Speech Enhancement with Fullband-Subband Cross-Attention Network

Jun Chen[1,†], Wei Rao[2,*], Zilin Wang[1], Zhiyong Wu[1,3,*], Yannan Wang[2],
Tao Yu[2], Shidong Shang[2], Helen Meng[1,3]

[1]Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[2]Tencent Ethereal Audio Lab, Shenzhen, China
[3]The Chinese University of Hong Kong, Hong Kong SAR, China

`y-chen21@mails.tsinghua.edu.cn, ellenwrao@tencent.com, zywu@sz.tsinghua.edu.cn`

Abstract

FullSubNet has shown its promising performance on speech enhancement by utilizing both fullband and subband information. However, the relationship between fullband and subband in FullSubNet is achieved by simply concatenating the output of fullband model and subband units. It only supplements the subband units with a small quantity of global information and has not considered the interaction between fullband and subband. This paper proposes a fullband-subband cross-attention (FSCA) module to interactively fuse the global and local information and applies it to FullSubNet. This new framework is called as FS-CANet. Moreover, different from FullSubNet, the proposed FS-CANet optimize the fullband extractor by temporal convolutional network (TCN) blocks to further reduce the model size. Experimental results on DNS Challenge - Interspeech 2021 dataset show that the proposed FS-CANet outperforms other state-of-the-art speech enhancement approaches, and demonstrate the effectiveness of fullband-subband cross-attention.

Index Terms: interactive fusion, fullband-subband cross-attention, speech enhancement

1. Introduction

The interference of environmental noise is one of the main factors that hinder the speech communication. Single-channel speech enhancement methods remove background noise from single-channel noisy audio signals, aiming to improve the quality and intelligibility of the speech, and have significant applications in hearing aids, audio communication and automatic speech recognition. Traditional speech enhancement methods use statistical signal theory to effectively suppress stationary noise, but they do not perform well under conditions of low signal-to-noise ratio (SNR). In the past few years, deep learning-based methods have achieved promising results, especially in dealing with non-stationary noise in complex acoustic environments. The neural networks can enhance the noisy speech either in frequency-domain or directly in time-domain. The frequency-domain approaches [1–4] generally take the noisy spectral feature as the input, and their learning target is the clean spectral feature or the mask such as Ideal Ratio Mask [5] and complex Ideal Ratio Mask (cIRM) [6]. On the other hand, The time-domain approaches [7–9] predict a clean speech waveform from the corresponding noisy speech waveform. Overall, the frequency domain approaches is more preferable considering the robustness of system and computational complexity [10].

The subband model [11] is a previously proposed work for frequency-domain speech enhancement. The work is performed in a subband style: the input of the model consists of one frequency together with several contextual frequencies. Thus the subband model learns the frequency-wise signal stationarity to discriminate between speech and stationary noise. However, since it can not model the global spectral pattern and exploit the long distance cross-band dependencies, it is difficult to recover clean speech on subband with low SNR. To solve this problem, the FullSubNet [12] introduces a fullband model extracting the global spectral information on the basis of the subband model, and performs joint optimization after connecting the two in series. By means of this way, the FullSubNet can capture the global spectral context [1,2] while retaining the ability to model signal stationarity and attend the local spectral patterns. Consequently, FullSubNet achieves excellent results on DNS Challenge dataset [13].

The success of FullSubNet effectively demonstrates the importance of global spectral information for the subband model. However, in FullSubNet, the relationship between fullband model and subband model is achieved by simply concatenating the output of fullband model and subband units. This concatenation method only supplements a small amount of global information for the subband units. It lacks the interaction between fullband information and subband information, which limits the potential of the FullSubNet.

To address the above issue of FullSubNet, this paper proposes the fullband-subband cross-attention (FSCA) module to interactively fuse the fullband and subband information and applies it to FullSubNet. Furthermore, a fullband extractor [14] composed of TCN blocks [15] is used to instead the fullband model in FullSubNet for reducing model size. This new network is called as FS-CANet. Experimental results on the DNS Challenge dataset show that our FS-CANet outperforms FullSubNet+ [14] and FullSubNet in terms of both number of parameters and performance. Furthermore, the FS-CANet also exceeds other state-of-the-art speech enhancement methods on the DNS Challenge - interspeech 2021 dataset. These experimental results also demonstrate that the fullband-subband cross-attention is an effective interactive fusion method.

