FullSubNet+: CHANNEL ATTENTION FULLSUBNET WITH COMPLEX SPECTROGRAMS FOR SPEECH ENHANCEMENT

Jun Chen[1,†], *Zilin Wang*[1], *Deyi Tuo*[2], *Zhiyong Wu*[1,3,∗], *Shiyin Kang*[2], *Helen Meng*[1,3]

[1] Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[2] Huya Inc., Guangzhou, China
[3] The Chinese University of Hong Kong, Hong Kong SAR, China

{y-chen21, wangzl21}@mails.tsinghua.edu.cn, {tuodeyi, kangshiyin}@huya.com, {zywu,hmmeng}@se.cuhk.edu.hk

ABSTRACT

Previously proposed FullSubNet has achieved outstanding performance in Deep Noise Suppression (DNS) Challenge and attracted much attention. However, it still encounters issues such as input-output mismatch and coarse processing for frequency bands. In this paper, we propose an extended single-channel real-time speech enhancement framework called FullSubNet+ with following significant improvements. First, we design a lightweight multi-scale time sensitive channel attention (MulCA) module which adopts multi-scale convolution and channel attention mechanism to help the network focus on more discriminative frequency bands for noise reduction. Then, to make full use of the phase information in noisy speech, our model takes all the magnitude, real and imaginary spectrograms as inputs. Moreover, by replacing the long short-term memory (LSTM) layers in original full-band model with stacked temporal convolutional network (TCN) blocks, we design a more efficient full-band module called full-band extractor. The experimental results in DNS Challenge dataset show the superior performance of our FullSubNet+, which reaches the state-of-the-art (SOTA) performance and outperforms other existing speech enhancement approaches.

Index Terms— speech enhancement, multi-scale time sensitive channel attention, phase information, full-band extractor

1. INTRODUCTION

Single-channel (monaural) speech enhancement methods remove background noise from single-channel noisy audio signals, aiming to improve the quality and intelligibility of the speech. Recently, deep learning-based methods have achieved promising results, especially in dealing with non-stationary noise under challenging conditions (e.g., low signal-to-noise ratio (SNR), reverberation). Noisy speech can be enhanced by neural networks either in time-domain or frequency-domain. The time-domain approaches [1–3] predict a clean speech waveform directly from the corresponding noisy speech waveform. As another main-stream, the frequency-domain approaches [4–7] generally take the noisy spectral feature as the input, and their learning target is the clean spectral feature or the mask (e.g., Ideal Binary Mask [8], Ideal Ratio Mask [9], complex Ideal Ratio Mask (cIRM) [10]). Overall, since the frequency domain signal has more distinguishable features for model learning [11], the frequency-domain approaches dominate the most SOTA speech enhancement methods.

---

† Work conducted when the first author was intern at Huya Inc.
∗ Corresponding author.

FullSubNet [12] has been proposed for single-channel frequency-domain speech enhancement. It consists of a full-band model and a sub-band model [13], and performs joint optimization after connecting the two in series. The full-band model extracts the global spectral information and the long-distance cross-band dependencies. Subsequently, taking the concatenation of the sub-band units and the output of full-band model as the input, the sub-band model focuses on the local spectral pattern together with the extracted global information and models signal stationarity. By means of this way, the FullSubNet can capture the global context [4, 5] while retaining the ability to model signal stationarity and attend the local spectral patterns. As a result, FullSubNet achieves excellent results on DNS Challenge dataset [14].

Despite the competitive performance, FullSubNet is still not perfect. It has been pointed out that local patterns in the spectrogram are often different in each frequency band [15]. However, in FullSubNet, the network takes the original whole frequency band as the input of full-band model, which degrades network's discriminative ability among different frequency bands in the input spectrogram. Moreover, FullSubNet takes cIRM as its predicting target while being fed with only the magnitude spectrum which contains only magnitude information and lacks of the phase information. We argue in this work that the mismatching of input and output information is unreasonable, which will cause that the phase of enhanced speech deviates significantly with serious interference and limit the upper bound of performance.

