IACAS

• 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] and Chengshi Zheng[2]

[1]State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing, China

[2]Key Laboratory of Noise and Vibration Research, Institute of Acoustics, Chinese Academy of Sciences, Beijing, China

[3]Bytedance, Beijing, China

IACAS

OUTLINE

IACAS

Introduction

In real acoustic environment, speech quality and intelligibility can be severely degraded by background noise.

Applications of speech enhancement [1]:

1：Remote communication

2：The front-ends for automatic speech recognition (ASR) systems

3：Hearing assistant devices

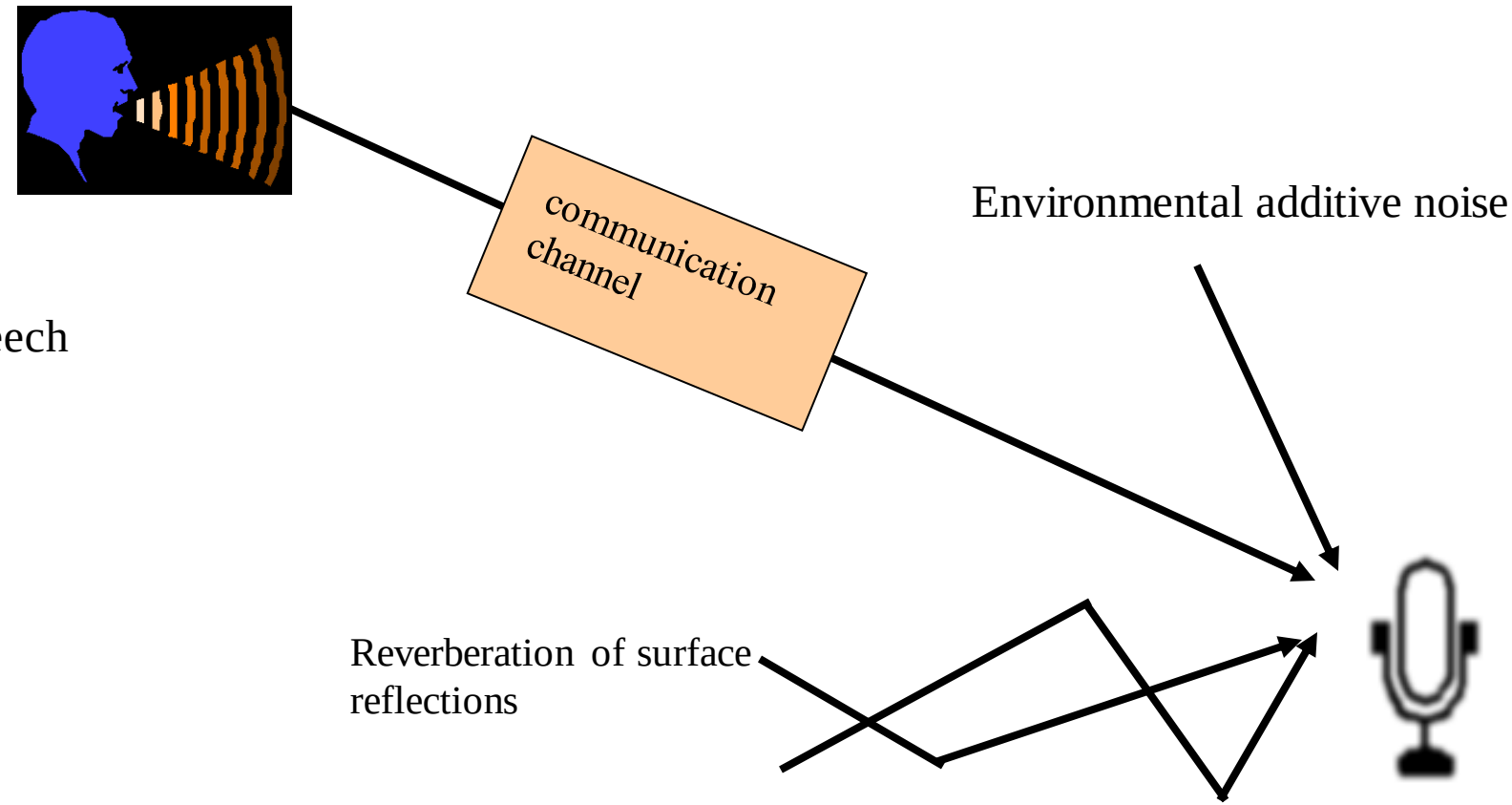


Fig1：Noise interferences in the real environment

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

IACAS

Introduction

Traditional supervised SE methods

1）Masking-based [2]: IRM, IBM, CRM, PSM

2）Feature-mapping based [3]: magnitude spectrum, complex spectrum, waveform

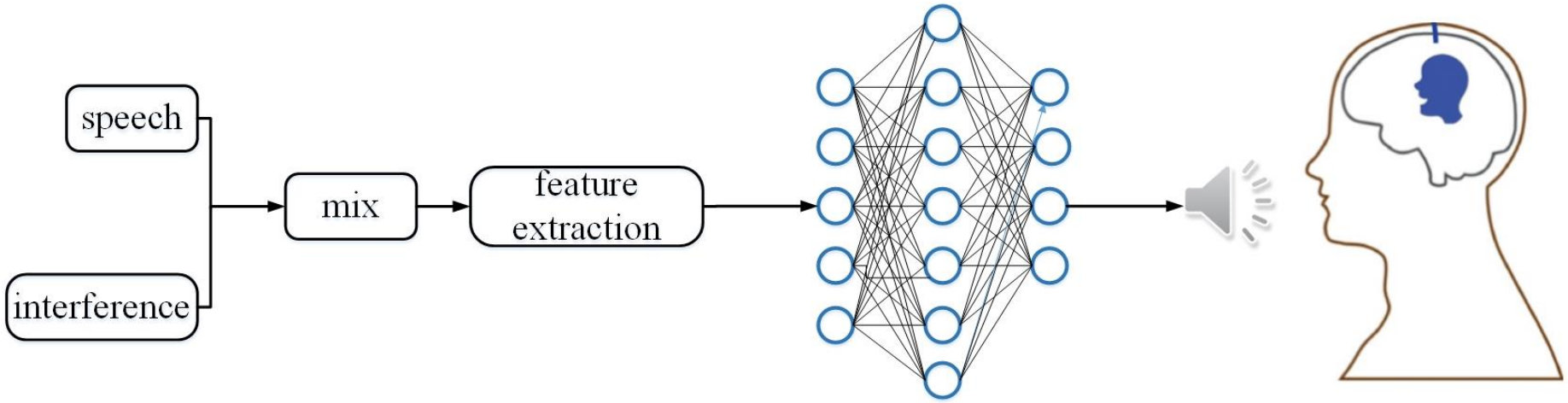


