Interspeech 2024
1-5 September 2024, Kos, Greece

Sub-PNWR: Speech Enhancement Based on Signal Sub-Band Splitting and Pseudo Noisy Waveform Reconstruction Loss

Yuewei Zhang[1,*], *Huanbin Zou*[2,*], *Jie Zhu*[1,†]

[1]Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China
[2]Tencent Video Cloud, Shanghai, China

yueweizhang@sjtu.edu.cn, 1518854421@alumni.sjtu.edu.cn, zhujie@sjtu.edu.cn

Abstract

Existing deep learning-based speech enhancement (SE) methods typically entail high computational complexity. In this paper, we propose to split the input audio into adjacent equally spaced sub-band signals by an analysis filter bank, and feed these sub-band signals into a SE model to recover the denoised sub-band signals. These denoised sub-band signals are then reconstructed back to the full-band signal by a synthesis filter bank. Meanwhile, we design a full-band information fusion module to complement the sub-band feature with full-band spectral information. We also devise a full-band spectrum prediction module to predict the target full-band spectrum, which assists model training. Additionally, a pseudo noisy waveform reconstruction (PNWR) loss is introduced for better SE performance. Experiments demonstrate that the proposed scheme reduces the computational volume by about half with nearly no performance loss. The final SE system (Sub-PNWR) outperforms the current advanced methods.

Index Terms: speech enhancement, signal sub-band splitting, full-band feature fusion, pseudo noisy waveform reconstruction

1. Introduction

Recently, the deep learning (DL) techniques have been extensively employed in the speech enhancement (SE) task. Mainstream DL-based SE methods commonly transform the time domain noisy signal into the time-frequency (TF) domain and then perform noise reduction on the noisy spectrum. Representative works comprise the approaches based on convolutional neural network (CNN) [1], recurrent neural network (RNN) [2], convolutional recurrent network (CRN) [3, 4], and transformer [5, 6].

However, during the practical deployment process of algorithms, the computational resources allocated to the SE task are often limited. Therefore, exploring strategies to make the current DL-based SE methods more computationally lightweight is quite valuable. Notably, the sub-band processing-based methods have advantages in terms of computational complexity. In [7, 8], the input noisy spectrum is divided into multiple overlapped sub-bands, which are subsequently processed in parallel by a sub-band model with shared parameters. This effectively reduces the model size and computational complexity. Meanwhile, considering that the sub-band model cannot capture the global spectral features, another full-band model is introduced and serially connected with the sub-band model. Nevertheless, the utilization of the full-band model results in a significant computational burden for [7, 8]. As a resolution, Inter-SubNet [9] discards the full-band model. It proposes a lightweight module to extract global spectral information and incorporate it into the sub-band model. Additionally, [10] analyzes the shortcomings of employing one unique sub-band model to process all sub-band spectrum, and proposes to separately process the sub-bands using multiple lightweight sub-band models.

By analyzing existing methods based on sub-band processing, it becomes evident that they all have to separately process each input sub-band spectrum in parallel. This inevitably becomes a critical computational bottleneck for the overall model. In this paper, we propose a computationally lightweight approach to process all sub-bands simultaneously. Unlike the conventional overlapped sub-band splitting method in the TF spectrum domain, we split the audio waveform into several shorter time domain sub-band signals. Specifically, we employ an analysis filter bank to divide the raw input audio into multiple adjacent equally spaced signals. Consequently, these sub-band signals have shorter sampling point lengths, respectively containing the spectral information of the original full-band spectrum distributed across different frequency bands. We concatenate the respective sub-band spectrum of these sub-band signals along the channel dimension and feed the concatenated results into a CRN-based SE network. Hence, the decreased size of the sub-band spectrum input in the frequency dimension can effectively decrease the computational complexity of the SE model. In addition, we modify the prediction target of the SE model to include the estimation for all the sub-band target signals. Compared to the previous sub-band processing schemes that only focused on reconstructing the target full-band spectrum, this further promotes the performance of sub-band processing.

