DUAL-BRANCH ATTENTION-IN-ATTENTION TRANSFORMER FOR SINGLE-CHANNEL SPEECH ENHANCEMENT

Guochen Yu†, Andong Li†, Chengshi Zheng†, Yinuo Guo†, Yutian Wang†, Hui Wang⋆

⋆ State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing, China
† Institute of Acoustics, Chinese Academy of Sciences, Beijing, China
† Bytedance, Beijing, China
{yuguochen, wangyutian, hwang}@cuc.edu.cn, {liandong, cszheng}@mail.ioa.ac.cn

ABSTRACT
Curriculum learning begins to thrive in the speech enhancement area, which decouples the original spectrum estimation task into multiple easier sub-tasks to achieve better performance. Motivated by that, we propose a 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 elaborately designed to compensate for the missing spectral details and implicitly derive phase information. 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. Specifically, the proposed attention-in-attention transformer consists of adaptive temporal-frequency attention transformer blocks and an adaptive hierarchical attention module, aiming to capture long-term temporal-frequency dependencies and further aggregate global hierarchical contextual information. 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).

Index Terms— Speech enhancement, dual-branch, attention-in-attention, transformer

1. INTRODUCTION
In real acoustic scenarios, various types of environmental interference may severely degrade the performance of telecommunication and hearing aids. Monaural speech enhancement (SE) technique aims at recovering clean speech from its noise-corrupted mixture to improve speech quality and/or intelligibility when only one channel recording is available [1]. Recently, deep neural networks (DNNs) have ignited the development of SE algorithms for their more powerful capability in dealing with non-stationary noise than conventional statistical signal-processing based approaches [2].

In a typical supervised SE paradigm, DNNs are usually leveraged to estimate the mask functions or directly predict the magnitude spectra of clean speech in the time-frequency (TF) domain [3, 4], where the noisy phase is unchanged and involved in waveform reconstruction. Recently, the importance of phase has been fully illustrated in improving the speech perceptual quality, especially under low signal-to-noise ratio (SNR) conditions [5]. In this regard, more recent approaches attempt to recover the phase information either explicitly or implicitly [6, 7, 8, 9, 10, 11]. For the first class, the network is employed to estimate the complex ratio masks (CRMs) or the real and imaginary (RI) spectra, which facilitate both magnitude and phase information recovery simultaneously in the TF domain. For the latter, the time-domain waveform is directly regenerated, which diverts around the phase estimation process. More recently, decoupling-style phase-aware methods are proposed, where the original complex-spectrum estimation problem is decomposed into two sub-stages [2, 12, 13]. Specifically, only the magnitude estimation is involved in the first stage, followed by the spectrum refinement with residual learning in the later stage. In this way, the optimization w.r.t. magnitude and phase can be effectively decoupled, which alleviates the implicit compensation effect between two targets [14].

In this paper, we propose a dual-branch SE structure dubbed DB-AIAT, to explore the complex spectrum recovery from the complementary perspective. Specifically, two core branches are elaborately designed in parallel, namely a magnitude masking branch (MMB) and a complex refining branch (CRB). In MMB, we seek to construct the filtering system which only applies to the magnitude domain. In this branch, most of the noise can be effectively suppressed. In CRB, it is established as the decorating system to compensate for the lost spectral details and phase mismatch effect. Two branches work collaboratively to facilitate the overall spectrum recovery. Besides, despite temporal convolutional networks (TCNs) [15] and LSTM layers are widely adopted for long-range sequence modeling in the SE area, they still lack sufficient capacity to capture the global context information [10, 16]. In addition, they usually only apply in the time axis, which neglects the correlations among different frequency sub-bands. To this end, we propose an attention-in-attention transformer (AIAT) to funnel the global sequence modeling process, which captures long-range dependencies along both time and frequency axes, and meanwhile, aggregates global hierarchical contextual information. 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.

The remainder of the paper is organized as follows. In Section2, the proposed framework is described in detail. The experimental setup is presented in Section3 while Section4 gives the results and analysis. Finally, some conclusions are drawn in Section5.

2. METHODOLOGY

2.1. Dual-branch magnitude masking and complex refining
The overall diagram of the proposed system is illustrated in Fig.1. It is mainly comprised of two branches, namely a magnitude spectrum...
masking branch (MMB) and a complex spectrum refining branch (CRB), which aim at collaboratively estimating the magnitude and phase information of clean speech in parallel. To be specific, in the MMB path, the input is the magnitude of the noisy spectrum, and the network estimates the magnitude mask $M^{mmb}$ to coarsely recover the magnitude of the target speech, i.e., $|\hat{S}^{mmb}|$. Subsequently, the coarsely estimated spectral magnitude is coupled with the noisy phase to derive the coarse-denoised complex spectrum.

