Band-Split Inter-SubNet: Band-Split with Subband Interaction for Monaural Speech Enhancement

Yen-Chou Pan*, Yih-Liang Shen*, Yuan-Fu Liao†, and Tai-Shih Chi*
* Institute of Communications Engineering, National Yang Ming Chiao Tung University, Taiwan
E-mail: yenchoupan@gmail.com; dennis831209@gmail.com; tschi@nycu.edu.tw
† Institute of Artificial Intelligence Innovation, Industry-Academia Innovation School,
National Yang Ming Chiao Tung University, Taiwan
E-mail: yfliao@nycu.edu.tw

***Abstract*—Speech enhancement models are developed to improve quality and intelligibility of speech for numerous daily applications. With the rapid development of technology, the neural network based speech enhancement models show significantly improved performance. The subband-based models focus on local spectral patterns and achieve outstanding results with fewer parameters. In this paper, we propose a subband-based composite model named Band-Split Inter-SubNet. It adopts the new constant-Q band-split setting to mimic human auditory perception. The proposed model demonstrates superior performance to other state-of-the-art models on the DNS Challenge - Interspeech 2021 dataset. Detailed analyses on experimental results demonstrate that the proposed band-split setting is effective, and the influence of neighboring frequency bins on the center-frequency bin across different frequency bands varies slightly.**

I. Introduction

In daily life, noise interference is one major factor hindering human speech communication and automatic speech recognition. It degrades both speech quality and intelligibility. The goal of speech enhancement models is to eliminate the interference of noise on speech. Conventional speech enhancement methods primarily adopt statistical theories to suppress stationary noise. With the advancement of technology, neural network (NN) based methods have shown significantly improved results in many complex and challenging situations, such as in low signal-to-noise ratios (SNRs) and reverberant conditions. At present, NN-based speech enhancement models are mainly implemented in the time domain or the frequency domain. The time-domain models [1], [2] directly generate a clean audio waveform from the noisy audio waveform. The frequency-domain models either directly predict the spectrogram of clean speech [3] or predict relevant masks, such as the ideal binary mask (IBM) [4], the ideal ratio mask (IRM) [5], and the complex ideal ratio mask (cIRM) [6][7], from the noisy spectrogram. Considering system robustness and computational complexity, frequency-domain models are more widely used than time-domain models.

Recently, the frequency-domain subband-based speech enhancement models [8], [9] have shown excellent performance in monaural speech enhancement. This approach is based on the idea that nearby frequency components are more influential and effective in distinguishing speech from noise than far-away frequency components. There are mainly two types of subband speech enhancement models, i.e., band-split [10], [11] and neighboring-frequency [12] models, as shown in Fig. 1. The band-split models divide the full frequency band into multiple subbands, with each subband processed independently. This allows for the capture of more local information. However, this type of model is limited by treating each band as a separate channel, preventing sharing information across bands. Consequently, various modules have been developed to enhance the interaction between channels to alleviate this limitation.

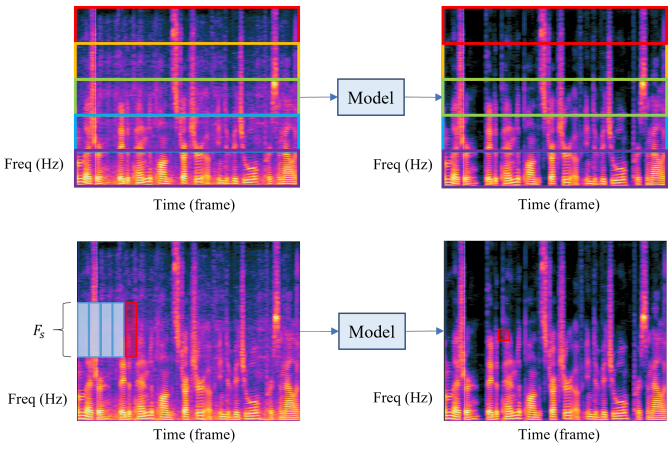

Fig. 1. Subband models: (Top) Band-split model; (bottom) neighboring-frequency model.

