Interspeech 2022
18-22 September 2022, Incheon, Korea

Lightweight Full-band and Sub-band Fusion Network for Real Time Speech Enhancement

Zhuangqi Chen, Pingjian Zhang

School of Software Engineering, South China University of Technology, Guangzhou, China

sezqchen@mail.scut.edu.cn, pjzhang@scut.edu.cn

Abstract

Recent studies in deep learning based real-time speech enhancement have proven the advantage of sub-band processing in parameter reduction. However, most sub-band based methods utilize the same model for all sub-bands, which limits the upper bound of performance, giving the fact that the spectral patterns in each sub-band are different. In this paper, we take into account this fact and propose a lightweight full-band and sub-band fusion network, where dual-branch based architecture is employed for modeling local and global spectral pattern simultaneously. A simple yet effective sub-band module, the weighted progressive convolutional module, is designed with a small number of parameters, which captures clean features progressively from local perspective. Each sub-band is handled by one module. A novel asymmetric convolutional recurrent network is also proposed to focus on full-band context and extract more robust global features, which is complementary to the sub-band module. We have conducted extensive experiments on both the VoiceBank+Demand and the DNS Challenge datasets, and the experimental results show that our proposed method has achieved superior performance to other state-of-the-art approaches with smaller model size and lower latency.

Index Terms: deep learning, speech enhancement, real-time, sub-band, full-band

1. Introduction

Speech enhancement, often referred as the noise suppression problem, aims to improve the perceptual quality and intelligibility of speech that has been degraded by additive noise [1]. There are a wide range of real-world scenarios in which it is extremely desired to enhance speech, such as voice communication, teleconferencing system, hearing aids, and automatic speech recognition. Most of the applications also require a real-time speech enhancement, which is causal and low-delay.

Because of the importance of speech enhancement, numerous attempts have been made to solve this problem over the decades. The traditional methods [1] are first introduced to deal with this problem. However, those methods usually only work for stationary noise situations. More recently, lots of data-driven models with deep learning (DL) technologies are proposed to effectively promote performance. Since the recurrent neural network (RNN) is inherently capable of modeling sequence, RNN and its variants (e.g., LSTM, GRU) are commonly used for speech enhancement [2, 3, 4]. The convolution neural network (CNN) is also employed for speech enhancement [5, 6], in which the model size is smaller and the computational speed is faster. However, It's difficult for CNN-based models to capture long-term dependence. To make full use the advantages of both RNN and CNN, several convolutional recurrent network (CRN) based enhancers [7, 8, 9, 10, 11, 12] are proposed. Recently, due to the success of Transformer in many areas, some studies [13, 14] also introduce Transformer to perform speech enhancement and achieve impressive improvement. However, the computational cost of Transformer is very high and thus it is not suitable for many applications that required low-resource.

The recent deep noise suppression (DNS) challenge [15, 16, 17] have fostered several state-of-the-art DL-based models in the area of real-time noise suppression. Some sub-band based methods are designed to process each sub-band independently to reduce the number of parameters. In [18], a delayed sub-band LSTM is proposed to model local spectral pattern at each frequency and achieves a considerable performance. Since the sub-band method processing each frequency independently and several neighbor frequencies are used, the computational cost is very high. In [10], neural network based analysis filters are employed for band splitting and only four sub-bands are generated. Although the above sub-band methods help to reduce the number of parameters, they use the same model for all sub-bands which implies that each sub-band can be processed with the same schema. Nevertheless, sub-bands differ from each other. For instance, the low frequency band has higher energy density. By taking into account the difference, we propose to process different sub-bands with different models. However, the strategy of "one model for one sub-band" will results in the increase of model size. To this end, a lightweight sub-band model is considered in this paper. We design a convolution based module with a small number of parameters to model local spectral dependence and progressively predict target magnitude.

