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

Guochen Yu\textsuperscript{1,2}, Andong Li\textsuperscript{2}, Chengshi Zheng\textsuperscript{2}, Yinuo Guo\textsuperscript{3}, Yutian Wang\textsuperscript{1},

and Hui Wang\textsuperscript{1}

\textsuperscript{1}State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing, China

\textsuperscript{2}Key Laboratory of Noise and Vibration Research, Institute of Acoustics, Chinese Academy of Sciences, Beijing, China

\textsuperscript{3}Bytedance, Beijing, China
OUTLINE

01 Background

02 Related works

03 Proposed Method

04 Experiments and Analysis

05 Conclusion
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

Background

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

Complex spectrum based SE:

\[ Y_{m,f}^{(r)} + iY_{m,f}^{(i)} = \left( S_{m,f}^{(r)} + N_{m,f}^{(r)} \right) + i \left( S_{m,f}^{(i)} + N_{m,f}^{(i)} \right), \]

1) complex ration mask (CRM) [5]

\[ CRM = \frac{X_r S_r + X_i S_i}{X_r^2 + X_i^2} + j \frac{X_r S_i - X_i S_r}{X_r^2 + X_i^2} = \tilde{M}_r + j\tilde{M}_i \]

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

01 Introduction

02 Related works

03 Proposed Method

04 Experiments and Analysis

05 Conclusion
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].

\[ e^{j\theta_X} \]
\[ \{X_r, X_i\} \]
\[ \text{Decouple Layer} \]
\[ X \]
\[ e^{j\theta_X} \]
\[ \text{ME-Net} \]
\[ \tilde{S}_m \]
\[ \text{Global Residual Connection} \]
\[ \text{CS-Net} \]
\[ \{\tilde{S}_r, \tilde{S}_i\} \]

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)

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

Proposed Method

IACAS

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

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.
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:

\[
\begin{align*}
|\tilde{S}_{\text{mmb}}| &= |X_{t,f}| \otimes M_{\text{mmb}}^\text{mmb} \\
\tilde{S}_r^\text{mmb} &= |\tilde{S}_{\text{mmb}}| \otimes \cos(\theta_X) \\
\tilde{S}_i^\text{mmb} &= |\tilde{S}_{\text{mmb}}| \otimes \sin(\theta_X) \\
\tilde{S}_r &= \tilde{S}_r^\text{mmb} + \tilde{S}_r^\text{crb} \\
\tilde{S}_i &= \tilde{S}_i^\text{mmb} + \tilde{S}_i^\text{crb}
\end{align*}
\]
**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.

**Fig 5:** The diagram of ATFA-T blocks

**Fig 6:** The diagram of AHA module
Proposed Method

The loss function of the proposed dual-branch model:

\[
L^{\text{Mag}} = \left\| \sqrt{|\tilde{S}_r|^2 + |\tilde{S}_i|^2} - \sqrt{|S_r|^2 + |S_i|^2} \right\|_F^2
\]  
(6)

\[
L^{\text{RI}} = \left\| \tilde{S}_r - S_r \right\|_F^2 + \left\| \tilde{S}_i - S_i \right\|_F^2
\]  
(7)

\[
L_{\text{FULL}} = L^{\text{Mag}} + L^{\text{RI}}
\]  
(8)
01 Introduction
02 Related works
03 Proposed Method
04 Experiments and Analysis
05 Conclusion
Experiments and Analysis

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.


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 $\eta$ is set to 0.5 towards magnitude. Input feature:

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

- $\beta_1=0.9, \ \beta_2=0.999$ in Adam[12] with the learning rate of 5e-4.