In this paper, to address the above issues of FullSubNet, we propose a framework with channel attention on multiple input spectrograms, that is called FullSubNet+. We propose a lightweight MulCA module to give different attention degrees to different frequency bands of the input spectrogram. At the same time, we design a full-band extractor composed of TCN blocks [16] instead of the full-band model to reduce model size and save computation cost. Furthermore, we deploy two additional independent full-band extractors to process the real and imaginary spectrograms to make full use of the phase information in noisy speech. Experimental results show that FullSubNet+ prominently outperforms FullSubNet and other SOTA methods on the DNS challenge dataset, which indicates the effectiveness of MulCA module and the use of noisy speech phase information for speech enhancement.

2. METHODOLOGY

This paper focuses only on the denoising task in the short-time fourier transform (STFT) domain, and the target is to suppress noise and recover the speech signal (the reverberant signal received at

ICASSP 2022

(a) FullSubNet+ diagram

(b) System flowchart on one branch of the model
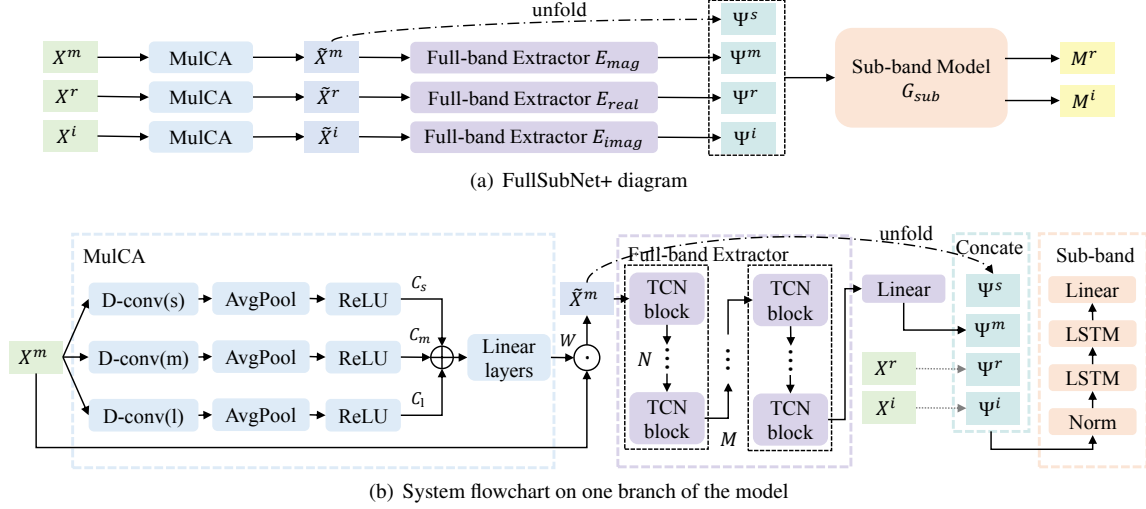
Fig. 1. (a) The overall diagram of the proposed FullSubNet+. The model contains three branches consisting of an MCA module and a Full-band Extractor, a concatenation operation, and a Sub-band Model. (b) The System flowchart on one branch of the model, which contains specific details of the MulCA module, Full-band Extractor, Concatenation operation, and Sub-band Model.

the microphone). To accomplish this task, we propose a framework called FullSubNet+ with channel attention on multiple input spectrograms. The model architecture is shown in Fig.1(a), which mainly consists of three main parts: three parallel MulCA modules, three parallel full-band extractors called E_{mag}, E_{real}, E_{imag} respectively, and a sub-band model G_{sub}.

