DPT-FSNET: DUAL-PATH TRANSFORMER BASED FULL-BAND AND SUB-BAND FUSION NETWORK FOR SPEECH ENHANCEMENT

Feng Dang*† Hangting Chen*† Pengyuan Zhang*†

* Key Laboratory of Speech Acoustics & Content Understanding, Institute of Acoustics, CAS, China
† University of Chinese Academy of Sciences, Beijing, China

ABSTRACT

Sub-band models have achieved promising results due to their ability to model local patterns in the spectrogram. Some studies further improve the performance by fusing sub-band and full-band information. However, the structure for the full-band and sub-band fusion model was not fully explored. This paper proposes a dual-path transformer-based full-band and sub-band fusion network (DPT-FSNet) for speech enhancement in the frequency domain. The intra and inter parts of the dual-path transformer model sub-band and full-band information, respectively. The features utilized by our proposed method are more interpretable than those utilized by the time-domain dual-path transformer. We conducted experiments on the Voice Bank + DEMAND and Interspeech 2020 Deep Noise Suppression (DNS) datasets to evaluate the proposed method. Experimental results show that the proposed method outperforms the current state-of-the-art.

Index Terms— speech enhancement, frequency domain, dual-path transformer, full-band and sub-band fusion

1. INTRODUCTION

Speech enhancement (SE) is a speech processing method that aims to improve the quality and intelligibility of noisy speech by removing noise [1]. It is commonly used as a front-end task for automatic speech recognition, hearing aids, and telecommunications. In recent years, the application of deep neural networks (DNNs) in SE research has received increasing interest.

In general, DNN-based methods can be divided into two major categories: time-domain methods [2, 3, 4] and time-frequency domain (T-F) methods [5, 6, 7, 8]. Time-domain methods estimate clean waveforms directly from the noisy raw data in the time domain. Traditional T-F domain methods usually transform the noisy input waveform into a Fourier magnitude spectrum by short-time Fourier transform, modify the spectrum by T-F mask, and reconstruct the enhanced spectrum into enhanced waveform by inverse short-time Fourier transform. They usually use the phase of the noisy mixture, which limits the upper bound of the denoising performance. Recent T-F domain methods using complex spectra as features can preserve phase information and have achieved promising performance [8].

Sub-band processing is a common method in audio processing [9, 10, 11], which takes sub-band spectral features as input and output. Previous work [12] pointed out that the local patterns in the spectrum tend to be different in each frequency band. The sub-band model handles each frequency independently, which allows the sub-band model to focus on the local patterns in the spectrum and therefore achieve good results in SE tasks. In [13] further improves the performance by fusing sub-band and full-band information.

Recently, dual-path networks [13, 14, 15, 16] have achieved exceptional performance due to their ability to model local and global features of the input sequence. Some studies [14, 15] have introduced transformer structures [17] into dual-path networks, where input elements can interact directly based on self-attention mechanism, to further improve the performance of dual-path networks. However, these studies were based on simple time-domain features and did not further investigate the effect of the input of the dual-path network on the enhancement performance. In [16] they tried to change the structure of encoder and decoder to extract more effective inputs for the dual-path network, but still limited to time-domain features.

Inspired by the above problems, we propose a dual-path transformer based full-band and sub-band fusion network (DPT-FSNet) for speech enhancement. Specifically, our proposed model consists of an encoder, decoder, and dual-path transformer. We utilize a convolutional encoder-decoder (CED) structure to extract an efficient latent feature space for the dual-path transformer. Both the encoder and decoder consist of a 1x1 convolutional layer and a dense block [18], where dilated convolutions are utilized inside the dense block for context aggregation. The dual-path transformer is composed of two parts, intra-transformers and inter-transformers. The intra-transformers model sub-band information and the inter-transformer merges the sub-band information from the intra-transformer to model the full-band information. We evaluated our model on the VoiceBank+DEMAND (VCTK+DEMAND) dataset [19] and Interspeech 2020 Deep Noise Suppression (DNS) dataset [20]. The experimental results show that the proposed model achieves better results than other speech enhancement models.