2. FS-CANet

This paper only focuses on the denoising task in the STFT domain, and the target is to suppress noise and recover the reverberant speech signal (the reverberant signal received at the microphone). FS-CANet is proposed to interactively fuse the fullband and subband information by fullband-subband cross-attention module.

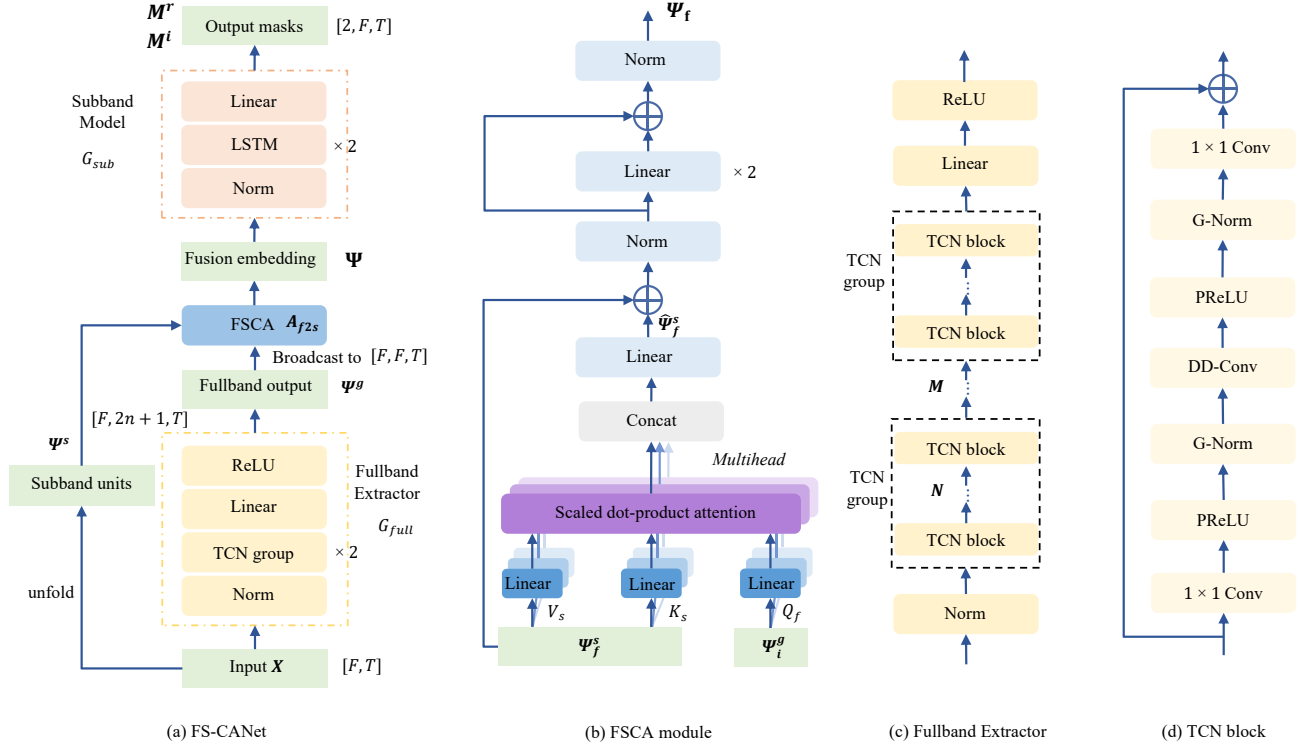The architecture of FS-CANet is shown in Fig.1(a), which

