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

Paper ID: 2333

Feng Dang [1,2,3] Hangting Chen [1] Pengyuan Zhang [1]

[1]Key Laboratory of Speech Acoustics & Content Understanding, Institute of Acoustics, CAS, China
[2]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[3]School of Cyber Security, 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.

Motivations

Our model can be seen as a combination of "full-band and sub-band" feature modeling and a dual-path structure.

1. **"Full-band and sub-band" feature modeling:** Inspired by FullSubNet (Hao et al. 2021), we thought it would be helpful to use full-band and sub-band feature modeling.
2. **Dual-path structure:** Recently, dual-path networks (Luo et al. 2020; Chen et al. 2020; Wang et al. 2021) have achieved exceptional performance due to their ability to model local and global features of the input sequence.
3. **A combination of the above two methods:** Inspired by FullSubNet's "full-band and sub-band" feature modeling, we used a dual-path structure suitable for modeling such features, in which the intra-transformer models sub-band information and the inter-transformer merges the sub-band information from the intra-transformer to model the full-band information.

Framework

Our proposed model consists of an encoder, a dual-path transformer processing module (DPTPM), and a decoder.

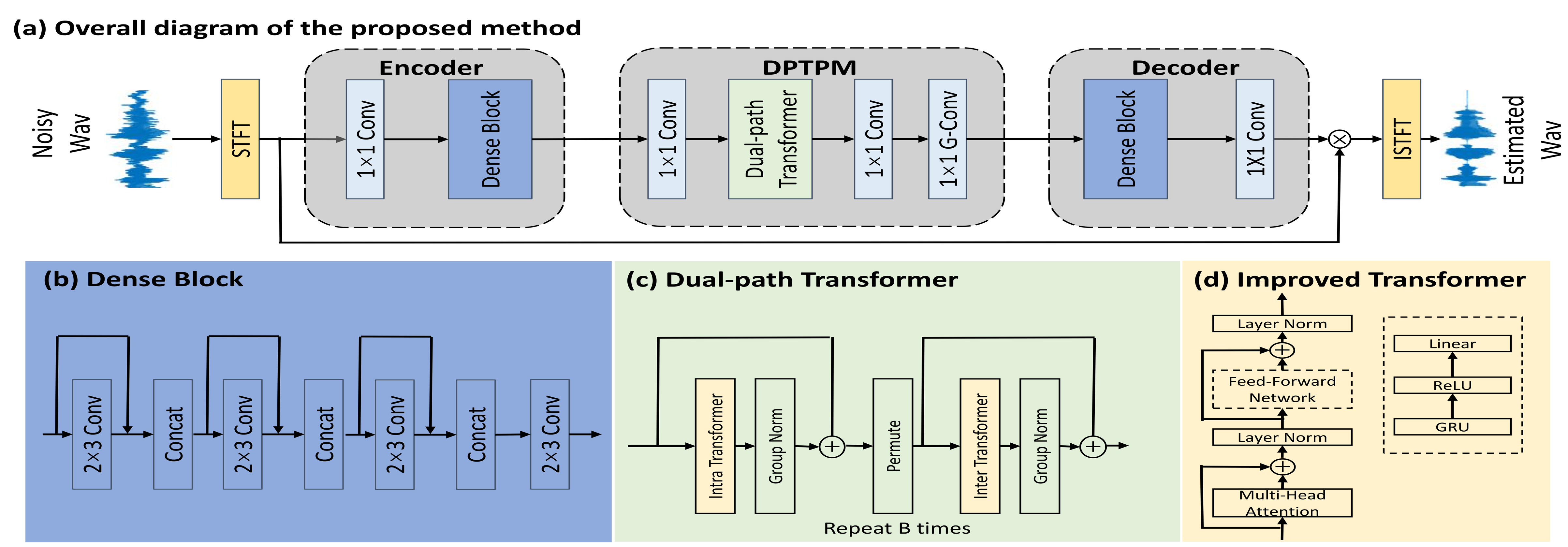

Figure 1. Architecture of the proposed DPT-FSNet. a) The overall diagram of the proposed method b) The detail of the dense block. c) The detail of the dual-path transformer. d) The detail of the improved transformer.

Workflow

Encoder

The input to the encoder is the complex spectrum $X \in \mathbb{R}^{2\times T\times F}$, and the output is a high-dimensional representation $U \in \mathbb{R}^{C\times T\times F}$.