To incorporate the full-band spectral information for better performance, we design a lightweight full-band information fusion (FIF) module, complementing the sub-band model with full-band features. Correspondingly, we add another lightweight full-band spectrum prediction (FSP) module at the end of the SE model, which generates an enhanced full-band spectrum to assist the optimization of the SE model. Moreover, inspired by the batch shuffle operation in RemixIT [11, 12], we design a pseudo noisy waveform reconstruction (PNWR) loss. As a part of the total loss function, the inclusion of PNWR loss enhances the robustness of SE model to various noisy signals, thereby yielding improved SE performance.

We take the discrete cosine transform-based CRN [13] with our previously proposed time-frequency sequence modeling (TFSM) blocks [14] as the network backbone for SE, which is abbreviated as DTFCRN. By incorporating the SE backbone with the proposed sub-band processing scheme and PNWR loss, we name the final SE system as Sub-PNWR. Experimental results show that the proposed sub-band processing approach significantly reduces the computational complexity without obvious performance degradation. With a computational volume of only 2.16 GMACs/s, the proposed Sub-PNWR exhibits superior

*Equal contribution.
†Corresponding author.

10.21437/Interspeech.2024-1927

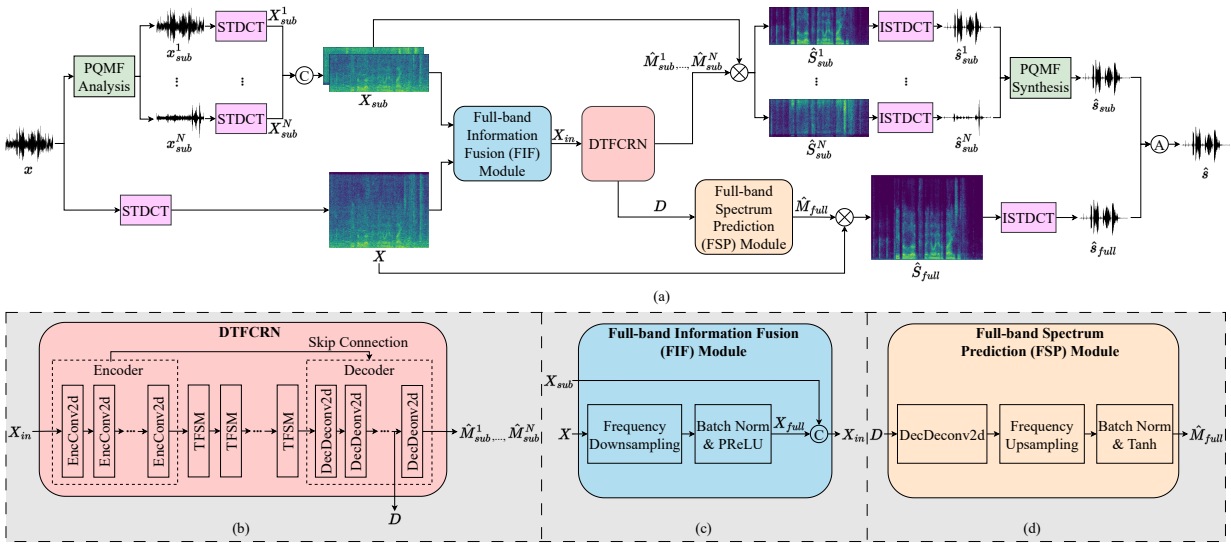

Figure 1: *(a) The overall framework of the proposed Sub-PNWR. (b) Details of the DTFCRN model. (c) Details of the FIF module. (d) Details of the FSP module.*

SE performance compared to existing advanced models on both VoiceBank+DEMAND [15] and DNS-Challenge [16] datasets.

2. Methodology

2.1. Overall framework

The overall framework of Sub-PNWR is shown in Figure 1 (a). It aims to estimate the clean speech $s \in \mathbb{R}^L$ from the noisy speech $x = s + z \in \mathbb{R}^L$, where z denotes the noise signal, and L is the audio length. DTFCRN serves as the network backbone for noise reduction. To reduce the computational operations, we divide the input audio x into N sub-band signals by the analysis filter bank of pseudo quadrature mirror filter (PQMF) [17] as:

$$x^i_{sub} = \mathrm{PQMF}(x), i = 1, ..., N \tag{1}$$

where $x^i_{sub} \in \mathbb{R}^{L/N}$ is the sub-band signal. Taking N sub-band noisy signals as input, the corresponding prediction targets of Sub-PNWR include N sub-band signals $s^i_{sub} \in \mathbb{R}^{L/N}$ of the clean speech s, where $s^i_{sub} = \mathrm{PQMF}(s), i = 1, ..., N$.