The proposed model takes the magnitude, real and imaginary spectrograms $\mathbf{X}^m$, $\mathbf{X}^r$, $\mathbf{X}^i \in \mathbb{R}^{F \times T}$ as inputs, where F and T represent the total number of frequency bins and frames respectively. The three parallel MulCA modules weight the above input spectrograms individually. Then the outputs of MulCA are fed to the next Full-band extractor and the weighted magnitude spectrogram $\widetilde{\mathbf{X}}^m$ will also be used to unfold for sub-band units $\mathbf{\Psi}^s$. Next, the full-band extractors E_{mag}, E_{real} and E_{imag} extract the global spectral information from the weighted spectrograms, and then their outputs $\mathbf{\Psi}^m$, $\mathbf{\Psi}^r$ and $\mathbf{\Psi}^i$ are concatenated with the sub-band units $\mathbf{\Psi}^s$. Finally, taking the concatenation as input, the sub-band model G_{sub} predicts the learning target cIRM $\mathbf{M}^r$ and $\mathbf{M}^i$. In addition, the proposed model also supports output delay: to infer $M^r_{f,t-\tau}$, $M^i_{f,t-\tau}$, the future time steps are provided in the input sequence. In the following these modules are described in detail.

2.1. Multi-scale Time Sensitive Channel Attention

It has been pointed out in the previous work [15] that the local patterns in the spectrogram are often different in each frequency band: the lower frequency band tends to contain high energies, tonalities as well as long duration sounds, while the higher frequency band may have low energies, noise and rapidly decaying sounds. Inspired by this, we regard different frequency bins as different channels, and give them different weights by using Channel Attention. Specifically, we use a vector with dimension F to represent the weight assigned to each frequency bin. Then we perform a dot product operation on this weight vector with the spectrogram after a python broadcast to achieve the weighting:

$$\begin{aligned} \mathbf{W} &= [W_0, \cdots, W_f, \cdots, W_{F-1}]^T \in \mathbb{R}^F, \\ \widetilde{\mathbf{X}} &= \mathbf{X} \odot \mathbf{W}. \end{aligned} \quad (1)$$

where $\mathbf{W}$ represents the weight vector, and $\mathbf{X}$, $\widetilde{\mathbf{X}} \in \mathbb{R}^{F \times T}$ denote the spectrogram before and after weighting respectively. $\odot$ denotes the dot product. Accordingly, the model will focus on frequency bands that play more significant roles in noise reduction.
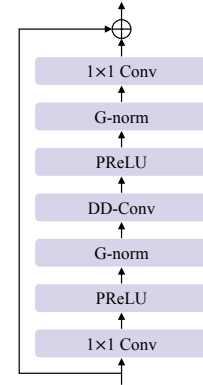
Fig. 2. The details of TCN block. The "DD-Conv" indicates a dilated depth-wise separable convolution. The "G-norm" is a global layer normalization [17].

A lightweight MulCA module is proposed to calculate the weights assigned to different frequency bins. Details of the MulCA module are shown in Fig.1(b). Motivated by SpEx [18], we use several parallel 1-D depthwise convolutions with different kernel size, k_1(small), k_2(middle) and k_3(large), to extract the different time-scale features of each frequency bin along the time axis. The convolutions are followed by average pooling layer and the rectified linear unit (ReLU) activation function to produce the different time-scale features $\mathbf{C}_s$, $\mathbf{C}_m$, $\mathbf{C}_l \in \mathbb{R}^F$. Then, a fully connected layer is adopted to fuse three features into the fusion feature $\mathbf{C}$. Eventually, following the operation in [19], we stack another two fully connected layers to learn the weights $\mathbf{W}$ from the fusion feature $\mathbf{C}$.

In FullSubNet+, three MulCA modules are adopted to weight

$\mathbf{X}^m$, $\mathbf{X}^r$, $\mathbf{X}^i$ respectively. Then the weighted spectrograms $\widetilde{\mathbf{X}}^m$, $\widetilde{\mathbf{X}}^r$, $\widetilde{\mathbf{X}}^i \in \mathbb{R}^{F \times T}$ are fed to full-band extractors E_{mag}, E_{real}, E_{imag} respectively.

