ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10888272

LiSenNet: Lightweight Sub-band and Dual-Path Modeling for Real-Time Speech Enhancement

Haoyin Yan*, Jie Zhang*, Cunhang Fan†, Yeping Zhou‡, Peiqi Liu‡
*NERC-SLIP, University of Science and Technology of China (USTC), Hefei, China
†Anhui Province Key Laboratory of Multimodal Cognitive Computation, Anhui University, Hefei, China
‡China Merchants Bank (CMB), Shenzhen, China

***Abstract*—Speech enhancement (SE) aims to extract the clean waveform from noise-contaminated measurements to improve the speech quality and intelligibility. Although learning-based methods can perform much better than traditional counterparts, the large computational complexity and model size heavily limit the deployment on latency-sensitive and low-resource edge devices. In this work, we propose a lightweight SE network (LiSenNet) for real-time applications. We design sub-band downsampling and upsampling blocks and a dual-path recurrent module to capture band-aware features and time-frequency patterns, respectively. A noise detector is developed to detect noisy regions in order to perform SE adaptively and save computational costs. Compared to recent higher-resource-dependent baseline models, the proposed LiSenNet can achieve a competitive performance with only 37k parameters (half of the state-of-the-art model) and 56M multiply-accumulate (MAC) operations per second.**

***Index Terms*—Speech enhancement, lightweight network, low complexity, real-time applications.**

I. Introduction

Speech enhancement (SE) aims to improve the instrumental quality and intelligibility of speech signals that are contaminated by ambient noises. Nowadays, speech-based applications are prevalent as a front-end module in systems, e.g., human-machine interaction, telephony, video-conferencing [1]–[3]. Traditional signal processing techniques are based on well-established theories and reasonable assumptions but often suffer from non-stationary acoustic scenes [4]. With the development of deep learning, deep neural network (DNN) based models can recover the clean speech in an end-to-end manner, even showing a superiority in the non-stationary case.

Formulated as a supervised learning problem, many DNN-based SE methods with impressive performance have been proposed recently. For example, convolution recurrent network (CRN) [5] utilizes convolution encoder-decoder and recurrent neural network (RNN) to capture local features and model temporal patterns. Deep complex convolution recurrent network (DCCRN) [6] improves the CRN by complex-valued operations. FullSubNet [7] is a cascade of the full-band and sub-band models, allowing to capture the global context while retaining the ability to model signal stationarity.

This work is supported by the National Natural Science Foundation of China (62101523), the joint AI laboratory of CMB-USTC (FTIT2022058), Anhui Province Major Science and Technology Research and Development Project (S2023Z20004) and USTC Research Funds of the Double First-Class Initiative (YD2100002008). (Correspondence: jzhang6@ustc.edu.cn)

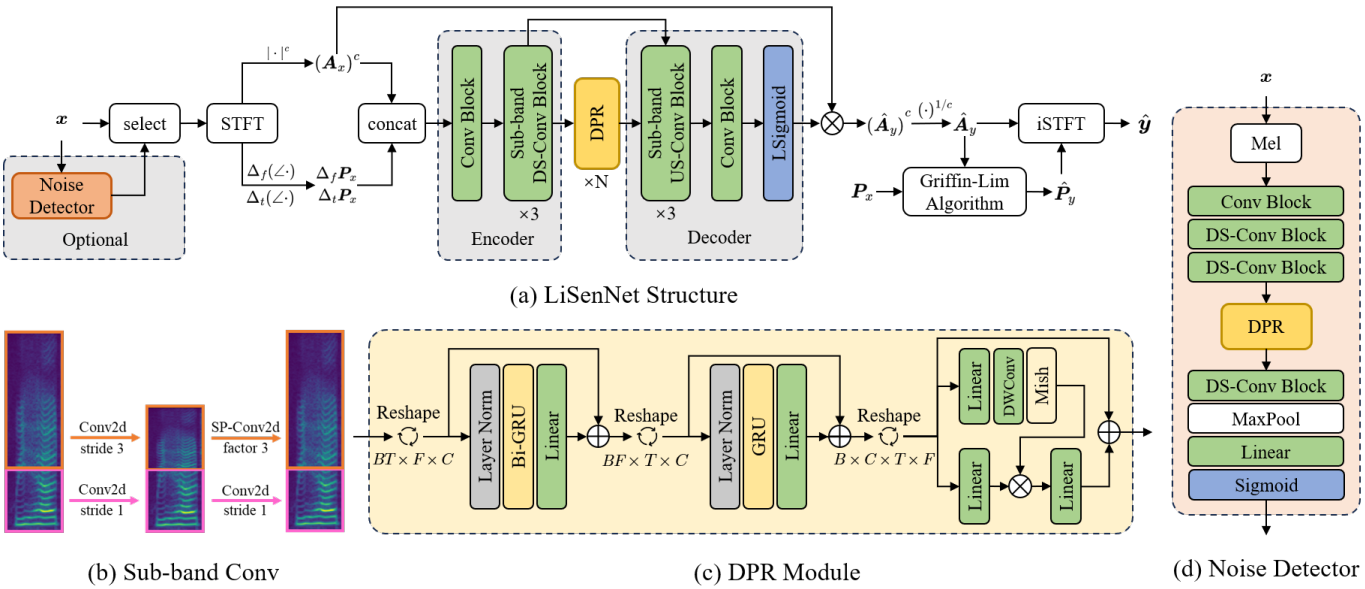
In spite of the promising performance, the large number of parameters and heavy computational burden hinder the deployment of these models on low-resource devices, e.g., hearing aids, headphones, IoT devices. This promotes the necessity of reducing the network complexity while maintaining performance. RNNoise [8] adopts a lightweight network to predict ideal critical band gains, and the coarse resolution of the bands is addressed by a pitch filter. In [9], a two-stage model is proposed to reduce the model size and computational complexity, which combines coarse-grained full-band mask estimation and low-frequency refinement. In [10], a lightweight model called FSPEN employs sub-band and full-band encoders to extract speech features. The FSPEN also utilizes an inter-frame path extension method to improve the dual-path architecture [11]. These methods can reach a balance between performance and complexity to some extent, yet it is still meaningful to develop a more lightweight SE model with comparable performance, which must be more appropriate for the deployment on low-resource edge devices.

