INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Improved Speech Enhancement using a Complex-Domain GAN with Fused Time-Domain and Time-frequency Domain Constraints

Feng Dang[1,2], *Pengyuan Zhang*[1,2], *Hangting Chen*[1,2]

[1]Key Laboratory of Speech Acoustics & Content Understanding, Institute of Acoustics, CAS, China
[2]University of Chinese Academy of Sciences, Beijing, China
{dangfeng, zhangpengyuan, chenhangting}@hccl.ioa.ac.cn

Abstract

Complex-domain models have achieved promising results for speech enhancement (SE) tasks. Some complex-domain models consider only time-frequency (T-F) domain constraints and do not take advantage of the information at the time-domain waveform level. Some complex-domain models consider only time-domain constraints and do not take into account T-F domain constraints that have rich harmonic structure information. Indeed some complex-domain models consider both time-domain and T-F domain constraints but only use the simple mean square loss as time-frequency-domain constraints. This paper proposes a complex-domain-based speech enhancement method that integrates time-domain constraints and T-F domain constraints into a unified framework using a Generative Adversarial Network (GAN). The proposed framework captures information at the time-domain waveform level features while paying attention to the harmonic structure by time-domain and T-F domain constraints. We conducted experiments on the Voice Bank + DEMAND dataset to evaluate the proposed method. Experimental results show that the proposed method improves the PESQ score by 0.09 and the STOI score by 1% over the strong baseline deep complex convolution recurrent network (DCCRN) and outperforms the state-of-the-art GAN-based SE systems.

Index Terms: speech enhancement, generative adversarial network, complex network, time-domain and T-F domain constraints

1. Introduction

Speech enhancement (SE) is a speech processing method that aims to improve the quality and intelligibility of noisy speech via removing noise [1]. Apart from traditional SE algorithms like spectral subtraction [2] and Wiener filtering [3], recent years have witnessed the rapid development of deep neural networks (DNNs) on SE research [4, 5].

DNN-based approaches are often trained in a supervised setting and can be divided into two main categories: time-domain methods and time-frequency domain (T-F) methods. Time-domain methods can be further divided into two categories - direct regression [6, 7] and adaptive front-end methods [8, 9]. The former learns the regression function directly from the speech-noise mixture of waveforms to the target speech without an explicit signal front-end. Taking time-domain signals in and out, the latter adaptive front-end approach usually employs a convolutional encoder-decoder or u-network framework, which is similar to the short-time Fourier transform (STFT) and its inversion (iSTFT). The enhancement networks are inserted between the encoder and decoder. The T-F domain methods generally transform input noisy waveforms into Fourier spectra (e.g., power spectrum, complex spectrum) by short-time Fourier transform, modify the spectra typically by T-F mask [10, 11] and transform enhanced spectra back into enhanced waveforms by inverse short-time Fourier transform.

Generative adversarial networks (GANs) [12] have shown their superiority in many different applications. SEGAN [13] is the first approach to apply GAN to SE task, which models a function that maps noisy waveform directly to clean waveform in an end-to-end way. In [14, 15], GAN-based algorithms are proposed to operate on spectral features rather than time-domain waveforms. In [16], an enhancement framework based on T-F masks is introduced which implicitly learns the mask using GAN while predicting a clean T-F representation. MetricGAN [17] optimizes the generator concerning one or multiple speech enhancement evaluation metrics. μ-law SGAN [18] takes the magnitude spectrum as input and passes the output through a trainable u-law compression layer, to make the output magnitude spectrum closer to the magnitude of the reference speech.