2. IMPROVED TRANSFORMER

Generally speaking, a transformer consists of an encoder and a decoder [17]. In this paper, we choose the transformer encoder as our basic block. To avoid confusion, the reference to the transformer in this paper refers to the encoder part of the transformer. The original transformer encoder usually contains three modules: positional encoding, multi-head self-attention, and position-wise feed-forward network. In this paper, our transformer consists of two modules as in [14]: multi-head self-attention and modified position-wise feed-forward network.

2.1. Multi-head self-attention

We used the multi-headed self-attention from [17]. The multi-headed self-attention module can be formulated as:

\[ Q_i = ZW_i^Q, K_i = ZW_i^K, V_i = ZW_i^V, i \in [1, h] \]  

\[ \text{head}_i = \text{Attention}(Q_i, K_i, V_i) = \text{SoftMax}(\frac{Q_i K_i}{\sqrt{d}}) V_i \]
In this section, we propose a frequency-domain dual-path transformer processing module (DPTPM), and a decoder.

### 3.1. Encoder

The encoder consists of a 1x1 convolutional layer and a dilated-dense block, where the dilated-dense block consists of four dilated convolutional layers. The input to the encoder is the complex spectrum $X \in \mathbb{R}^{2 \times T \times F}$ resulted from short-time Fourier transform (STFT), and the output is a high-dimensional representation $U \in \mathbb{R}^{C' \times T' \times F}$ with C-T-F spectral feature maps.

### 3.2. Dual-path transformer processing module

The DPTPM consists of two 1x1 convolutional layers, $B$ dual-path transformers (DPTs), and a gated 1x1 convolutional layer. Before the DPTs, we use a 1x1 convolutional layer to halve the channel dimension of the encoder output features to form a new 3-D tensor $D \in \mathbb{R}^{C' \times T' \times F}$ ($C' = C/2$), and use $D$ as the input to the DPTs, as presented in Fig.1. Each DPT consists of an intra-transformer and an inter-transformer, where the intra-transformer models sub-band information and the inter-transformer models full-band information. Different from [22], the DPT handles time and frequency paths alternatively instead of parallely.

The intra-transformer processing block models the sub-band of the input features, which acts on the second dimension of $D$

$$D_{b \text{ intra}}^{DPT} = \text{IntraTransformer}_b[D_{b-1 \text{ intra}}]$$

where $D_{b \text{ intra}}^{DPT}$ is the output of IntraTransformer$_b$, $D_{b-1 \text{ intra}}$ is the mapping function defined by the transformer, $b = 1, 2, ..., B$ and $D_{b-1 \text{ intra}} = D, D_{b-1 \text{ intra}}^{DPT} \in \mathbb{R}^{C' \times T'}$ is the sequence defined by all the $T$ time step in the $i$-th sub-band. That is, the intra-transformer models the information of all time steps in each sub-band of the speech signal.

The inter-transformer processing block is used to summarize the information from each sub-band of the intra-transformer output to
learn the global information of the speech signal, which acts on the last dimension of $D$

\[ D_{\text{inter}} = \text{InterTransformer}_{b}[D_{\text{intra}}] = [f_{\text{inter}}(D_{\text{intra}}[:,j:], j = 1, \ldots, T)] \]  

(8)

where $D_{\text{inter}}$ is the output of $\text{InterTransformer}_b$, $f_{\text{inter}}$ is the mapping function defined by the transformer, and $D_{\text{intra}}[:,j:] \in \mathbb{R}^{C \times F}$ is the sequence defined by the $j$-th time step in all $F$ sub-bands. That is, the inter-transformer models the information of all sub-bands of the speech signal at each time step. With the intra-transformer, each time step in $D_{\text{intra}}$ contains all the information of the corresponding sub-band, which allows the inter-transformer to model the global (i.e., full-band) information of the speech signal.

The final output of the transformer $D_{\text{inter}}$ is passed through a 1x1 convolutional layer to double the channel dimension of the output feature and then through a gated convolutional layer to smooth the output value of the DPTPM.