On the other hand, the neighboring-frequency models predict the center-frequency bin by considering its neighboring frequency bins as additional information. However, due to the lack of fullband information, this type of models can be less effective in conditions where local time-frequency features are not well-observed. To tackle this problem, neighboring-frequency models incorporating fullband information have been developed [13], [14], [15], enabling models to capture both local and global features. Not surprisingly, this combination results in increased number of model parameters and complexity. To reduce the complexity, Chen et al. [16] proposed a novel lightweight framework named Inter-SubNet, which comprises an interactive SubInter module to effectively captures global and local spectral patterns. Inspired by concepts of both types of subband models, we propose

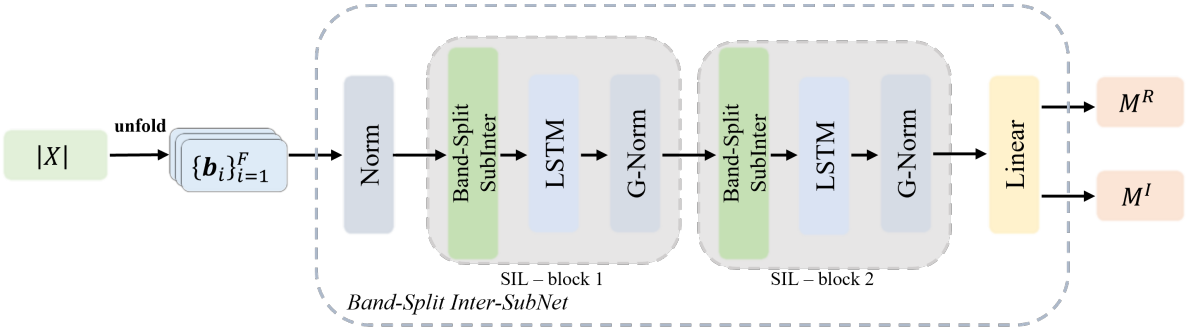

Fig. 2. The architecture of the proposed Band-Split Inter-SubNet model, which is extended from Inter-SubNet [16].

a combinational model, named Band-Split Inter-SubNet, to leverage the influence of neighboring frequency bins on the center-frequency bin across different frequency bands. The contributions of this paper can be summarized as follows:

1) We propose a combinational subband model from two types of models to achieve better performance.
2) Rather than the uniform splitting setting, we use the constant-Q band-split setting by mimicking human auditory perception to achieve better performance.

The rest of the paper is organized as follows. In Section 2, we introduce related work and the proposed Band-Split Inter-SubNet model. The experimental setup including datasets, training setup and baseline models are introduced in Section 3. In Section 4, we present results and discussions. Lastly, we conclude the paper in Section 5.

II. Proposed Model

A. Problem formulation

The signal model in the short-time Fourier transform (STFT) domain for monaural speech can be written as

$$X(t,f) = S(t,f) + N(t,f) \tag{1}$$

where t and f are the indexes of time frame and frequency bin, respectively. $X(t,f)$, $S(t,f)$ and $N(t,f)$ denote the complex spectrograms of noisy speech, clean speech, and interference noise.

The architecture of the proposed Band-Split Inter-SubNet model is depicted in Fig. 2. It is composed of two stacked SubInter-LSTM (SIL) blocks and one linear layer. The SIL block has been shown effective in Inter-SubNet [16]. It can capture global spectral information while maintaining the ability to focus on local spectral patterns. The "G-norm" indicates group normalization [17]. The magnitude spectrogram $|X| \in \mathbb{R}^{F\times T}$ and the real part and the imaginary part of cIRM (M^R and M^I) are the input and the output of the model, respectively, where F and T denote the numbers of overall frequency bins and time frames.

We concatenate the i-th frequency bin $|X_i| \in \mathbb{R}^{1\times T}$ with the $2\times n$ neighboring frequency bins to form a subband unit b_i by the "unfold" operation:

$$b_i = [|X_{i-n}|, \cdots, |X_i|, \cdots, |X_{i+n}|] \in \mathbb{R}^{F_s\times T} \tag{2}$$

where $F_s = 2n+1$.