Although GAN has achieved promising performance on SE tasks, the importance of generator network structure for SE tasks has not been fully explored. Complex-domain models have recently achieved state-of-the-art (SOTA) performance on speech enhancement tasks [19, 20]. In this paper, we use the complex-domain model deep complex convolution recurrent network (DCCRN) [19] as our generator. However, the original DCCRN model only considers time-domain constraints and does not consider T-F domain constraints that are rich in harmonic structure information. Therefore, we add a T-F domain constraint to the DCCRN, i.e., the L1 loss between the estimated amplitude spectrum and the clean speech amplitude spectrum. We then employ adversarial training to make the output of the generator network better preserve the harmonic structure of speech by narrowing the gap between the estimated and reference log amplitude spectrum. We take a step further by using a trainable compression layer instead of logarithmic compression to compress the dynamic range of the spectrum so that the compressed amplitude spectrum can show more detailed information and further improve the speech quality. We evaluate the proposed method on the Voice Bank + DEMAND dataset [21]. Experimental results show that the proposed method outperforms the baseline and the state-of-the-art GAN-based SE systems. In addition, we also investigate the performance of our proposed method at different signal-to-noise ratios (SNRs), and the results show that better performance is achieved at low SNRs.

The remainder of the paper is organized as follows: In Section 2, we first describe GAN-based speech enhancement and then explain our proposed model architecture in Section 3. Sections 4 and 5 report the experimental setup and results, respectively, and Section 6 concludes the paper with some conclu-

http://dx.doi.org/10.21437/Interspeech.2021-1134

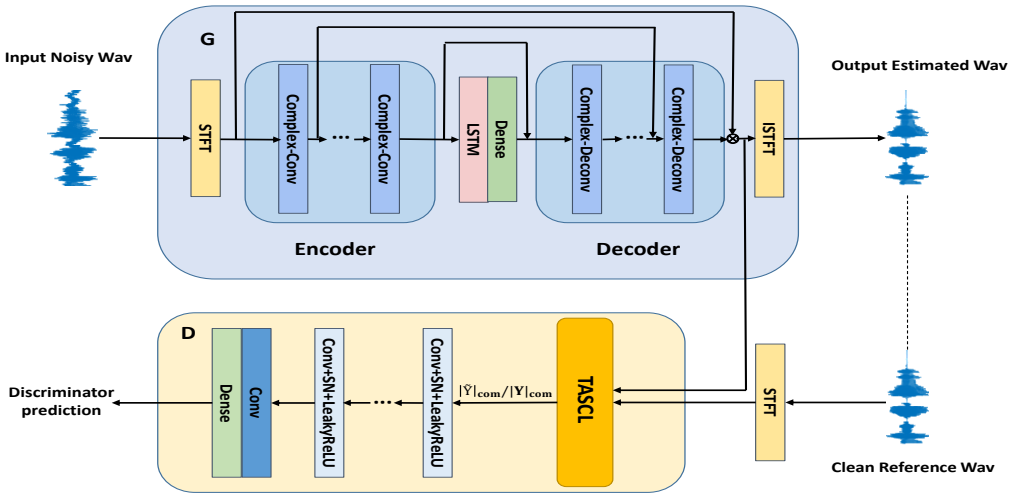

Figure 1: *An overview of the proposed method.*

sions.

2. GAN-based speech enhancement

A generalized adversarial network (GAN) consists of a generator network (G) and a discriminator network (D). In the applications of GAN on SE, G is trained to map a noisy speech signal z from a prior distribution $p_z(z)$ to an estimated speech signal $G(z)$ whose distribution is as close as possible to the distribution $p_{\text{data}}(x)$ of the clean speech signals. D is a binary classifier that regards the clean speech signals as true and the estimated speech signals generated by G as fake. By employing an alternative mini-max training scheme between G and D, where G is updated to fool D into treating its estimated speech signals as real, and D is updated to more confidently classify the speech signals as clean or estimated. The loss function of this mini-max game between G and D is defined as follows:

$$\min_G \max_D L(G,D) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[1-\log D(G(z))] \quad (1)$$

However, traditional GAN generators ignore an important piece of a prior information: for the discriminator, half of the data actually comes from real samples and a half from forged samples. For this reason, if the probability of a forged sample being judged true increases, then the probability of a real sample being judged true should decrease accordingly. Therefore, we use the relativistic GAN (RGAN) proposed in [22] and used in [7, 23]. RGAN's discriminator estimates the probability that the clean speech signal is more realistic than the estimated speech signal, while RGAN's generator in turn estimates the probability that the estimated speech signal is more realistic than the clean speech signal. The RGAN loss functions are as follows:

$$\min_G L(G) = -\mathbb{E}_{x\sim p_{\text{data}}(x),z\sim p_z(z)}[\log(\sigma(D(G(z))-D(x)))] \quad (2)$$

$$\min_D L(D) = -\mathbb{E}_{x\sim p_{\text{data}}(x),z\sim p_z(z)}[\log(\sigma(D(x)-D(G(z))))] \quad (3)$$