Since the local information is inadequate for discriminating speech from noise, a full-band model focusing on global dependence is required. In [19], a prominent improvement of the sub-band model[18] is made on full-band dependence, where a full-band model is first employed to predict a rough magnitude from global perspective and the sub-band model is then used to further refine the features. In this paper, we consider a parallel structure to captures local and global features simultaneously.

The main contributions of this paper are three-folds:

- We propose a lightweight full-band and sub-band fusion network for real-time speech enhancement. We first project input features into two latent representations with a simple preprocessing module. Two parallel full-band and sub-band modules are then employed to capture complementary features which helps to reduce noise.
- We propose a simple but powerful sub-band based method, which treats each sub-band differently and captures local spectral features progressively.
- We propose a novel asymmetric CRN, where a encoder cascaded with a dual-path RNN extracts basic features first and a weighted progressive full-band decoder is then introduced to capture more robust global features.

10.21437/Interspeech.2022-153

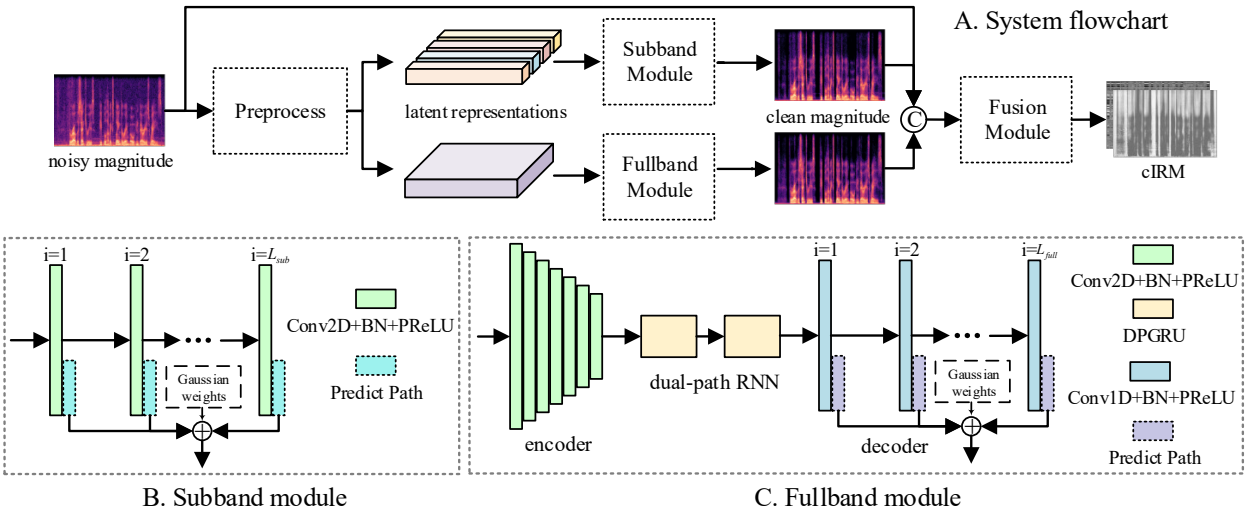

Figure 1: *(A): A flowchart of the proposed system. (B): The design of the sub-band module, which is a weighted progressive convolutional module. Each sub-band is processed by one module individually. (C): The design of full-band module, which is an asymmetric convolutional recurrent network. A weighted progressive full-band decoder is designed to model global dependence over full-band.*

2. Lightweight full-band and sub-band fusion network

2.1. Problem formulation

The noisy speech can be formulated in the time-frequency (TF) domain using short-time Fourier transform (STFT):

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

where $Y(f,t), X(f,t), N(f,t) \in \mathbb{C}^{F\times T}$ denote the TF representations of noisy speech, clean speech and noise signal, respectively, where F, T are the number of frequency and time dimensions respectively. In the real time situation, a causal and low delay speech enhancer is required, where only current frame with several past frames of noisy speech can be used to estimate target signal of current frame.