Fig 2：DNN-based SE framework

[2] 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.

[3] 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.

IACAS

Introduction

Limitations of supervised SE topology:

1）These methods need numerous paired clean-noisy samples to conduct supervised training and improve the generalization.

2）In real scenarios, we can only obtain the clean recordings that mismatch the source noisy data.

3）The performance of these supervised SE methods always degrades under unseen speaker and noise conditions.

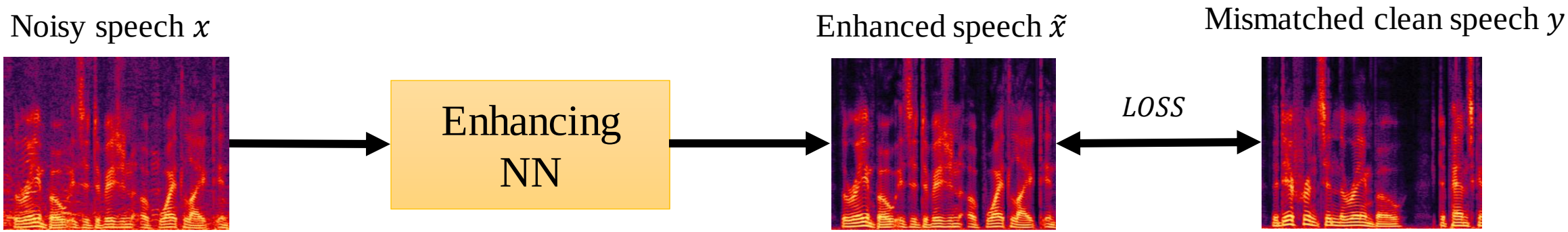


Fig 3：Non-parallel speech enhancement topology

IACAS

OUTLINE

IACAS

Related works

Generative adversarial networks (GAN) can produce SE by adversarial training [4]:

1) Generators conduct the enhancement process.

2) Discriminators can classify the target speech features as real and the generated speech features as fake (can substitute traditional supervised loss (MSE, MAE, etc)).