3. Proposed GAN with fusion constraints

In this section, we propose a GAN-based network with time-domain and T-F domain fusion constraints for speech enhancement. As shown in Fig. 1, the proposed model consists of a generator and a discriminator with a trainable amplitude spectral compression layer.

3.1. Generator

The generator is an encoder-decoder network that takes the noisy waveform as input. After short-time Fourier transform (STFT) using a convolutional layer, the resulting complex spectrum is fed to the encoder, and high-level features are extracted by multiple complex convolutional layers of the encoder. There are two LSTM layers between the encoder and decoder for extracting long-term information. The decoder is the inverse process of the encoder. The skip connections are used to connect each encoder layer to the homologous decoder layer. The feature dimension is restored to the same size as the output of the conv-STFT layer after the last complex decoding layer. The decoder estimates the complex ratio mask (CRM) [24] and applies it to the input of the encoder, $\otimes$ representing element multiplication. The estimated waveform is reconstructed by iSTFT operation using a convolutional layer.

Given the estimated mask M and the complex-valued STFT spectrogram of noisy speech Y, the estimated clean speech $\tilde{S}$ can be calculated as below:

$$\tilde{S} = (Y_r \cdot \tilde{M}_r - Y_i \cdot \tilde{M}_i) + j(Y_r \cdot \tilde{M}_i + Y_i \cdot \tilde{M}_r) \quad (4)$$

3.2. Discriminator

The discriminator network consists of trainable amplitude spectrum compression layer (TASCL) and convolutional blocks with good pattern recognition capability. In the training phase of D, TASCL is optimized to better distinguish between the target spectra and estimated enhanced spectra. Thus, the discriminator network plays a role in guiding the TASCL to learn a better compression function. A better performing TASCL can find more differences in geometric structure details between the

clean spectra and the estimated spectra, which can enable the generator network to generate a spectrum closer to the target spectrum.

3.3. Trainable amplitude spectrum compression layer

We argue that the amplitude spectrum rich in geometric structure can make the output of the trained model better preserve speech quality. Therefore, we first use log compression and add trainable parameters to log compression to make the compressed amplitude spectrum show more details. Give the amplitude spectrum $|X|$, the compressed amplitude spectrum can be expressed as

$$|X|_{com} = \frac{\ln\left(1+\alpha_1|X|\right)}{\ln\left(1+\alpha_2\right)} \tag{5}$$

where α_1, α_2 are trainable parameters. By adjusting α_1, the compression degree of the amplitude spectrum can be controlled, while by adjusting α_2, the amplitude range of the compressed amplitude spectrum can be controlled.

3.4. Loss fuction

In order to make full use of the time-domain waveform level features and the T-F domain amplitude spectrum features, our loss function combines both time-domain and T-F domain losses. The loss function is as follows

$$L_D = -\mathbb{E}_{x\sim p_{\mathrm{x}},Y\sim p_{\mathrm{Y}}}[\log\left(\sigma(D(|Y|_{com}) - D(|\tilde{Y}|_{com}))\right)] \tag{6}$$

$$\begin{aligned} L_G = &-\lambda_1\mathbb{E}_{x\sim p_{\mathrm{x}},Y\sim p_{\mathrm{Y}}}[\log\left(\sigma(D(|\tilde{Y}|_{com}) - D(|Y|_{com}))\right)] \\ &+ \lambda_2\left\|G(x)-y\right\| + \lambda_3\left\||\tilde{Y}|_{com} - |Y|_{com}\right\| \end{aligned} \tag{7}$$