As the proposed Sub-PNWR performs noise suppression in the TF domain, we first acquire the sub-band spectrum $X^i_{sub} \in \mathbb{R}^{F\times T}$ by applying F-point short-time discrete cosine transform (STDCT) [18] to the sub-band noisy signals x^i_{sub}, where T represents the number of frames. Subsequently, all N sub-band spectrum are then concatenated along the channel dimension to obtain $X_{sub} \in \mathbb{R}^{N\times F\times T}$.

To complement the sub-band input X_{sub} with full-band spectral information, a full-band information fusion (FIF) module is proposed. This module extracts global spectral feature from the full-band noisy spectrum $X \in \mathbb{R}^{NF\times T}$ derived by NF-point STDCT, and fuses it with the concatenated sub-band spectrum X_{sub} to generate the sub-band and full-band fused feature X_{in}.

Afterwards, the DTFCRN receives the fused feature X_{in} and predicts N sub-band spectrum masks $\hat{M}^i_{sub} \in \mathbb{R}^{F\times T}, i = 1, ..., N$. These masks are respectively multiplied with N input sub-band noisy spectrum X^i_{sub} to derive the denoised sub-band spectrum $\hat{S}^i_{sub}$ as:

$$\hat{S}^i_{sub} = \hat{M}^i_{sub} \otimes X^i_{sub} \tag{2}$$

where $\hat{S}^i_{sub} \in \mathbb{R}^{F\times T}, i = 1, ..., N$. $\otimes$ denotes the element-wise multiplication. Applying inverse STDCT (ISTDCT) to each $\hat{S}^i_{sub}$, we obtain N enhanced sub-band signal $\hat{s}^i_{sub} \in \mathbb{R}^{L/N}$. Therefore, the enhanced full-band audio $\hat{s}_{sub}$ can be obtained using the synthesis filter bank of PQMF.

In addition, we devise a full-band spectrum prediction (FSP) module and integrate it at the end of the DTFCRN. This module estimates a full-band spectrum mask $\hat{M}_{full} \in \mathbb{R}^{NF\times T}$, which is multiplied with the full-band noisy spectrum X to reconstruct the enhanced full-band spectrum $\hat{S}_{full} \in \mathbb{R}^{NF\times T}$ as:

$$\hat{S}_{full} = \hat{M}_{full} \otimes X \tag{3}$$

Consequently, another enhanced full-band audio $\hat{s}_{full}$ can be derived after applying ISTDCT to $\hat{S}_{full}$.

The final enhanced speech, denoted as $\hat{s}$, is obtained by averaging $\hat{s}_{sub}$ and $\hat{s}_{full}$ as follows:

$$\hat{s} = \frac{\hat{s}_{sub} + \hat{s}_{full}}{2} \tag{4}$$

The details of the DTFCRN, FIF module, FSP module, and loss function are described in the following contents.

2.2. Model structure

2.2.1. Discrete cosine transform-based CRN with time-frequency sequence modeling blocks (DTFCRN)

As illustrated in Figure 1 (b), the DTFCRN follows the encoder-decoder (ED) structure. The encoder encodes the fused feature X_{in} into a compressed representation with high-level features. There are multiple 2D convolutional (EncConv2d) blocks in the encoder, and each block is composed of a 2D convolutional layer, a batch normalization (BN) layer, and a PReLU layer.

The encoded result is subsequently fed into multiple time-frequency sequence modeling (TFSM) blocks [14]. Each TFSM block serially employs a bidirectional gated recurrent unit (Bi-GRU) layer and unidirectional gated recurrent unit (GRU) layer for the sequence modeling in the frequency and time dimensions, respectively. The utilization of GRU layer for temporal correlation modeling ensures causal inference.

