INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

MFT-CRN:Multi-scale Fourier Transform for Monaural Speech Enhancement

Yulong Wang[1], *Xueliang Zhang*[1]

[1]College of Computer Science,Inner Mongolia University,China

32109010@mail.imu.edu.cn, cszxl@imu.edu.cn

Abstract

Convolutional recurrent networks (CRN) that combine a convolutional encoder-decoder (CED) structure with a recurrent structure have shown promising results in monaural speech enhancement. However, the commonly used short-time Fourier transform fails to balance the needs of frequency and time resolution effectively, which is crucial for accurate speech estimation. To address this issue, we propose MFT-CRN, a multi-scale short-time Fourier transform fusion model. We process the input speech signal through short-time Fourier transforms with different window functions, and add them layer by layer in the encoder and decoder of the network to achieve feature fusion with different window functions, effectively balancing frequency and temporal resolution. Comprehensive experiments on the WSJ0 dataset show that MFT-CRN significantly outperforms the method using only a single window function in terms of short-time intelligibility and perceptual evaluation of speech quality.

Index Terms: monaural speech enhancement, frequency domain, short-time fourier transform,multi-scale fusion

1. Introduction

With the great popularity of COVID-19, people's working style has changed from an offline office to an online office, and video conferencing and teleconferencing have become an indispensable tool. However, there are always various kinds of noise in the home office, such as the noise of children, the sound of renovation downstairs, the sound of car sirens, and so on. All of these can seriously affect our work and communication efficiency, and pose new challenges for voice enhancement.

Speech enhancement is widely used in scenarios such as voice communication systems, speech recognition, and hearing aids. According to the different signals processed by speech enhancement, it can be divided into monaural speech enhancement and multichannel speech enhancement. Monaural speech data is easier to obtain compared to multichannel data, and it has low hardware requirements and a lower acquisition cost, so monaural speech enhancement technology is more widely used. Earlier, digital signal processing techniques were mainly used to cope with speech enhancement tasks, but such methods usually assume that noise is a smooth invariant signal, an assumption that is clearly not valid in realistic situations. Therefore, the performance of speech enhancement algorithms based on traditional digital signal processing is limited.With the development of deep learning techniques, researchers have started to explore deep learning-based speech enhancement methods. Thanks to the strong learning ability, wide coverage, and good generalization of deep learning, speech enhancement based on deep learning has surpassed traditional digital signal processing methods at this stage.

With the advancement of deep learning, many researchers regard speech enhancement as a supervised learning problem [1] [2] and have achieved outstanding performance. In [1] [3], the author investigates the issue of noise reduction in the short-time Fourier transform (STFT) domain. The STFT approach aligns more closely with human auditory perception and makes the speech characteristics clearer. The concept of Convolutional Encoder-Decoder (CED) for speech enhancement was first introduced in [4]. They proposed a Redundant CED network (RCED) made up of repeated convolution, batch normalization (BN) [5], and ReLU activation [6] layers. The RCED architecture also includes skip connections to aid optimization. These skip connections link each layer in the encoder to its corresponding layer in the decoder. Many subsequent studies have adopted the CED architecture [7–16] . Despite the significant achievements of deep learning in speech enhancement technology, there are still numerous challenges in input signal feature extraction and utilization.

For instance, in the frequency domain, research often involves Fourier transforms, and time-frequency analysis is achieved by adding a window function. However, the Fourier transform employs a fixed window function. Once this function is selected, its shape remains unchanged, and the resolution of the Fourier transform is determined. To alter the resolution, a different window function must be chosen. The Fourier transform is useful for analyzing piecewise stationary signals or approximating stationary signals, but it cannot handle non-stationary signals.The short-time Fourier transform [17] can solve non-stationary signals. When the signal changes rapidly, the window function needs to have a higher time resolution, while for signals that change gently, with mostly low-frequency components, the window function requires a higher frequency resolution. In this paper, to address both frequency and time resolution, we present a novel multi-scale short-time Fourier transform fusion algorithm, MFT-CRN, for the first time. This algorithm obtains multi-scale short-time Fourier transform features from input signals using short-time Fourier transforms with different window functions, then integrates the features into the encoder and decoder of the network layer by layer. The experimental results demonstrate that this method delivers good results on the WSJ0 SI-84 data sets.

