JOINT MAGNITUDE ESTIMATION AND PHASE RECOVERY USING CYCLE-IN-CYCLE GAN FOR NON-PARALLEL SPEECH ENHANCEMENT

Guochen Yu[*†]*, Andong Li*[†]*, Yutian Wang*[*]*, Yinuo Guo*[*]*, Hui Wang*[*]*, Chengshi Zheng*[†]

[*] State Key Laboratory of Media Convergence and Communication, Communication University
of China, Beijing, China
[†] Institute of Acoustics, Chinese Academy of Sciences, Beijing, China
[*] Bytedance, Beijing, China
{yuguochen, wangyutian, hwang}@cuc.edu.cn, {liandong, cszheng}@mail.ioa.ac.cn

ABSTRACT

For the lack of adequate paired noisy-clean speech corpus in many real scenarios, non-parallel training is a promising task for DNN-based speech enhancement methods. However, because of the severe mismatch between input and target speeches, many previous studies only focus on the magnitude spectrum estimation and remain the phase unaltered, resulting in the degraded speech quality under low signal-to-noise ratio conditions. To tackle this problem, we decouple the difficult target *w.r.t.* original spectrum optimization into spectral magnitude and phase, and a novel Cycle-in-Cycle generative adversarial network (dubbed CinCGAN) is proposed to jointly estimate the spectral magnitude and phase information stage by stage under unpaired data. In the first stage, we pretrain a magnitude Cycle-GAN to coarsely estimate the spectral magnitude of clean speech. In the second stage, we incorporate the pretrained CycleGAN with a complex-valued CycleGAN as a cycle-in-cycle structure to simultaneously recover phase information and refine the overall spectrum. Experimental results demonstrate that the proposed approach significantly outperforms previous baselines under non-parallel training. The evaluation on training the models with standard paired data also shows that CinCGAN achieves remarkable performance especially in reducing background noise and speech distortion.

Index Terms— speech enhancement, non-parallel, Cycle-in-Cycle GAN, magnitude spectrum estimation, phase recovery

1. INTRODUCTION

Speech is often corrupted by various types of background noise in real-world environments. These interferences severely degrade the performance of many speech-related tasks such as automatic speech recognition (ASR) systems or hearing aids [1]. To cope with these degradations, speech enhancement (SE) can be applied to remove or attenuate the background noise for better speech quality and intelligibility. Recent SE approaches based on deep neural networks (DNNs) have demonstrated their superior capability in dealing with non-stationary noise under low signal-to-noise ratio (SNR) conditions [2]. These DNN-based methods can be roughly categorized into masking-based approaches [3] and mapping-based approaches [4]. More recently, generative adversarial networks (GANs) [5] begin to gain development in the single-channel SE area [6, 7, 8, 9, 10, 11, 12], in which a generator (G) aims to conduct

the enhancement process and a discriminator (D) aims to distinguish between the generated speech features and real clean ones. For these standard supervised methods, the networks are trained to minimize the discrepancy between the enhanced output and the clean target.

Despite the impressive performance of the above DNN-based SE approaches, they always need numerous paired clean-noisy samples to conduct supervised training and improve the generalization. However, in some real scenarios, it is troublesome to record parallel clean-noisy pairs, and we can only obtain clean speech that mismatches the source noisy speech. In this case, standard supervised SE methods usually cannot produce satisfactory performance. To tackle this problem, cycle-consistent GAN (CycleGAN) has been developed to conduct unsupervised SE, which was originally used for unpaired image-to-image translation [13]. In the SE area, CycleGAN-based approaches demonstrate the remarkable ability in preserving the speech structure and improving speech quality in both paired and unpaired cases [14, 15, 16, 17, 18]. Nevertheless, these conventional CycleGAN-based approaches have several limitations for non-parallel SE. First, to ensure cycle consistency of the original noisy speech and target clean speech for unpaired data, the enhanced signal always contains the original noise information. In other words, cycle-consistency-based methods remain audible residual noise, which is challenging to be eliminated completely. Second, previous CycleGAN-based SE methods only focus on magnitude spectrum estimation and remain the noisy phase unaltered. This is because simultaneously restoring the clean spectral magnitude and phase information still remains to be a tough task, let alone when noisy-clean pairs are mismatched. However, recent studies revealed the importance of the accurate phase for improving perceptual speech quality, especially in low SNR cases [19].