B. Band-Split SubInter module

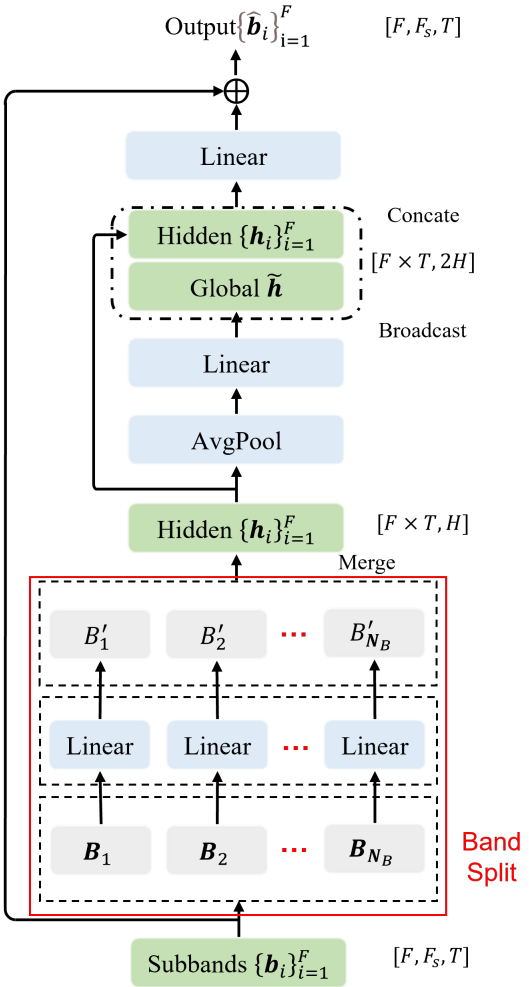

Fig. 3. Detailed structure of the Band-Split SubInter module.

The proposed Band-Split SubInter module shown in Fig. 2 is detailed in Fig. 3, which is slightly different from the original SubInter module in the Inter-SubNet framework [16]. Instead of using the same linear layer for all subband units,

the subband units of all frequencies are firstly divided into N_B bands and the subband units of each band are processed by each linear layer separately and merged together as the subband features $\{\boldsymbol{h}_i\}_{i=1}^{F}$, which contain local information from neighboring frequencies. After that, the subband features $\{\boldsymbol{h}_i\}_{i=1}^{F}$ are concatenated with global features $\tilde{\boldsymbol{h}}$, obtained by averaging local subband features followed by the second linear layer, and sent to the third linear layer. At the end, the final output subband features $\{\hat{\boldsymbol{b}}_i\}_{i=1}^{F}$ are obtained by applying a residual connection between the input subband features and the extracted subband features.

C. Setting for splitting bands

Typical band-split models split bands uniformly along the linear-frequency axis, i.e., all bands have the identical frequency bandwidth. In this paper, we propose a new band-split setting, as shown in Fig. 4 to mimic human auditory perception, which is more akin to a logarithmic frequency scale. Similar to the constant-Q transform, each bandwidth is half that of the previous band from high to low frequency. The first band is derived by splitting the fullband with the highest frequency of F_B Hz in half. And then, the second band is obtained by splitting the remaining frequencies ranging from 0 Hz to $\frac{1}{2}F_B$ Hz in half, and so on. The lowest two bands have the same bandwidth of $2^{-N_B+1} \times F_B$ Hz when we split the fullband into N_B bands. Compared with the uniform band-split setting, this new setting can effectively divide the low-frequency region into more bands without significantly increasing the number of model parameters.

