IACAS

Dual-branch Attention-In-Attention Transformer for single-channel speech enhancement

Guochen Yu[1,2], Andong Li[2], Chengshi Zheng[2] , Yinuo Guo[3], Yutian Wang[1], and Hui Wang[1]

[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

Background

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

- Supervised SE methods based on deep learning are mainly divided into time-frequency domain methods and time domain methods [1].

- The time-frequency domain methods mainly conduct masking and mapping on spectral magnitude or complex spectrum [2, 3].

- The time domain method directly map the clean waveform.

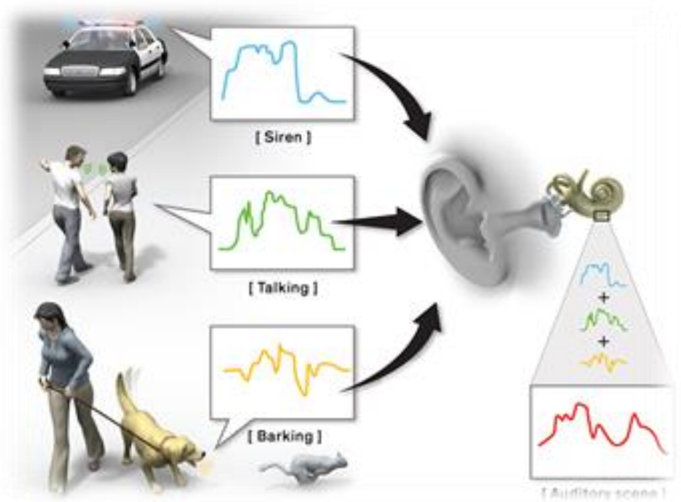


Figure 1: adverse acoustic environment

[1] 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

[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] 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.

IACAS

Background

The recovery of phase is important to improve speech perception quality. [4]

Complex spectrum based SE:
$$Y^{(r)}_{m,f} + iY^{(i)}_{m,f} = \left(S^{(r)}_{m,f} + N^{(r)}_{m,f}\right) + i\left(S^{(i)}_{m,f} + N^{(i)}_{m,f}\right),$$

1） complex ration mask （CRM）[5]

$$CRM = \frac{X_r S_r + X_i S_i}{X_r^2 + X_i^2} + j\frac{X_r S_i - X_i S_r}{X_r^2 + X_i^2} = \widetilde{M}_r + j\widetilde{M}_i$$

2） estimating real and imaginary components of complex spectrum [6]

[4] K. Paliwal, K. Wojcicki, and B. Shannon, “The importance of phase in speech enhancement,” Speech Commun, vol. 53, no.4, pp. 465–494, 2011.
[5] D. S. Williamson, Y. Wang, D. Wang, Complex ratio masking for monaural speech separation, IEEE/ACM Trans. Audio. Speech, Lang. Process., vol. 24, no. 3, pp. 483–492, 2015
[6] 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

OUTLINE

IACAS

Related works

Decoupling-style phase-aware SE methods:

Decouple the original complex spectrum optimization into magnitude and phase estimation, and two sub-network are utilized in a step-wise manner [7].

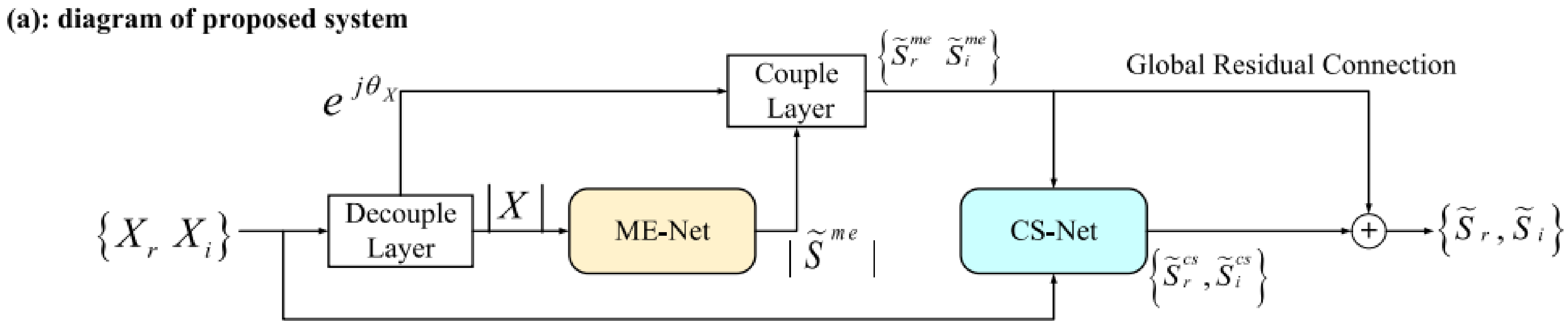


Fig 2：The diagram of CTS-Net [5], which consist of a magnitude estimation network (ME-Net) and a complex spectrum refine network (CS-Net)

[7] 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.