---

Figure 1: *(a) The overall diagram of the proposed FS-CANet. The model mainly contains a Fullband Extractor, a FSCA module and a Subband Model. (b) The details of the FSCA module, where "Multihead" refers to the multi-head attention. (c) The details of the Fullband Extractor. (d) The details of TCN block. The "DD-Conv" indicates a dilated depth-wise separable convolution. The "G-norm" is a global layer normalization [16].*

mainly consists of three main parts: a fullband extractor G_{full}, a FSCA module A_{f2s}, and a subband model G_{sub}. To begin with, the input $\mathbf{X} \in \mathbb{R}^{F \times T}$ is fed to the fullband extractor G_{full}, and additionally it is also the subband units $\mathbf{\Psi}^s$ after unfolding. Then the fullband extractor G_{full} extracts the global spectral information from the input spectrogram and outputs the fullband embedding $\mathbf{\Psi}^g$. After a python broadcast, $\mathbf{\Psi}^g$ is treated as the input of FSCA module A_{f2s} together with the subband units $\mathbf{\Psi}^s$. Next, A_{f2s} obtains the fusion embeddings $\mathbf{\Psi}$ after interactively fusing $\mathbf{\Psi}^s$ and $\mathbf{\Psi}^g$. Finally, taking the $\mathbf{\Psi}$ as input, the subband model G_{sub} predicts the learning target cIRM $\mathbf{M}^r$ and $\mathbf{M}^i$. In the following these modules are described in detail.

2.1. Fullband Extractor

The fullband extractor is an efficient fullband processing model, which is illustrated in [14]. Compared with the fullband model in FullSubNet, it has a better performance with a smaller number of parameters [14]. Therefore, the fullband extractor is used to replace the original fullband model in FullSubNet.

As a model with powerful temporal sequence modeling ability, TCN blocks have been widely used in speech separation [16–18] and target speaker extraction tasks [19] recently. Fig.1(d) shows the structure of the TCN block, which consists of three main components, namely input $1 \times 1\ Conv$, depthwise dilated convolution ($DD-Conv$) and output $1 \times 1\ Conv$. Parametric ReLU (PReLU) activation function and normalization layers are inserted between adjacent convolutions. Residual connection is applied to alleviate gradient vanishing problem. In fullband extractor, similar to Conv-TasNet [16], the TCN blocks are stacked by exponentially increasing the dila-

tion factor in each group to capture the features of speech signals with long-range dependence over the fullband. As shown in Fig.1(c), for the fullband extractor, there are M groups of N TCN blocks, where M and N are hyper-parameters. A fully connected layer and a ReLU activation function are deployed after these stacked TCN blocks.

The fullband extractor extracts fullband information and outputs the fullband embedding $\mathbf{\Psi}^g$ with the same size as its input $\mathbf{X}$, which is expected to provide complementary information for the subband units $\mathbf{\Psi}^s$. The fullband embedding and subband units together serve as the input of A_{f2s}.

2.2. Fullband-Subband Cross-Attention

Recently, the multi-head cross attention mechanism has shown a surprising potential in speech-related interactive fusion tasks like multi-modal active speaker detection [20] and speaker extraction [21]. Inspired by this, FSCA module is proposed for interactively fusing the fullband and subband information to replace the simple concatenation [12] in original FullSubNet.

2.2.1. The concatenation in FullSubNet

For each frequency f, we take a frequency bin vector $\mathbf{X}_f \in \mathbb{R}^T$ and the $2 \times n$ frequency bin vectors adjacent to it in the frequency domain from the weighted magnitude spectrogram $\mathbf{X}$ as a subband unit $\mathbf{\Psi}_f^s$:

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

In addition, circular Fourier frequencies are used for boundary frequencies.

According to the concatenation method of FullSubNet, each subband unit is concatenated with a frequency bin vector of the output of fullband extractor, denoted as $\mathbf{\Psi}_f^g \in \mathbb{R}^T$, to serve as the input $\widetilde{\mathbf{\Psi}}_f$ to the G_{sub}:

$$\widetilde{\mathbf{\Psi}}_f = [\mathbf{\Psi}_f^s, \mathbf{\Psi}_f^g] \in \mathbb{R}^{(2n+2)\times T}. \qquad (2)$$

In this way, each subband unit is actually supplemented with the output of fullband extractor in only one frequency domain dimension, which does not make an adequate use of the global information. Furthermore, it lacks the interaction of global and subband information. To cope with the above issues, we propose the FSCA module.

2.2.2. The details of Fullband-Subband Cross-Attention

Different from the existing work on frequency-domain self-attention [22], our FSCA module focuses on the interfusion of the output of the full-band model and the subband units to convey more fullband temporal and spectral information to the subband units. The structure of the FSCA module A_{f2s} is shown in Fig.1(b), which mainly consists of the multi-headed attention layer for interactive fusion and linear layers for further fusion. Besides, residual connections are applied in the FSCA module to alleviate the gradient vanishing problem.

