Joint magnitude estimation and phase recovery using Cycle-in-Cycle GAN for non-parallel speech enhancement

Guochen Yu[1,2], Andong Li[2], Yutian Wang[1], Yinuo Guo[3], Hui Wang[1], Chengshi Zheng[2]
[1]State Key Laboratory of Media Convergence and Communication, Communication University [2]Key Laboratory of Noise and Vibration Research, Institute of Acoustics, Chinese Academy of Sciences [3]Bytedance, Beijing
{yuguochen, wangyutian, hwang}@cuc.edu.cn, {liandong, cszheng}@mail.ioa.ac.cn

Summary

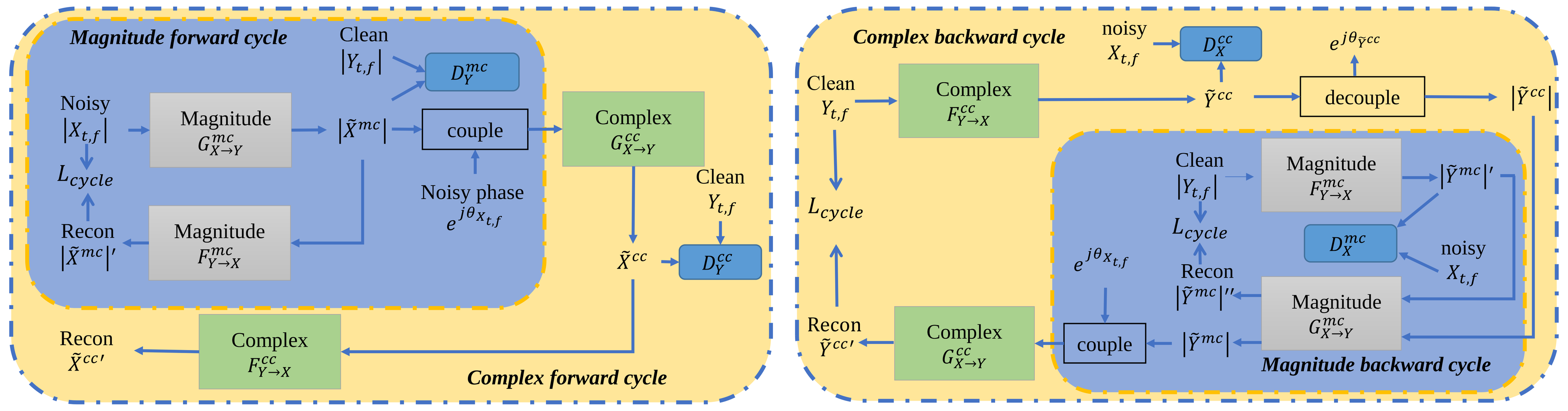
- We propose a novel system, dubbed Cycle-in-Cycle GAN, to handle speech enhancement when the training noisy-clean data pairs are mismatched.
- Inspired by the decoupling-style concept, we decouple the difficult target *w.r.t.* original spectrum optimization and use two CycleGANs to jointly estimate the spectral magnitude and phase information in a stage-wise manner.
- In the first stage, we pretrain a magnitude CycleGAN 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.
- Experimental results demonstrate that the proposed approach significantly outperforms previous baselines under non-parallel training.

Introduction

Non-parallel single-channel speech enhancement:

- Standard DNN-based supervised SE approaches always need numerous paired clean-noisy samples to conduct supervised training and improve the generalization.
- 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.
- cycle-consistent GAN (CycleGAN) has been developed to conduct unsupervised SE by using adversarial loss, cycle-consistency loss and identity-mapping loss.
- Due to the severe mismatch between input and target, previous CylceGAN based methods only focus on magnitude spectrum estimation and remain the noisy phase unaltered.