IACAS

Related works

Transformer-based SE approaches:

Dual-path transformer has been developed for sequence modelling in speech area [8].

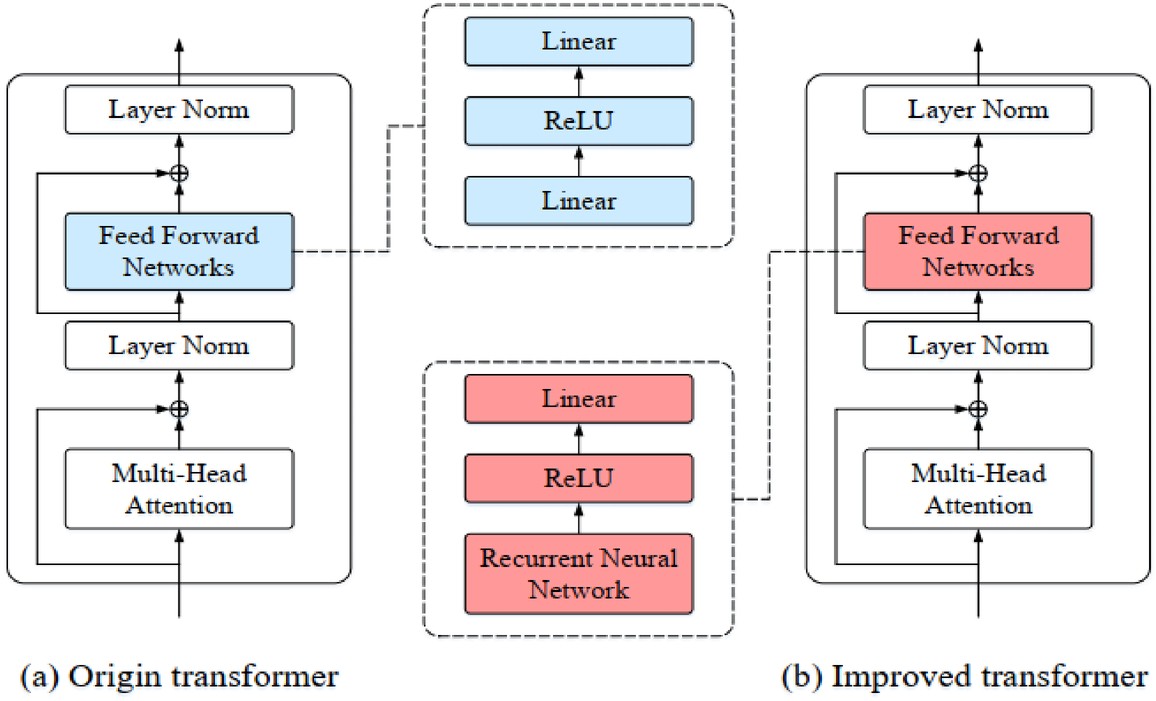


Fig 3: The diagram of dual-path transformer for speech separation

[8] J. Chen, Q. Mao, and D. Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," arXiv preprint arXiv:2007.13975, 2020.

