Valentini 2017 dataset show that our method performs better than other baseline methods and it can obtain good performance at low SNR conditions. However, the noise types in our experiment are limited and the G network still lost part of fine-grained information. Therefore, in the future work we will improve the generalization ability of the model for unknown noise types and keep more fine-grained information.

ACKNOWLEDGEMENTS

This work was partly supported by the National Key Research and Development Program of China under Grant 2018YFC0807205, the National Natural Science Foundation of China (NSFC) grant No.61563030, No.61763028, No.61773071, No.61922015 and No.61906080, the Natural Science Foundation of Gansu Province, China, grant No.17JR5RA125, the Beijing Nova Program Interdisciplinary Cooperation Project No. Z181100006218137, Beijing Nova Program No. Z171100001117049 and by the Hong-liu Outstanding Youth Talents Foundation of Lanzhou University of Technology.

REFERENCES

[1] Taha, Tayseer MF, Ahsan Adeel, and Amir Hussain, "Survey on Techniques for Enhancing Speech," *International Journal of Computer Applications,* vol. 179, no.17, pp. 1-14, 2018.

[2] Xu Yong, Research on speech enhancement based on deep neural network, Ph. D Thesis, University of Science and Technology of China, 2015.

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

[4] 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 (ICASS),* vol. 4, pp. 208-211, 1979.

[5] Y. Ephraim, "Statistical-model-based speech enhancement systems," in *Proceedings of the IEEE,* vol. 80, no. 10, pp. 1526–1555, 1992.

[6] Lu X, Tsao Y, Matsuda S et al., "Speech enhancement based on deep denoising autoencoder" in *INTERSPPECH,* pp. 436-440, 2013.

[7] Xu Y, Du J, Dai L R et al., " regression approach to speech enhancement based on deep neu-ral networks," *IEEE/ACM Transactions on Audio, Speech, and Language Pro-cessing,* vol. 22, no. 1, pp. 7-19, 2014.

[8] Kumar A, Florencio D, "Speech enhancement in multiple-noise conditions using deep neural networks," in *INTERSPEECH,* pp. 3738-3742, 2016.

[9] Weninger F, Erdogan H, Watanabe S et al., "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *International Conference on Latent Variable Analysis and Signal Separation,* pp. 91-99, 2015.

[10] I. Goodfellow, J. Pouget-Abadie et al., "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NIPS),* pp. 2672–2680, 2014.

[11] D. Michelsanti and Z. H. Tan, "Conditional generative adversarial networks for speech enhance-ment and noise-robust speaker verification," in *INTERSPEECH,* pp. 3642–3646, 2017.

[12] Arjovsky M, Chintala S, Bottou L, "Wasserstein generative adversarial networks," in *International Conference on Machine Learning,* pp. 214-223, 2017.

[13] Qin S, Jiang T, "Improved Wasserstein conditional generative adversarial network speech enhancement," in *EURASIP Journal on Wireless Communications and Networking,* 2018.

[14] M. Mirza and S.Osindero, "Conditional generative adversarial nets," *CoRR,* vol.abs/1411.1784, 2014.

[15] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech Enhancement Generative Adversarial Network," in *INTERSPEECH,* pp. 3624-3646, 2017.

[16] Baby D, Verhulst S, "SERGAN: Speech enhancement using relativistic generative adversarial networks with gradient penalty," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* pp. 106-110, 2019.

[17] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *NIPS,* pp. 5769-5779, 2017.

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

[19] Mao X, Li Q, Xie H, et al., "Least squares generative ad versarial networks," in *Proceedings of the IEEE International Conference on Computer Vision,* pp. 2794-2802, 2017.

[20] A. Jolicoeur-Martineau, "The relativistic discriminator: a key element missing from stand-ard GAN," *CoRR,* vol.abs/1807.00734, 2018.

[21] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *International Conference on Learning Representations (ICLR),* pp. 1–16, 2016.

[22] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating rnn-based speech enhancement methods for noiserobust text-to-speech" in *9th ISCA Speech Synthesis Workshop,* pp. 146–152.

[23] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multichannel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings," in *Proceedings of Meetings on Acoustics,* vol. 19, no. 1, p. 035081, 2013.

[24] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," pp. 2226-2234, 2016.

[25] T. Tieleman and G. Hinton, "Lecture 6.5-RMSprop: divide the gradient by a running average of its recent magnitude," COURSERA: Neural Networks for Machine Learning, 2012.

[26] *P.862.2: Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs,* ITU-T Std.P862.2, 2007.

[27] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. on Audio Speech, and Language Processing,* vol. 16, pp. 229-238, 2008.

[28] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *IEEE ICASSP,* Texas, USA, pp. 4214-4217, 2010.