The rest of the paper is structured as follows. Section 2 describes the proposed MFT-CRN structure. Section 3 outlines the experimental setup. Section 4 presents the experimental results, and Section 5 summarizes the paper.

10.21437/Interspeech.2023-865

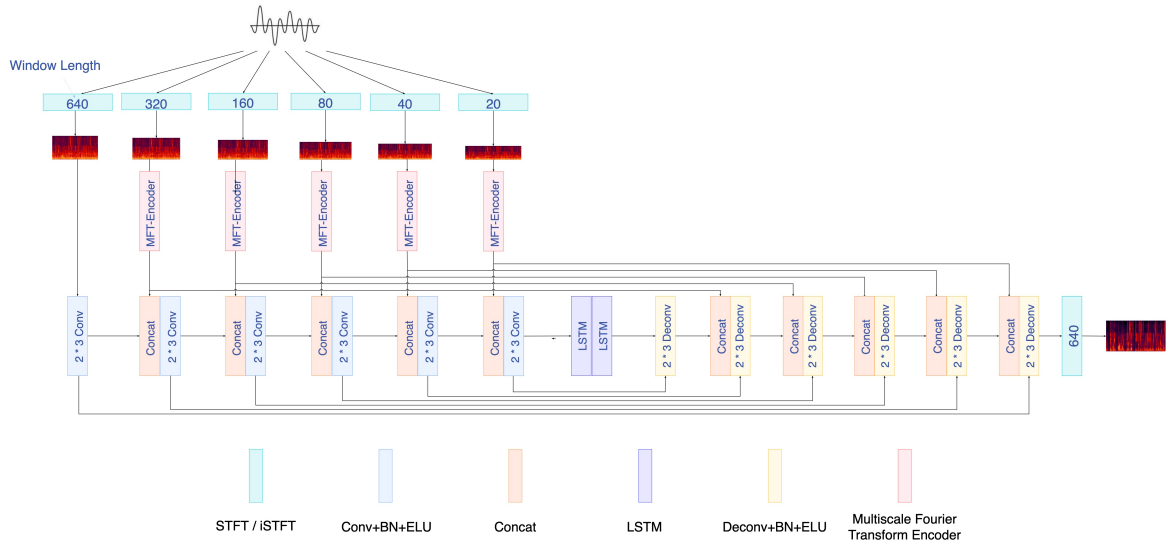

Figure 1: *Schematic diagram of the proposed MFT-CRN.*

2. Model Description

2.1. Overview

CRN, which was first proposed in [7], is essentially a regular CED architecture that employs two LSTM [18] layers in between the encoder and decoder. LSTMs are specifically used here to model temporal dependencies. The encoder contains five 2D convolutional blocks, which serve to extract high-dimensional features from the input or reduce its resolution. Subsequently, the decoder reconstructs the low-resolution features to the scale of the original input, resulting in a symmetric design for the codec's structure. Specifically, the Conv2D module of the codec consists of a series of convolutional/deconvolutional layers, as well as BN [5] and activation functions. Skip-connections are used to facilitate the transfer of gradients between codecs.

Figure 1 illustrates a schematic diagram of the proposed MTF-CRN, which follows an encoder-decoder scheme and utilizes a sequence of downsampling and upsampling blocks to make predictions. As shown in Figure 1, the proposed MTF-CRN comprises an encoder part, a fusion part, a multi-scale Fourier transform encoder part, and a decoder part. In this study, we used the short-time Fourier transform (STFT) magnitude spectrum of noisy speech in 321 dimensions as the input features of the encoder, and clean speech as the training target. The multi-scale Fourier transform encoder uses the STFT magnitude spectrum of noisy speech in dimensions 161, 81, 41, 21, 11.

The encoder is designed to use magnitude tiles as input to convolutional layers, and each layer of the encoder is followed by batch normalization [5] and exponential linear unit (ELU) [6]. A multi-scale Fourier encoder is used to process the input embedding after the short-time Fourier transform with different window functions through multiple layers of convolutional blocks. It is important to note that the dimensionality of the input feature vector after a convolutional layer must be the same as the corresponding layer-wise output feature vector, since these two vectors go through the fusion part of the encoding stage at each encoding layer. The encoder's output features are then fed into two LSTM blocks to aggregate temporal context. The decoder part is reconstructed by deconvolution, followed by batch normalization and ELU again.