2.2. Full-band Extractor

As a model with powerful temporal sequence modeling ability, TCN blocks have been widely used in speech separation [17, 20, 21] and target speaker extraction tasks [18] recently. Fig.2 shows the structure of the TCN block, which consists of three main components, namely input 1×1 $Conv$, depthwise dilated convolution ($DD - Conv$) and output 1×1 $Conv$. Parametric ReLU (PReLU) activation function and normalization layers are inserted between adjacent convolutions. Residual connection is applied to alleviate gradient vanishing problem.

In order to further reduce model size and save computation cost, we use stacked TCN blocks to replace LSTMs in full-band model from the original FullSubNet, and name it as full-band extractor. In detail, a group of stacked TCN blocks is used to replace a layer of LSTM. Similar to Conv-TasNet [17], we stack the TCN blocks by exponentially increasing the dilation factor in each group to capture the features of speech signals with long-range dependence over the full-band. As shown in Fig.1(b), for each full-band extractor, there are M groups of N TCN blocks, where M and N are hyperparameters. A fully connected layer is deployed after these stacked TCN blocks.

Full-band extractors extract global context information and output three spectral embeddings $\mathbf{\Psi}^m$, $\mathbf{\Psi}^r$ and $\mathbf{\Psi}^i$ with the same size as their inputs $\widetilde{\mathbf{X}}^m$, $\widetilde{\mathbf{X}}^r$ and $\widetilde{\mathbf{X}}^i$, which are expected to provide complementary information for the sub-band units $\mathbf{\Psi}^s$. Three spectral embeddings and sub-band units together serve as the input of G_{sub}.

2.3. Sub-band Model

For the purpose of learning the frequency-wise signal stationarity and maintaining the stability in model training, as shown in Fig.1(b), the sub-band model G_{sub} uses a structure consisting of two stacked unidirectional LSTM layers and one fully connected layer instead of stacked TCN blocks.

The sub-band model G_{sub} predicts the cIRM based on the weighted noisy sub-band units and the full-band extractors outputs. In detail, for each frequency f, we take a frequency bin vector $\widetilde{\mathbf{X}}_f^m \in \mathbb{R}^T$ and the $2 \times n$ frequency bin vectors adjacent to it in the frequency domain from the weighted magnitude spectrogram $\widetilde{\mathbf{X}}^m$ as a sub-band unit $\mathbf{\Psi}_f^s$:

$$\mathbf{\Psi}_f^s = [\widetilde{\mathbf{X}}_{f-n}^m, \cdots, \widetilde{\mathbf{X}}_f^m, \cdots, \widetilde{\mathbf{X}}_{f+n}^m] \in \mathbb{R}^{(2n+1) \times T}. \quad (2)$$

In addition, circular Fourier frequencies are used for boundary frequencies. The sub-band unit $\mathbf{\Psi}_f^s$ and the outputs of the full-band extractors, denoted as $\mathbf{\Psi}_f^m$, $\mathbf{\Psi}_f^r$, $\mathbf{\Psi}_f^i \in \mathbb{R}^T$ are concatenated with the sub-band unit $\mathbf{\Psi}_f^s$ to serve as the input $\mathbf{\Psi}_f$ to the G_{sub}:

$$\mathbf{\Psi}_f = [\mathbf{\Psi}_f^s, \mathbf{\Psi}_f^m, \mathbf{\Psi}_f^r, \mathbf{\Psi}_f^i]^T \in \mathbb{R}^{T \times (2n+4)}. \quad (3)$$

We have a total of F such independent input sequences with a dimension of $(T, 2n + 4)$. Each sequence contains: 1) time-varying signal stationarity and local spectrum patterns in sub-band units; 2) global information and phase information extracted from the magnitude, real and imaginary spectrograms. They complement each other. The sub-band model finally makes predictions based on these sequences.