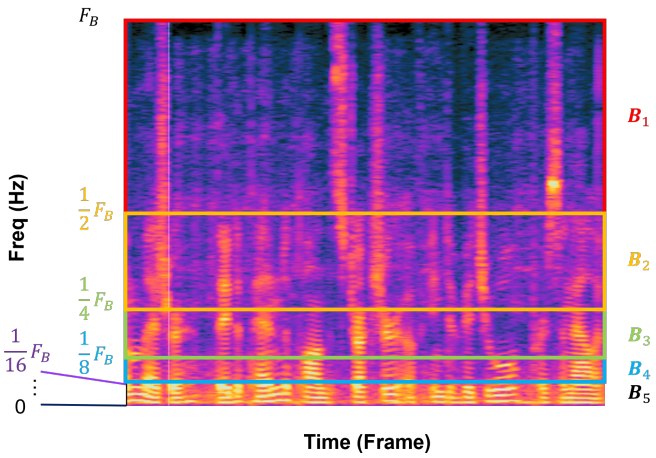

Fig. 4. Setting for splitting bands. The fullband is split in a way similar to the constant-Q transform.

III. Experimental Setup

A. Datasets

The proposed model was trained and evaluated on a subset of the Deep Noise Suppression Challenge (DNS Challenge) Interspeech 2021 dataset [18]. The clean speech dataset contains 562.72 hours of clips from 2150 speakers. The noise dataset includes 181 hours of 65302 clips from over 150 classes. Moreover, we randomly select and add room impulse responses from the openSLR26 and openSLR28 datasets [19] to 75% of the clean speech. Then, we mix clean or reverberant speech with noise at SNRs ranging from -5 to 20 dB to generate noisy speech. To evaluate models' performance, we used the public test set from DNS Challenge, which contains synthesized clips of two classes, namely with and without reverberations. It includes 150 noisy clips with SNRs ranging from 0 to 20 dB.

B. Training setup

We followed the settings of DNS Challenge for model training. The speech waveforms with 16 kHz sampling rate were converted into STFT spectrograms using the Hanning window with the window length of 32 ms and the frame shift of 16 ms. Adam optimizer was adopted with the learning rate of $1e^{-3}$ and the mean square error (MSE) and PReLU were used as the loss function and the activation function. The number of neighboring frequency bins on each side n was set to 15 and the number of bands N_B was set to 5. The number of frames of input and output sequences was set to 192 frames (approximately 3 s) in training. Compared models were evaluated by the common speech enhancement metrics WB-PESQ [20], NB-PESQ [21], STOI [22], and SI-SDR [23]. The following statements describe particular settings for the proposed model, inspired from the baseline Inter-SubNet model [16], and its variants.

Band-Split Inter-SubNet (prop): This is the basic version of the proposed model. The Band-Split SubInter module contains 93 and 307 hidden units in the first and the second SIL blocks, respectively. Both LSTMs in the first and the second SIL blocks are built with 384 hidden units.

Band-Split Inter-SubNet (lin): This is the proposed model with the uniform band-split setting in the linear-frequency axis. The performance of this variant is examined to see if the proposed constant-Q band-split setting offers benefits.

Band-Split Inter-SubNet (+): This is the plus version of the proposed model. In the basic version, the frequency band splitting is only applied in the first linear layer of the Band-Split SubInter module. In the plus version, the frequency band splitting is applied to all linear layers.

IV. Results and Discussions

As shown in Table I, the performance of the plus version is not significantly different from that of the basic version with $N_B = 5$. Therefore, we conclude that splitting the frequency band only in the first linear layer of the SubInter module is enough to achieve the optimal performance. Furthermore, we observe the proposed constant-Q band-split setting provides more benefits to performance than the uniform band-split setting under all conditions in terms of all metrics. The score differences under non-reverberant conditions are quite noticeable.

Comparison results between the baseline model and the proposed model indicate that band splitting is effective but not significantly so, suggesting that the influence of neighboring