4. Experiment setup

4.1. Datasets

The proposed method is evaluated on an open dataset released by Valentini et al. [21] for a fair comparison. This dataset is generated using the speech corpus Voice Bank [25] and the noise database DEMAND [26]. Voice Bank corpus is divided into a training set composed of 28 speakers and a test set composed of 2 speakers. The noisy speech is generated with 10 types of noises (8 from DEMAND database and 2 artificially generated) at SNRs of 0 dB, 5 dB, 10 dB, and 15 dB for training, and with 5 types of unseen noises at SNRS of 2.5 dB, 7.5 dB, 12.5 dB and 17.5 dB for testing. All utterances are resampled from 48kHz to 16 kHz and sliced by a sliding window of length 16000 with 50% overlap.

We also investigate the performance of our proposed method at different signal-to-noise ratios (SNRs) using another dataset. We choose clean utterances randomly from the TIMIT corpus [27] which includes 630 speakers. 5670 utterances are chosen to form the training set and the other 630 utterances serve as a testing set. Both training and test sets are created by adding 100 types of noises (all from NOISE-100 corpus [28]) to the clean utterances segmented into one-second pieces at SNRs of 15 dB, 10 dB,5 dB, 0 dB, and 5 dB, resulting in 118-hour training set and 13-hour test set. The utterances are downsampled from 48 kHz to 16 kHz.

4.2. Model setup

We use DCCRN-C as our G network. We set the window length and hop size to 25 ms and 6.25 ms, respectively, and a 512 FFT length in the conv-STFT layer and the conv-ISTFT layer of the G network. The encoder has 6 complex convolutional layers with 16,32,64,128,256,256 number of channels. The stride and kernel size of each layer are set to (2, 1) and (5, 2) respectively. The LSTM layer consists of 2 layers of 256 cells. The channel number of the decoder follows the reverse setting of the encoder. In the last decoder layer, the complex batch normalization and activation functions are not used in training.

The D network consists of 6 two-dimensional convolutional layers with the number of channels 64, 128, 256, 512, 1024, 1024 and each layer has a filter length of (5, 2) and a step size of (2, 1). LeakyReLU with α = 0.3 is applied after each two-dimensional layer. After the last activation layer, there is an additional two-dimensional convolutional layer with a filter length of (1, 1) and a step size of (1, 1). The output is fed into a linearly activated fully connected layer for binary classification. We use spectral normalization at each layer of D to produce better performance. Spectral normalization is used to control the range of weights in D and to avoid exploding or vanishing gradients.

We trained all the models for 150 epochs using a batch size of 128 with Adam optimizer [29]. The learning is set to 0.001 and then decay 0.98 for every two epochs. The hyperparameters $\lambda_1, \lambda_2, \lambda_3$ in equation (7) are set to 0.05, 5, 1 respectively.

4.3. Evaluation metrics

For evaluation metrics, we provide a perceptual evaluation of speech quality (PESQ) scores and Short-Time Objective Intelligence (STOI) scores. Speech enhancement is usually measured using PESQ scores, see [30, 31], and STOI scores, see [32]. PESQ scores are highly correlated with subjective evaluation scores and are mostly used for compressed objective measures. PESQ scores are calculated by comparing enhanced speech to clean reference speech and range from -0.5 to 4.5. STOI scores are highly correlated with human speech intelligibility and range from 0 to 1. Higher scores indicate better performance for all metrics.

5. Experiment results

5.1. Results on the Voice Bank + DEMAND dataset

Our method proposed in this paper is compared with other methods which also employ the same open dataset Voice Bank + DEMAND introduced in subsection 5.1. Table I lists the speech enhancement performance of related studies recently trained and tested on this dataset.

As can be seen from Table 1, our proposed achieves the state-of-the-art performance among the GAN-based models compared with SEGAN, MMSE-GAN, CNN-GAN, MetricGAN, and u-law SGAN. Furthermore, the proposed method achieves higher scores of PESQ and STOI than other time-domain methods (such as SerGAN) and T-F domain methods (such as DCT-UNet).