The FSCA module A_{f2s} takes the fullband embeddings and subband units together as the input. Following section 2.2.1, we get a total of F subband units. For the purpose of making each subband unit interact with a fullband embedding, we obtain F fullband embeddings $\mathbf{\Psi}^g$ by python broadcast. These subband units and fullband embeddings serve as the input to the A_{f2s}. The FSCA module A_{f2s} processes these F pairs of subband units and fullband embeddings in parallel. For each subband unit $\mathbf{\Psi}_f^s$ and fullband embedding $\mathbf{\Psi}^g$, the $\mathbf{\Psi}_f^s$ is linearly converted to $\mathbf{V}_s, \mathbf{K}_s \in \mathbb{R}^{F\times T}$ while the $\mathbf{\Psi}^g$ is linearly converted to $\mathbf{Q}_g \in \mathbb{R}^{F\times T}$. A multi-head attention mechanism is carried out to achieve the interaction between fullband and subband information:

$$\widehat{\mathbf{\Psi}}_f^s = Multihead(\mathbf{Q}_g, \mathbf{K}_s, \mathbf{V}_s) \qquad (3)$$

where $Multihead(\cdot)$ denotes the multi-head attention and the $\widehat{\mathbf{\Psi}}_f^s$ represents the output with the same size as $\mathbf{\Psi}_f^s$. Through this interactive fusion, $\widehat{\mathbf{\Psi}}_f^s$ contains global temporal and spectral information at multiple dimensions that $\mathbf{\Psi}_f^s$ focus on. Then, the $\widehat{\mathbf{\Psi}}_f^s$ is added to $\mathbf{\Psi}_f^s$. Eventually, we stack another two linear layers with residual connection to further fuse them into the fusion embedding $\mathbf{\Psi}_f \in \mathbb{R}^{(2n+1)\times T}$. The total of F fusion embeddings output by the FSCA module are then fed to the subband model G_{sub}.

2.3. Subband Model

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

We obtain a total of F fusion embeddings after the interactive fusion of the FSCA module. Each fusion embedding contains the local spectral patterns as well as the global spectral information, both of which complement each other. All of the fusion embeddings are fed into the subband model with shared parameters in parallel. In subband model, the stacked LSTM layers learn the global and local frequency-wise signal stationarity

based on these fusion embeddings. Finally, the fully-connected layer outputs the cIRM as our learning target.

3. Experiments

3.1. Datasets

We trained and evaluated FS-CANet on a subset of the DNS Challenge - Interspeech 2021 dataset with 16 kHz sampling rate, which will be called as DNS Challenge dataset in the following sections. The clean speech set includes 562.72 hours of clips from 2150 speakers. The noise dataset includes 181 hours of 65302 clips from over 150 classes. During training, we used dynamic mixing to simulate speech-noise mixture as noisy speech to make full use of the dataset. To be specific, 75% of the clean speeches were mixed with the randomly selected room impulse response (RIR) from openSLR26 and openSLR28 [29] datasets before the start of each training epoch. After that, the speech-noise mixtures were dynamically generated by mixing clean speeches and noise at a random SNR between -5 and 20 dB. The DNS Challenge also provides a publicly available test dataset consisting of two categories of synthetic clips [30], namely without and with reverberations. Each category has 150 noise clips with a SNR distributed between 0 dB to 20 dB. We used this test set to evaluate the effectiveness of the model.

3.2. Training setup and baselines

We used Hanning window with frame length of 32 ms and frame shift of 16 ms to transform the signals to the STFT domain. Adam optimizer was used with a learning rate of 1e-3. For the subband units, we set $n = 15$ as in [11], which means 15 neighbor frequencies are taken on each side of each input frequency bin. During model training, the input-target sequence pairs were generated as constant-length sequences, with the sequence length set to $T = 192$ frames (approximately 3 s).

In order to verify the effectiveness of the proposed model, we compared the following models. All of the models used the same experimental settings as well as learning target (cIRM).