Multi-stage algorithms based on curriculum learning begin to thrive in the SE area [20, 21], where the original difficult task is decomposed into multiple simpler sub-tasks and a better result can be induced progressively. Motivated by these studies, we couple a magnitude spectrum estimation CycleGAN (MCGAN) and a complex spectrum refined CycleGAN (CCGAN) as a Cycle-in-Cycle GAN (CinCGAN) paradigm to estimate the clean spectral magnitude and phase information step-by-step under non-parallel training. In the first step, the complex original spectrum is decoupled into the spectral magnitude and phase, and only the magnitude of the clean spectrum is estimated by MCGAN. Subsequently, a complex-valued CCGAN is integrated with the pretrained MCGAN to jointly recover the clean spectral magnitude and phase information. To be specific, we couple the estimated magnitude spectrum by MCGAN and original phase as a coarse-estimated complex spectrum, and feed it into
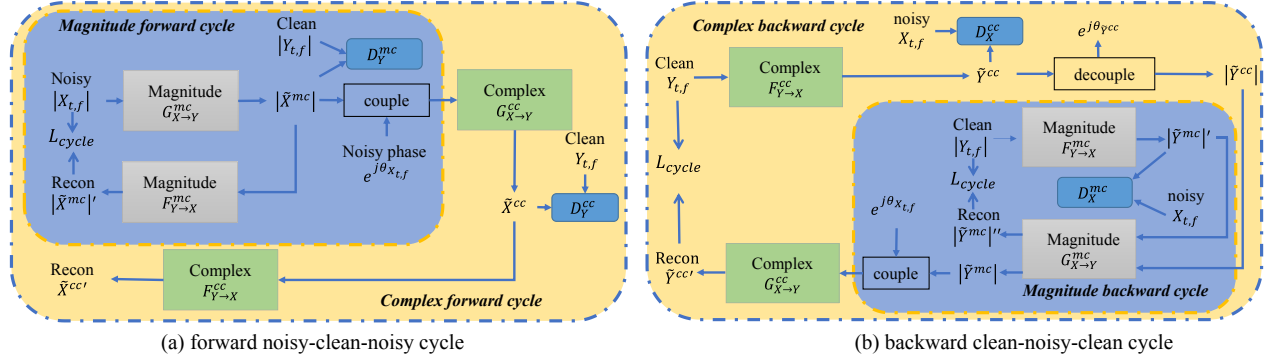
6967

Fig. 1. System flowchart of CinCGAN. The magnitude and complex cycle are shown in the yellow and blue dotted boxes, respectively. (a) The diagram of the forward cycle in CinCGAN. (b) The diagram of the backward cycle in CinCGAN.

CCGAN to estimate both real and imaginary (RI) components of the clean spectrum.

The remainder of the paper is organized as follows. In Section 2, the overall diagram is described in detail. The experimental setup is presented in Section 3, while the results and analysis are given in Section 4. Some conclusions are drawn in Section 5.

2. ALGORITHM DESCRIPTION

2.1. Cycle-in-Cycle Scheme

To simultaneously estimate magnitude and phase information of clean spectrum under non-parallel training conditions, we integrate a magnitude spectrum estimation CycleGAN (MCGAN) with a complex spectrum restoration CycleGAN (CCGAN) as a Cycle-in-Cycle GAN (CinCGAN). The proposed CinCGAN consists of a forward noisy-clean-noisy cycle and a backward clean-noisy-clean cycle, as shown in Fig. 1. In the forward cycle, the enhancement procedure can be divided into two steps. First, we decouple the complex spectrum into spectral magnitude and phase, and only the amplitude is processed. A magnitude generator $G_{X \to Y}^{mc}$ is utilized to convert the compressed noisy spectral magnitude $|X_{t,f}|$ to the "clean" one $|\widetilde{X}^{mc}|$, while a magnitude discriminator D_Y^{mc} is used to distinguish the generated spectral magnitude (i.e., $|\widetilde{X}_{t,f}|$) from the clean one (i.e., $|Y_{t,f}|$), where t and f denote the time and frequency indices. To constrain the cycle consistency under unpaired data, an inverse generator $F_{Y \to X}^{mc}$ is employed to reconstruct the noisy spectral magnitude from the generated one. Subsequently, we couple the estimated spectral magnitude and the original phase (i.e., $e^{j\theta_{X_{t,f}}}$) into a coarsely estimated complex spectrum and feed it into CCGAN. By estimating both RI parts of the clean spectrum, CCGAN can refine the spectral magnitude and recover the clean phase. Specifically, a complex-valued generator $G_{X \to Y}^{cc}$ attempts to generate the enhanced complex spectrum $\widetilde{X}^{cc}$ as clean one, and a complex discriminator D_Y^{cc} attempts to discriminate between the generated and clean complex spectrum (i.e., $Y_{t,f}$). Similarly, an inverse complex generator $F_{Y \to X}^{cc}$ brings the enhanced complex spectrum $\widetilde{X}^{cc}$ back to the original noisy domain. To better constrain the mapping optimization space, a backward clean-noisy-clean cycle is also designed and jointly trained with the forward cycle. In the testing stage, the whole enhancement procedure can be formulated as:

$$|\widetilde{X}^{mc}| = G_{X \to Y}^{mc}\left(|X_{t,f}|; \Phi_1\right), \tag{1}$$

$$\widetilde{X}^{mc} = |\widetilde{X}^{mc}| \exp(j\theta_{X_{t,f}}), \tag{2}$$

$$\widetilde{X}^{cc} = G_{X \to Y}^{cc}\left(\widetilde{X}^{mc}; \Phi_2\right), \tag{3}$$
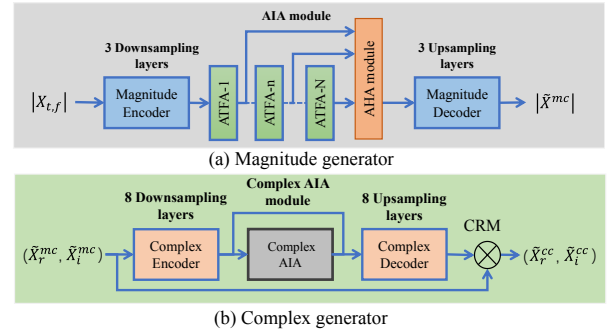

Fig. 2. (a) The diagram of the magnitude generators in MCGAN. (b) The diagram of the complex generators in CCGAN.

where $|X_{t,f}|$ and $\theta_{X_{t,f}}$ denote the noisy spectral magnitude and phase, respectively. $|\widetilde{X}^{mc}|$ and $\widetilde{X}^{cc}$ denote the outputs of the noisy-to-clean magnitude generator $G_{X \to Y}^{mc}$ and complex generator $G_{X \to Y}^{cc}$ with parameter set $\Phi_{(\cdot)}$.

2.2. Network Architecture

Following the same configuration as [17], the magnitude generator in MCGAN consists of three major parts, including three downsampling layers as the encoder, an adaptive attention-in-attention (AIA) module and three upsampling layers as the decoder, as shown in Fig. 2 (a). Each downsampling/upsampling layer block is composed of a 2D (de)convolution layer, followed by instance normalization (IN), Parametric Relu activation function (PReLU) and gated liner units (GLUs). The proposed AIA module consists of six adaptive time-frequency attention (ATFA) modules and an adaptive hierarchical attention (AHA) module, as proposed in [17]. More specifically, ATFA consists of two branches: an adaptive time attention branch (ATAB) and an adaptive frequency attention branch (AFAB). The two branches cooperate to capture the global dependencies along temporal and spectral axes in parallel. The AHA module integrates different hierarchical feature maps given all outputs of ATFA modules and fuses the global context of all intermediate outputs with different weights, thus helping to guide the feature learning progressively. The output channels of each layer in the encoder are (16, 32, 64), while the kernel size and strides are set to (3, 5) and (1, 2) along the time and frequency axes, respectively.

We also design the pure complex-valued generators of CC-GAN similar as [18], which consists of eight complex-valued encoder/decoder blocks, and an complex-valued AIA module, as shown in Fig. 2(b). The complex encoder/decoder blocks include complex-valued 2D (de)convolutions, complex IN and real-valued

PReLU. The complex-valued convolution simulates the correlation between real and imaginary parts of feature maps by complex multiplication, which is composed of four traditional convolution operations [22]. To further capture long-term dependency with complex-valued features, we utilize the above attention-in-attention mechanism in the complex-valued manner. The number of the channels for each complex encoder layer is (32, 32, 64, 64, 128, 128, 256, 256), with kernel size and strides set to (3, 3) and (1, 2), respectively.