TABLE I
DENOISING PERFORMANCE COMPARISON AMONG BAND-SPLIT MODELS

Model	Band-Split (Scale)	N_B	#Para (M)	With Reverb				Without Reverb			
				WB-PESQ	NB-PESQ	STOI	SI-SDR	WB-PESQ	NB-PESQ	STOI	SI-SDR
Noisy	-	-	-	1.822	2.753	86.62	9.03	1.582	2.454	91.52	9.07
Inter-SubNet [16]	-	-	2.29	3.207	3.659	93.98	16.76	2.997	3.504	96.61	18.05
Band-Split Inter-SubNet (prop)	log	2	2.41	3.216	3.656	93.90	16.64	2.995	3.497	96.53	17.98
Band-Split Inter-SubNet (prop)	log	3	2.54	3.223	3.651	93.97	16.66	3.005	3.501	96.63	18.10
Band-Split Inter-SubNet (prop)	log	4	2.66	3.221	**3.666**	93.92	16.77	2.987	3.495	96.54	17.99
Band-Split Inter-SubNet (prop)	log	5	2.78	**3.230**	3.665	94.03	16.80	**3.023**	**3.509**	**96.77**	18.23
Band-Split Inter-SubNet (prop)	log	6	2.90	3.229	3.659	93.95	16.71	3.005	3.502	96.71	18.19
Band-Split Inter-SubNet (lin)	linear	5	2.78	3.228	3.657	93.91	16.80	2.989	3.494	96.54	18.13
Band-Split Inter-SubNet (+)	log	5	3.75	3.222	3.662	**94.04**	**16.84**	2.995	3.506	96.72	**18.27**

frequencies on the center frequency bin does not vary greatly across different frequency ranges. In addition, the primary distinction between the uniform splitting and the proposed constant-Q splitting lies in the low-frequency range. Specifically, the uniform splitting produces the lowest frequency band from 0 to 1600 Hz, while the constant-Q splitting produces the band from 0 to 500 Hz. The superior performance from the constant-Q splitting suggests that lower-frequency bands have a more profound impact on model performance. The poorer performance observed with $N_B = 6$ compared to $N_B = 5$ may be attributed to overly meticulous division of the low-frequency range, leading to sub-optimal results.

Additionally, we conducted a detailed analysis on results of the proposed model with $N_B = 5$ across different SNRs. The box plots of four scores under reverberant and non-reverberant conditions are illustrated in Fig. 5 and Fig. 6, respectively. These box plots show that the proposed model makes an improvement from the baseline model in overall performance under both reverberant and non-reverberant conditions at low SNRs (< 5 dB). However, there is a trade-off, as scores of some metrics exhibit slight compromises at high SNRs. We speculate that by logarithmically splitting the frequency bands, the model can effectively learn the characteristics of each band, especially the low-frequency band, thereby enhancing its robustness in low SNR environments. However, this band-split approach also sacrifices some overall coherence across the full frequency band, which explains why the proposed model does not perform as well at high SNRs compared to the baseline