FullSubNet: The model consisted of a fullband model and a subband model, each containing two layers of stacked LSTMs and one fully-connected layer. The fullband model had 512 hidden units per LSTM layer, while the subband model had 384 hidden units per LSTM layer. In addition, compared with the subband model, the fullband model had an additional ReLU activation function after the fully-connected layer.

FullSubNet+: To further demonstrate the advantages of our model with a small number of parameters and excellent performance, we introduced FullSubNet+ [14] for comparison. Following the configuration in [14], the number of channels in the MulCA module was 257 and the sizes of the kernels of the parallel 1-D depthwise convolutions were {3,5,10} respectively. For each fullband extractor, 2 groups of TCN blocks were deployed, each containing 4 TCN blocks with kernel size 3 and dilation rate {1,2,5,9}. The subband model contained 2 LSTM layers with 384 hidden units for each layer.

FS-CANet: There were 2 groups of TCN blocks deployed in the fullband extractor, each containing 4 TCN blocks with kernel size 3 and dilation rate {1,2,5,9}. The multi-head attention in FSCA module had 8 attention heads. The subband model contained 2 LSTM layers with 384 hidden units per layer.

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

Model	Year	# 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	9.033	1.582	2.454	91.52	9.07
DCCRN-E [23]	2020	3.7	-	3.077	-	-	-	3.266	-	-
Conv-TasNet [24]	2020	5.08	2.750	-	-	-	2.730	-	-	-
PoCoNet [25]	2020	50	2.832	-	-	-	2.748	-	-	-
DCCRN+ [26]	2021	3.3	-	3.300	-	-	-	3.330	-	-
TRU-Net [27]	2021	0.38	2.740	3.350	91.29	14.87	2.860	3.360	96.32	17.55
CTS-Net [28]	2021	4.99	3.020	3.470	92.70	15.58	2.940	3.420	96.66	17.99
FullSubNet [12]	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	2022	4.21	**3.218**	**3.665**	**93.93**	**16.82**	**3.017**	**3.513**	**96.74**	**18.08**

Table 2: *Performance of WB-PESQ [MOS], STOI [%] and SI-SDR [dB] in the comparative study 3.4 using the test set without reverberation.*

Models	FSCA	Concat	WB-PESQ	STOI	SI-SDR
FS-CANet$_a$	✔	✗	**3.017**	**96.74**	**18.08**
FS-CANet$_c$	✗	✔	2.900	96.52	17.73
FS-CANet$_{ac}$	✔	✔	2.966	96.57	17.96

3.3. Performance Comparison

Table 1 shows the performance of different speech enhancement models on the DNS challenge dataset. In the table, "With Reverb" and "Without Reverb" refer to test sets with and without reverberation respectively. "# Para" represents the parameter amount of the model, which is measured in millions.

In the last three rows of Table 1, we compare the performances of FullSubNet, FullSubNet+, and the proposed FS-CANet. According to Table 1, FS-CANet outperforms the baseline FullSubNet in all evaluation metrics with a smaller number of parameters. In addition, Table 1 shows that the FullSubNet+, which is a improved version of FullSubNet, has better performance than FullSubNet. However, the FullSubNet+ also suffers from a large number of parameters and a complex model structure. In contrast, the FS-CANet not only outperforms FullSubNet+, but also is achieved with a smaller number of parameters and a simpler architecture.

To further analyse the proposed model, this paper also compares it with other state-of-the-art time domain and frequency domain speech enhancement methods [23, 25–28] on the DNS Challenge - Interspeech 2021 dataset in Table 1. It can be concluded that comparing with the latest methods, the proposed FS-CANet shows the superior performance on noise reduction tasks without reverberation and even more prominent performance improvement with reverberation. This indicates that the proposed FS-CANet inherits the excellent capabilities of the FullSubNet for reverberation effects described in [12], and greatly improves the noise reduction ability by fullband-subband cross-attention module.

3.4. Investigation of FSCA module

In order to investigate fullband-subband cross-attention (FSCA) method, we use FS-CANet as the backbone and conduct experiments with different fusion methods. FS-CANet$_a$ means original FS-CANet with FSCA module for interactive fusion. FS-

CANet$_c$ represents the FS-CANet with the concatenation between the fullband embedding and subband units as described in FullSubNet [12], and it does not contain FSCA module. FS-CANet$_{ac}$ denotes the FS-CANet that uses the above concatenation between fullband embedding and the fusion embeddings outputted by the FSCA module. Table 2 shows the results of the model on the test set without reverberation in three cases, where "FSCA" refers to the use of FSCA module while "Concat" refers to the use of the above concatenation approach.