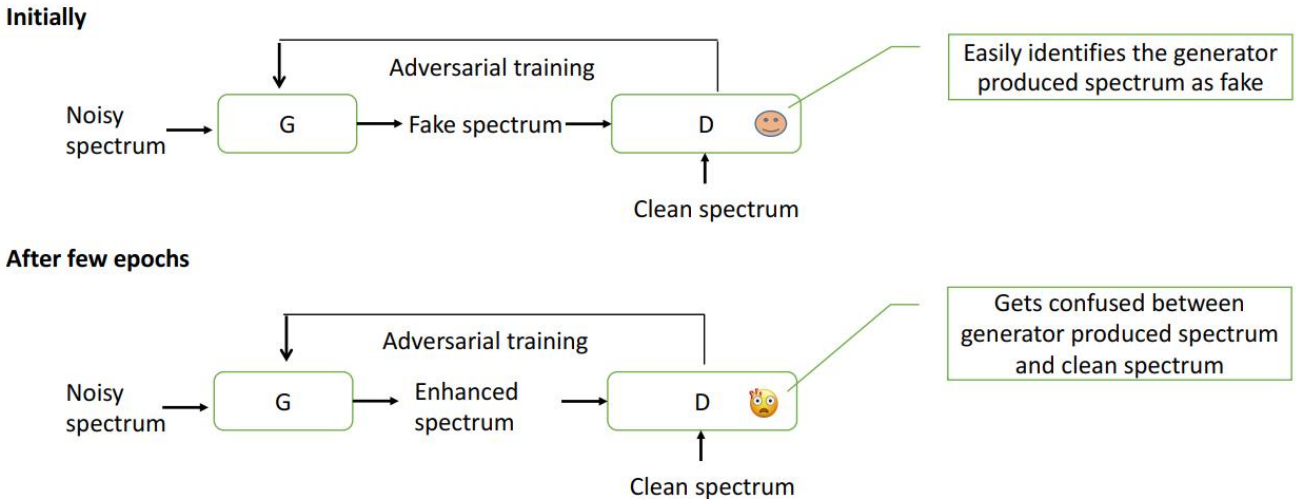


Fig 4: GAN-based SE framework [5]

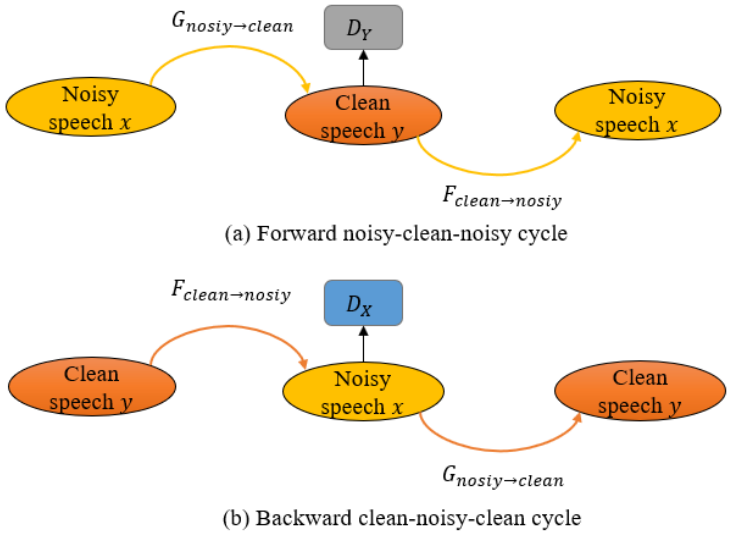


Fig 5: Cycle consistent GAN based SE framework [5]

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

[5] 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 Communication, vol. 134, pp. 42–54, 2021.

IACAS

Related works

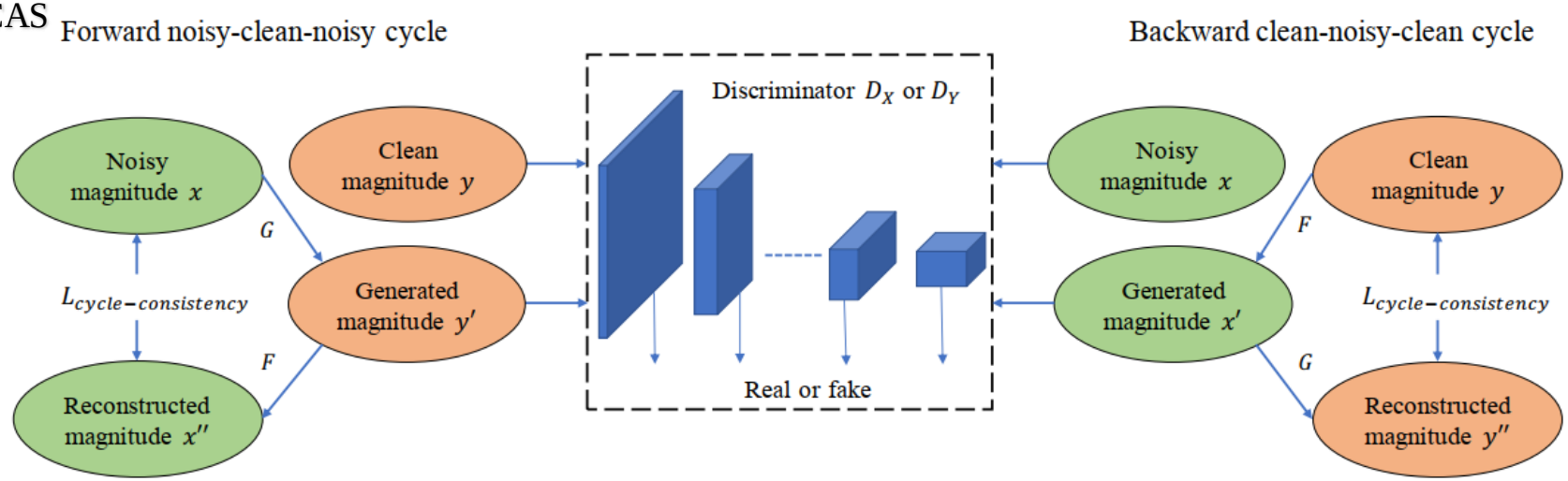


Fig 6: Non-parallel CycleGAN-based SE framework [6]