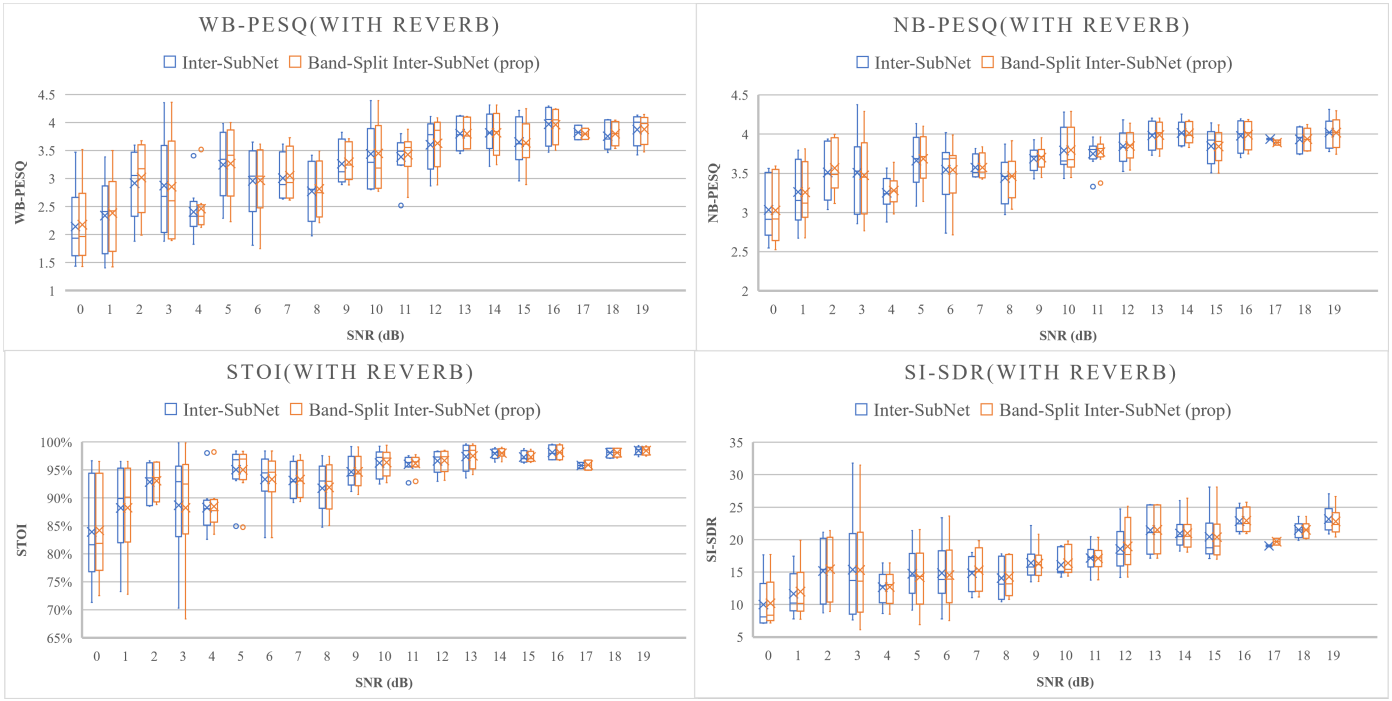

Fig. 5. Box plots of denoising scores of the proposed and baseline models across different SNRs under reverberant conditions.

