# 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], Hui Wang[1]

[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 dual-branch attention-in-attention transformer dubbed DB-AIAT to handle both coarse- and fine-grained regions of the spectrum in parallel.
- From a complementary perspective, a magnitude masking branch is proposed to coarsely estimate the overall magnitude spectrum, and simultaneously a complex refining branch is designed to compensate for the missing spectral details.
- Within each branch, we propose a novel attention-in-attention transformer-based module to replace the conventional RNNs and temporal convolutional networks for temporal sequence modeling.
- Experimental results on Voice Bank + DEMAND demonstrate that DB-AIAT yields state-of-the-art performance (*e.g.*, 3.31 PESQ, 95.6% STOI and 10.79dB SSNR) over previous advanced systems with a relatively small model size (2.81M).


## Introduction

**Decoupling-style phase-aware speech enhancement:**

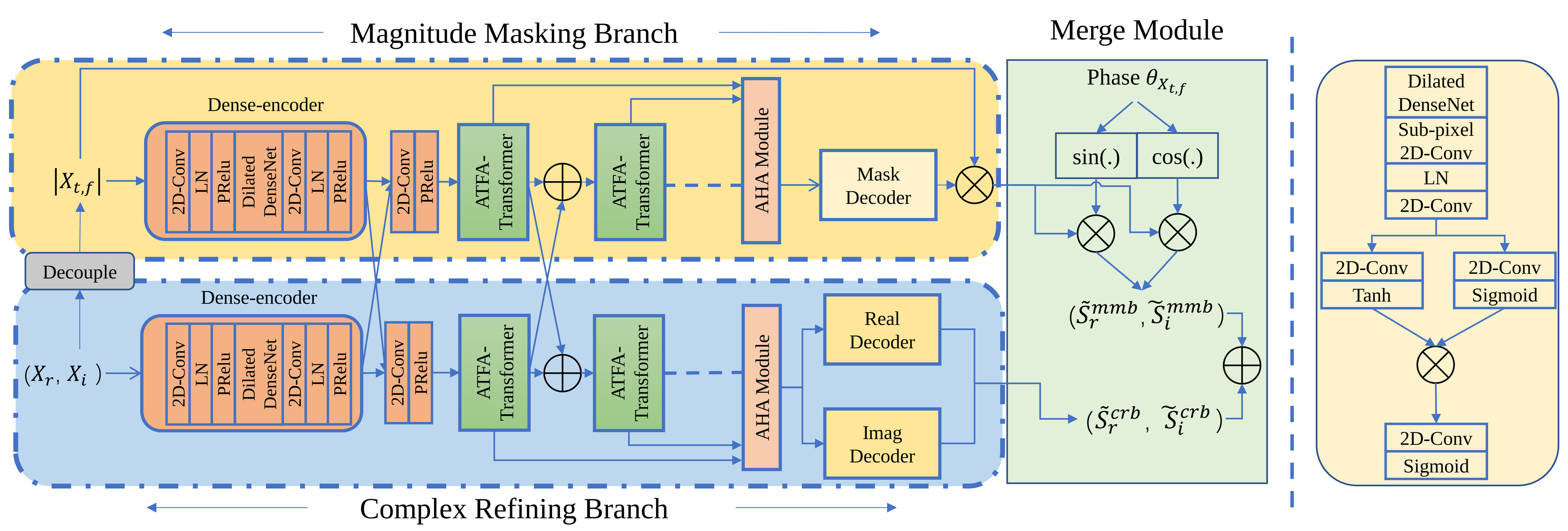
- The importance of phase has been illustrated in improving the speech perceptual quality, especially under low SNR conditions.
- Decoupling-style phase-aware methods decouple the original complex spectrum estimation into magnitude and phase stage by stage, and alleviate the implicit compensation effect between two targets.

**Transformer-based speech sequence modeling:**

- Convolutional recurrent networks (CRNs) and temporal convolutional networks (TCNs) still lack sufficient capacity to capture the global contextual information.
- In the speech separation and enhancement task, dual-path transformer-based networks are employed for extracting contextual information along both the time and frequency axes.