As the complement, CRB receives noisy RI component $\{X_r, X_i\}$ as the input and focuses on the fine-grained spectral structures which may be lost in the MMB path and further suppressing the residual noise components. Note that we only estimate the residual details instead of explicitly estimating the whole complex spectrum, which alleviates the overall burden of the network. The alternate interconnections are adopted to exchange information between the two branches, enabling better feature representation. Finally, we sum the coarse-denoised complex spectrum and the fine-grained complex spectral details together to reconstruct the clean complex spectrum. In a nutshell, the whole procedure can be formulated as:

$$|\hat{S}^{mmb}| = |X_{t,f}| \otimes M^{mmb},$$

$$\hat{S}_{mmb}^{r} = |S^{mmb}| \otimes \cos(\theta_X),$$

$$\hat{S}_{mmb}^{i} = |S^{mmb}| \otimes \sin(\theta_X),$$

$$\hat{S}_{r} = \hat{S}_{mmb}^{r} + \hat{S}_{r}^{\text{crb}},$$

$$\hat{S}_{i} = \hat{S}_{mmb}^{i} + \hat{S}_{i}^{\text{crb}},$$

where $\{\hat{S}_{r}^{\text{crb}}, \hat{S}_{i}^{\text{crb}}\}$ denote the output RI components of CRB and $\{\hat{S}_{r}, \hat{S}_{i}\}$ denote the final merged clean RI components. Tilde denotes the estimated variable. $\theta_X$ denotes the noisy phase and $\otimes$ is the element-wise multiplication operator. The input features of MMB and CRB are denoted as $|X_{t,f}| \in \mathbb{R}^{T \times F \times 1}$ and $\text{C}_\text{outer}(X_r, X_i) \in \mathbb{R}^{T \times F \times 2}$, respectively. Here $T$ is the number of frames and $F$ is the number of frequency bins. As shown in Fig. 1, MMB consists of a densely-connected convolutional encoder, an AIAT and a mask decoder. Analogously, CRB is composed of a dense-encoder, an AIAT and two complex decoders.

### 2.2. Densely convolutional encoder

The dense-encoder in each branch is composed of two 2-D convolutional layers, followed by layer normalization (LN) and parametric ReLU (PReLU) activation. A densenet with four dilated convolutional layers is employed between the above 2-D convolutional layers, in which the dilation rates are $\{1, 2, 4, 8\}$. The output channel of the first 2-D convolutional layer is set to 64 and keeps unaltered, with kernel size and stride being $(1, 1)$, while the second 2-D convolutional layer halves the dimension of the frequency axis, with kernel size and stride being $(1, 3)$ and $(1, 2)$, respectively.

### 2.3. Attention-in-attention transformer

The proposed AIAT module consists of four adaptive time-frequency attention transformer-based (ATFAT) modules and an adaptive hierarchical attention (AHA) module, as illustrated in Fig. 2. Each ATFAT can strengthen the long-range temporal-frequency dependencies with relatively low computational cost and the AHA module can aggregate different intermediate features to capture multi-scale contextual information by adaptive weights, as pointed out in [17]. Before feeding the compressed output into AIAT, we concatenate the output of two dense-encoders and a 2-D convolution layer with kernel size $(1, 1)$ is used to halve the number of channels (i.e., from $128$...
to 64), followed by PreLU activation. As shown in Fig. 2(a), ATFAT is composed of an adaptive temporal attention branch (ATAB) and an adaptive frequency attention branch (AFAB), which can capture long-term dependencies along temporal and frequency axes with two adaptive weights α and β. In each branch, an improved transformer [10] is employed, which is comprised of a multi-head self-attention (MHSA) module and a GRU-based feed-forward network, followed by residual connections and LN. The feed-forward network employs a bi-directional GRU (Bi-GRU) to replace the first fully connected layer in the traditional transformer and yields better performance [19]. After each ATFAT module, we employ PreLU and a 2-D convolution with kernel size (1, 1) to retain the channel number.

The AHA module aims at integrating different hierarchical feature maps given all the ATFAT modules’ outputs $F = \{F_n\}_{n=1}^N$, $F_n \in \mathbb{R}^{B \times T \times F' \times C}$, where $B$ denotes the batch size, $F'$ denotes the halved frequency dimension and $N$ is the number of ATFAT (set $N=4$ in this paper). We first employ an average pooling layer $Pool_{Avg}$ and a $1 \times 1$ convolutional layer $W_n$ to squeeze the output feature of each ATFAT into a global representation: $P_n = Pool_{Avg}(F_n) + W_n \in \mathbb{R}^{B \times 1 \times 1 \times N \times 1}$. Then we concatenate all the pooled outputs as $P = \{P_n\} \in \mathbb{R}^{B \times 1 \times 1 \times N \times 1}$, which is subsequently fed into a softmax function to derive the hierarchical attention map $W^h = \mathbb{R}^{B \times 1 \times 1 \times N \times 1}$. After that we concatenate all the intermediate outputs of ATFAT (i.e., $\{F_n\}_{n=1}^N$) to obtain a global feature map $F^h \in \mathbb{R}^{B \times T \times F' \times C \times N}$. Subsequently, we obtain the multi-scale contextual information by performing a matrix multiplication between $F^h$ and $W^h$, which can be calculated as:

$$G_N = W^h F^h = \sum_{n=1}^{N} \frac{\exp(Pool_{Avg}(F_n) + W_n)}{\sum_{n=1}^{N} \exp(Pool_{Avg}(F_n) + W_n)} F_n,$$  \hspace{1cm} (6)

where $F_n$ denotes the $n^{th}$ intermediate output of ATFAT and $G_N \in \mathbb{R}^{B \times T \times F' \times C}$ denotes the global contextual feature map. Finally, we perform an element-wise sum operation between the last ATFAT module’s output feature map $F_N$ and the global contextual feature map $G_N$ to obtain the final output, i.e., $Out_{AHA} \in \mathbb{R}^{T \times F' \times C}$:

$$Out_{AHA} = F_N + \gamma G_N,$$  \hspace{1cm} (7)

where γ is a learnable scalar coefficient and initialized to 0. This adaptive learning weight automatically learns to assign a suitable value to merge global contextual information.

2.4. Mask/Complex decoder

The mask decoder consists of a dilated dense block the same as in dense-encoder, a sub-pixel 2-D convolution module, and a dual-path mask module. The sub-pixel convolution layer is used to upsample both RI components in parallel, which are also composed of a dilated dense block and a sub-pixel 2-D convolution. The sub-pixel 2-D convolution in the mask/complex decoders sets the upsampling factor to 2, with kernel size set to (1, 3).

2.5. Loss function

The loss function of the proposed two-branch model is calculated by the RI components and the magnitude of the estimated spectrum, which can be expressed as:

$$\mathcal{L}^{Mag} = \left\| \sqrt{|S_i|^2} + \sqrt{|\tilde{S}_i|^2} - \sqrt{|S_i|^2 + |\tilde{S}_i|^2} \right\|_F^2,$$  \hspace{1cm} (8)

$$\mathcal{L}^{RI} = \left\| S_i - S_i' \right\|_F^2 + \left\| \tilde{S}_i - \tilde{S}_i' \right\|_F^2,$$  \hspace{1cm} (9)

$$\mathcal{L}_{Full} = \mu \mathcal{L}^{RI} + (1 - \mu) \mathcal{L}^{Mag},$$  \hspace{1cm} (10)

where $\mathcal{L}^{Mag}$ and $\mathcal{L}^{RI}$ denote the loss function toward magnitude and RI, respectively. Here, $\{S_i, \tilde{S}_i\}$ represent the RI components of the estimated speech spectrum, while $\{S_i, \tilde{S}_i\}$ represent the target RI components of the clean speech spectrum. With the internal trial, we empirically set $\mu = 0.5$ in all the following experiments.

3. EXPERIMENTS

3.1. Datasets

The dataset used in this work is publicly available as proposed in [19], which is a selection of Voice Bank corpus [20] with 28 speakers for training and another 2 unseen speakers for testing. The training set includes 11,572 noisy-clean pairs, while the test set contains 824 pairs. For training, the audio samples are mixed with one of the 10 noise types, i.e., two artificial (babble and speech shaped) and eight real recording noise processes taken from the DEMAND database [21] at four SNRs, i.e., 0dB, 5dB, 10dB, 15dB. The test utterances are created with 5 unseen noise taken from the Demand database at SNRs of {2.5dB, 7.5dB, 12.5dB, 17.5dB}.

3.2. Implementation setup

All the utterances are resampled at 16 kHz and chunked to 3 seconds. The Hanning window of length 20 ms is selected, with 50% overlap between consecutive frames. The 320-point STFT is utilized and 161-dimension spectral features are obtained. Due to the efficacy of the compressed spectrum in dereverberation and denoising task [12], we conduct the power compression toward the magnitude while remaining the phase unaltered, and the optimal compression coefficient is set to 0.5, i.e., $Cat(|X|^0.5 \cos(\theta_x), |X|^0.5 \sin(\theta_x))$ as input, $Cat(|s|^0.5 \cos(\theta_s), |s|^0.5 \sin(\theta_s))$ as target. All the models are optimized using Adam [23], with the learning rate of 5e-4. 80 epochs are conducted for network training in total, and the batch size is set to 4 at the utterance level. The processed samples are available online, where the source code is also released.