IACAS

OUTLINE

Proposed Method

Dual-branch Attention-In-Attention Transformer for single-channel SE

IACAS

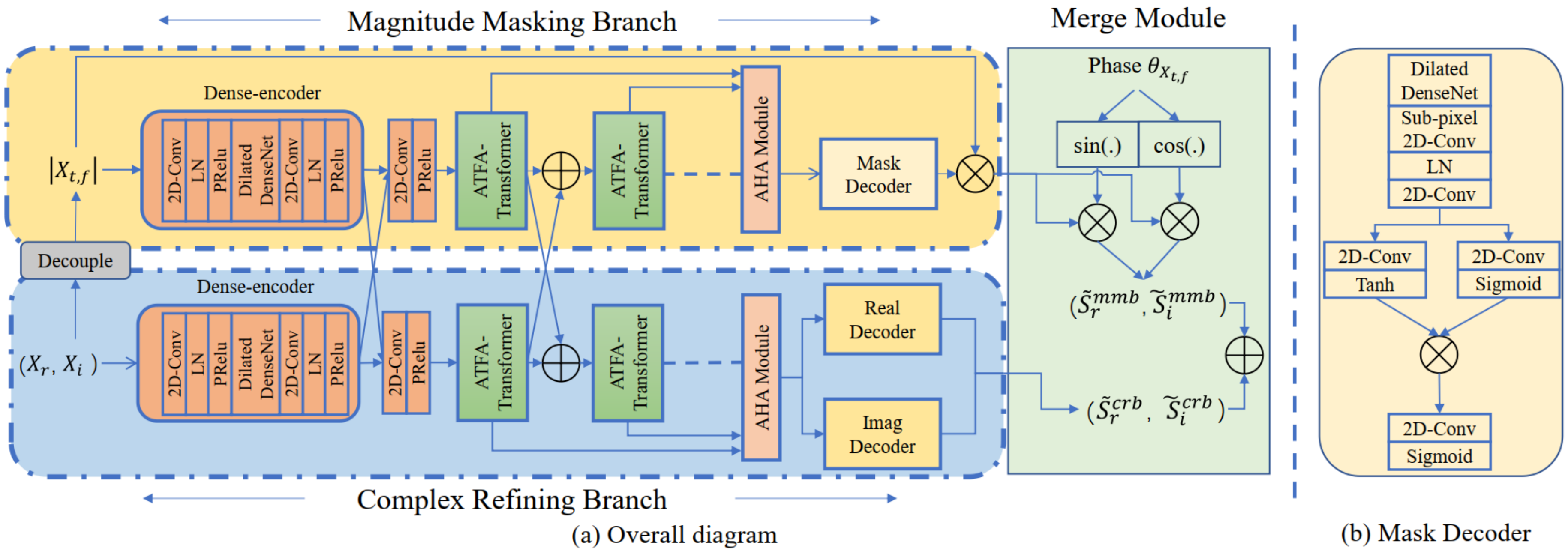


Fig 4： Proposed dual-branch system flowchart

- Two core branches are elaborately designed in parallel:
  - A magnitude masking branch (MMB): filtering out most of the noise in the magnitude domain.
  - A complex refining branch (CRB): compensate for the lost spectral details and implicitly recover phase in the complex domain

IACAS

Proposed Method

- MMB path estimates the magnitude mask to coarsely recover the magnitude of the target speech, and the coarsely estimated spectral magnitude is coupled with the noisy phase.

- CRB path receives noisy real and imaginary (RI) components as the input and focuses on the residual fine-grained spectral structures which is lost in MMB.

- Finally, we sum the coarse-denoised complex spectrum in MMB and the fine-grained complex spectral details in CPB together to reconstruct the clean complex spectrum

- The training procedure can be expressed as:

$$\left|\tilde{S}^{mmb}\right| = \left|X_{t,f}\right| \otimes M^{mmb} \tag{1}$$

$$\tilde{S}_r^{mmb} = \left|\tilde{S}^{mmb}\right| \otimes \cos(\theta_X) \tag{2}$$

$$\tilde{S}_i^{mmb} = \left|\tilde{S}^{mmb}\right| \otimes \sin(\theta_X) \tag{3}$$

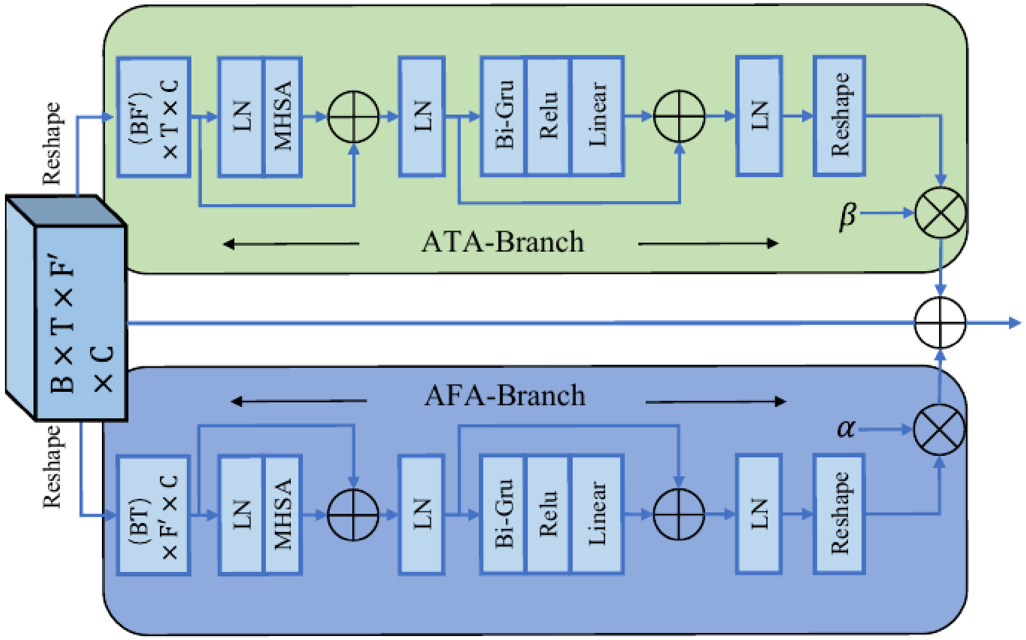
$$\tilde{S}_r = \tilde{S}_r^{mmb} + \tilde{S}_r^{crb} \tag{4}$$