Adversarial loss: $\mathcal{L}_{adv}(G_{X\to Y}, D_Y) = \mathbb{E}_{y\sim P_Y(y)}[\log D_Y(y)] + \mathbb{E}_{x\sim P_X(x)}[\log(1 - D_Y(G_{X\to Y}(x)))]$

cycle consistency loss: $\mathcal{L}_{cycle}(G_{X\to Y}, F_{Y\to X}) = \mathbb{E}_{x\sim P_X(x)}[\| F_{Y\to X}(G_{X\to Y}(x)) - x \|_1] + \mathbb{E}_{y\sim P_Y(y)}[\| G_{X\to Y}(F_{Y\to X}(y)) - y \|_1]$

Identity-mapping loss: $\mathcal{L}_{identity}(G_{X\to Y}, F_{Y\to X}) = \mathbb{E}_{x\sim P_X(x)}[\| F_{Y\to X}(x) - x \|_1] + \mathbb{E}_{y\sim P_Y(y)}[G_{X\to Y}(y) - y \|_1]$

Full loss: $\mathcal{L}_{Full} = \mathcal{L}_{adv}(G_{X\to Y}, D_Y) + \mathcal{L}_{adv}(F_{Y\to X}, D_X) + \lambda_{cycle}\mathcal{L}_{cycle}(G_{X\to Y}, F_{Y\to X}) + \lambda_{id}\mathcal{L}_{identity}(G_{X\to Y}, F_{Y\to X})$

[6] 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. APSIPA-ASC, pp. 523–529, 2021.



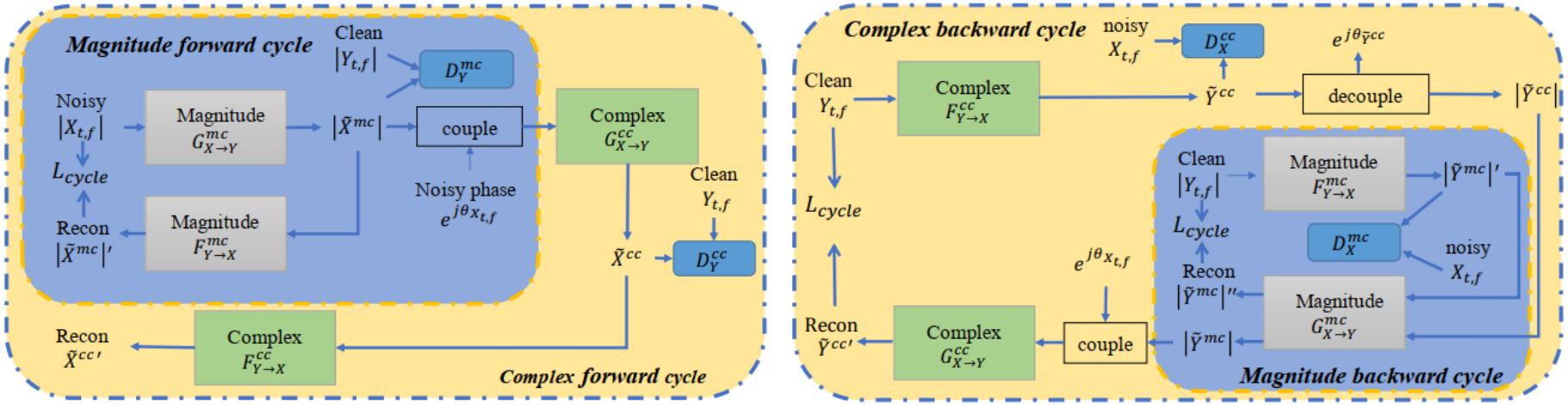
IACAS

OUTLINE

Proposed Method

IACAS

Joint magnitude estimation and phase recovery by Cycle-in-Cycle GAN



(a) (b)

Fig 5: Proposed Cycle-in-Cycle GAN

- Decoupling the complex-spectrum optimization into two Cycles:
  - Magnitude CycleGAN (MCGAN): coarse estimation toward the spectral magnitude
  - Complex CycleGAN (CCGAN): refine the complex spectrum and implicitly recover phase

Inference procedure:

$$|\tilde{X}^{mc}| = G^{mc}_{X\to Y}\left(|X_{t,f}|;\Phi_1\right), \quad (1)$$

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

$$\tilde{X}^{cc} = G^{cc}_{X\to Y}\left(\tilde{X}^{mc};\Phi_2\right), \quad (3)$$

Proposed Method

IACAS

Forward noisy-clean-noisy cycle:

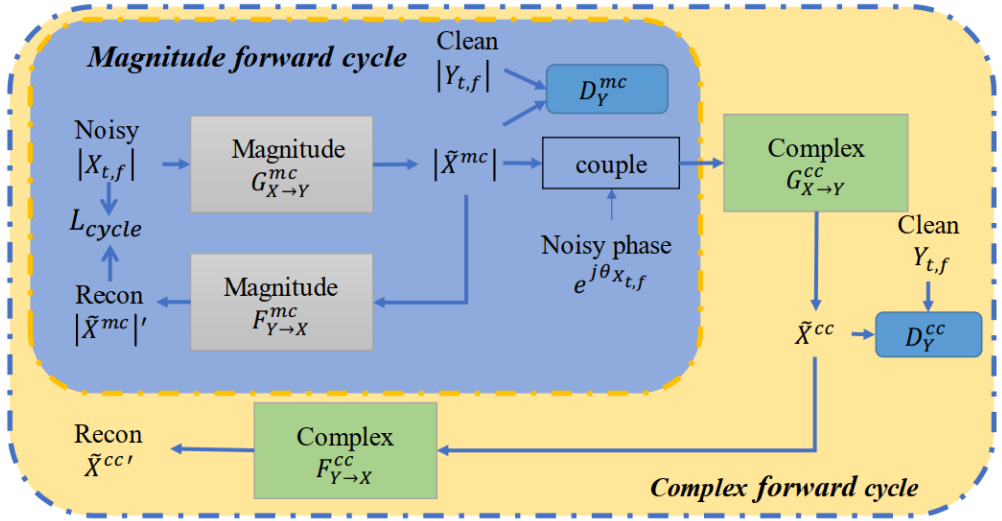


Fig 5: Forward noisy-clean-noisy cycle

Two steps: noisy-clean mapping and clean-noisy reconstruction

- Given noisy spectrum, estimate clean magnitude spectrum $|\tilde{X}^{mc}|$ and clean complex spectrum $\tilde{X}^{cc}$
- Reconstruct noisy spectral magnitude $|\tilde{X}^{mc}|'$ and noisy complex spectrum $\tilde{X}^{cc}$

Proposed Method

IACAS

Backward clean-noisy-clean cycle:

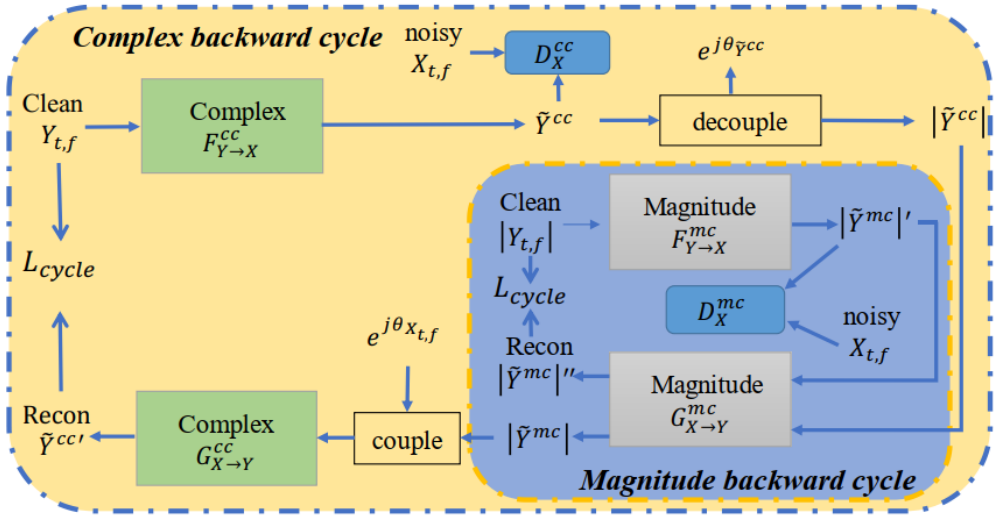


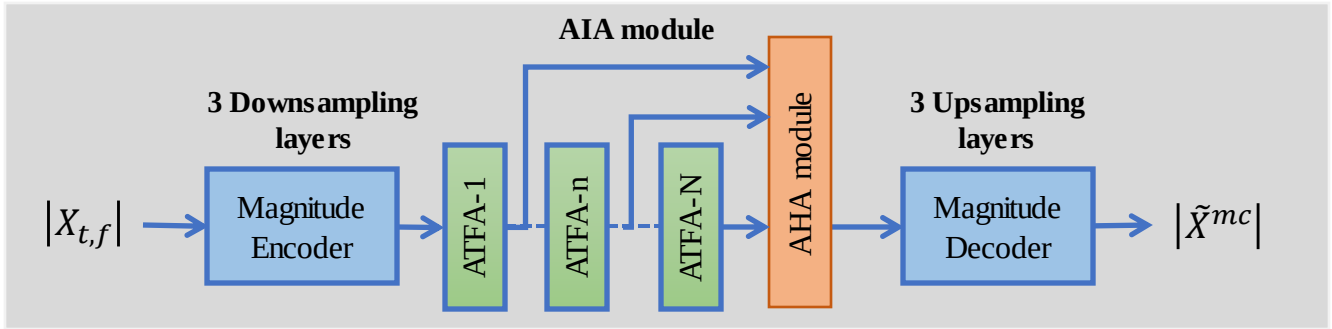
Fig 5: Backward clean-noisy-clean cycle

Two steps: clean-noisy mapping and noisy-clean reconstruction

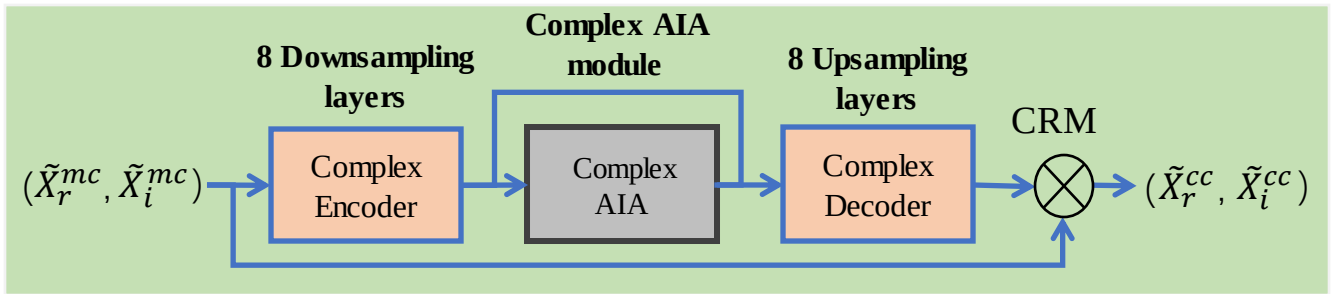
- Given clean maginitude and complex spectrum, estimate clean complex spectrum $\tilde{Y}^{cc}$ and clean magnitude spectrum $\tilde{X}^{cc}$
- Reconstruct clean spectral magnitude $|\tilde{Y}^{mc}|''$ and clean complex spectrum $\tilde{Y}^{cc\prime}$

Proposed Method

The architecture of generators



(a)



(b)

Fig 7: (a) The diagram of the magnitude generator. (b) The diagram of the complex generator.

- ✓ Each generator follows encoder-decoder topology: real-valued magnitude generators [6] and complex-valued complex generators [6]
- ✓ Adaptive attention-in-attention (AIA) module for sequence modelling: four adaptive time-frequency attention (ATFA) modules and adaptive hierarchical attention (AHA) module [6]

[5] 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 Communication, vol. 134, pp. 42–54, 2021.

[6] 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. APSIPA-ASC, pp. 523–529, 2021.

Proposed Method

IACAS

Adaptive time-frequency attention (ATFA) module

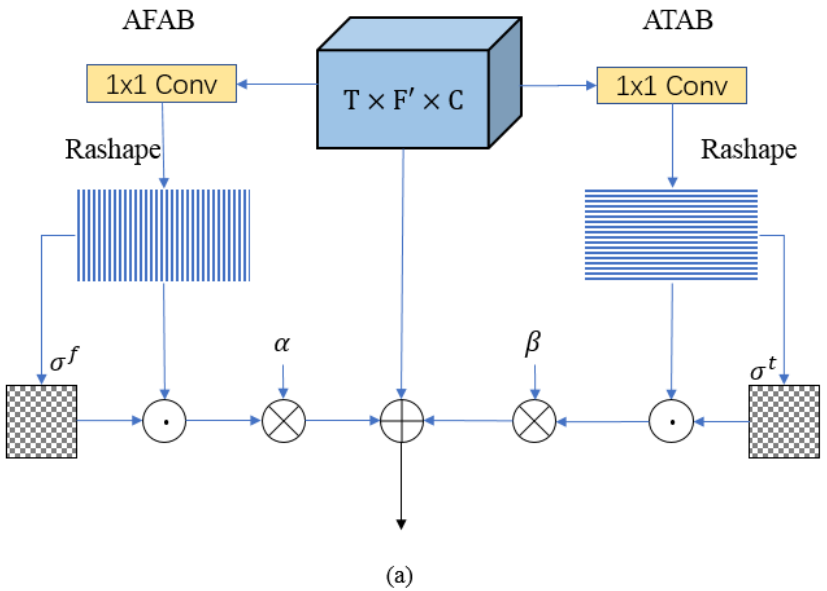


Fig 8: The diagram of adaptive time-frequency attention (ATFA) modules. ⊙ and ⊗ denote the matrix multiplication and element-wise multiplication, respectively.

- ✓ ATFA module consists of two branches: an adaptive time attention branch (ATAB) and an adaptive frequency attention branch (AFAB) .
- ✓ ATAB and AFAB cooperate to capture the global dependencies along temporal and frequency dimensions in parallel, with lower computational cost than original attention.

ATAB:

$$Q^t = Conv_Q^t(F_{in}), K^t = Conv_K^t(F_{in}), V^t = Conv_V^t(F_{in}),$$
$$Q_{Res}^t, K_{Res}^t, V_{Res}^t = Reshape^t(Q^t, K^t, V^t),$$
$$\sigma^t = softmax((Q_{Res}^t) \cdot (K_{Res}^t)^T),$$
$$Out_{ATAB} = Reshape^{t'}(\sigma^t \cdot V_{Res}^t),$$