DPTPM

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

$$D_b^{intra} = [f_b^{intra}(D_{b-1}^{inter}[:,:,i], i = 1,\dots,F)] \tag{1}$$

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

$$D_b^{inter} = [f_b^{inter}(D_b^{intra}[:,j,:], j = 1,\dots,T)] \tag{2}$$

Decoder

The feature from the DPTPM output is passed through the decoder to obtain the estimated complex ratio mask. The enhanced complex spectrum is obtained by the element-wise multiplication between encoder's input and the mask.

Results & Discussion

We use a small-scale (Voicebank+DEMAND) and a large-scale dataset (DNS dataset) to evaluate the proposed model.In both of the above datasets, we compared our proposed algorithm with the current state-of-the-art, as shown in Tables 1 and 2.

Method	WB-PESQ	STOI	CSIG	CBAK	COVL	Para. (M)
Noisy	1.97	0.91	3.34	2.44	2.63	-
MetricGAN	2.86	-	3.99	3.18	3.42	1.90
TSTNN	2.96	0.95	4.33	3.53	3.67	0.92
T-GSA	3.06	-	4.18	3.59	3.62	-
DEMUCS	3.07	0.95	4.31	3.40	3.63	33.5
SE-Conformer	3.13	0.95	4.45	3.55	3.82	-
Learnable Loss Mixup	3.26	-	4.49	3.27	3.91	20.32
DPT-FSNet	**3.33**	**0.96**	**4.58**	**3.72**	**4.00**	**0.88**

Table 1. Comparison with other state-of-the-art systems on the VCTK+DEMAND dataset.

Method	WB-PESQ	STOI (%)	SI-SDR (dB)
Noisy	1.82 (1.58)	86.62 (91.52)	9.03 (9.07)
NSNet	2.37 (2.15)	90.43 (94.47)	14.72 (15.61)
DTLN	- (-)	84.68 (94.76)	10.53 (16.34)
PoCoNet	2.83 (2.75)	- (-)	- (-)
FullSubNet	2.97 (2.78)	92.62 (96.11)	15.75 (17.29)
CTS-Net	3.02 (2.94)	92.70 (96.66)	15.58 (17.99)
GaGNet	- (3.17)	- (97.13)	- (18.91)
DPT-FSNet	**3.53 (3.26)**	**95.23 (97.68)**	**18.14 (20.36)**

Table 2. Comparison with other state-of-the-art systems on the DNS *with_reverb* (*no_reverb*) test sets.

To further validate the effectiveness of our method, we performed two ablation analysis experiments.

Method	WB-PESQ	STOI
CED + Dual-path former	2.97	0.95
STFT + CED + Sub-band former	3.20	0.95
STFT + CED + Full-sub former	**3.33**	**0.96**

Table 3. Ablation analysis results in terms of feature modeling on the VBD dataset.

Method	WB-PESQ	STOI
STFT + CED + Original Transfomer	3.04	0.95
STFT + Improved Transformer	3.11	0.95
STFT + CED + Improved Transformer	**3.33**	**0.96**

Table 4. Ablation analysis results in terms of model structure on the VBD dataset.

In Table 3: By comparing *exp.3* and *exp.2*,it can be seen that using two transformers to model sub-band and full-band information separately improves the performance over using two identical transformers to model only sub-band information. Moreover, the results of *exp.3* are much better than *exp.1*, which proves that the frequency domain feature is more effective than the time domain feature for the dual-path transformer.

In Table 4: By comparing *exp.3* and *exp.1*, we can find that the performance of the improved transformer is much better than that of the original transformer. Performance can be improved by combining a convolutional encoder/decoder with the transformer as shown in *exp.3* and *exp.2*.

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.

References

- Hao, X., X. Su, R. Horaud, and X. Li (2021). "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, pp. 6633–6637.
- Wang, K., B. He, and W.-P. Zhu (2021). "TSTNN: Two-Stage Transformer Based Neural Network for Speech Enhancement in the Time Domain". In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 7098–7102.
- Chen, J., Q. Mao, and D. Liu (2020). "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation". In: *arXiv preprint arXiv:2007.13975*.
- Luo, Y., Z. Chen, and T. Yoshioka (2020). "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 46–50.