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

---


Experiments and Analysis

• 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]

Experimental Results

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

<table>
<thead>
<tr>
<th>Methods</th>
<th>Year</th>
<th>Feature type</th>
<th>Param.</th>
<th>PESQ</th>
<th>STOI(%)</th>
<th>CSIG</th>
<th>CBAK</th>
<th>COVL</th>
<th>SSNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noisy</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>1.97</td>
<td>92.1</td>
<td>3.35</td>
<td>2.44</td>
<td>2.63</td>
<td>1.68</td>
</tr>
</tbody>
</table>

**SOTA time and T-F Domain approaches**

<table>
<thead>
<tr>
<th>Methods</th>
<th>Year</th>
<th>Feature type</th>
<th>Param.</th>
<th>PESQ</th>
<th>STOI(%)</th>
<th>CSIG</th>
<th>CBAK</th>
<th>COVL</th>
<th>SSNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEGAN [24]</td>
<td>2017</td>
<td>Waveform</td>
<td>43.2 M</td>
<td>2.16</td>
<td>92.5</td>
<td>3.48</td>
<td>2.94</td>
<td>2.80</td>
<td>7.73</td>
</tr>
<tr>
<td>CRGAN [27]</td>
<td>2020</td>
<td>Magnitude</td>
<td>–</td>
<td>2.92</td>
<td>94.0</td>
<td>4.16</td>
<td>3.24</td>
<td>3.54</td>
<td>–</td>
</tr>
<tr>
<td>DCCRN [8]</td>
<td>2020</td>
<td>RI components</td>
<td>3.7 M</td>
<td>2.68</td>
<td>93.7</td>
<td>3.88</td>
<td>3.18</td>
<td>3.27</td>
<td>8.62</td>
</tr>
<tr>
<td>RDL-Net [28]</td>
<td>2020</td>
<td>Magnitude</td>
<td>3.91 M</td>
<td>3.02</td>
<td>93.8</td>
<td>4.38</td>
<td>3.34</td>
<td>3.72</td>
<td>–</td>
</tr>
<tr>
<td>T-GSA [31]</td>
<td>2020</td>
<td>RI components</td>
<td>–</td>
<td>3.06</td>
<td>93.7</td>
<td>4.18</td>
<td>3.59</td>
<td>3.62</td>
<td>10.78</td>
</tr>
<tr>
<td>TSTNN [10]</td>
<td>2021</td>
<td>Waveform</td>
<td>0.92 M</td>
<td>2.96</td>
<td>95.0</td>
<td>4.17</td>
<td>3.53</td>
<td>3.49</td>
<td>9.70</td>
</tr>
<tr>
<td>SE-Conformer [33]</td>
<td>2021</td>
<td>Waveform</td>
<td>3.13 M</td>
<td>3.13</td>
<td>95.0</td>
<td>4.45</td>
<td>3.55</td>
<td>3.82</td>
<td>–</td>
</tr>
</tbody>
</table>

**Proposed approaches**

<table>
<thead>
<tr>
<th>Methods</th>
<th>Year</th>
<th>Feature type</th>
<th>Param.</th>
<th>PESQ</th>
<th>STOI(%)</th>
<th>CSIG</th>
<th>CBAK</th>
<th>COVL</th>
<th>SSNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMB-AIAT</td>
<td>2021</td>
<td>Magnitude</td>
<td>0.90 M</td>
<td>3.11</td>
<td>94.9</td>
<td>4.45</td>
<td>3.60</td>
<td>3.79</td>
<td>9.74</td>
</tr>
<tr>
<td>CRB-AIAT</td>
<td>2021</td>
<td>RI components</td>
<td>1.17 M</td>
<td>3.15</td>
<td>94.7</td>
<td>4.48</td>
<td>3.54</td>
<td>3.81</td>
<td>8.81</td>
</tr>
<tr>
<td>DB-AIAT</td>
<td>2021</td>
<td>Magnitude+RI</td>
<td>2.81 M</td>
<td><strong>3.31</strong></td>
<td><strong>95.6</strong></td>
<td><strong>4.61</strong></td>
<td><strong>3.75</strong></td>
<td><strong>3.96</strong></td>
<td><strong>10.79</strong></td>
</tr>
</tbody>
</table>

- 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**

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

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

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

Fig 7: Visualization of the spectrograms.

(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)
01 Introduction
02 Related works
03 Proposed Method
04 Experiments and Analysis
05 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).