Output:

$$Out_{ATFA} = F_{in} + \alpha Out_{ATAB} + \beta Out_{AFAB}$$

Proposed Method

IACAS

Adaptive adaptive hierarchical attention (AHA) module

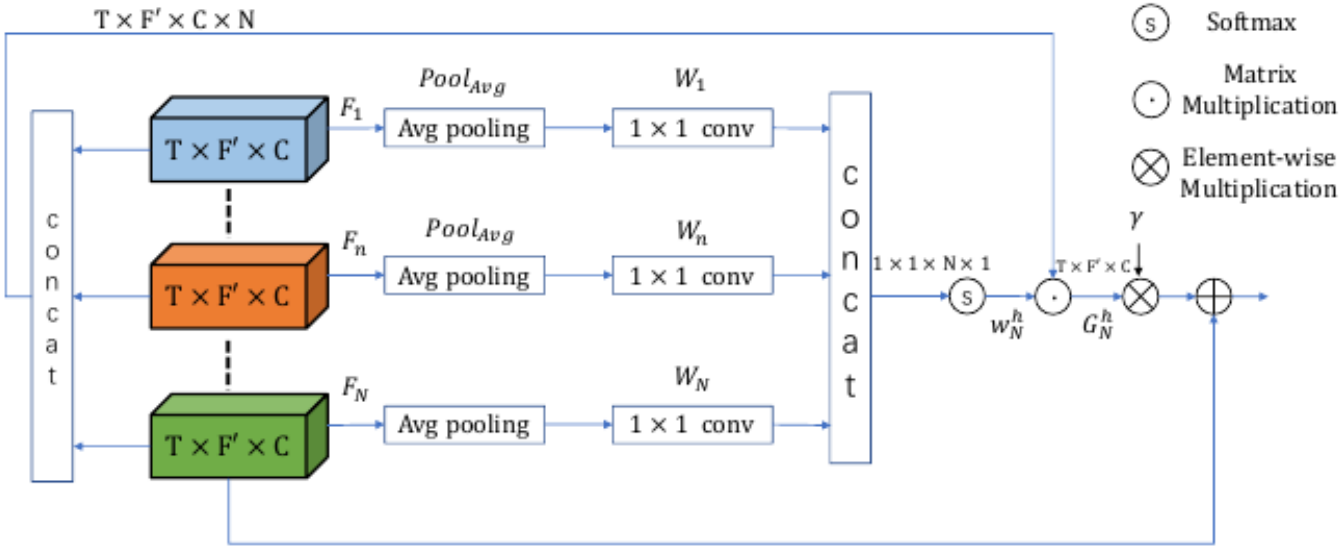


Fig 9: The diagram of the adaptive hierarchical attention (AHA) module.

- ✓ AHA module aims at integrating the different hierarchical feature maps given a set of ATFA modules' outputs.
- ✓ It is designed to fuse the global context of all intermediate feature maps with different weights and progressively guide the enhancement procedure.

$$Out_{AHA} = F_N + \gamma \sum_{i=1}^{N} W_i^h F_i^h$$

$$= F_N + \gamma \sum_{i=1}^{N} \frac{\exp(Pool_{Avg}(F_i) * W_i)}{\sum_{n=1}^{N} \exp(Pool_{Avg}(F_n) * W_n)} F_i.$$

IACAS

OUTLINE

IACAS

Experiments and Analysis

- Corpus: Voice Bank [7], which includes 30 speakers.
- Training set
  - ✓ 11572 utterances from 28 speakers (14 male and 14 female)
  - ✓ ten environmental noise from DEMAND [8], mixed at 0, 5,10, 15 dB.
- Test set :
  - ✓ 824 utterances from 2 unseen speakers
  - ✓ SNRs and Noises: five unseen environmental mixed at 2.5, 7.5, 12.5, 17.5 dB.
- Unpaired noisy-clean data pairs:
  - ✓ Shuffle the noisy-clean pairs
  - ✓ we randomly crop a fixed-length segment (i.e., 108 frames) from a randomly selected noisy audio file (different from the target)

[7] 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.
[8] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multichannel acoustic noise database: A database of multichannel environmental noise recordings," Acoustical Society of America Journal, vol. 133, no. 5, pp. 3591, 2013.

Experiments and Analysis

IACAS

Experimental setup:

- Sampling rate: 16kHz
- STFT Window size: 512 samples (32ms), Overlap: 384 samples (24ms), 257-dimensional STFT spectrum
- Loss criterion: Relativistic average least-square adversarial loss (RsLS)[9]
- Power compression [10]: compression coefficient η is set to 0.5 towards magnitude, i.e., $\left|\tilde{X}_{t,f}\right|^{\eta} = G_{X\to Y}(|X_{t,f}|^{\eta};\phi_G),$
- When pretraining MCGAN, the initialized learning rate (LR) is set to 0.0005 for G and 0.0002 for D; When MCGAN and CCGAN are jointly trained, LRs are set to 0.0002 for all G and D
- Full Loss function: $\mathcal{L}_{CinCGAN} = \gamma\mathcal{L}_{MCGAN} + \mathcal{L}_{CCGAN}$

[9] A. Jolicoeur-Martineau, "The relativistic discriminator: a key elementmissing from standard gan," arXiv preprint arXiv:1807.00734, 2018
[10] A. Li, C. Zheng, R. Peng, and X. Li, "On the importance of power compression and phase estimation in monaural speech dereverberation," JASA Express Letters, vol. 1, no. 1, pp. 014802, 2021

Experiments and Analysis

IACAS

- Baselines:

  Non-parallel ablation study:

  ✓ Original maginitude and complex GAN (MGAN and CGAN) and + forward cycle (fc) (MGAN + fc and CGAN + fc)

  ✓ Magnitude and Complex CycleGAN (MCGAN and CCGAN)

  ✓ Cycle-in-Cycle GAN with different fc and backward cycle (bc)

  Comparsion with SOTA under standard parallel training:

  ✓ GAN-based: SEGAN, MMSEGAN, SERGAN, MetricGAN, CP-GAN, SASEGAN and CRGAN

  ✓ Non-GAN based: Wave-U-net, DFL-SE, CRN-MSE, GCRN, DCCRN and TFSNN.