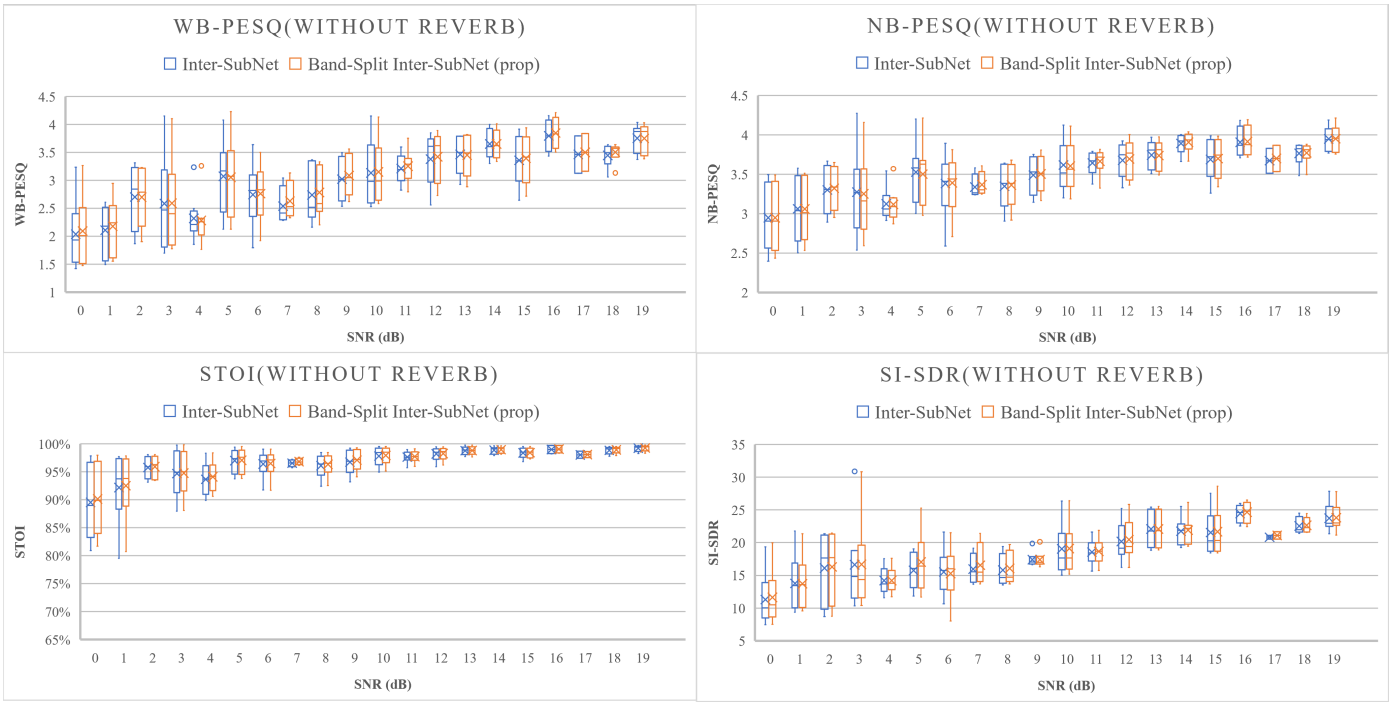

Fig. 6. Box plots of denoising scores of the proposed and baseline models across different SNRs under non-reverberant conditions.

model.

Finally, Table II shows the performance of the proposed model and other state-of-the-art models developed in recent years. Compared with other models, the proposed model shows superb performance in the denoising task with fewer model parameters. It is because the proposed model with the constant-Q band-split setting effectively highlights the distinctions between bands, resulting in superior model performance.

V. Conclusion and Future Work

In this paper, we propose a composite model named Band-Split Inter-SubNet. In addition, we adopt a new constant-Q band-split setting to mimic human auditory perception. Results demonstrate that the proposed model outperforms the original Inter-SubNet model. The new band-split setting enhances the robustness at low SNR conditions with slightly increased number of model parameters. The Band-Split SubInter module effectively learns the characteristics of each band and captures

TABLE II
Denoising performance comparison between the proposed model and other state-of-the-art models

Model	Year	#Para	With Reverb				Without Reverb			
		(M)	WB-PESQ	NB-PESQ	STOI	SI-SDR	WB-PESQ	NB-PESQ	STOI	SI-SDR
Noisy		-	1.822	2.753	86.62	9.03	1.582	2.454	91.52	9.07
DCCRN-E [24]	2020	3.7	-	3.077	-	-	-	3.27	-	-
Conv-TasNet [2]	2020	5.08	2.750	-	-	-	2.730	-	-	-
PoCoNet [25]	2021	50	2.832	-	-	-	2.748	-	-	-
DCCRN+ [8]	2021	3.3	-	3.300	-	-	-	3.330	-	-
TRU-Net [26]	2021	0.38	2.740	3.350	91.29	14.87	2.860	3.360	96.32	17.55
CTS-Net [27]	2021	4.99	3.020	3.470	92.70	15.58	2.940	3.420	96.66	17.99
FullSubNet [13]	2021	5.64	3.057	3.584	92.11	16.04	2.882	3.428	96.32	17.30
FullSubNet+ [14]	2022	8.67	3.177	3.648	93.64	16.44	3.002	3.503	96.67	18.00
FS-CANet [15]	2022	4.21	3.218	**3.665**	93.93	**16.82**	3.017	**3.513**	96.74	18.08
Inter-SubNet [16]	2023	2.29	3.207	3.659	93.98	16.76	2.997	3.504	96.61	18.05
Band-Split Inter-SubNet (prop)	2024	2.78	**3.230**	**3.665**	**94.03**	16.80	**3.023**	3.509	**96.77**	**18.23**

local patterns. Compared with other state-of-the-art models, the proposed model shows better performance on denoising under both reverberant and non-reverberant conditions. However, the proposed model demands high computations such that developing simpler modules and more efficient band-split settings still require further effort. In the future, we will explore new band-split settings and refine our modules to reduce model parameters and computations. Nevertheless, applying the model to real-time noise reduction remains a challenging task.