Next, the decoder receives the output from the last TFSM block to reconstruct the spectrum masks. It includes multiple 2D deconvolutional (DecDeconv2d) blocks, where each DecDeconv2d block consists of a 2D deconvolutional layer, a BN layer, and a PReLU layer. Note that the activation function of the last DecDeconv2d block is adjusted to Tanh. Meanwhile, the skip connection is used between the encoder and decoder for better performance. The decoder predicts N sub-band spectrum masks $\hat{M}^i_{sub}, i = 1, ..., N$ for target sub-band spectrum estimation using Equation 2. In addition, the feature map D before the last DecDeconv2d block is extracted and sent to the FSP module to predict the full-band spectrum mask $\hat{M}_{full}$.

2.2.2. Full-band information fusion (FIF) module

The details of the proposed FIF module are depicted in Figure 1 (c). It takes both the sub-band feature $X_{sub} \in \mathbb{R}^{N\times F\times T}$ and full-band spectrum $X \in \mathbb{R}^{NF\times T}$ as inputs. The FIF module employs a frequency downsampling 2D convolutional layer to process the full-band spectrum X. The convolutional stride in the frequency dimension is set as N, aiming to extract a global spectral feature $X_{full} \in \mathbb{R}^{K\times F\times T}$, where K is the number of output channels in the feature map. Therefore, the fused result X_{in} can be obtained as:

$$X_{in} = [X_{sub}; X_{full}] \tag{5}$$

where $X_{in} \in \mathbb{R}^{(N+K)\times F\times T}$, and $[;]$ denotes the concatenation operation.

2.2.3. Full-band spectrum prediction (FSP) module

To further enhance the SE performance, we proposed a lightweight FSP module for simultaneous full-band spectrum mask prediction. The internal structure of the FSP module is shown in Figure 1 (d). It receives the intermediate feature map D from the DTFCRN, and sequentially adopts a DecDeconv2d block, a frequency upsampling 2D deconvolutional layer, a BN layer, and a Tanh layer to predict a full-band spectrum mask $\hat{M}_{full} \in \mathbb{R}^{NF\times T}$. Symmetrically to the frequency downsampling layer in the FIF module, the frequency upsampling deconvolution layer here has a stride of N in the frequency dimension.

2.3. Loss function

Since the proposed Sub-PNWR model predicts both the sub-band and full-band spectrum masks, we define a hybrid TF domain loss $\mathcal{L}_{\mathrm{TF}}$ containing both the sub-band spectral loss $\mathcal{L}_{sub}$ and full-band spectral loss $\mathcal{L}_{full}$ as:

$$\begin{aligned}\mathcal{L}_{\mathrm{TF}} &= \mathcal{L}_{sub} + \mathcal{L}_{full} \\ &= \frac{1}{N}\sum_{i=1}^{N}\|\hat{M}^i_{sub} - M^i_{sub}\|_F^2 + \|\hat{M}_{full} - M_{full}\|_F^2\end{aligned} \tag{6}$$

where $(\hat{M}^i_{sub}, M^i_{sub}), i = 1, ..., N$ denote the estimated and target sub-band spectrum masks, and $(\hat{M}_{full}, M_{full})$ denote the estimated and target full-band spectrum masks. $\|\cdot\|_F^2$ represents the mean square error (MSE) loss.

We also consider a time domain loss L_{T} as shown below:

$$\mathcal{L}_{\mathrm{T}} = \|\hat{s} - s\|_1 \tag{7}$$

where $(\hat{s}, s)$ represent the enhanced and clean speech. $\|\cdot\|_1$ denotes the L1 loss.

Additionally, we design a pseudo noisy waveform reconstruction (PNWR) loss to further optimize the training process. In the feedforward process, the SE model $\mathcal{F}(\cdot)$ estimates the enhanced audios for a given noisy batch $x_b = s_b + z_b \in \mathbb{R}^{B\times L}$ (B denotes the batch size) as follows:

$$\hat{s}_b, \hat{z}_b = \mathcal{F}(x_b), \quad x_b = \hat{s}_b + \hat{z}_b \tag{8}$$

where $(\hat{s}_b, \hat{z}_b)$ denote estimated speech and noise signals. To improve the robustness of model to various noisy mixtures, we use these estimated sources to generate new noisy mixtures as:

$$\hat{x}^{\Pi}_b = \hat{s}_b + \hat{z}^{\Pi}_b, \quad \hat{z}^{\Pi}_b = \Pi\hat{z}_b, \quad \Pi \sim \mathcal{P}_{B\times B} \tag{9}$$