2.2. Overview

In causal system, performance degradation is inevitable, since there is no future information to be used. Fortunately, there are various acoustic cues that can be utilized to alleviate the problem. First, the energy distribution varies between different frequency bands, where the magnitude spectra of most phonemes have a low-frequency dominance. In addition, the voiced sounds are periodic and reflected as harmonic in the frequency domain. The unvoiced sounds are more noise-like but can be discriminated based on local rapid spectral changes. To make full use of the acoustic properties, we proposed a lightweight neural network to capture the harmonic dependency and local spectral pattern.

As shown in figure 1:(A), the proposed model, a lightweight full-band and sub-band fusion network (LFSFNet), consists of four processing stages: preprocessing module for projecting input features into two hidden spaces, sub-band module for separately modeling local spectral patterns in each frequency band, full-band module for capturing global dependency (e.g., harmonic), and features fusion module for fusing the outputs of full-band and sub-band modules and estimating a mask for target speech.

2.3. Preprocessing

Motivated by [20], we take the compressed magnitude $\tilde{Y}_{cprs}(f,t) \in \mathbb{R}^{F\times T}$ as input features, which has lower variance and is defined as following.

$$\tilde{Y}_{cprs}(f,t) = |Y(f,t)|^{0.5} \tag{2}$$

To better extract the inherent information of speech, we fist project input features into two different latent representations. The projection can be formulated as following (the index f and t are dropped from now on):

$$\tilde{Y}_{pre} = \Gamma(\tilde{Y}_{cprs}; \theta_{pre}) \tag{3}$$

$$[Y_{sub}, Y_{full}] = [W_{sub}, W_{full}] \circledast \tilde{Y}_{pre} \tag{4}$$

where $\tilde{Y}_{pre} \in \mathbb{R}^{C_{pre}\times F\times T}$ is the primary features extracted by a primary handling unit Γ with parameters set θ_{pre} where the C_{pre} denotes the feature dimension, $W_{sub}, W_{full} \in \mathbb{R}^{C_{in}\times 1\times 1\times C_{pre}}$ are two 2D convolution kernels of size $(1,1)$, $Y_{sub}, Y_{full} \in \mathbb{R}^{C_{in}\times F\times T}$ denote two latent features for sub-band and full-band processing respectively, $\circledast$ represents the convolution operation. The primary handling unit contains of N_{pre} Conv2D blocks, and each of them has a Conv2D layer followed by a batch normalization (BN) layer and a parametric rectified linear unit (PReLU) activation function.

2.4. Sub-band module

Unlike most existing sub-band processing methods in which single model is employed for all sub-bands, we aim to design one module for each sub-band individually. To further reduce parameters, we design a simple yet powerful sub-band processing module, the weighted progressive convolutional module (WPCM). Figure 1:(B) shows the design of WPCM. The WPCM consists of stacked local convolutional blocks (LCBs), and each of them contains a Conv2D block and a prediction path. The prediction path take the output of the Conv2D block as input and predict a intermediate magnitude, which is a *1x1*-Conv2D

layer . Those intermediate magnitudes are then summed up with Gaussian weights which defined as following:

$$w_i = \alpha e^{-\frac{(i-L_{sub})^2}{2\sigma^2}}, i = 1, 2, ..., L_{sub} \quad (5)$$

where L_{sub} indicates the depth of WPCM, w_i is the weight applied for the output of i-th LCB, and α, σ are scalar factors controlling the range of weights. The motivation of weighted sum is that we wish the WPCM would progressively refine the features. Since the shallow features are coarser, we place less emphasis on shallow layers and focus more on deeper layers.

Specifically, given a latent representation Y_{sub}, we first split it into K sub-bands $Y_{sub}^{(k)} \in \mathbb{R}^{C_{in} \times F' \times T}, k = 1, 2, ..., K$ where $F' = F/K$. For each $Y_{sub}^{(k)}$, a WPCM is employed to predict rough sub-band magnitude $O_{sub}^{(k)} \in \mathbb{R}^{F' \times T}$ of clean speech from local perspective. Outputs of all WPCMs are then concatenated along frequency dimension to generate final rough magnitude $O_{sub} \in \mathbb{R}^{F \times T}$.