3.3. Decoder

The decoder consists of a 1x1 convolutional layer and a dilated-dense block, where the dilated-dense block is the same as in the encoder. The feature from the DPTPM output is passed through the decoder to obtain the estimated complex ratio mask [23]. The enhanced complex spectrum is obtained by the element-wise multiplication between encoder’s input and the mask, which is passed through the ISTFT to obtain the enhanced speech waveform.

3.4. Loss function

In order to make full use of the time-domain waveform-level features and the T-F domain spectrum features, our loss function combines both time-domain and T-F domain losses. The loss function is as follows:

\[ L = \alpha_1 \times L_{\text{audio}} + \alpha_2 \times L_{\text{spectral}} \]  

(9)

$L_{\text{audio}}$ is mean square error (MSE) loss:

\[ L_{\text{audio}} = \frac{1}{N} \sum_{i=0}^{N-1} (y_i - \hat{y}_i)^2 \]  

(10)

where $y$ and $\hat{y}$ are the sample of the clean speech and the enhanced speech, respectively, and $N$ denotes the number of samples in the waveform. $L_{\text{spectral}}$ is L1 loss, which is defined as:

\[ L_{\text{spectral}} = \frac{1}{TF} \sum_{t=0}^{T-1} \sum_{f=0}^{F-1} [(|Y_r(t,f)| - |\hat{Y}_r(t,f)|) + (|Y_i(t,f)| - |\hat{Y}_i(t,f)|)] \]  

(11)

where $Y$ and $\hat{Y}$ denote the spectrum of the clean speech and the spectrum of the enhanced speech, respectively. $r$ and $i$ are the real and imaginary parts of the complex spectrogram. $T$ and $F$ are the number of frames and the number of frequency bins, respectively.

4. EXPERIMENTS

4.1. Dataset

We use a small-scale and a large-scale dataset to evaluate the proposed model. For the small-scale dataset, we use the VCTK+DEMAND dataset, which is widely used in SE research. This dataset contains pre-mixed noisy speech and its paired clean speech. The clean sets are selected from the VoiceBank corpus [25], where the training set contains 11,572 utterances from 28 speakers, and the test set contains 872 utterances from 2 speakers. For the noise set, the training set contains 40 different noise conditions with 10 types of noises (8 from DEMAND [26] and 2 artificially generated) at SNRs of 0, 5, 10, and 15 dB. The test set contains 20 different noise conditions with 5 types of unseen noise from the DEMAND database at SNRs of 2.5, 7.5, 12.5, and 17.5 dB. All the utterances are downsampled to 16kHz. We use 4-second long segments. If an utterance is longer than 4 seconds, a random 4-second slice will be selected from that utterance.

For the large-scale dataset, we use the DNS dataset. The DNS dataset contains over 500 hours of clean clips from 2150 speakers and over 180 hours of noise clips from 150 classes. We simulate the noisy-clean pairs with dynamic mixing during training stage. Specifically, before the start of each training epoch, 75% of the clean speeches are mixed with randomly selected room impulse responses (RIR) provided by [27]. By mixing the clean speech (75% of them are reverberant) and noise with a random SNR in between -5 and 20 dB, we generate the speech-noise mixtures. For evaluation, the DNS dataset has two non-blind test sets named with_reverb and no_reverb, both of which contain 150 noisy-clean pairs.

4.2. Experimental setup

The window length and frame shift of STFT and ISTFT are 25ms and 6.25ms, respectively, and the FFT length is 512. The number of frames and the number of frequency bins, respectively. For the VCTK+DEMAND dataset, we also employ the three most commonly used metrics in the VCTK+DEMAND dataset, which are CSIG for signal distortion, CBAK for noise distortion, and COVL for overall quality evaluation [31]. CSIG, CBAK, and COVL are mean opinion score (MOS) predictors, with a score range from 1 to 5. For the DNS dataset, we also employ SI-SDR as evaluation metrics. Higher scores indicate better performance for all metrics.