- Evaluation metrics:

  ✓ PESQ, STOI, segmental signal-to-noise ratio (SegSNR)

  ✓ The MOS prediction of speech distortion (CSIG), background noise (CBAK) and overall effect (COVL).[10]

  ✓ Subjective non-intrusive perceptual speech quality metric (DNSMOS) [11]

[10] 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.

[11] C. K. 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.

Experimental Results

IACAS

Ablation study

Table 1: Ablation study among different models including GAN-based methods and the proposed CycleGAN-based methods 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**

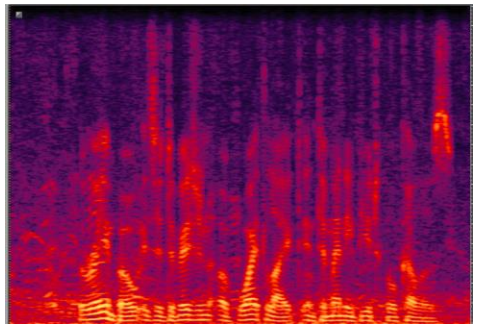
- CycleGAN-based methods surpass original GAN-based methods under non-parallel training
- For the difficulty of estimating magnitude and phase simultaneously, MCGAN outperforms CCGAN.
- CinCGAN-based methods achieve better performance than single magnitude and complex CycleGAN.
- When both forward and backward cycle are applied, performance in all metrics are increased.

Experimental Results

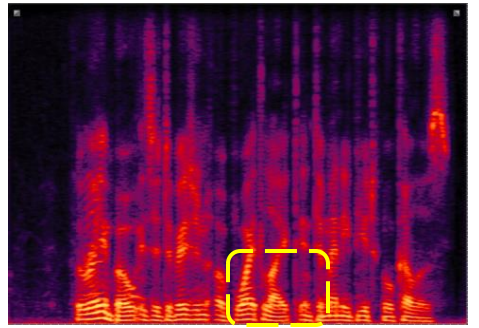
IACAS

Table 2: Comparison with other GAN and Non-GAN based systems under standard paired data.

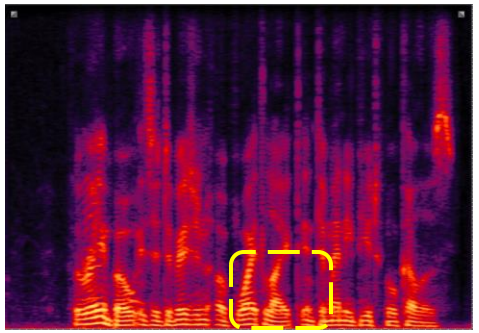
Methods	Feature type	PESQ	STOI(%)	CSIG	CBAK	COVL
Unprocessed	–	1.97	92.1	3.35	2.44	2.63
GAN-based Systems						
SEGAN [6]	Waveform	2.16	92.5	3.48	2.94	2.80
MMSEGAN [9]	Gammatone	2.53	93.0	3.80	3.12	3.14
SERGAN [7]	Waveform	2.51	93.7	3.78	3.23	3.16
CP-GAN [10]	Waveform	2.64	94.0	3.93	3.29	3.28
MetricGAN [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



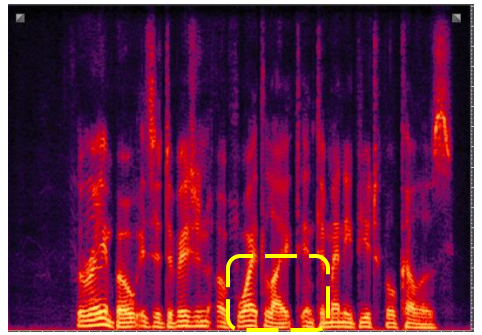
(a) Noisy (pesq= 1.55)



(b) MCGAN (pesq= 2.56)



(c) CCGAN (pesq= 2.44)



(d) CinCGAN (pesq= 2.72)

Comparison with SOTA under parallel training:

- Compared with SOTA GAN-based and Non-GAN methods, CinCGAN provides better performance on STOI, CSIG and CBAK scores.

IACAS

OUTLINE

IACAS

Experimental Results

- We propose a Cycle-in-Cycle GAN framework dubbed CinCGAN for non-parallel speech enhancement in the T-F domain.
- By coupling the magnitude and complex cycle, CinCGAN aims to jointly recover the spectral magnitude and phase information.
- The proposed system surpasses previous state-of-the-art non-parallel GAN based 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.