$$\tilde{S}_i = \tilde{S}_i^{mmb} + \tilde{S}_i^{crb} \tag{5}$$

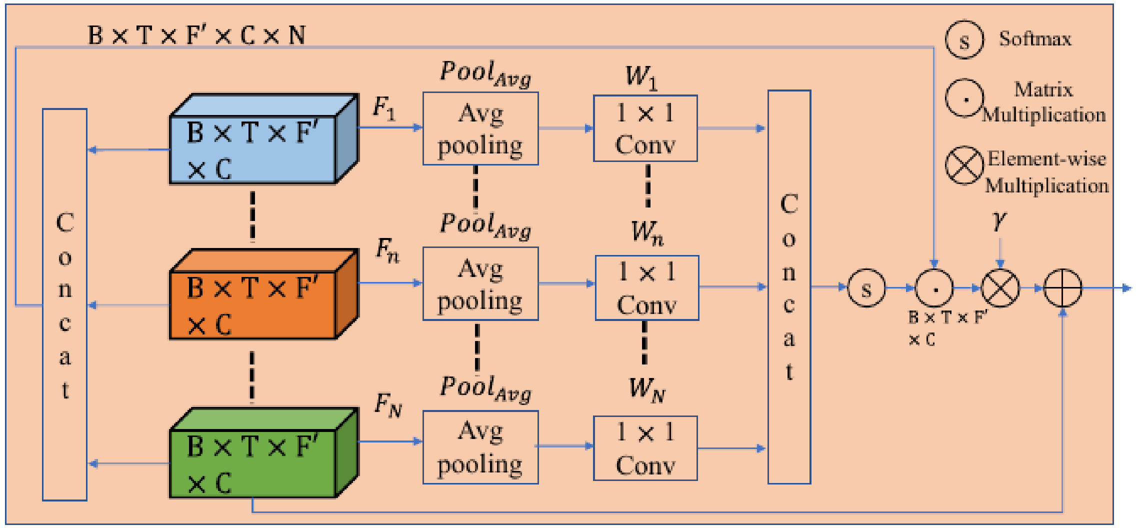
Proposed Method

Attention-in-attention transformer:

consists of four adaptive time-frequency attention transformer-based (ATFA-T) blocks and an adaptive hierarchical attention (AHA) module.



(a)



(b)

Fig 5: The diagram of ATFA-T blocks

Fig 6: The diagram of AHA module

IACAS

Proposed Method

- The loss function of the proposed dual-branch model:

$$L^{Mag} = \left\| \sqrt{\left|\tilde{S}_r\right|^2 + \left|\tilde{S}_i\right|^2} - \sqrt{\left|S_r\right|^2 + \left|S_i\right|^2} \right\|_F^2 \tag{6}$$

$$L^{RI} = \left\| \tilde{S}_r - S_r \right\|_F^2 + \left\| \tilde{S}_i - S_i \right\|_F^2 \tag{7}$$

$$L_{FULL} = L^{Mag} + L^{RI} \tag{8}$$

IACAS

OUTLINE

Experiments and Analysis

IACAS

Dataset

- Corpus: Voice Bank [9], which includes 28 speakers for training and 2 unseen speakers for testing.

- Training set

  - ✓ 11572 utterances from 28 speakers (14 male and 14 female)

  - ✓ ten environmental noise from DEMAND database [10], 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.

[9] 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.
[10] 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.