3. EXPERIMENTS

3.1. Datasets

We trained and evaluated FullSubNet+ on a subset of the Interspeech 2021 DNS Challenge dataset. The clean speech set includes 562.72 hours of clips from 2150 speakers. The noise dataset includes 181 hours of 60000 clips from 150 classes. During model training, we use dynamic mixing [12] to simulate speech-noise mixture as noisy speech. Specifically, before the start of each training epoch, 75% of the clean speeches are mixed with the randomly selected room impulse response (RIR) from openSLR26 and openSLR28 [27] datasets. After that, the speech-noise mixtures are dynamically generated by mixing clean speeches and noise at a random SNR between -5 and 20 dB. The DNS Challenge provides a publicly available test dataset consisting of two categories of synthetic clips, namely without and with reverberations. Each category has 150 noise clips with a SNR distributed between 0 dB to 20 dB. We use this test set to evaluate the effectiveness of the model.

3.2. Training setup and baselines

We use Hanning window with window length of 32 ms (512-sample) and frame shift of 16 ms to transform the signals to the STFT domain. Adam optimizer is used with a learning rate of 1e-3. We set the τ described above to 2, which is equivalent to exploiting 32 ms information of the next two frames while enhancing the current frame. For the sub-band units, we set $n = 15$ as in [13]. During training, the input-target sequence pairs are generated as constant-length sequences, with the sequence length set to $T = 192$ frames.

For a fair comparison, the following two models use the same experimental settings as well as learning target (cIRM).

FullSubNet: The model consists of a full-band model and a sub-band model, each containing two layers of stacked LSTMs and one fully connected layer. The full-band model has 512 hidden units per LSTM layer, while the sub-band model has 384 hidden units per LSTM layer.

FullSubNet+: The number of channels in the MulCA module is 257 and the sizes of the kernels of the parallel 1-D depthwise convolutions are {3,5,10} respectively. For each full-band extractor, 2 groups of TCN blocks are deployed, each containing 4 TCN blocks with kernel size 3 and dilation rate {1,2,5,9}. The sub-band model contains 384 hidden units per LSTM layer. The total number of parameters in the model is 8.67 M.

3.3. Comparison with baseline and SOTA methods

Table 1 shows the performance of each model on the DNS challenge test dataset. In the table, "With Reverb" and "Without Reverb" refer to test sets with and without reverberation respectively. "Look Ahead" represents the length of future information. In the last two rows of Table 1, we compare the performance of FullSubNet+ with FullSubNet as a baseline. From the table, we can observe that FullSubNet+ has a significant performance improvement over FullSubNet in all metrics. This improvement shows that the use of the MulCA module to help the network focus on more discriminative frequency bands and the use of independent full-band extractor to extract real as well as imaginary spectrograms information are essential for speech enhancement.

Regarding the computational complexity, after using TCN blocks, the parameters of full-band extractor are reduced by 1.64 M. Moreover, compared with FullSubNet, the training and inference speed of FullSubNet+ on GPU has increased by nearly 18%. The

Table 1. The performance in terms of WB-PESQ [MOS], NB-PESQ [MOS], STOI [%], and SI-SDR [dB] on the DNS challege test dataset.

Model	Year	Look Ahead (ms)	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	9.033	1.582	2.454	91.52	9.07
DCCRN-E [22]	2020	37.5	-	3.077	-	-	-	3.266	-	-
PoCoNet [23]	2020	-	2.832	-	-	-	2.748	-	-	-
DCCRN+ [24]	2021	10	-	3.300	-	-	-	3.330	-	-
TRU-Net [25]	2021	0	2.740	3.350	91.29	14.87	2.860	3.360	96.32	17.55
CTS-Net [26]	2021	-	3.020	3.470	92.70	15.58	2.940	3.420	96.66	17.99
FullSubNet [12]	2021	32	3.063	3.581	92.93	16.09	2.813	3.403	96.17	17.44
FullSubNet+	2021	32	**3.218**	**3.666**	**93.84**	**16.81**	**2.982**	**3.504**	**96.69**	**18.34**

Table 2. Performance of WB-PESQ and SI-SDR in the ablation study 3.4.1 using the test set without reverberation. Here the backbone is the FullSubNet.