5.2. Ablation analysis

The experimental results in the previous subsection prove that our method improves speech enhancement performance. To verify the validity of our method, we performed an ablation analysis. We utilize the DCCRN which denotes generator G

Table 1: *Evaluation results of our proposed model compared with other methods on the same dataset.*

Method	Domain	PESQ	STOI
Noisy	-	1.97	0.91
SEGAN, 2017[13]	T	2.16	0.93
MMSE-GAN, 2018[16]	F	2.53	0.93
Wave U-Net, 2018[8]	T	2.40	-
MetricGAN, 2019[17]	F	2.86	0.92
DCT-UNet, 2019[33]	F	2.70	-
SerGAN, 2019 [7]	T	2.62	-
μ-law SGAN, 2020[18]	F	2.86	0.94
Proposed method	F	**2.87**	**0.95**

trained alone as the baseline, and then improve the baseline step by step until the best performance is reached.We designed three experiments labeled DCCRN, DCCRN+LOG+pD, DCCRN+TASC+pD. Experimental results of ablation analysis are given in Table 2. Note that the discriminator module is not involved during inference. The same DCCRN configuration was used for the three experiments.

Table 2: *Ablation analysis results.*

Method	PESQ	STOI
Noisy	1.97	0.91
DCCRN	2.78	0.94
DCCRN+LOG+pD	2.81	0.94
DCCRN+TASC+pD	**2.87**	**0.95**

By comparing experiments DCCRN and DCCRN+LOG+pD, it can be seen that the PESQ and STOI scores are improved a little with the addition of the logarithmic amplitude spectral loss and discriminator. By comparing trials DCCRN+LOG+pD and DCCRN+TASC+pD it can be seen that replacing logarithmic amplitude spectral compression with trainable amplitude spectral compression can further improve the scores of PESQ and STOI. This is because the original DCCRN has only time-domain loss, whereas the amplitude spectrum of speech is rich in harmonic structure information. The addition of logarithmic amplitude loss allows the model to learn more information and thus improve the speech quality. Further by adding trainable amplitude spectrum compression, the resulting amplitude spectrum can be made to show more details, thus further improving the speech quality.

5.3. Performance at different SNRs

We also investigated the performance of our proposed method with different SNRs on the second dataset introduced in subsection 5.1. We use DCCRN as the baseline.

As shown in Table 3, our model achieves the most improvement in PESQ scores when the signal-to-noise ratio is 0 compared to DCCRN. The relative boost in PESQ becomes less when the noise component or speech component becomes larger. This is probably because PESQ is an indicator of the degree of distortion in speech and is strongly related to the harmonic structure of speech. It is difficult to distinguish between noise and speech at the waveform level when the speech and noise components have the same energy, whereas the amplitude spectrum of the noisy speech provides information on the harmonic structure, which facilitates the distinction between noise and speech, thus reducing speech distortion. When the signal-to-noise ratio is relatively high, the noise component is small and the output speech waveform of the model is as close to the reference clean speech waveform as possible to reduce the distortion of the speech, the amplitude spectrum of the noisy speech provides limited information for distortion reduction, while when the noise component dominates, the harmonic structure of the speech is contaminated by the noise, and the PESQ boost, therefore, becomes smaller.

Table 3: *Performance at different SNRs.*

Method	DCCRN		Proposed method	
SNRs	PESQ	STOI(%)	PESQ	STOI(%)
15dB	3.77	**98.90**	**3.80**	98.88
10dB	3.47	**98.08**	**3.52**	98.07
5dB	3.10	96.71	**3.16**	**96.71**
0dB	2.68	94.31	**2.75**	**94.36**
-5dB	2.29	90.51	**2.33**	**90.74**
Average	2.91	94.89	**2.96**	**94.95**

We can also find that the smaller the SNR, the larger the STOI boost. This may be due to the fact that the score of STOI is related to the proportion of noise. When the proportion of noise is larger, the noise structure of the noisy speech amplitude spectrum is more obvious and can provide more information for noise reduction. We note that the proposed method has a slightly smaller boost on this dataset compared to baseline than on Voice Bank + DEMAND, which may be because the noise type and SNRs of the test set on this dataset are the same as the training set, while the Voice Bank +DEMAND dataset uses different noise and SNRs for the test and training sets. This proves that our proposed method can increase the generalization ability of the complex-domain-based model.