We utilize multi-scale discriminators that can force the generator to produce feature maps with global consistency and finer details, as pointed out in [17]. Both the magnitude and complex discriminators are composed of six real-valued 2D convolutions, spectral normalization (SN) and PReLU. The kernel size for each 2D convolutions layer is (3, 5) except (1, 1) for the last layer in the temporal and frequency axis, respectively. The number of channels throughout the 2D convolutions is (32, 32, 64, 64, 128, 1), with stride set to (1, 2). Note that we concatenate RI components along the channel axis and feed them into the complex discriminator.

2.3. Loss Function

In both MCGAN and CCGAN, we use the following three losses to jointly optimize the forward and backward cycle, namely relativistic adversarial loss, cycle-consistency loss, and an identity mapping loss. Taking the magnitude CycleGAN (MCGAN) as an example, for the forward noisy-to-clean mapping, the relativistic average least-square (RaLS) adversarial loss [23] is used to enforce the enhanced magnitude spectra indistinguishable from the clean ones, which is calculated as follows.

$$\mathcal{L}_{Radv}(D_Y^{mc}) = \mathbb{E}_y \left[(D_Y^{mc}(y) - \mathbb{E}_x D_Y^{mc}(G_{X \to Y}(x)) - 1)^2 \right] +$$
$$\mathbb{E}_x \left[(D_Y^{mc}(G_{X \to Y}(x)) - \mathbb{E}_y (D_Y^{mc}(y)) + 1)^2 \right],$$
$$\mathcal{L}_{Radv}(G_{X \to Y}^{mc}) = \mathbb{E}_x \left[(D_Y^{mc}(G_{X \to Y}(x))) - \mathbb{E}_y D_Y^{mc}(y) - 1)^2 \right]$$
$$+ \mathbb{E}_y \left[(D_Y^{mc}(y) - \mathbb{E}_x (D_Y^{mc}(G_{X \to Y}(x))) + 1)^2 \right], \quad (4)$$

Analogously, $\mathcal{L}_{Radv}(D_X^{mc})$ and $\mathcal{L}_{Radv}(F_{Y \to X}^{mc})$ are employed in the inverse magnitude clean-to-noisy mapping. The cycle-consistency loss and identity mapping loss are utilized to constrain the mapping functions under non-parallel training conditions and to further preserve the linguistic composition of the original speech, which can be defined as:

$$\mathcal{L}_{cycle}(G_{X \to Y}^{mc}, F_{Y \to X}^{mc}) = \mathbb{E}_x \left[\|F_{Y \to X}(G_{X \to Y}^{mc}(x)) - x\|_1 \right]$$
$$+ \mathbb{E}_y \left[\|G_{X \to Y}^{mc}(F_{Y \to X}^{mc}(y)) - y\|_1 \right], \quad (5)$$
$$\mathcal{L}_{id}(G_{X \to Y}^{mc}, F_{Y \to X}^{mc}) = \mathbb{E}_x \left[\|F_{Y \to X}^{mc}(x) - x\|_1 \right] +$$
$$\mathbb{E}_y \left[\|G_{X \to Y}^{mc}(y) - y\|_1 \right], \quad (6)$$

In the first step, we pretrain MCGAN alone with $\mathcal{L}_{MCGAN}$ until convergence, which is calculated as:

$$\mathcal{L}_{MCGAN} = \mathcal{L}_{Radv}(G_{X \to Y}^{mc}, D_Y^{mc}) + \mathcal{L}_{Radv}(F_{Y \to X}^{mc}, D_X^{mc})$$
$$+ \lambda_{cycle} \mathcal{L}_{cycle}(G_{X \to Y}^{mc}, F_{Y \to X}^{mc}) + \lambda_{id} \mathcal{L}_{id}(G_{X \to Y}^{mc}, F_{Y \to X}^{mc}), \quad (7)$$

where λ_{cycle} and λ_{id} are empirically set to be 5 and 10, respectively. Then, we jointly fine-tune MCGAN and CCGAN with the same relativistic adversarial loss, cycle-consistency loss, and identity mapping loss, where the parameters of the first sub-network are initialized with the pretrained MCGAN. The overall loss functions of CinCGAN are expressed as:

$$\mathcal{L}_{CinCGAN} = \gamma \mathcal{L}_{MCGAN} + \mathcal{L}_{CCGAN} \quad (8)$$

where $\mathcal{L}_{CCGAN}$ denotes the loss function for the complex cycle. With some trials, we empirically set $\gamma = 0.1$ in all the following experiments .