Models	Para (M)	MulCA	Full-input	Sub-input	WB-PESQ	SI-SDR
FullSubNet$^{\spadesuit}$	5.64	✗	✗	✗	2.813	17.44
FullSubNet$^{\heartsuit}$	5.75	✔	✔	✗	2.900	17.71
FullSubNet$^{\clubsuit}$	5.75	✔	✗	✔	2.912	17.78
FullSubNet$^{\diamondsuit}$	5.75	✔	✔	✔	**2.932**	**17.90**

Table 3. Performance of NB-PESQ, WB-PESQ and SI-SDR in the ablation study 3.4.2 using the test set without reverberation.

Models	MulCA	Phase-info	NB-PESQ	WB-PESQ	SI-SDR
FullSubNet+	✔	✔	**3.504**	**2.982**	**18.34**
− Phase Information	✔	✗	3.475	2.955	17.92
− MulCA Module	✗	✔	3.477	2.960	17.98
− Both	✗	✗	3.453	2.900	17.73

processing time for one STFT frame (32 ms) of FullSubNet+ on a 2.50 GHz Intel(R) Xeon(R) Platinum 8163 CPU is 18.33 ms, which clearly meets the real-time requirement.

In addition to showing the effectiveness of the improvements for the FullSubNet, we compare FullSubNet+ with some current SOTA methods [22–26] in Table 1. It can be seen that compared with the latest methods, our proposed FullSubNet+ shows superior performance on noise reduction tasks without reverberation and even more prominent performance improvement with reverberation. This indicates that, as an improvement of the FullSubNet, FullSubNet+ inherits the excellent modeling ability of the sub-band model for reverberation effects described in [12], and greatly improves the noise reduction ability on this basis.

3.4. Ablation Study

3.4.1. The effect of the MulCA

We first explore the impact of the MulCA module. Here the backbone is the FullSubNet and the MulCA module is used to weight. We note $\spadesuit$ as original FullSubNet with no spectrogram weighting, $\heartsuit$

as weighting only the input spectrogram of the full-band model, $\clubsuit$ as weighting only the input spectrogram of the sub-band model and $\diamondsuit$ as weighting both the input spectrograms of the full-band model and the sub-band model. Table 2 shows the effect of the model on the test set without reverberation in four cases, where "Full-input" and "Sub-input" refer to weighting the input of the full-band model and weighting the input of the sub-band model, respectively. As we can see from the table, after the weighting of MulCA, both the full-band model and the sub-band model have a gain in performance, and the best performance is obtained when both are used together. It can be concluded that the MulCA module can have a significant effect on the speech enhancement model with only a small increase in the number of parameters.

3.4.2. The effect of using phase information

We then study the effect of adding phase information on the model. From Table 3, we can observe the results of four variation models on the test set without reverberation, where "Phase-info" refers to the use of real and imaginary spectrograms branches to exploit the phase information of noisy speech. It can be found that the model gains in performance after using the features extracted from the real and imaginary spectrograms branches. In addition, this improvement does not conflict with the MulCA module, and the combination of both the use of phase information and MulCA module gives the best results. This experimental result confirms that the use of phase information is of great significance to speech enhancement based on sub-band methods, and its combination with MulCA module can achieve excellent results.

4. CONCLUSIONS

In this paper, we have proposed a new single-channel real-time speech enhancement framework named FullSubNet+. It adopts the multi-scale time sensitive channel attention (MulCA) to help the network focus on more discriminative frequency bands for noise reduction. We also design a full-band extractor which replaces the LSTMs in full-band model from original FullSubNet with stacked TCN blocks to extract suitable feature from all the magnitude, real and imaginary spectrograms. The effectiveness of our modules is demonstrated by ablation studies. We compared FullSubNet+ with other top-ranked methods on the DNS challenge, and the results show that the proposed FullSubNet+ outperforms other nowadays speech enhancement methods. Some of the enhanced audio clips are available at: https://hit-thusz-rookiecj.github.io/fullsubnet-plus.github.io/.