IACAS

Experiments and Analysis

Experimental setup:

- Sampling rate: 16kHz

- STFT Window size: 320 samples (20ms), Overlap: 160 samples (10ms), 161-dimensional STFT spectrum

- Power compression [11]: compression coefficient η is set to 0.5 towards magnitude. Input feature:

$$Cat\left(|X|^{0.5}\cos\left(\theta_X\right), |X|^{0.5}\sin\left(\theta_X\right)\right)$$

- β_1=0.9， β_2=0.999 in Adam[12] with with the learning rate of 5e-4.

- 80 epochs for training, and the batch size is set to 4.

[11] 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
[12] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

Experiments and Analysis

IACAS

- Baselines:

  Magnitude domain baselines:

  - MMSE-GAN, MetriGAN, CRGAN, RDL-Net, MetriGAN+

  Time domain baselines:

  - SEGAN, SERGAN, MHSA-SPK, TSTNN, Demucs, SE-Conformer

  Complex domain baselines:

  - DCCRN, TGSA

  Decoupling-style baselines:

  - GAG-NET, PHASEN

- Evaluation metrics:
  - PESQ, STOI, segmental signal-to-noise ratio (SSNR)
  - The MOS prediction of speech distortion (CSIG), background noise (CBAK) and overall effect (COVL).[13]

[13] 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.

IACAS

Experimental Results

Table 1: Comparison with other state-of-the-art methods including time and T-F domain methods.

Methods	Year	Feature type	Param.	PESQ	STOI(%)	CSIG	CBAK	COVL	SSNR
Noisy	–	–	–	1.97	92.1	3.35	2.44	2.63	1.68
SOTA time and T-F Domain approaches									
SEGAN [24]	2017	Waveform	43.2 M	2.16	92.5	3.48	2.94	2.80	7.73
MMSEGAN [25]	2018	Gammatone	–	2.53	93.0	3.80	3.12	3.14	–
MetricGAN [26]	2019	Magnitude	1.86 M	2.86	–	3.99	3.18	3.42	–
CRGAN [27]	2020	Magnitude	–	2.92	94.0	4.16	3.24	3.54	–
DCCRN [8]	2020	RI components	3.7 M	2.68	93.7	3.88	3.18	3.27	8.62
RDL-Net [28]	2020	Magnitude	3.91 M	3.02	93.8	4.38	3.43	3.72	–
PHASEN [29]	2020	Magnitude+Phase	–	2.99	–	4.21	3.55	3.62	10.18
MHSA-SPK [30]	2020	Waveform	–	2.99	–	4.15	3.42	3.53	–
T-GSA [31]	2020	RI components	–	3.06	93.7	4.18	3.59	3.62	10.78
TSTNN [10]	2021	Waveform	0.92 M	2.96	95.0	4.17	3.53	3.49	9.70
DEMUCS [11]	2021	Waveform	128 M	3.07	95.0	4.31	3.40	3.63	–
GaGNet [13]	2021	Magnitude+RI	5.94 M	2.94	94.7	4.26	3.45	3.59	9.24
MetricGAN+ [32]	2021	Magnitude	–	3.15	–	4.14	3.16	3.64	–
SE-Conformer [33]	2021	Waveform	–	3.13	95.0	4.45	3.55	3.82	–
Proposed approaches									
MMB-AIAT	2021	Magnitude	0.90 M	3.11	94.9	4.45	3.60	3.79	9.74
CRB-AIAT	2021	RI components	1.17 M	3.15	94.7	4.48	3.54	3.81	8.81
DB-AIAT	2021	Magnitude+RI	2.81 M	**3.31**	**95.6**	**4.61**	**3.75**	**3.96**	**10.79**

- when only either single-path is adopted, MMB-AIAT and CRB-AIAT achieves competitive performance compared with advanced single-branch baselines.