3. EXPERIMENTS

3.1. Datasets

The dataset used in this work is publicly available as proposed in [24], which is a selection of the Voice Bank corpus [25] with 28 speakers for training and another 2 unseen speakers for testing. The training set consists of 11,572 mono audio samples, while the test set contains 824 utterances by 2 speakers (one male and one female). For the training set, audio samples are mixed together with one of the 10 noise types (*i.e.*, two artificial (babble and speech shaped) and eight real noise from the DEMAND database [26]) at four SNRs of {0dB, 5dB, 10dB, 15dB}. The testing utterances are created with 5 unseen test-noise types (all from the DEMAND database) at SNRs of {2.5dB, 7.5dB, 12.5dB, 17.5dB}. The original raw waveforms are downsampled from 48kHz to 16kHz for our task.

3.2. Implementation Setup

The Hanning window of length 32ms is applied, with 75% overlap between adjacent frames. The 512-point STFT is utilized, leading to a 257-D spectral feature. Due to the efficacy of power compression in dereverberation and denoising tasks [21], we conduct the power compression toward the spectral magnitude while leaving the phase unaltered, and the optimal compression coefficient is set to 0.5. For the non-parallel training strategy, we randomly crop a fixed-length segment (*i.e.*, 108 frames) from a randomly selected noisy audio file as the input, while the target is a randomly selected clean audio file whose content is totally different from the input audio. We adopt the Adam optimizer [27] with the momentum term $\beta_1 = 0.9$, $\beta_2 = 0.999$. We divide our training process into two steps. Firstly, we pretrain the MCGAN with an initial learning rate of 0.0002 for discriminators and 0.0005 for generators, respectively. Subsequently, we jointly fine-tune the pretrained MCGAN and CC-GAN as the proposed CinCGAN with a learning rate of 0.0002 for all generators and discriminators, while the learning rate is linearly decayed after 40 epochs. We use $\mathcal{L}_{id}$ in each cycle only for the first 20 epochs. The sound demo samples are available online [1].

4. RESULTS AND ANALYSIS

We use the following objective metrics to evaluate speech enhancement performance: the perceptual evaluation of speech quality (PESQ) [34], short-time objective intelligibility (STOI) [35], segmental signal-to-noise ratio (SegSNR), and three mean opinion score (MOS) prediction (*i.e.*, signal distortion evaluation (CISG), the intrusiveness of background noise (CBAK) and overall effect (COVL)) [36]. We also evaluate the subjective quality by DNS-MOS [37], which is a robust non-intrusive perceptual speech quality metric. Higher values of all metrics indicate better performance.

4.1. Comparison with GAN-based and CycleGAN-based methods under non-parallel Training

To investigate the effectiveness of our proposed method under a non-parallel training strategy, we re-implement two GAN based SE approaches with spectral magnitude and complex spectrum as input (*i.e.*, MGAN and CGAN), and their corresponding models with a forward cycle (fc) (*i.e.*, MGAN+fc and CGAN+fc). We then investigate the performance of different cycle structures, including MCGAN, CCGAN fed with the original noisy complex spectrum, and CinCGANs with different cycles (*i.e.*, CinCGAN (I)-(IV)). To be specific, CinCGAN (I) employs the magnitude and complex fc, while CinCGAN (II) and CinCGAN (III) add the backward cycle in MCGAN and CCGAN, respectively. Furthermore, CinCGAN (IV) employs all the forward and backward cycle in the magnitude and

---

Table 1. Experimental results among different models including GAN-based methods and the proposed CycleGAN-based methods under unpaired data. Best scores are highlighted in **bold**.