2.5. Full-band module

The preceding sub-band module is mainly focus on local dependence, while the global dependence (e.g., harmonic) is also critical for noise reduction. Motivated by the commonly used convolutional recurrent network (CRN) structure, we propose a novel asymmetric CRN which consists of three module: encoder, dual-path RNN and decoder. The encoder consists of stacked Conv2D blocks, which is used to encode the features in higher dimensional space and reduces the dimension of frequency. To capture long-term dependence over frequency and time dimension, a gated recurrent unit (GRU) based dual-path RNN is employed, which has two dual-path GRU (DPGRU) blocks, and each of them is followed by a BN layer and a PReLU. Each DPGRU block contains a bi-directional GRU for modeling full-band dependence and a unidirectional GRU for capturing long-term dependence over time dimension. Unlike the existing CRN in which the decoder is symmetrical to the encoder, a weighted progressive full-band decoder (WPFD) is proposed to model global dependence over full-band.

The motivation of WPFD is that we wish the model could focus more on global features. In that sense, based on the extracted features by encoder and dual-path RNN, the target speech can be progressively predicted during decoder stage from global perspective. As shown in Figure 1:(C), the WPFD consists of stacked global convolutional blocks (GCBs), and each of them contains a Conv1D block and a predict path. The Conv1D block has a Conv1D layer followed by a BN layer and a PReLU, and the predict path is a *1x1*-Conv1D layer. The preceding Conv1D operations treat the frequency dimension as the features dimension and thus model the full-band pattern. Similar to the sub-band module, outputs of all predict paths are summed up with the Gaussian weights defined as Equation 5.

Specifically, given a latent representation Y_{full}, an encoder cascaded with dual-path RNN is employed to capture global dependence and outputs basic full-band features $Fea_{full} \in \mathbb{R}^{(C_{mid} \times F_{mid}) \times T}$. The WPFD takes Fea_{full} as input and progressively estimates the clean magnitude $O_{full} \in \mathbb{R}^{F \times T}$.

2.6. Features fusion

To reconstruct the target speech waveform, both magnitude and phase are required. However, both sub-band module and full-band module perform speech enhancement in the magnitude domain without considering phase. To this end, we design a simple fusion module, which consists of stacked Conv2D blocks, to reconstruct phase information. Specifically, We first concatenate O_{sub}, O_{sull} and input features $\tilde{Y}_{cprs}$ and pass the features into fusion module to predict a complex ideal ratio mask (cIRM [21]) $M \in \mathbb{C}^{F \times T}$. The mask is then applied to the compressed TF representations of noisy speech. We finally uncompress the masked TF representations and reconstruct the target waveform using inverse-STFT (iSTFT).

Table 1: *The effect of the number of sub-bands K.*

K	#para(M)	PESQ	CSIG	CBAK	COVL
noisy	-	1.99	3.37	2.49	2.66
1	2.903	2.995	4.314	3.481	3.662
2	2.964	**3.080**	4.374	**3.551**	3.755
4	3.088	3.066	4.399	3.527	3.761
8	3.339	3.072	**4.409**	3.538	**3.768**
16	3.862	3.053	4.395	3.515	3.749