2.2. Multi-scale Fourier Transform Encoder

The main purpose of the multi-scale Fourier transform encoder is to align the time dimension features of different window functions. It has the same convolutional layer structure as the encoder, with a BN layer and an ELU layer following each convolutional layer, except for the difference in the parameters of the convolutional kernel. The kernel size is set to (3, 2), and the time dimension is downsampled to achieve alignment of the encoder and multi-scale encoder features in the time dimension, which are then fused into the network. The number of multi-scale encoder layers is increased based on the gap in the time dimension of the input features from the encoder, where K represents the number of multi-scale encoder layers.

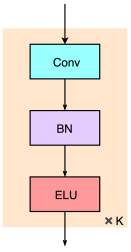

Figure 2: *Schematic diagram of Multi-scale Fourier Transform Encoder.*

2.3. Parameters Setting

The encoder has six layers, each consisting of a convolutional layer followed by batch normalization and ELU nonlinearity. Except for the first layer, each layer adds a module to change the shape of the feature representation for different window functions. The input size of the model is [batch size, 1, seq len, features]. The encoder down-samples the input layer by layer, and the number of channels is amplified by the same multiple.

The number of channels is [1, 8, 16, 32, 64, 128, 256] in turn, and the newly added multi-scale Fourier transform encoder of each layer changes different windows. The shape of the feature function's window is changed, and the number of channels is changed to [8, 16, 32, 32, 32] in turn. The window length used by the multi-scale Fourier transform is [640, 320, 160, 80, 40, 20], and the window shift is [320, 160, 80, 40, 20, 10], respectively. Selecting these scales minimizes the number of doown-sampling operations, which in turn reduces information loss. The output of the previous layer and the features of different window functions are connected along the channel dimension and sent to the convolutional layer for processing. The convolution kernel and stride are set to (2, 3) and (1, 2), respectively. The final output size of the encoder is [batch size, 256, seq len, 4]. The output is fed into an LSTM layer to extract long-term relationships of frequency-domain features separately.

The decoder also consists of six layers. Each layer receives skip connections from the corresponding layer of the encoder, in addition to features from different window functions. These skip connections are connected with the output of the previous layer along the channel axis. The decoder uses deconvolution to double the feature dimension layer by layer, ultimately reconstructing the signal to its original size. The final layer of the decoder boosts the signal into one channel, which is then converted to speech using an overlap-add operation.

Table 1: *Architecture of backbone of our proposed MFT-CRN. Here T denotes the number of time frames in the STFT magnitude spectrum.*

layername	input size	hyperparameters	output size
reshape_1	T × 321	-	1 × T × 321
conv2d_1	1 × T × 321	2 × 3 , (1 , 2) , 8	8 × T × 160
conv2d_2	16 × T × 160	2 × 3 , (1 , 2) , 16	16 × T × 79
conv2d_3	32 × T × 79	2 × 3 , (1 , 2) , 32	32 × T × 39
conv2d_4	64 × T × 39	2 × 3 , (1 , 2) , 64	64 × T × 19
conv2d_5	96 × T × 19	2 × 3 , (1 , 2) , 128	128 × T × 9
conv2d_6	160 × T × 9	2 × 3 , (1 , 2) , 256	256 × T × 4
reshape_2	256 × T × 4	-	T × 1024
lstm_1	T × 1024	1024	T × 1024
lstm_2	T × 1024	1024	T × 1024
reshape_3	T × 1024	-	256 × T × 4
deconv2d_6	512 × T × 4	2 × 3 , (1 , 2) , 128	128 × T × 9
deconv2d_5	288 × T × 9	2 × 3 , (1 , 2) , 64	64 × T × 19
deconv2d_4	160 × T × 19	2 × 3 , (1 , 2) , 32	32 × T × 39
deconv2d_3	96 × T × 39	2 × 3 , (1 , 2) , 16	16 × T × 79
deconv2d_2	48 × T × 79	2 × 3 , (1 , 2) , 8	8 × T × 160
deconv2d_1	24 × T × 160	2 × 3 , (1 , 2) , 1	1 × T × 321
reshape_4	1 × T × 321	-	T × 321

Table 1 provides a detailed description of the backbone of our proposed network structure. The input and output sizes of each layer are specified in the format of featureMaps × timeSteps × frequencyChannels. The hyperparameters of the layers are given in the format (kernelSize, strides, outChannels). For all convolution and deconvolution operations, we apply zero padding in the time direction but not in the frequency direction. For causal convolution, we use a kernel size of 2 × 3 (time × frequency). It's worth noting that the number of feature maps in each decoder layer is more than doubled by the skip connections and the skip connections of the multi-scale Fourier transform encoder.