Methods	Feature type	Magnitude		Complex		PESQ	STOI(%)	CSIG	CBAK	COVL	SegSNR	DNSMOS
		fc	bc	fc	bc							
Unprocessed	–	–	–	–	–	1.97	92.1	3.35	2.44	2.63	1.68	3.02
GAN-based methods												
MGAN	Magnitude	×	×	×	×	2.03	91.6	3.54	2.78	2.72	5.28	2.72
MGAN+fc	Magnitude	✓	×	×	×	2.58	92.8	3.81	3.03	3.19	5.28	3.26
CGAN	RI components	×	×	×	×	1.86	88.9	3.17	2.62	2.64	2.98	2.63
CGAN+fc	RI components	×	×	✓	×	2.32	91.2	3.48	2.74	3.18	4.67	3.04
Proposed CycleGAN-based Systems												
MCGAN	Magnitude	✓	✓	×	×	2.67	93.2	3.86	3.20	3.21	7.23	3.47
CCGAN	RI components	×	×	✓	✓	2.56	92.1	3.67	3.10	3.16	5.38	3.42
CinCGAN (I)	Magnitude + RI	✓	×	✓	×	2.70	93.4	3.93	3.24	3.25	7.34	3.44
CinCGAN (II)	Magnitude + RI	✓	✓	✓	×	2.77	93.6	3.96	3.02	3.30	4.49	3.49
CinCGAN (III)	Magnitude + RI	✓	×	✓	✓	2.73	93.5	3.94	3.27	3.29	7.98	3.51
CinCGAN (IV)	Magnitude + RI	✓	✓	✓	✓	**2.84**	**94.1**	**4.10**	**3.36**	**3.37**	**8.91**	**3.53**

Table 2. Comparison with other GAN and Non-GAN based systems under standard paired data. "−" denotes that the result is not provided in the original paper.

Methods	Feature type	PESQ	STOI(%)	CSIG	CBAK	COVL
Unprocessed	–	1.97	92.1	3.35	2.44	2.63
GAN-based Systems						
SEGAN [6]	Waveform	2.16	92.5	3.48	2.94	2.80
MMSEGAN [9]	Gammatone	2.53	93.0	3.80	3.12	3.14
SERGAN [7]	Waveform	2.51	93.7	3.78	3.23	3.16
CP-GAN [10]	Waveform	2.64	94.0	3.93	3.29	3.28
MetricGAN [8]	Magnitude	2.86	–	3.99	3.18	3.42
CRGAN [11]	Magnitude	**2.92**	94.0	4.16	3.24	**3.54**
SASEGAN [12]	Waveform	2.36	93.5	3.54	3.08	2.93
Non-GAN based Systems						
Wave-U-net [28]	Waveform	2.64	–	3.56	3.08	3.09
DFL-SE [29]	Waveform	–	–	3.86	3.33	3.22
CRN-MSE [30]	Magnitude	2.61	93.8	3.78	3.11	3.24
GCRN [31]	RI components	2.51	94.0	3.71	3.24	3.09
DCCRN [32]	RI components	2.68	93.9	3.88	3.18	3.27
TFSNN [33]	Waveform	2.79	–	4.17	3.27	3.49
Proposed CycleGAN-based approaches						
MCGAN	Magnitude	2.74	93.6	3.96	3.25	3.29
CCGAN	RI components	2.60	92.8	3.82	3.12	3.20
CinCGAN	Magnitude+ RI	2.86	**94.4**	**4.18**	**3.38**	3.42

complex domain. From Table 1, one can have the following observations. First, MGAN+fc outperforms other GAN-based methods in terms of all metrics. This fully validates the efficacy of fc under a non-parallel training strategy. Second, when adding a backward cycle (bc) as standard CycleGANs, MGAN and CCGAN consistently surpass MGAN+fc and CGAN+fc, respectively. Even only estimating the spectral magnitude and remaining the noisy phase unaltered, MCGAN achieves better performance than CCGAN, which aims to estimate both RI components of the clean spectrum. The possible reason is that it seems quite difficult to simultaneously restore the magnitude spectrum and phase information under unpaired data by a one-stage strategy, due to the severe mismatch between input and target. Third, by integrating CCGAN with pretrained MCGAN, CinCGANs with different cycles improve speech performance at different degrees. When only using spectral magnitude fc and complex fc as shown in Figure.1 (a), CinCGAN (I) outperforms MCGAN and CCGAN in terms of most metrics. This phenomenon reveals the significance of simultaneously recovering spectral magnitude and phase information under non-parallel training conditions. That is to say, when the phase is incorporated in the complex cycle, better speech quality can be derived. When adding the spectral magnitude bc, CinCGAN (II) further achieves average 0.07 PESQ, 0.2% STOI, and 0.05 DNSMOS improvements than CinCGAN (I), which demonstrates the merit of bc in improving perceptual speech quality and intelligibility. However, the spectral magnitude bc forces the noisy-to-clean generators in MCGAN to preserve original speech information, thus leading to residual noise. We also notice that using both spectral magnitude bc and complex bc can facilitate all gen-