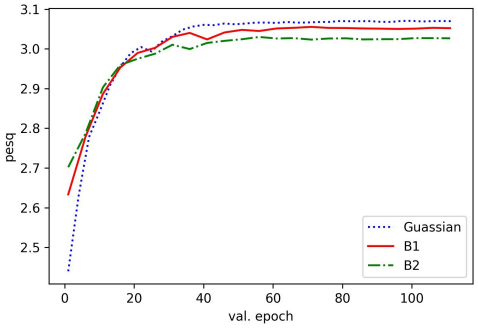

Figure 2: *The training procedure in terms of PESQ-WB.*

3. Experiments

3.1. Datasets

We evaluate our proposed model on the VoiceBank+Demand dataset [22] and the DNS Challenge dataset [15, 16]. The VoiceBank+Demand dataset contains 11572 training noisy and clean speech pairs and 824 test speech pairs, which is relatively small but widely used. The DNS Challenge dataset is a large speech dataset for the task of real-time speech enhancement. We totally generate 560 hours of training data and 16 hours of validation data (with 30% reverb). All waveform are resampled at 16 kHz.

3.2. Experimental setups

The input waveform is converted into STFT domain using a 20ms Hanning window with 10ms hop-size. Adam optimizer is used. In the preprocessing module, the number of filters (NOF) of the Conv2D layers are {16,16,32,32} with kernel size (KS) {(5,1),(1,5),(5,5),(3,3)} and stride 1, and the feature dimension C_{pre} is set to 16. In the sub-band module, the NOF of the Conv2D layers are {32,32,64,64} with KS {(3,1),(1,3),(3,3),(3,3)} and stride 1. In the full-band module, the NOF of the Conv2D layers in encoder are {32,32,64,64,64,96,128} with KS {(5,1),(1,5),(5,5),(4,3),(5,3),(3,5),(3,3)} and stride {(1,1),(1,1),(2,1),(2,1),(2,1),(2,1),(1,1)}, the NOF of the Conv1D layers in decoder are {512, 256, 256} with KS {1, 3, 3} and stride 1, and the hidden sizes of two DPGRU are {128,128} and {256,128} over frequency and time dimension respectively. The Gaussian

Table 2: *The effect of the sub-band and full-band modules.*

model	#para(M)	PESQ-WB	SSNR
B3	0.260	2.623	8.286
B4	3.001	3.065	9.088
ours(K=2)	2.964	**3.080**	**9.348**

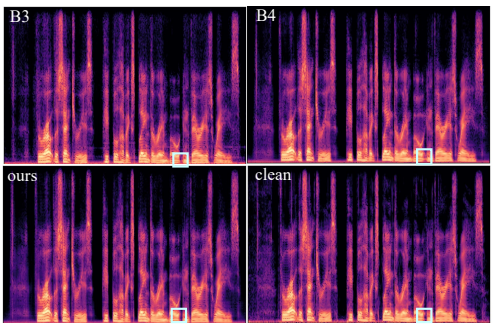

Figure 3: *The spectral difference between sub-band and full-band modules.*

weights are bounded to [0.12, 0.88]. In the feature fusion stage, the NOF of the Conv2D layers are {16,16,32,16,2} with KS {(5,1),(1,5),(3,3),(3,3),(1,1)} and stride 1. All neural network layers are causal and the BN is only performed during training stage. For better noise reduction, only 10ms future speech is seen in the last layer of the primary handling unit. In this case, the total algorithmic latency (frame length + stride time + any look ahead) is $20 + 10 + 10 = 40ms$.

We train the proposed model with the phase-aware loss to achieve noise suppression and perceptual quality promotion, which is defined as:

$$loss = \frac{1}{2}(||X_{cprs}| - |\hat{X}_{cprs}|| + |X_{cprs} - \hat{X}_{cprs}|) \quad (6)$$

where $X_{cprs}, \hat{X}_{cprs}$ are the target and estimated compressed TF representations of clean speech respectively.

3.3. Results and analysis

We evaluate our proposed method on six commonly used metrics: PESQ, CSIG, SBAK, COVL, STOI, segmental SNR (SSNR). Each metric is better if the score is higher.

We have conducted extensive ablation experiments on the VoiceBank+Demand dataset to optimize the configuration of proposed model and verify its effectiveness. Since the non-causal setting indicates a upper bound of performance, we use the non-causal implementations for all ablation experiments.