Table 2 provides a more detailed description of our proposed multi-scale Fourier transform encoder. The input and output sizes of each layer are specified in the format of featureMaps × timeSteps × frequencyChannels. The hyperparameters of the layers are given in the format (kernelSize, strides, outChannels). For all convolution and deconvolution operations, we apply zero padding in the frequency direction but not in the time direction. For causal convolution, we use a kernel size of 3 × 2 (time × frequency).

Table 2: *Architecture of Multi-scale Fourier Transform Encoder of our proposed MFT-CRN. Here T denotes the number of time frames in the STFT magnitude spectrum.*

layername	input size	hyperparameters	output size
conv320_1	1 × 2T × 161	3 × 2 , (2 , 1) , 8	8 × T × 160
conv160_1	1 × 4T × 81	3 × 2 , (2 , 1) , 8	8 × 2T × 80
conv160_2	8 × 2T × 80	3 × 2 , (2 , 1) , 16	16 × T × 79
conv80_1	1 × 8T × 41	3 × 2 , (2 , 1) , 8	8 × 4T × 40
conv80_2	8 × 4T × 40	3 × 2 , (2 , 1) , 16	16 × 2T × 40
conv80_3	16 × 2T × 40	3 × 2 , (2 , 1) , 32	32 × T × 39
conv40_1	1 × 16T × 21	3 × 2 , (2 , 1) , 4	4 × 8T × 20
conv40_2	4 × 8T × 20	3 × 2 , (2 , 1) , 8	8 × 4T × 20
conv40_3	8 × 4T × 20	3 × 2 , (2 , 1) , 16	16 × 2T × 20
conv40_4	16 × 2T × 20	3 × 2 , (2 , 1) , 32	32 × T × 19
conv20_1	1 × 32T × 11	3 × 2 , (2 , 1) , 2	2 × 16T × 10
conv20_2	2 × 16T × 10	3 × 2 , (2 , 1) , 4	4 × 8T × 10
conv20_3	4 × 8T × 10	3 × 2 , (2 , 1) , 8	8 × 4T × 10
conv20_4	8 × 4T × 10	3 × 2 , (2 , 1) , 16	16 × 2T × 10
conv20_5	16 × 2T × 9	3 × 2 , (2 , 1) , 32	32 × T × 9

2.4. Loss Function

In our MFT-CRN, we use MSE to train the MFT-CRN until convergence. The loss function value only includes the magnitude spread estimate, expressed as:

$$\mathcal{L}^{Mag} = |||\tilde{X}| - |S|\,||_F^2 \tag{1}$$

where $\mathcal{L}^{Mag}$ denotes the loss function of magnitude.$|S|$ denotes the target spectral magnitude, $|\tilde{X}|$ denotes the estimate spectral magnitude.In this paper, only the magnitude spectrum with a window function of 640 is used as a target, and the window functions of other scales are not used as targets.

3. Experiments

3.1. Datasets

In this study, we evaluated the performance of our proposed model on the WSJ0 SI-84 dataset [7] which includes 7138 utterances from 83 speakers (42 males and 41 females). We used the utterances of 77 speakers for training and the rest for test. We used 10000 non-speech sounds from a sound effect library (available at http://www.sound-ideas.com) [13] and generated 320000 and 3000 utterances at the SNRs uniformly sampled from -5dB, -4dB, -3dB, -2dB, -1dB, -0dB for training and validation, respectively. For the test set, two noises (babble and cafeteria) from Auditec CD (available at http://www.auditec.com) are used to generate 900 mixtures at each SNR of -5dB, 0dB, and 5dB.