erators, achieving notable improvements. For example, going from CinCGAN (I) to CinCGAN (IV), average 0.14, 0.7%, 1.57dB and 0.09 improvements are observed in terms of PESQ, STOI, SegSNR and DNSMOS, respectively. This indicates that CinCGAN with both cycles in MCGAN and CCGAN consistently improves speech quality, while effectively reducing residual noise and speech distortion.

4.2. Comparison with other competitive methods under standard parallel training

As presented in Table 2, our proposed methods are compared with other state-of-the-art (SOTA) SE approaches under standard paired data, including GAN-based methods (*e.g.*, SEGAN [6], MM-SEGAN [9], SERGAN [7], MetricGAN [8], SASEGAN [12] and CRGAN [11]) and Non-GAN based methods (*i.e.*, Wave-U-net [28], DFL-SE [29], CRN-MSE [30], GCRN [31], DCCRN [32], and TFSNN [33]). Note that we re-implement CRN-MSE, GCRN and DCCRN on Voice Bank + DEMAND dataset, and directly use the reported scores of other methods in their original papers. One can have the following observations. Compared with the existing competitive GAN-based methods, CinCGAN provides better performance on STOI, CSIG and CBAK scores, which verifies the merit of CinCGAN in reducing speech distortion and suppressing residual noise components. Although CRGAN achieves a better PESQ score than CinCGAN, which is because that PESQ is directly used as the metric loss for CRGAN, CinCGAN provides average 0.4%, 0.02 and 0.14 improvements than CRGAN with respect to STOI, CSIG and CBAK, respectively. When compared with Non-GAN based methods, CinCGAN also provides competitive performance in terms of all metrics. For example, CinCGAN achieves average 0.18, 0.5%, 0.30, 0.20 and 0.15 improvements than DCCRN in PESQ, STOI, CSIG, CBAK and COVL, respectively.

5. CONCLUSIONS

In this paper, we propose a novel Cycle-in-Cycle GAN dubbed CinCGAN to jointly recover the spectral magnitude and phase information of clean speech for non-parallel speech enhancement. In the first stage, a standard magnitude spectrum estimation CycleGAN is pretained to only estimate the spectral magnitude under unpaired data. In the second stage, we couple the pretrained CycleGAN with a complex-valued CycleGAN as a cycle-in-cycle structure to further refine the magnitude spectrum and recover phase information. More specifically, the CinCGAN aims at estimating both real and imaginary components of the clean speech complex spectrum. Comprehensive experiments on Voice Bank + DEMAND dataset demonstrate that the proposed approach consistently outperforms original GAN-based and CycleGAN-based baselines under mismatched noisy-clean pairs. When experiments are conducted on standard parallel data, the proposed approach also shows its effectiveness in improving speech quality and reducing speech distortion.

6. REFERENCES

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

[2] D. L. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio. Speech, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, 2018.

[3] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio. Speech, Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, 2014.

[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 Trans. Audio. Speech, Lang. Process.*, vol. 23, no. 1, pp. 7–19, 2014.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NIPS*, 2014, pp. 2672–2680.

[6] S. Pascual, A. Bonafonte, and J. Serra, "Segan: Speech enhancement generative adversarial network," in *Proc. Interspeech*, 2017, pp. 3642–3646.

[7] D. Baby and S. Verhulst, "Sergan: Speech enhancement using relativistic generative adversarial networks with gradient penalty," in *Proc. ICASSP*. IEEE, 2019, pp. 106–110.

[8] 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 *Proc. ICML*. PMLR, 2019, pp. 2031–2041.

[9] M. H. Soni, N. Shah, and H. A. Patil, "Time-frequency masking-based speech enhancement using generative adversarial network," in *Proc. ICASSP*. IEEE, 2018, pp. 5039–5043.