## Dual-branch Attention-in-Attention Transformer

**Figure 1: The diagram of the proposed DB-AIAT. (a) The overall diagram of the proposed system.**




(a) Overall diagram (b) Mask Decoder

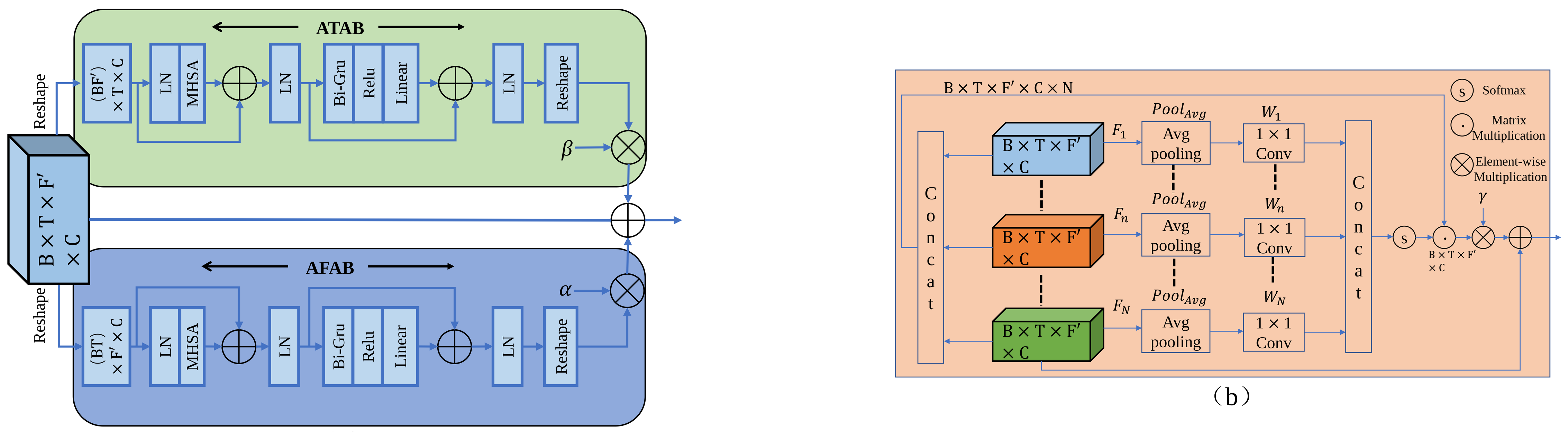
- As Figure. 1(a) and (b) show, MMB path estimates the magnitude mask to coarsely recover the magnitude of the target speech, and the coarsely estimated spectral magnitude is coupled with the noisy phase, while CRB path receives noisy real and imaginary (RI) components as the input and focuses on the residual fine-grained spectral structures.

$$|\widetilde{S}^{mmb}| = |X_{t,f}| \otimes M^{mmb}, \quad (1)$$

$$\widetilde{S}_r^{mmb} = |\widetilde{S}^{mmb}| \otimes \cos(\theta_X), \widetilde{S}_i^{mmb} = |\widetilde{S}^{mmb}| \otimes \sin(\theta_X), \quad (2)$$

$$\widetilde{S}_r = \widetilde{S}_r^{mmb} + \widetilde{S}_r^{crb}, \widetilde{S}_i = \widetilde{S}_i^{mmb} + \widetilde{S}_i^{crb} \quad (3)$$

**Figure 2: (a) The diagram of ATFAT blocks. (b) The diagram of the AHA module.**




(a) (b)

## 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 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 20ms is applied, with 50% overlap between adjacent frames. The 320-point STFT is utilized, leading to a 161-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, *i.e.*, $Cat\left(|X|^{0.5}\cos(\theta_X), |X|^{0.5}\sin(\theta_X)\right)$ as input, $Cat\left(|S|^{0.5}\cos(\theta_S), |S|^{0.5}\sin(\theta_S)\right)$ as target.
- Adam optimizer is utilized with the learning rate of 5e-4. 80 epochs are conducted for training in total, and the batch size is set to 4 at the utterance level.

## Comparison results & analysis

**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 | 2017 | Waveform | 43.2 M | 2.16 | 92.5 | 3.48 | 2.94 | 2.80 | 7.73 |
| MMSEGAN | 2018 | Gammatone | – | 2.53 | 93.0 | 3.80 | 3.12 | 3.14 | – |
| MetricGAN | 2019 | Magnitude | 1.86 M | 2.86 | – | 3.99 | 3.18 | 3.42 | – |
| CRGAN | 2020 | Magnitude | – | 2.92 | 94.0 | 4.16 | 3.24 | 3.54 | – |
| DCCRN | 2020 | RI components | 3.7 M | 2.68 | 93.7 | 3.88 | 3.18 | 3.27 | 8.62 |
| RDL-Net | 2020 | Magnitude | 3.91 M | 3.02 | 93.8 | 4.38 | 3.43 | 3.72 | – |
| PHASEN | 2020 | Magnitude+Phase | – | 2.99 | – | 4.21 | 3.55 | 3.62 | 10.18 |
| MHSA-SPK | 2020 | Waveform | – | 2.99 | – | 4.15 | 3.42 | 3.53 | – |
| T-GSA | 2020 | RI components | – | 3.06 | 93.7 | 4.18 | 3.59 | 3.62 | 10.78 |
| TSTNN | 2021 | Waveform | 0.92 M | 2.96 | 95.0 | 4.17 | 3.53 | 3.49 | 9.70 |
| DEMUCS | 2021 | Waveform | 128 M | 3.07 | 95.0 | 4.31 | 3.40 | 3.63 | – |
| GaGNet | 2021 | Magnitude+RI | 5.94 M | 2.94 | 94.7 | 4.26 | 3.45 | 3.59 | 9.24 |
| MetricGAN+ | 2021 | Magnitude | – | 3.15 | – | 4.14 | 3.16 | 3.64 | – |
| SE-Conformer | 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** |

**Table 2: Ablation study *w.r.t.* 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 |

## Conclusions

- This paper propose a dual-branch transformer-based framework to collaboratively facilitate the clean spectrum estimation from the complementary perspective.
- A magnitude masking branch (MMB) is designed to coarsely filter out the dominant noise components in the magnitude domain, while the residual spectral details are derived by a complex refining branch (CRB) in parallel.
- Experimental results on Voice Bank + DEMAND dataset show that DB-AIAT achieves remarkable results and consistently outperforms state-of-the-art baselines with a relatively light model size.