3.2. System Settings

All utterances are 16kHz samples. Frames were extracted using rectangular windows and Hamming windows of sizes 40ms,

Table 3: *STOI and PESQ comparisons between different window functions of CRN.*

Metrics	STOI								PESQ								MACs (G/s)
Test Noise	Babble				Cafeteria				Babble				Cafeteria				
Test SNR	-5db	0db	5db	AVG	-5db	0db	5db	AVG	-5db	0db	5db	AVG	-5db	0db	5db	AVG	
Mixture	59.04	71.95	83.82	71.60	57.70	70.83	83.16	70.56	1.49	1.77	2.06	1.77	1.34	1.69	2.04	1.69	-
CRN80	73.38	84.99	91.89	83.42	70.55	83.07	90.76	81.46	1.68	2.16	2.56	2.13	1.68	2.16	2.56	2.13	7.73
CRN160	73.98	85.55	92.16	83.89	70.80	83.67	91.04	81.83	1.71	2.21	2.62	2.18	1.73	2.21	2.61	2.18	4.84
CRN320	75.46	86.87	92.81	85.05	73.56	85.30	91.99	83.62	1.84	2.37	2.76	2.32	1.87	2.35	2.74	2.32	2.56
CRN640	**75.46**	**87.27**	**93.10**	**85.28**	**74.46**	**86.05**	**92.33**	**84.28**	**1.93**	**2.48**	**2.86**	**2.42**	**1.99**	**2.46**	**2.82**	**2.42**	1.30
CRN1280	73.01	85.73	92.35	83.70	72.99	84.75	91.45	83.06	1.84	2.38	2.74	2.32	1.95	2.40	2.73	2.36	0.67

Table 4: *STOI and PESQ comparisons between MFT-CRN and the baseline models.*

Metrics	STOI								PESQ							
Test Noise	Babble				Cafeteria				Babble				Cafeteria			
Test SNR	-5db	0db	5db	AVG	-5db	0db	5db	AVG	-5db	0db	5db	AVG	-5db	0db	5db	AVG
Mixture	59.04	71.95	83.82	71.60	57.70	70.83	83.16	70.56	1.49	1.77	2.06	1.77	1.34	1.69	2.04	1.69
CRN	75.46	87.27	93.10	85.28	74.46	86.05	92.33	84.28	1.93	2.48	2.86	2.42	1.99	2.46	2.82	2.42
MFT-CRN	**76.99**	**88.06**	**93.53**	**86.19**	**75.61**	**86.81**	**92.76**	**85.06**	**1.98**	**2.53**	**2.92**	**2.48**	**2.03**	**2.51**	**2.88**	**2.47**
w/o Decoder-MFT	76.79	88.04	93.50	86.11	75.48	86.75	92.74	84.99	1.97	2.52	2.91	2.47	2.02	2.51	2.87	2.47

Table 5: *Parameter efficiency comparison of different models.*

Models	Para(M)	STOI	PESQ
Noisy	-	58.37	1.42
CRN	17.58	74.96	1.96
MFT-CRN	17.63	76.30	2.01

20ms, 10ms, 5ms, 2.5ms, 1.25ms, respectively. The overlap is 50%. The models are trained using the Adam optimizer [19] with a learning rate of 0.001. The batch size is set to 32 at the utterance level. Note that if the speech is longer than 7 seconds, a random 7-second segment will be cut from the speech. Smaller utterances are zero-padded to match the size of the largest utterance in the batch.

3.3. Evaluate Metrics

Performance is evaluated by two objective metrics: short-time objective intelligibility (STOI) [20] and perceptual evaluation of speech quality (PESQ) [21]. STOI values usually range from 0 to 1 and can be roughly interpreted as a correct percentage. PESQ values range from -0.5 to 4.5. For the STOI and PESQ metrics, higher numbers indicate better performance.

4. Result and Analysis

4.1. Ablation study

We first conducted an ablation study to investigate the effect of different window functions on the network performance, with window lengths of [1280, 640, 320, 160, 80] and window shifts of [640, 320, 160, 80, 40], with names corresponding to CRN1280,CRN640,CRN320,CRN160,CRN80. And the quantitative results are shown in Table 3 , we can obtain the following observations. First, in the comparison of CRN320 and CRN1280, the difference between the two focusing on the frequency dimension and the time dimension is not large. This reveals that both frequency and time dimensions are effective in improving speech quality. Secondly, in the case of low window length, the low metrics are due to the low frequency resolution, which is not conducive to estimating speech in the frequency domain environment. Thirdly, CRN640 has the best overall result because it is the closest in frequency and temporal dimension in the case of training samples of 7s speech, compromising frequency and temporal resolution. The best results are achieved on a single window function.Therefore, 640 is used as the window length for CRN in the following experiments.