On both datasets, we use wide-band PESQ (dubbed WB-PESQ) [29] and STOI [30] as evaluation metrics. WB-PESQ and STOI quantify the perceptual quality and the intelligibility of a speech signal, respectively. For the VCTK+DEMAND dataset, we also employ SI-SDR as evaluation metrics. Higher scores indicate better performance for all metrics.
5. EXPERIMENTAL RESULTS

5.1. Results on the VCTK+DEMAND dataset

The proposed method is compared with other methods which also employ the same VCTK dataset. As shown in Table 1, our proposed model outperforms other transformer-based models such as TSTNN, T-GSA, SE-Conformer, and achieves state-of-the-art performance in terms of WB-PESQ, STOI, CSIG, CBAK, COVL with the least parameters.

5.2. Ablation analysis

The experimental results in the previous subsection demonstrate that our method improves the SE performance. To further validate the effectiveness of our method, we performed an ablation analysis. We designed four experiments labeled CED+Dual-path former, STFT+CED+BLSTM, STFT+CED+Sub-band former, STFT+CED+Full-sub former, which are abbreviated as exp.1, exp.2, exp.3, exp.4 in the following. Exp.4 is our proposed method. The difference between exp.1 and exp.4 is that exp.1 takes time-domain features as input, which replaces STFT/ISTFT with segmentation and overlap-add stage as in [16]. The difference between exp.3 and exp.4 is that the intra part and inter part of the dual-path transformer in exp.3 both model sub-band information as in Eq.7 while the inter part of the dual-path transformer in exp.4 model full-band information as in Eq.8. Same as exp.3, the BLSTM in exp.2 only models sub-band information. For a fair comparison, the number of parameters and computational complexity of the models in the four experiments was essentially the same, and all use a window length of 25 ms and a frame shift of 6.25 ms to extract frames. Therefore, the system latency is also essentially the same for all four experiments.

| Table 2 | Ablation analysis results on the VCTK+DEMAND dataset |
|---------|-----------------------------------------------|
| Method | WB-PESQ | STOI |
| CED + Dual-path former | 2.97 | 0.95 |
| STFT + CED + BLSTM | 3.05 | 0.95 |
| STFT + CED + Sub-band former | 3.20 | 0.95 |
| STFT + CED + Full-sub former | **3.33** | **0.96** |

By comparing exp.2 and exp.3, we can see that the improved transformer performs better than BLSTM, which demonstrates the effectiveness of the improved transformer. Furthermore, exp.4 outperforms exp.3, proving the advantages of fusing sub-band information and full-band information. In both exp.1 and exp.4, the intra transformer and inter transformer in the dual-path transformer model local and global information, respectively. However, the evaluation results of exp.4 is much better than those of exp.1, which proves that the frequency domain feature is more effective than the time domain feature for the dual-path transformer.

5.3. Results on the DNS dataset

Table 3 compares the metric scores of the proposed model with those of other architectures on the DNS dataset. We can see that our method outperforms the baseline. We noticed that compared with the full-band model, the proposed model has a more significant performance improvement on the reverberation data. We also see a similar trend in FullSubNet. The possible reason is that the full-band and sub-band fusion models include a sub-band model, and the sub-band model helps to model reverberation effects by focusing on the temporal evolution of the narrow-band spectrum.

| Table 3 | Comparison with other state-of-the-art systems on the DNS with_reverb (no_reverb) test sets. |
|---------|-----------------------------------------------|
| Method | WB-PESQ | STOI (%) | SI-SDR (dB) |
| Noisy | 1.82 (1.58) | 86.62 (91.52) | 9.03 (9.07) |
| NSNet [20] | 2.37 (2.15) | 90.43 (94.47) | 14.72 (15.61) |
| DTLN [32] | - (-) | 84.68 (94.76) | 10.53 (16.34) |
| PoCoNet [33] | 2.83 (2.75) | - (-) | - (-) |
| FullSubNet [7] | 2.97 (2.78) | 92.62 (96.11) | 15.75 (17.29) |
| CTS-Net [8] | 3.02 (2.94) | 92.70 (96.66) | 15.58 (17.99) |
| GaGNet [34] | - (3.17) | - (97.13) | - (18.91) |
| DPT-FSNet (Proposed) | **3.53 (3.26)** | **95.23 (97.68)** | **18.14 (20.36)** |