where Π is uniformly sampled from the set of all $B \times B$ permutation matrices [1] [11, 12]. Correspondingly, the target new mixtures x^{Π}_b is:

$$x^{\Pi}_b = s_b + z^{\Pi}_b, \quad z^{\Pi}_b = \Pi z_b \tag{10}$$

Then, we utilize the remixed noisy audios to calculate the PNWR loss $\mathcal{L}_{\mathrm{PNWR}}$ as:

$$\mathcal{L}_{\mathrm{PNWR}} = \|\hat{x}^{\Pi}_b - x^{\Pi}_b\|_1 \tag{11}$$

Eventually, the total loss $\mathcal{L}$ is the sum of the above losses:

$$\mathcal{L} = \mathcal{L}_{\mathrm{TF}} + \mathcal{L}_{\mathrm{T}} + \mathcal{L}_{\mathrm{PNWR}} \tag{12}$$

3. Experiments

3.1. Datasets and experimental setup

We adopt the VoiceBank+DEMAND [15] and DNS-Challenge [16] datasets for our experiments. The VoiceBank+DEMAND dataset consists of a training set of 11,572 audio clips from 28 speakers and a test set of 824 audio clips from 2 unseen speakers. The clean audios in the training set are mixed with 10 types of noise under 4 signal-to-noise ratios (SNRs), i.e., {0,5,10,15}dB. In the test set, the noisy audio contain 5 other types of noise, and the mixed SNRs are {2.5,7.5,12.5,17.5}dB. The DNS-challenge dataset contains over 500 hours of clean audios and over 180 hours of noise audios for training. We generate 3,000 hours of noisy-clean audio pairs as the training set, with the mixed SNRs ranging from -5dB to 15dB. Besides, the DNS-Challenge dataset provides a non-blind validation set for model evaluation, which comprises 150 noisy-clean audio pairs.

All the audios are sampled at 16 kHz. When calculating the full-band spectrum, the DCT point number, Hamming window size, and hop size are set as 512, 32 ms, and 8 ms, respectively. The sub-band number N is set to 2 when using the PQMF. Therefore, the DCT point number F for sub-band spectrum computation is 256. For the DTFCRN, it consists of 5 EncConv2d blocks, whose convolutional output channels are {16,32,64,128,256}. Correspondingly, the output channels of the deconvolutional layers in the DTFCRN decoder are {128,64,32,16,2}. The kernel size and stride of all the (de)convolutional layers are {5,2} and {2,1} in frequency and time dimensions. 3 TFSM modules are inserted between the encoder and decoder, whose GRU&BiGRU hidden sizes are {128,64,32}. The channel number of the full-band spectral feature is set as $K = 2$. Note that we guarantee the causality of model by using causal (de)convolutional layers. The models are trained using RMSprop optimizer for 100 epochs. The learning rate is initially set as 0.0002 and halved if the performance does not improve for 8 consecutive epochs. The batch size is 16.

[1] A permutation matrix reorders the signals within each batch, where each row and column contains only one 1 and the remaining elements are zeros, indicating the new order of signals.

Table 1: *Results of different SE systems and the proposed method on the VoiceBank+DEMAND dataset.*

	Causal	MACs(G/s)	Param. (M)	WB-PESQ	CSIG	CBAK	COVL
noisy	-	-	-	1.97	3.35	2.44	2.63
DCCRN [4]	✔	14.06	3.67	2.68	3.88	3.18	3.27
FullSubNet+ [8]	✔	30.06	8.67	2.88	3.86	3.42	3.57
CompNet [19]	✔	5.92	4.26	2.90	4.16	3.37	3.53
CTS-Net [20]	✔	5.57	4.35	2.92	4.25	3.46	3.59
DEMUCS [21]	✔	4.32	18.87	2.93	4.22	3.25	3.52
PHASEN [22]	✗	6.12	8.76	2.99	4.21	**3.55**	3.62
DTFCRN (Baseline)	✔	8.54	2.58	2.96	4.22	3.51	3.60
DTFCRN-CL1	✔	4.29	2.58	2.88	4.16	3.45	3.53
DTFCRN-CL2	✔	4.23	1.13	2.92	4.06	3.49	3.49
DTFCRN+Sub-band	✔	4.31	2.57	2.94	4.23	3.49	3.60
DTFCRN-CL2+Sub-band	✔	2.16	1.12	2.91	4.13	3.47	3.53
DTFCRN-CL2+Sub-band+PNWR Loss (**Sub-PNWR**)	✔	2.16	1.12	**3.03**	**4.26**	3.49	**3.65**