4. RESULTS AND ANALYSIS

We use the following objective metrics to evaluate speech enhancement performance: the perceptual evaluation of speech quality (PESQ) [34], short-time objective intelligibility (STOI) [35], segmental signal-to-noise ratio (SSNR), the mean opinion score (MOS) prediction of the speech signal distortion (CSIG) [46], the MOS prediction of the intrusiveness of background noise (CBAK) and the MOS prediction of the overall effect (COVL) [56] to evaluate speech enhancement performance. Higher values of all metrics indicate better performance.

4.1. Comparison with previous advanced baselines

We first compare the objective performance of the proposed methods with other state-of-the-art (SOTA) baselines, whose results are presented in Table 1. The baselines include five time-domain methods (e.g., SEGAN [24], MHSA-SPK [30], TSTN [10] and DEMUCS [11]) and nine T-F domain methods (e.g., CRGAN [27], PHASEN [29], T-GSA [31], GaGNet [13] and MetricGAN+ [32]).
We then investigate the effects of the proposed dual-branch strategy and AIAT structure, as shown in Table 2. From the results, one can have the following observations. First, when adopting the single-branch topology, i.e., MMB-AIAT and CRB-AIAT, we can find that CRB-AIAT yields better performance in PESQ, STOI, CBAK and COVL than MMB-AIAT, while MMB-AIAT achieves a higher score in CBAK. This indicates that MMB can better eliminate noise and provide higher STOI scores, while CRB conducts better speech over-all quality. Second, when both CRB and MMB are employed in parallel, DB-AIAT yields significant improvements in terms of all metrics than the single-branch methods. This verifies that merging two branches can collaboratively facilitate the spectrum recovery from the complementary perspective. Third, we investigate the effectiveness of our proposed attention-in-attention transformer (AIAT) model. When combining ATAB and AFAB to capture both temporal-frequency dependency, DB-AIAT yields significant improvements in terms of all metrics than the single-branch methods. This verifies that merging two branches can collaboratively facilitate the spectrum recovery from the complementary perspective.
6. REFERENCES

[1] P. C. Loizou, *Speech enhancement: theory and practice*, CRC press, 2013.
[2] 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.
[3] 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.
[4] Y. Xu, J. Du, L.-R. D. 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.
[5] K. Paliwal, K. Wójcicki, and B. Shannon, “The importance of phase in speech enhancement,” *Speech Commun.*, vol. 53, no. 4, pp. 465–494, 2011.
[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.
[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.
[8] Y. Hu, Y. Liu, S. Lv, M. Xing, and L. Xie, “Decrm: Deep complex convolution recurrent network for phase-aware speech enhancement,” in *Proc. Interspeech*, 2020, pp. 2472–2476.
[9] H. S. Choi, J. H. Kim, J. Huh, A. Kim, J. W. Ha, and K. Lee, “Phase-aware Speech Enhancement with Deep Complex U-Net,” *arXiv preprint arXiv:1903.03107*, 2019.
[10] K. Wang, B. He, and W. P. Zhu, “TSTNN: Two-Stage Transformer Based Neural Network for Speech Enhancement in the Time Domain,” in *Proc. ICASSP*. IEEE, 2021, pp. 7098–7102.
[11] A. Defossez, G. Synnaeve, and Y. Adi, “Real time speech enhancement in the waveform domain,” *arXiv preprint arXiv:2006.12847*, 2020.
[12] A. Li, W. Liu, X. Luo, G. Yu, C. Zheng, and X. Li, “A simultaneous denoising and dereverberation framework with target decoupling,” in *Proc. Interspeech*, 2021, pp. 2801–2805.
[13] A. Li, C. Zheng, L. Zhang, and X. Li, “Glance and Gaze: A Collaborative Learning Framework for Single-channel Speech Enhancement,” *arXiv preprint arXiv:2106.11789*, 2021.
[14] Z.-Q. Wang, G. Wichern, and J. Le Roux, “On The Compensation Between Magnitude and Phase in Speech Separation,” *arXiv preprint arXiv:2108.05470*, 2021.
[15] A. Pandey and D. Wang, “TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain,” in *Proc. ICASSP*. IEEE, 2019, pp. 6875–6879.
[16] J. Chen, Q. Mao, and D. Liu, “Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation,” in *Proc. Interspeech*, 2020, pp. 2642–2646.
[17] G. Yu, Y. Wang, C. Zheng, H. Wang, and Q. Zhang, “Cyrclegan-based non-parallel speech enhancement with an adaptive attention-in-attention mechanism,” in *Proc. Asia-Pacific Signal Inf. Process. Assoc. (APSIPA)*, 2021, pp. 523–529.
[18] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, D. Bishop, R. and Rueckert, and Z. Wang, “Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network,” in *Proc. CVPR*, 2016, pp. 1874–1883.