4.2. Objective Comparisons

Initially, we compared our multi-scale Fourier transform fusion model with the STOI and PESQ baselines, and the results, displayed in Table 4, show the best outcomes in bold. Across all SNR levels and noise types, MTF-CRN outperformed all baselines in terms of STOI, and compared to CRN, babble and cafeteria noise showed average improvements of 0.91% and 0.78%, respectively. For PESQ, the average improvement was 0.06 and 0.05 for babble and cafeteria noise, respectively.

We also conducted a comparison between our model and CRN in a -5dB environment, as presented in Table 5. Although the multi-scale Fourier transform fusion algorithm added a few parameters, the significant improvement in the low signal-to-noise ratio and the manageable increase in parameter count remained within an acceptable range.

Overall, our proposed multi-scale Fourier transform fusion model outperforms the Fourier transform model based on a single window function. Our results highlight the critical importance of considering information in both time and frequency dimensions to improve model performance.

5. Conclusions

In this study, we propose a novel multi-scale Fourier transform fusion system that considers both the time and frequency dimensions. Our results demonstrate that the proposed model outperforms the baseline models in terms of objective intelligibility and quality scores. We attribute the excellent performance of our model to the fact that features obtained from different window functions have distinct frequency and time dimensions: larger window functions provide richer frequency information, while smaller window functions provide richer time information. By fusing the features of different window functions, the proposed model can combine diverse learning focuses to obtain richer features.

Acknowledgments: This research was partly supported by the China National Nature Science Foundation (No.61876214).

6. References

[1] K. Tan and D. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 380–390, 2020.

[2] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, 2018.

[3] 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.

[4] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," *arXiv preprint arXiv:1609.07132*, 2016.

[5] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. pmlr, 2015, pp. 448–456.

[6] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv preprint arXiv:1511.07289*, 2015.

[7] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement." in *Interspeech*, vol. 2018, 2018, pp. 3229–3233.

[8] Y. Fu, Y. Liu, J. Li, D. Luo, S. Lv, Y. Jv, and L. Xie, "Uformer: A unet based dilated complex & real dual-path conformer network for simultaneous speech enhancement and dereverberation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7417–7421.

[9] K. Zhang, S. He, H. Li, and X. Zhang, "Dbnet: A dual-branch network architecture processing on spectrum and waveform for single-channel speech enhancement," *arXiv preprint arXiv:2105.02436*, 2021.

[10] X. Xu, Y. Wang, J. Jia, B. Chen, and D. Li, "Improving visual speech enhancement network by learning audio-visual affinity with multi-head attention," *arXiv preprint arXiv:2206.14964*, 2022.

[11] S. Zhao, B. Ma, K. N. Watcharasupat, and W.-S. Gan, "Frcrn: Boosting feature representation using frequency recurrence for monaural speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9281–9285.

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

[13] S. Lv, Y. Fu, M. Xing, J. Sun, L. Xie, J. Huang, Y. Wang, and T. Yu, "S-dccrn: Super wide band dccrn with learnable complex feature for speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7767–7771.

[14] G. Yu, Y. Guan, W. Meng, C. Zheng, and H. Wang, "Dmf-net: A decoupling-style multi-band fusion model for real-time full-band speech enhancement," *arXiv preprint arXiv:2203.00472*, 2022.

[15] G. Yu, A. Li, W. Liu, C. Zheng, Y. Wang, and H. Wang, "Optimizing shoulder to shoulder: A coordinated sub-band fusion model for real-time full-band speech enhancement," *arXiv preprint arXiv:2203.16033*, 2022.

[16] T. Wang, W. Zhu, Y. Gao, Y. Chen, J. Feng, and S. Zhang, "Harmonic gated compensation network plus for icassp 2022 dns challenge," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9286–9290.

[17] S. Nawab, T. Quatieri, and J. Lim, "Signal reconstruction from short-time fourier transform magnitude," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 31, no. 4, pp. 986–998, 1983.

[18] A. Graves and A. Graves, "Long short-term memory," *Supervised sequence labelling with recurrent neural networks*, pp. 37–45, 2012.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[20] 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.

[21] A. W. Rix, M. P. Hollier, A. P. Hekstra, and J. G. Beerends, "Perceptual evaluation of speech quality (pesq) the new itu standard for end-to-end speech quality assessment part i–time-delay compensation," *Journal of the Audio Engineering Society*, vol. 50, no. 10, pp. 755–764, 2002.