6. Conclusions

In this paper, we propose a novel speech enhancement method based on complex-domain GAN with fused time-domain and time-frequency domain constraints. The experimental results show that the proposed method achieves better performance in both seen and unseen conditions without increasing the number of parameters compared to the baseline system. The proposed method also achieved the state-of-the-art performance among the GAN-based models. We analyze the possible reasons why the proposed method works by comparing the performance of the model at different signal-to-noise ratios and show the importance of adding more detailed amplitude spectral constraints to the time-domain constraints for the complex-domain models.

7. Acknowledgements

This work is partially supported by National Natural Science Foundation of China (Nos. 62071461,11774380).

8. References

[1] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

[2] S. Boll, "Suppression of acoustic noise in speech using spectral

subtraction," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113–120, 1979.

[3] P. Scalart *et al.*, "Speech enhancement based on a priori signal to noise estimation," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 2. IEEE, 1996, pp. 629–632.

[4] 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, 2014.

[5] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[6] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5069–5073.

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

[8] C. Macartney and T. Weyde, "Improved speech enhancement with the wave-u-net," *arXiv preprint arXiv:1811.11307*, 2018.

[9] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[10] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[11] D. Baby and S. Verhulst, "Biophysically-inspired features improve the generalizability of neural network-based speech enhancement systems," in *19th Annual Conference of the International-Speech-Communication-Association (INTERSPEECH 2018)*. ISCA, 2018, pp. 3264–3268.

[12] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *arXiv preprint arXiv:1406.2661*, 2014.

[13] S. Pascual, A. Bonafonte, and J. Serra, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[14] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5024–5028.

[15] J. Lin, S. Niu, Z. Wei, X. Lan, A. J. Wijngaarden, M. C. Smith, and K.-C. Wang, "Speech enhancement using forked generative adversarial networks with spectral subtraction," *Proceedings of Interspeech 2019*, 2019.

[16] M. H. Soni, N. Shah, and H. A. Patil, "Time-frequency masking-based speech enhancement using generative adversarial network," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5039–5043.

[17] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2031–2041.

[18] H. Li, Y. Xu, D. Ke, and K. Su, "μ-law sgan for generating spectra with more details in speech enhancement," *Neural Networks*, vol. 136, pp. 17–27, 2021.

[19] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement," *arXiv preprint arXiv:2008.00264*, 2020.

[20] A. Li, W. Liu, X. Luo, C. Zheng, and X. Li, "Icassp 2021 deep noise suppression challenge: Decoupling magnitude and phase optimization with a two-stage deep network," *arXiv preprint arXiv:2102.04198*, 2021.

[21] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating rnn-based speech enhancement methods for noise-robust text-to-speech." in *SSW*, 2016, pp. 146–152.

[22] A. Jolicoeur-Martineau, "The relativistic discriminator: a key element missing from standard gan," *arXiv preprint arXiv:1807.00734*, 2018.

[23] Z. Zhang, C. Deng, Y. Shen, D. S. Williamson, Y. Sha, Y. Zhang, H. Song, and X. Li, "On loss functions and recurrency training for gan-based speech enhancement systems," *arXiv preprint arXiv:2007.14974*, 2020.

[24] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, no. 3, pp. 483–492, 2015.

[25] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *2013 international conference oriental COCOSDA held jointly with 2013 conference on Asian spoken language research and evaluation (O-COCOSDA/CASLRE)*. IEEE, 2013, pp. 1–4.

[26] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Proceedings of Meetings on Acoustics ICA2013*, vol. 19, no. 1. Acoustical Society of America, 2013, p. 035081.

[27] 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, p. 27403, 1993.

[28] G. Hu and D. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2067–2079, 2010.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[30] 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 *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[31] R. P. ITU-T, "862,"," *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," February*, 2001.

[32] 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.

[33] C. Geng and L. Wang, "End-to-end speech enhancement based on discrete cosine transform," in *2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*. IEEE, 2020, pp. 379–383.