3.2. Ablation study

To verify the benefits of the proposed improvements, we conduct ablation studies on the VoiceBank+DEMAND dataset, and the results are presented in Table 1. Four objective evaluation metrics are utilized for performance evaluation, including wide-band perceptual evaluation of speech quality (WB-PESQ) [23] and three composite metrics [24] measuring the mean opinion score (MOS) on signal distortion (CSIG), background noise intrusiveness (CBAK), and overall audio quality (COVL).

We take the original DTFCRN as baseline, which receives the noisy full-band spectrum and predicts a full-band spectrum mask. As we combine the proposed sub-band processing scheme with the baseline model (i.e., DTFCRN+Sub-band), the computational operations are approximately halved. Meanwhile, the performance of DTFCRN+Sub-band has almost no decrease compared to the baseline model. Additionally, we have also tried to reduce the computational complexity of the baseline by directly adjusting its model hyperparameters. Specifically, we increase the hop size to 16 ms (i.e., DTFCRN-CL1). We have also attempted to decrease the convolutional output channels to {16,32,48,96,128} and synchronously decrease the deconvolutional output channels (i.e., DTFCRN-CL2). Both DTFCRN-CL1 and DTFCRN-CL2 have similar computational volumes to DTFCRN+Sub-band, but their performance is obviously worse. This demonstrates the effectiveness of the proposed sub-band processing method. Moreover, we further incorporate the sub-band processing scheme with DTFCRN-CL2 (i.e., DTFCRN-CL2+Sub-band) for fewer computational operations, while this still ensures no significant performance degradation compared to DTFCRN-CL2.

The loss function used to train the above models does not include the proposed PNWR loss. Thus, we further combine the PNWR loss with DTFCRN-CL2+Sub-band to obtain the final Sub-PNWR model. Experimental results show that the introduction of the PNWR loss results in further performance gains.

3.3. Comparison with other advanced systems

We also compare the proposed Sub-PNWR with other advanced systems on both VoiceBank+DEMAND and DNS-Challenge datasets. As shown in Table 1, the proposed Sub-PNWR outperforms the previous methods on WB-PESQ, CSIG, and COVL, except that PHASEN [22] with non-causal inference achieves a higher CBAK score than our Sub-PNWR. For the DNS-Challenge dataset, four metrics are used for model evaluation, including WB-PESQ, narrow-band perceptual evaluation of speech quality (NB-PESQ) [27], short-time objective intelligibility (STOI) [28], and scale-invariant signal-to-noise ratio (SI-SNR) [29]. The comparison results in Table 2 illustrate that the proposed Sub-PNWR generally achieves superior performance over the previous advanced systems.

Table 2: *Performance comparison with advanced methods on DNS-Challenge non-blind test set under causal implementation.*

	MACs (G/s)	WB-PESQ	NB-PESQ	STOI (%)	SI-SNR (dB)
noisy	-	1.58	2.45	91.52	9.07
FullSubNet [7]	30.85	2.88	3.43	96.32	17.30
CTS-Net [20]	5.57	2.94	3.42	96.66	17.99
FullSubNet+ [8]	30.06	2.98	3.50	96.69	18.34
Inter-SubNet [9]	36.65	3.00	3.50	96.61	18.05
FS-CANet [25]	-	3.02	3.51	96.74	18.08
GaGNet [26]	1.64	3.17	3.56	97.13	**18.91**
Sub-PNWR	2.16	**3.28**	**3.60**	**97.31**	18.35

In addition to the excellent model performance, the computational complexity and model size of Sub-PNWR is also advantageous compared to most existing methods. This makes its deployment on devices with limited resources more feasible.