6. CONCLUSIONS

In this paper, we propose a dual-path transformer-based full-band and sub-band fusion network for speech enhancement in the frequency domain. Inspired by the full-band and sub-band fusion models, we explore features that are more efficient for dual-path structures with the intra part in the dual-path transformer models the sub-band information, and the inter part models the full-band information. Experimental results on the Voice Bank + DEMAND dataset and DNS dataset show that the proposed method outperforms the current state of the art at a relatively small model size.
7. REFERENCES

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

[2] Ashutosh Pandey and DeLiang Wang, “A new framework for cnn-based speech enhancement in the time domain,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 7, pp. 1179–1188, 2019.

[3] Alexandre Défossez, Gabriel Synnaeve, and Yossi Adi, “Real Time Speech Enhancement in the Waveform Domain,” in Proc. Interspeech 2020, 2020, pp. 3291–3295.

[4] Eesung Kim and Hyeji Seo, “Se-conformer: Time-domain speech enhancement using conformer,” Proc. Interspeech 2021, pp. 2736–2740, 2021.

[5] Szu-Wei Fu, Chien-Feng Liao, Yu Tsao, and Shou-De Lin, “Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement,” in International Conference on Machine Learning, PMLR, 2019, pp. 2031–2041.

[6] Jaeyong Kim, Mostafa El-Khamy, and Jungwon Lee, “F-gsa: Transformer with gaussian-weighted self-attention for speech enhancement,” in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 6649–6653.

[7] Xiang Hao, Xiangdong Su, Ruda Horaud, and Xiaohei Li, “Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement,” in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2021, pp. 6633–6637.

[8] Andong Li, Wenzhe Liu, Chengshi Zhung, Cunhang Fan, and Xiaodong Li, “Two heads are better than one: A two-stage complex spectral mapping approach for monaural speech enhancement,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 829–843, 2021.

[9] Haohe Liu, Lei Xie, Jian Wu, and Geng Yang, “Channel-Wise Subband Input for Better Voice and Accompaniment Separation on High Resolution Music,” in Proc. Interspeech 2020, 2020, pp. 1241–1245.

[10] Geng Yang, Shan Yang, Kai Liu, Peng Fang, Wei Chen, and Lei Xie, “Multi-band melgan: Faster waveform generation for high-quality text-to-speech,” in 2021 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2021, pp. 492–498.

[11] Shubo Lv, Yanxin Hu, Shimin Zhang, and Lei Xie, “DCCRNe: Channel-Wise Subband DCRNe with SNR Estimation for Speech Enhancement,” in Proc. Interspeech 2021, 2021, pp. 2816–2820.

[12] Naoya Takahashi and Yuki Mitsufuji, “Self-attentional acoustic models,” arXiv preprint arXiv:1803.09519, 2018.

[13] Matthias Sperber, Jan Niewies, Graham Neubig, Sebastian Stüker, and Alex Waibel, “Self-attentional acoustic models,” arXiv preprint arXiv:2001.08662, 2020.

[14] vector output of the transformer is all you need,” in Advances in neural information processing systems, 2017, pp. 5998–6008.

[15] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, “Densely connected convolutional networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” in Advances in neural information processing systems, 2017, pp. 5998–6008.

[17] Pieter-Jan Kindermans and Alex Waibel, “Self-attentional acoustic models,” arXiv preprint arXiv:1803.09519, 2018.

[18] Matthias Sperber, Jan Niewies, Graham Neubig, Sebastian Stüker, and Alex Waibel, “Self-attentional acoustic models,” arXiv preprint arXiv:2001.08662, 2020.

[19] Matthias Sperber, Jan Niewies, Graham Neubig, Sebastian Stüker, and Alex Waibel, “Self-attentional acoustic models,” arXiv preprint arXiv:2001.08662, 2020.

[20] Matthias Sperber, Jan Niewies, Graham Neubig, Sebastian Stüker, and Alex Waibel, “Self-attentional acoustic models,” arXiv preprint arXiv:2001.08662, 2020.