5. REFERENCES

[1] Dario Rethage, Jordi Pons, and Xavier Serra, "A wavenet for speech denoising," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5069–5073.

[2] Daniel Stoller, Sebastian Ewert, and Simon Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *arXiv preprint arXiv:1806.03185*, 2018.

[3] Ashutosh Pandey and DeLiang Wang, "Tcnn: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6875–6879.

[4] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui 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, 2014.

[5] DeLiang Wang and Jitong 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.

[6] Jitong Chen and DeLiang Wang, "Long short-term memory for speaker generalization in supervised speech separation," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4705–4714, 2017.

[7] Hakan Erdogan, John R Hershey, Shinji Watanabe, and Jonathan Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 708–712.

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

[9] Lei Sun, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "Multiple-target deep learning for lstm-rnn based speech enhancement," in *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*. IEEE, 2017, pp. 136–140.

[10] Donald S Williamson, Yuxuan Wang, and DeLiang 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.

[11] Dacheng Yin, Chong Luo, Zhiwei Xiong, and Wenjun Zeng, "Phasen: A phase-and-harmonics-aware speech enhancement network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 9458–9465.

[12] Xiang Hao, Xiangdong Su, Radu Horaud, and Xiaofei 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.

[13] Xiaofei Li and Radu Horaud, "Online monaural speech enhancement using delayed subband lstm," in *Proceedings of INTERSPEECH*, 2020.

[14] Chandan KA Reddy, Harishchandra Dubey, Kazuhito Koishida, Arun Nair, Vishak Gopal, Ross Cutler, Sebastian Braun, Hannes Gamper, Robert Aichner, and Sriram Srinivasan, "Interspeech 2021 deep noise suppression challenge," *arXiv preprint arXiv:2101.01902*, 2021.

[15] Naoya Takahashi and Yuki Mitsufuji, "Multi-scale multi-band densenets for audio source separation," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, pp. 21–25.

[16] Shaojie Bai, J Zico Kolter, and Vladlen Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[17] Yi Luo and Nima 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.

[18] Chenglin Xu, Wei Rao, Eng Siong Chng, and Haizhou Li, "Spex: Multi-scale time domain speaker extraction network," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 28, pp. 1370–1384, 2020.

[19] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[20] Cunhang Fan, Jianhua Tao, Bin Liu, Jiangyan Yi, Zhengqi Wen, and Xuefei Liu, "End-to-end post-filter for speech separation with deep attention fusion features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1303–1314, 2020.

[21] Efthymios Tzinis, Shrikant Venkataramani, Zhepei Wang, Cem Subakan, and Paris Smaragdis, "Two-step sound source separation: Training on learned latent targets," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 31–35.

[22] Yanxin Hu, Yun Liu, Shubo Lv, Mengtao Xing, Shimin Zhang, Yihui Fu, Jian Wu, Bihong Zhang, and Lei Xie, "Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement," *Proc. Interspeech 2020*, pp. 2472–2476, 2020.

[23] Umut Isik, Ritwik Giri, Neerad Phansalkar, Jean-Marc Valin, Karim Helwani, and Arvindh Krishnaswamy, "Poconet: Better speech enhancement with frequency-positional embeddings, semi-supervised conversational data, and biased loss," *Proc. Interspeech 2020*, pp. 2487–2491, 2020.

[24] Shubo Lv, Yanxin Hu, Shimin Zhang, and Lei Xie, "Dccrn+: Channel-wise subband dccrn with snr estimation for speech enhancement," *arXiv preprint arXiv:2106.08672*, 2021.

[25] Hyeong-Seok Choi, Sungjin Park, Jie Hwan Lee, Hoon Heo, Dongsuk Jeon, and Kyogu 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.

[26] Andong Li, Wenzhe Liu, Chengshi Zheng, 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. 1829–1843, 2021.

[27] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev 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)*. IEEE, 2017, pp. 5220–5224.