4. Conclusions

In this paper, we propose a computationally lightweight SE model named Sub-PNWR. The Sub-PNWR combines the signal sub-band processing method to effectively reduce the computational complexity of the full-band baseline model, namely DTFCRN. Meanwhile, we design a FIF module and a FSP module for full-band spectrum information supplement and full-band spectrum mask prediction, respectively. Experimental results demonstrate that the above improvements significantly reduce the computational complexity by half with no obvious performance degradation. Additionally, a PNWR loss is proposed, which is proved to further improve the model performance. In the future, we will explore whether the proposed computationally lightweight approach is applicable to other SE frameworks.

5. Acknowledgements

This work was supported by the special funds of Shenzhen Science and Technology Innovation Commission under Grant No. CJGJZD20220517141400002.

6. References

[1] S. R. Park and J. W. Lee, "A Fully Convolutional Neural Network for Speech Enhancement," in *Proc. Interspeech 2017*, 2017, pp. 1993–1997.

[2] 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)*, 2020, pp. 851–855.

[3] K. Tan and D. Wang, "A Convolutional Recurrent Neural Network for Real-Time Speech Enhancement," in *Proc. Interspeech 2018*, 2018, pp. 3229–3233.

[4] 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.

[5] R. Cao, S. Abdulatif, and B. Yang, "CMGAN: Conformer-based Metric GAN for Speech Enhancement," in *Proc. Interspeech 2022*, 2022, pp. 936–940.

[6] G. Yu, A. Li, H. Wang, Y. Wang, Y. Ke, and C. Zheng, "Dbt-net: Dual-branch federative magnitude and phase estimation with attention-in-attention transformer for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2629–2644, 2022.

[7] 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)*, 2021, pp. 6633–6637.

[8] 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)*, 2022, pp. 7857–7861.

[9] 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.

[10] Z. Chen and P. Zhang, "Lightweight Full-band and Sub-band Fusion Network for Real Time Speech Enhancement," in *Proc. Interspeech 2022*, 2022, pp. 921–925.

[11] E. Tzinis, Y. Adi, V. K. Ithapu, B. Xu, and A. Kumar, "Continual self-training with bootstrapped remixing for speech enhancement," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6947–6951.

[12] E. Tzinis, Y. Adi, V. K. Ithapu, B. Xu, P. Smaragdis, and A. Kumar, "Remixit: Continual self-training of speech enhancement models via bootstrapped remixing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1329–1341, 2022.

[13] Q. Li, F. Gao, H. Guan, and K. Ma, "Real-time monaural speech enhancement with short-time discrete cosine transform," *arXiv preprint arXiv:2102.04629*, 2021.

[14] Y. Zhang, H. Zou, and J. Zhu, "A two-stage framework in cross-spectrum domain for real-time speech enhancement," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 12 587–12 591.

[15] 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.

[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] H. J. Nussbaumer and M. Vetterli, "Pseudo quadrature mirror filters," in *Proceedings International Conference on Digital Signal Processing*, no. CONF, 1984.

[18] N. Ahmed, T. Natarajan, and K. Rao, "Discrete cosine transform," *IEEE Transactions on Computers*, vol. C-23, no. 1, pp. 90–93, 1974.

[19] C. Fan, H. Zhang, A. Li, W. Xiang, C. Zheng, Z. Lv, and X. Wu, "Compnet: Complementary network for single-channel speech enhancement," *Neural Networks*, vol. 168, pp. 508–517, 2023.

[20] 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.

[21] A. Défossez, G. Synnaeve, and Y. Adi, "Real Time Speech Enhancement in the Waveform Domain," in *Proc. Interspeech 2020*, 2020, pp. 3291–3295.

[22] D. Yin, C. Luo, Z. Xiong, and W. Zeng, "Phasen: A phase-and-harmonics-aware speech enhancement network," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, pp. 9458–9465, Apr. 2020.

[23] 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*, 2005.

[24] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

[25] 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.

[26] A. Li, C. Zheng, L. Zhang, and X. Li, "Glance and gaze: A collaborative learning framework for single-channel speech enhancement," *Applied Acoustics*, vol. 187, p. 108499, 2022.

[27] A. Rix, J. Beerends, M. Hollier, and A. 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, 2001, pp. 749–752 vol.2.

[28] 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.

[29] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr – half-baked or well done?" in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630.