We first compare the models with different number of sub-bands and the results are shown in Table 1. The way of sub-band processing ($K > 1$) improves the performance. However, increasing the number of sub-bands does not bring a significant promotion but increase several parameters. Therefore, we select $K = 2$ as a trade-off between performance and model size. We also verify the effectiveness of the Gaussian weights used in the sub-band and full-band modules by comparing with two baselines: **B1**: the average weight is used; **B2**: only the last predict path is used. Figure 2 shows the training procedure in terms of the PESQ-WB on evaluation set. From the figure, we can find that the proposed Gaussian wights promote the upper bound of performance. In addition, we compare the proposed model with other two baselines: **B3**: sub-band module; **B4**: full-band module with more parameters, as shown in Table 2. The results show that the fusion of sub-band and full-band improve performance. As shown in Figure 3, we further compare the spectral difference between sub-band and full-band modules. The spectrogram of **B3** is rougher and the **B4** produces residual noise in some local region (see the region bounded by the rectangular box). The spectrogram of proposed model is more close to the target and has less residual noise.

Table 3: *Comparison with other state-of-the-art methods on the VoiceBank+Demand dataset under causal implementation.*

model	#para(M)	Causal	PESQ-WB
noisy	-	-	1.99
RNNoise[23]	0.06	√	2.29
PercepNet[24]	8	√	2.73
DCCRN[9]	3.7	√	2.68
DCCRN+[10]	3.3	√	2.84
ours	3.1	√	**2.905**

Table 4: *Comparison with other state-of-the-art methods on the DNS test set (Without Reverb) [15]. $\mathcal{L}$ denotes total algorithmic latency (frame length + stride time + any look ahead). PW and PN denotes PESQ-WB and PESQ-NB respectively.*

model	#para (M)	$\mathcal{L}$ (ms)	PW	PN	STOI
noisy	-	-	1.582	2.454	91.52
NSNet[25]	5.1	40	2.145	2.873	94.47
DTLN[26]	1.0	40	-	3.04	94.76
Subband Model[18]	1.3	80	2.369	3.052	94.24
FullSubNet[19]	5.6	80	2.777	3.305	96.11
DCCRN+[10]	3.3	40	-	3.33	-
PP_WO-MALE[27]	3.4	40	-	3.25	95.36
ours	3.1	40	**2.8004**	**3.354**	**96.44**

We also compare our proposed model with other state-of-the-art (SOTA) methods on both the VoiceBank+Demand and the DNS challenge datasets, and the results are shown in Table 3 and Table 4. It is obvious that our proposed method is superior to other methods with smaller model size. It is worth noting that the FullSubNet is also a full-band and sub-band fusion model. Our proposed model outperforms the FullSubNet with less parameters and less latency, which demonstrates the proposed fusion network is more lightweight and effective.

Regarding the computational complexity, the Real-Time Factor (RTF) of the proposed model is 0.180 on a single-core CPU clocked at 2.4GHz of Intel Xeon CPU E5-2620 v3, which obviously meets the real-time requirement.

4. Conclusion

In this paper, we proposed a lightweight full-band and sub-band fusion network for real-time speech enhancement which is simple and robust. The network first projects the input features into two latent spaces, and then performs speech enhancement progressively from local and global perspectives, respectively. Both local features and global features are finally merged with a lightweight fusion module. Experimental results show that the proposed method outperforms other state-of-the-art methods, especially other full-band and sub-band based models, with fewer parameters and lower latency.

5. References

[1] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2007.

[2] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, "Densely connected progressive learning for lstm-based speech enhancement," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5054–5058.

[3] D. Takeuchi, K. Yatabe, Y. Koizumi, Y. Oikawa, and N. Harada, "Real-time speech enhancement using equilibriated rnn," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 851–855.