ACKNOWLEDGMENT

This work was supported in part by the Ministry of Science and Technology, Taiwan under Grant MOST 110-2221-E-A49-115-MY3, and in part by the Co-creation Platform of the Speech-AI Research Center, Industry-Academia Innovation School, NYCU, under the framework of the National Key Fields Industry-University Cooperation and Skilled Personnel Training Act, from the Ministry of Education (MOE), the National Development Fund (NDF), and industry partners in Taiwan.

REFERENCES

[1] A. Pandey and D. Wang, "Densely connected neural network with dilated convolutions for real-time speech enhancement in the time domain," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6629–6633.

[2] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[3] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[4] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech separation by humans and machines*. Springer, 2005, pp. 181–197.

[5] L. Sun, J. Du, L.-R. Dai, and C.-H. Lee, "Multiple-target deep learning for lstm-rnn based speech enhancement," in *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*, 2017, pp. 136–140.

[6] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, no. 3, pp. 483–492, 2015.

[7] K. Tan and D. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 380–390, 2020.

[8] S. Lv, Y. Hu, S. Zhang, and L. Xie, "DCCRN+: Channel-Wise Subband DCCRN with SNR Estimation for Speech Enhancement," in *Proc. Interspeech 2021*, 2021, pp. 2816–2820.

[9] J. Li, D. Luo, Y. Liu, Y. Zhu, Z. Li, G. Cui, W. Tang, and W. Chen, "Densely connected multi-stage model with channel wise subband feature for real-time speech enhancement," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6638–6642.

[10] Y. Luo and J. Yu, "Music source separation with band-split rnn," *arXiv preprint arXiv:2209.15174*, 2022.

[11] J. Yu, H. Chen, Y. Luo, R. Gu, and C. Weng, "High Fidelity Speech Enhancement with Band-split RNN," in *Proc. Interspeech 2023*, 2023, pp. 2483–2487.

[12] X. Li and R. Horaud, "Online monaural speech enhancement using delayed subband lstm," *arXiv preprint arXiv:2005.05037*, 2020.

[13] X. Hao, X. Su, R. Horaud, and X. 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.

[14] J. Chen, Z. Wang, D. Tuo, Z. Wu, S. Kang, and H. Meng, "Fullsubnet+: Channel attention fullsubnet with complex spectrograms for speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7857–7861.

[15] J. Chen, W. Rao, Z. Wang, Z. Wu, Y. Wang, T. Yu, S. Shang, and H. Meng, "Speech Enhancement with Fullband-Subband Cross-Attention Network," in *Proc. Interspeech 2022*, 2022, pp. 976–980.

[16] J. Chen, W. Rao, Z. Wang, J. Lin, Z. Wu, Y. Wang, S. Shang, and H. Meng, "Inter-subnet: Speech enhancement with subband interaction," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[17] Y. Wu and K. He, "Group normalization," 2018. [Online]. Available: https://arxiv.org/abs/1803.08494

[18] C. K. Reddy, H. Dubey, K. Koishida, A. Nair, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "INTERSPEECH 2021 Deep Noise Suppression Challenge," in *Proc. Interspeech 2021*, 2021, pp. 2796–2800.

[19] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5220–5224.

[20] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[21] I. Rec, "P. 862.2: Wideband extension to recommendation p. 862 for the assessment of wideband telephone networks and speech codecs," *International Telecommunication Union, CH–Geneva*, vol. 41, pp. 48–60, 2005.

[22] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[23] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[24] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement," in *Proc. Interspeech 2020*, 2020, pp. 2472–2476.

[25] U. Isik, R. Giri, N. Phansalkar, J.-M. Valin, K. Helwani, and A. Krishnaswamy, "PoCoNet: Better Speech Enhancement with Frequency-Positional Embeddings, Semi-Supervised Conversational Data, and Biased Loss," in *Proc. Interspeech 2020*, 2020, pp. 2487–2491.

[26] H.-S. Choi, S. Park, J. H. Lee, H. Heo, D. Jeon, and K. Lee, "Real-time denoising and dereverberation wtih tiny recurrent u-net," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5789–5793.

[27] 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 Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1829–1843, 2021.