According to Table 2, the performance of FS-CANet$_a$, FS-CANet$_{ac}$ are better than FS-CANet$_c$. This indicates that the proposed FSCA module can effectively improve the performance of the model in terms of both PESQ and SNR. In addition, the performance of FS-CANet$_a$ is better than that of FS-CANet$_c$, which shows that the FSCA module can effectively replace the role of the concatenation approach, and reduce the burden of the subsequent subband model by decreasing the input dimension. Moreover, It can be found that the performance of FS-CANet$_{ac}$ is not as good as FS-CANet$_a$. This is probably because the fusion embeddings has conflicts with the information contained in the original fullband embedding.

4. Conclusion and Future Work

This paper proposes a new single-channel speech enhancement framework named FS-CANet. It adopts a fullband-subband cross-attention (FSCA) module to interactively fuse the global and local information. This paper also deploys a fullband extractor composed of stacked TCN blocks to instead the fullband model in FullSubNet for reducing model size. Experimental results[1] demonstrate the effectiveness of the fullband-subband cross-attention module. This paper also compare FS-CANet with other state-of-the-art methods on the DNS challenge dataset, which shows that the superior performance of proposed FS-CANet.

This paper only considers incorporating fullband information into the subband information. In the future, we will explore bi-directional fusion of the subband and fullband features.

---

[1]Demo page: https://hit-thusz-rookiecj.github.io/FS-CANet

5. References

[1] 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, 2014.

[2] 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.

[3] J. Chen and D. 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.

[4] H. Erdogan, J. R. Hershey, S. Watanabe, and J. 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.

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

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

[9] A. Pandey and D. 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.

[10] D. Yin, C. Luo, Z. Xiong, and W. Zeng, "Phasen: A phase-and-harmonics-aware speech enhancement network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 9458–9465.

[11] X. Li and R. Horaud, "Online monaural speech enhancement using delayed subband lstm," in *Proceedings of INTERSPEECH*, 2020.

[12] 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.

[13] 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," *arXiv preprint arXiv:2101.01902*, 2021.

[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. [Online]. Available: https://arxiv.org/abs/2203.12188

[15] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[16] 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.

[17] C. Fan, J. Tao, B. Liu, J. Yi, Z. Wen, and X. 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.

[18] E. Tzinis, S. Venkataramani, Z. Wang, C. Subakan, and P. 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.

[19] C. Xu, W. Rao, E. S. Chng, and H. Li, "Spex: Multi-scale time domain speaker extraction network," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 28, pp. 1370–1384, 2020.

[20] R. Tao, Z. Pan, R. K. Das, X. Qian, M. Z. Shou, and H. Li, "Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3927–3935.

[21] W. Wang, C. Xu, M. Ge, and H. Li, "Neural speaker extraction with speaker-speech cross-attention network," *Proc. Interspeech 2021*, pp. 3535–3539, 2021.

[22] F. Dang, H. Chen, and P. Zhang, "Dpt-fsnet: Dual-path transformer based full-band and sub-band fusion network for speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6857–6861.

[23] 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," *Proc. Interspeech 2020*, pp. 2472–2476, 2020.

[24] Y. Koyama, T. Vuong, S. Uhlich, and B. Raj, "Exploring the best loss function for dnn-based low-latency speech enhancement with temporal convolutional networks," *arXiv preprint arXiv:2005.11611*, 2020.

[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," *Proc. Interspeech 2020*, pp. 2487–2491, 2020.

[26] S. Lv, Y. Hu, S. Zhang, and L. Xie, "Dccrn+: Channel-wise sub-band dccrn with snr estimation for speech enhancement," *arXiv preprint arXiv:2106.08672*, 2021.

[27] 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.

[28] 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.

[29] 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)*. IEEE, 2017, pp. 5220–5224.

[30] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun *et al.*, "The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," *arXiv preprint arXiv:2005.13981*, 2020.