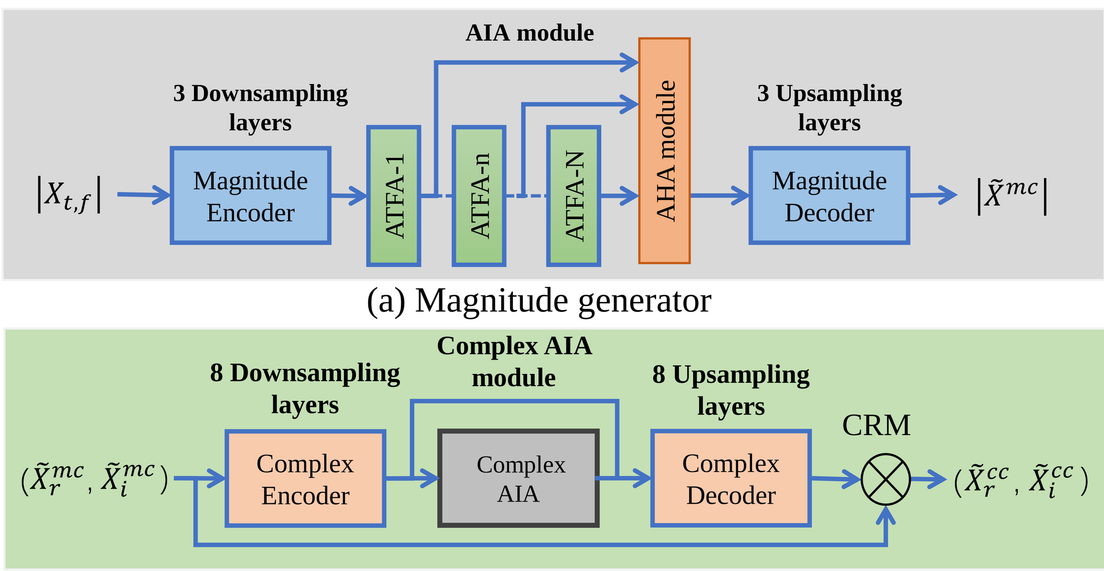
Cycle-in-Cycle GAN for non-parallel SE



(a) forward noisy-clean-noisy cycle (b) backward clean-noisy-clean cycle

Figure 1: System flowchart of CinCGAN. The magnitude and complex cycle are shown in the yellow and blue dotted boxes, respectively.

- As Figure. 1(a) and (b) show, the proposed CinCGAN consists of a forward noisy-clean-noisy cycle and a backward clean-noisy-clean cycle. 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. Subsequently, the estimated spectral magnitude and the original phase are fed it into CCGAN to estimate both real and imaginary parts.



(a) Magnitude generator
(b) Complex generator

Figure 1: (a) The diagram of the magnitude generators in MCGAN. (b) The diagram of the complex generators in CCGAN.

- As Figure 2(a) and (b) illustrate, both magnitude and complex generators adopt convolutional encoder-decoder topology, and multiple adaptive time-frequency attention (ATFA) modules and an adaptive hierarchical attention (AHA) module are inserted for temporal modeling.
- In the first step, we pretrain MCGAN alone with the same relativistic adversarial loss, cycle-consistency loss, and identity mapping loss until convergence. Then, we jointly fine-tune MCGAN and CCGAN with the same losses, which can be expressed as:

$$\mathcal{L}_{MCGAN} = \mathcal{L}_{Radv}(G^{mc}_{X\rightarrow Y}, D^{mc}_{Y}) + \mathcal{L}_{Radv}(F^{mc}_{Y\rightarrow X}, D^{mc}_{X}) + \lambda_{cycle}\mathcal{L}_{cycle}(G^{mc}_{X\rightarrow Y}, F^{mc}_{Y\rightarrow X}) + \lambda_{id}\mathcal{L}_{id}(G^{mc}_{X\rightarrow Y}, F^{mc}_{Y\rightarrow X}) \quad (1)$$

Experiments

Datset:

- The dataset we chosen is a selection of the Voice Bank corpus 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.
- For the training set, audio samples are mixed together with one of the 10 noise types (*i.e.*, two artificial and eight real noise from the DEMAND database. The testing utterances are created with 5 unseen test-noise types from the DEMAND.

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.
- We conduct the power compression toward the spectral magnitude while leaving the phase unaltered, and the optimal compression coefficient is set to 0.5.
- 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 which is different from the input audio.
- The training process is divided in two steps and the overall loss function can be expressed as:

$$\mathcal{L}_{CinCGAN} = \gamma\mathcal{L}_{MCGAN} + \mathcal{L}_{CCGAN} \quad (2)$$

Comparison results & analysis

Table 1: Experimental results among different models under unpaired data.

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

Methods	Feature type	PESQ	STOI(%)	CSIG	CBAK	COVL
Unprocessed	–	1.97	92.1	3.35	2.44	2.63
GAN-based Systems						
SEGAN	Waveform	2.16	92.5	3.48	2.94	2.80
MMSEGAN	Gammatone	2.53	93.0	3.80	3.12	3.14
SERGAN	Waveform	2.51	93.7	3.78	3.23	3.16
CP-GAN	Waveform	2.64	94.0	3.93	3.29	3.28
MetricGAN	Magnitude	2.86	–	3.99	3.18	3.42
CRGAN	Magnitude	**2.92**	94.0	4.16	3.24	**3.54**
SASEGAN	Waveform	2.36	93.5	3.54	3.08	2.93
Non-GAN based Systems						
Wave-U-net	Waveform	2.64	–	3.56	3.08	3.09
DFL-SE	Waveform	–	–	3.86	3.33	3.22
CRN-MSE	Magnitude	2.61	93.8	3.78	3.11	3.24
GCRN	RI components	2.51	94.0	3.71	3.24	3.09
DCCRN	RI components	2.68	93.9	3.88	3.18	3.27
TFSNN	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

Conclusions

- This paper proposes 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.
- The proposed system surpasses previous state-of-the-art non-parallel GAN based speech enhancement systems, indicating the superiority of the cycle-in-cycle paradigm under mismatched noisy-clean pairs.
- When experiments are conducted on standard parallel data, the proposed approach also demonstrates its effectiveness in improving speech quality and reducing speech distortion.