[4] I. Fedorov, M. Stamenovic, C. Jensen, L.-C. Yang, A. Mandell, Y. Gan, M. Mattina, and P. N. Whatmough, "TinyLSTMs: Efficient Neural Speech Enhancement for Hearing Aids," in *Proc. Interspeech 2020*, 2020, pp. 4054–4058.

[5] N. Mamun, S. Khorram, and J. H. Hansen, "Convolutional Neural Network-Based Speech Enhancement for Cochlear Implant Recipients," in *Proc. Interspeech 2019*, 2019, pp. 4265–4269.

[6] A. Pandey and D. Wang, "Dense cnn with self-attention for time-domain speech enhancement," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 1270–1279, 2021.

[7] K. Tan and D. Wang, "Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6865–6869.

[8] K. Tan, X. Zhang, and D. Wang, "Real-time speech enhancement using an efficient convolutional recurrent network for dual-microphone mobile phones in close-talk scenarios," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5751–5755.

[9] 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.

[10] 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.

[11] J. Hou and S. Zhao, "A real-time speech enhancement algorithm based on convolutional recurrent network and wiener filter," in *2021 IEEE 6th International Conference on Computer and Communication Systems (ICCCS)*. IEEE, 2021, pp. 683–688.

[12] X. Le, H. Chen, K. Chen, and J. Lu, "DPCRN: Dual-Path Convolution Recurrent Network for Single Channel Speech Enhancement," in *Proc. Interspeech 2021*, 2021, pp. 2811–2815.

[13] K. Wang, B. He, and W.-P. Zhu, "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, 2021, pp. 7098–7102.

[14] K. Oostermeijer, Q. Wang, and J. Du, "Lightweight Causal Transformer with Local Self-Attention for Real-Time Speech Enhancement," in *Proc. Interspeech 2021*, 2021, pp. 2831–2835.

[15] C. K. Reddy, H. Dubey, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "Icassp 2021 deep noise suppression challenge," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6623–6627.

[16] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun, P. Rana, S. Srinivasan, and J. Gehrke, "The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Testing Framework, and Challenge Results," in *Proc. Interspeech 2020*, 2020, pp. 2492–2496.

[17] 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.

[18] X. Li and R. Horaud, "Online Monaural Speech Enhancement Using Delayed Subband LSTM," in *Proc. Interspeech 2020*, 2020, pp. 2462–2466.

[19] 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.

[20] A. Li, C. Zheng, R. Peng, and X. Li, "On the importance of power compression and phase estimation in monaural speech dereverberation," *JASA Express Letters*, vol. 1, no. 1, p. 014802, 2021.

[21] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[22] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating rnn-based speech enhancement methods for noise-robust text-to-speech." in *SSW*, 2016, pp. 146–152.

[23] J.-M. Valin, "A hybrid dsp/deep learning approach to real-time full-band speech enhancement," in *2018 IEEE 20th international workshop on multimedia signal processing (MMSP)*. IEEE, 2018, pp. 1–5.

[24] J.-M. Valin, U. Isik, N. Phansalkar, R. Giri, K. Helwani, and A. Krishnaswamy, "A Perceptually-Motivated Approach for Low-Complexity, Real-Time Enhancement of Fullband Speech," in *Proc. Interspeech*, 2020, pp. 2482–2486.

[25] Y. Xia, S. Braun, C. K. Reddy, H. Dubey, R. Cutler, and I. Tashev, "Weighted speech distortion losses for neural-network-based real-time speech enhancement," in *Proc. ICASSP*. IEEE, 2020, pp. 871–875.

[26] N. L. Westhausen and B. T. Meyer, "Dual-Signal Transformation LSTM Network for Real-Time Noise Suppression," in *Proc. Interspeech*, 2020, pp. 2477–2481.

[27] X. Zhang, X. Ren, X. Zheng, L. Chen, C. Zhang, L. Guo, and B. Yu, "Low-Delay Speech Enhancement Using Perceptually Motivated Target and Loss," in *Proc. Interspeech 2021*, 2021, pp. 2826–2830.