[10] G. Liu, K. Gong, X. Liang, and Z. Chen, "Cp-gan: Context pyramid generative adversarial network for speech enhancement," in *Proc. ICASSP*. IEEE, 2020, pp. 6624–6628.

[11] 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," in *Proc. Interspeech*, 2020, pp. 3266–3270.

[12] H. Phan, H. Le Nguyen, O. Y. Chén, P. Koch, N. Q. Duong, I. McLoughlin, and A. Mertins, "Self-attention generative adversarial network for speech enhancement," in *Proc. ICASSP*. IEEE, 2021, pp. 7103–7107.

[13] J. Y. Zhu, T. Park, P. Isola, and A. A Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. ICCV*, 2017, pp. 2223–2232.

[14] Z. Meng, J. Li, and Y. Gong, "Cycle-consistent speech enhancement," in *Proc. Interspeech 2018*, 2018, pp. 1165–1169.

[15] Y. Xiang and C. Bao, "A parallel-data-free speech enhancement method using multi-objective learning cycle-consistent generative adversarial network," *IEEE/ACM Trans. Audio. Speech, Lang. Process.*, vol. 28, pp. 1826–1838, 2020.

[16] Y. Wang, G. Yu, J. Wang, H. Wang, and Q. Zhang, "Improved relativistic cycle-consistent gan with dilated residual network and multi-attention for speech enhancement," *IEEE Access*, vol. 8, pp. 183272–183285, 2020.

[17] G. Yu, Y. Wang, C. Zheng, H. Wang, and Q. Zhang, "Cyclegan-based non-parallel speech enhancement with an adaptive attention-in-attention mechanism," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. (APSIPA)*, 2021, pp. 523–529.

[18] G. Yu, Y. Wang, H. Wang, Q. Zhang, and C. Zheng, "A two-stage complex network using cycle-consistent generative adversarial networks for speech enhancement," *Speech Commun.*, vol. 134, pp. 42–54, 2021.

[19] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Commun.*, vol. 53, no. 4, pp. 465–494, 2011.

[20] A. Li, W. Liu, C. Zheng, C. Fan, and X. Li, "Two heads are better than one: A two-stage complex spectral mapping approach for monaural speech enhancement," *IEEE/ACM Trans. Audio. Speech, Lang. Process.*, vol. 29, pp. 1829–1843, 2021.

[21] A. Li, W. Liu, X. Luo, G. Yu, C. Zheng, and X. Li, "A simultaneous denoising and dereverberation framework with target decoupling," in *Proc. Interspeech*, 2021, pp. 2801–2805.

[22] C. Trabelsi, O. Bilaniuk, Y. Zhang, D. Serdyuk, S. Subramanian, J. F. Santos, S. Mehri, N. Rostamzadeh, Y. Bengio, and C. J. Pal, "Deep complex networks," *arXiv preprint arXiv:1705.09792*, 2017.

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

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

[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 *Proc. O-COCOSDA/CASLRE*. IEEE, 2013, pp. 1–4.

[26] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings," *JASA*, vol. 133, no. 5, pp. 3591–3591, 2013.

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

[28] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *arXiv preprint arXiv:1806.03185*, 2018.

[29] F. Germain, Q. Chen, and V. Koltun, "Speech denoising with deep feature losses," in *Proc. Interspeech*, 2019, pp. 2723–2727.

[30] K. Tan and D. L. Wang, "A convolutional recurrent neural network for real-time speech enhancement.," in *Proc. Interspeech*, 2018, pp. 3229–3233.

[31] K. Tan and D. L. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Trans. Audio. Speech, Lang. Process.*, vol. 28, pp. 380–390, 2019.

[32] Y. Hu, Y. Liu, S. Lv, M. Xing, and L. Xie, "Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Proc. Interspeech*, 2020, pp. 2472–2476.

[33] W. Yuan, "A time–frequency smoothing neural network for speech enhancement," *Speech Commun*, vol. 124, pp. 75–84, 2020.

[34] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*. IEEE, 2001, vol. 2, pp. 749–752.

[35] 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. ICASSP*. IEEE, 2010, pp. 4214–4217.

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

[37] C. Reddy, V. Gopal, and R. Cutler, "Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," *arXiv preprint arXiv:2010.15258*, 2020.