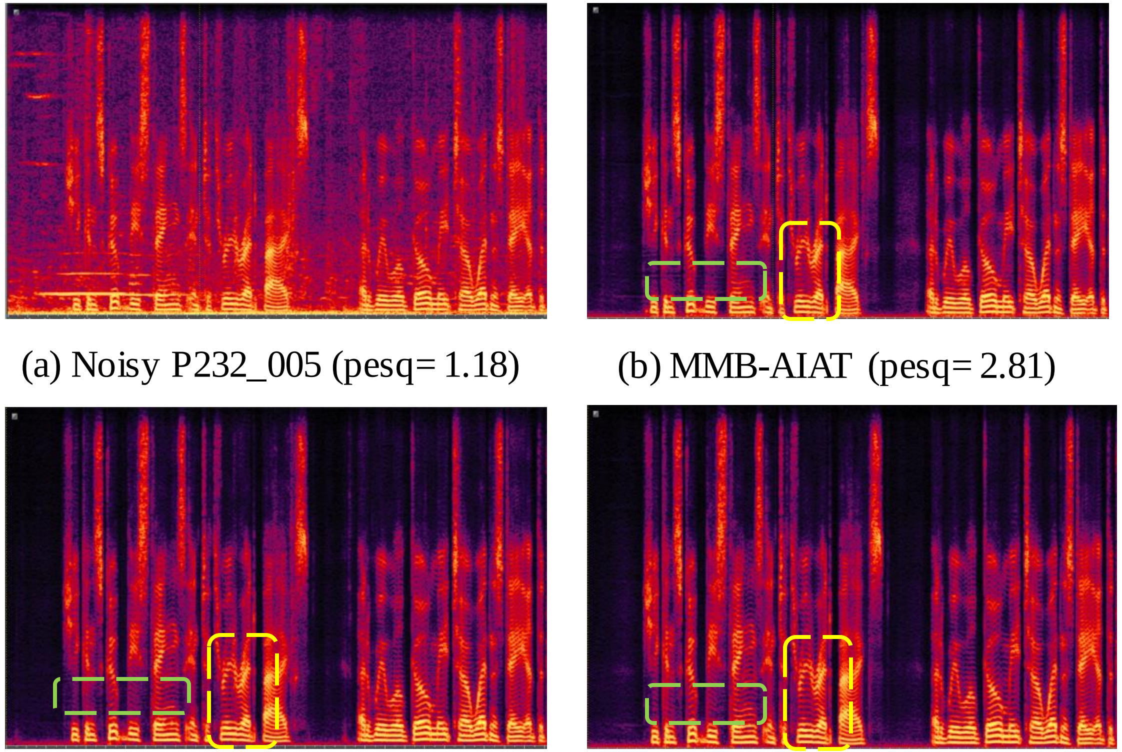
- By simultaneously adopting two branches in parallel, DB-AIAT consistently surpasses existing SOTA time and T-F domain methods in terms of most metrics.

Experimental Results

IACAS

Table 2： Ablation study on dual-branch strategy and attention-in-attention transformer structure.

Models	ATAB /AFAB	AHA	PESQ	STOI(%)	CSIG	CBAK	COVL
Unprocessed	–	–	1.97	92.1	3.35	2.44	2.63
Single-Branch approaches							
MMB-ATFAT	✓/✓	✗	3.05	94.6	4.37	3.53	3.71
MMB-AIAT	✓/✓	✓	3.11	94.9	4.45	3.60	3.79
CRB-ATFAT	✓/✓	✗	3.07	94.5	4.40	3.52	3.72
CRB-AIAT	✓/✓	✓	3.15	94.7	4.48	3.54	3.81
Dual-Branch approaches							
DB-ATAT	✓/✗	✗	2.82	94.2	4.17	3.29	3.47
DB-AFAT	✗/✓	✗	2.93	94.4	4.28	3.31	3.63
DB-ATFAT	✓/✓	✗	3.18	95.0	4.50	3.68	3.86
DB-AIAT	✓/✓	✓	3.31	95.6	4.61	3.75	3.96



(a) Noisy P232_005 (pesq= 1.18)

(b) MMB-AIAT (pesq= 2.81)

(c) CRB-AIAT (pesq= 2.85)

(d) DB-AIAT (pesq= 3.19)

Fig 7： Visualization of the spectrograms.

- The proposed attention-in-attention transformer significantlt improve speech quality.
- Merging two branches can collaboratively facilitate the spectrum recovery from the complementary perspective.

IACAS

OUTLINE

01 Introduction

02 Related works

03 Proposed Method

04 Experiments and Analysis

05 Conclusion

IACAS

Conclusion

- We propose a dual-branch transformer-based method to collaboratively recover the clean complex spectrum from the complementary perspective.
- A magnitude masking branch (MMB) is designed to coarsely estimate the magnitude spectrum of clean speech, and the residual spectral details are derived in parallel by a complex refining branch (CRB).
- We propose an attention-in-attention transformer (AIAT) to capture long-range temporal-frequency dependencies and aggregate global hierarchical contextual information
- Experimental results show that DB-AIAT yields state-of-the-art performance (3.31 PESQ, 95.6% STOI and 10.79dB SSNR) over previous advanced systems with a relatively small model size (2.81M).