In this work, we therefore propose a **Li**ghtweight **S**peech **en**hancement **Net**work, called **LiSenNet** for real-time speech applications. A sub-band downsampling-upsampling method is developed to capture band-aware features, and a dual-path recurrent module is designed to capture time and frequency dependencies efficiently. A post-processing step is then utilized to refine the spectral phase to improve the perceptual quality. In order to further reduce the computational burden, we apply a noise detector as an optional front-end module to detect noisy regions. This is rather benefitial for the inference efficiency in the instantaneous noise case. Experimental results show that our model achieves a competitive performance with only 37k parameters and 56M MACs per second. The noise detector can significantly reduce the real time factor (RTF) in case of low noise proportions in speech. The reproducible code and audio examples are available online[1].

II. METHODOLOGY

An overview of the proposed LiSenNet is shown in Fig. 1(a), which aims to recover the clean speech $\hat{\boldsymbol{y}} \in \mathbb{R}^L$ from the noisy input $\boldsymbol{x} \in \mathbb{R}^L$, where L denotes the waveform length. Next, we will introduce the model in detail.

[1]https://github.com/hyyan2k/LiSenNet



(a) LiSenNet Structure

(b) Sub-band Conv

(c) DPR Module

(d) Noise Detector

Fig. 1. (a) The framework of our proposed LiSenNet, where $|\cdot|$ and $\angle\cdot$ denote the magnitude and phase extractors, Δ_f and Δ_t the differential operators along frequency and time axis, $(\cdot)^c$ and $(\cdot)^{1/c}$ the power compress and decompress operations at a ratio of c, respectively; (b) an example of sub-band DS-Conv and sub-band US-Conv; (c) the proposed DPR module; (d) the optional noise detector, where Mel spectrogram is extracted as the input feature.

A. Model Input

First, we extract the complex spectrum $\boldsymbol{X} \in \mathbb{C}^{T\times F}$ of the noisy speech using the short-time Fourier transform (STFT) with T and F denoting the total time and frequency amounts, respectively. The power-compressed magnitude spectrum $(\boldsymbol{A}_x)^c \in \mathbb{R}^{T\times F}$ and phase spectrum $\boldsymbol{P}_x \in \mathbb{R}^{T\times F}$ can then be calculated, where c is the compression ratio. Similarly to [12], we set c = 0.3 to equalize the importance of quieter sounds relative to loud ones.

Since phase difference (PD) exhibits similar patterns as the magnitude spectrum [13], we also adopt the PD as input to help estimate the clean magnitude spectrum. The time-frequency (TF) domain PDs are computed as

$$\Delta_f \boldsymbol{P}_x(t,f) = \boldsymbol{P}_x(t,f) - \boldsymbol{P}_x(t,f-1), \tag{1}$$

$$\Delta_t \boldsymbol{P}_x(t,f) = \boldsymbol{P}_x(t,f) - \boldsymbol{P}_x(t-1,f) - 2\pi f \frac{Q}{M}, \tag{2}$$

where t and f are the frame and frequency indices, Q and M the frame shift and frame length, respectively. As the STFT is performed on overlapped frames, resulting in an offset in the phase, which would distort the structure of temporal PD, we thus subtract $2\pi f \frac{Q}{M}$ in (2) to obtain the baseband PD [14]. The power-compressed magnitude spectrum and noisy PD are concatenated as the model input.

B. Encoder and Decoder

In Fig. 1(a), both the encoder and decoder contain a series of convolution (Conv) blocks, where each block consists of a convolution layer, a layer normalization and a parametric rectified linear unit (PReLU) activation. To reduce the computational complexity, we exploit downsampling convolution (DS-Conv) blocks to halve the frequency-axis size in the encoder and upsampling convolution (US-Conv) blocks to recover the frequency resolution in the decoder. Skip connection is conducive to integrating low-level and high-level features.

Due to the fact that in general low-frequency bands play a more important role in human hearing than high-frequency bands [15], we propose the sub-band DS-Conv and sub-band US-Conv to maintain low-frequency resolution, e.g. see the illustrative example in Fig. 1(b). For low-frequency bands, Conv2d with a stride of 1 is utilized. For high-frequency bands, we apply Conv2d with a stride of 3 in the downsampling and sub-pixel convolution (SP-Conv2d) [16] with factor 3 in the upsampling. Therefore, the overall downsampling and upsampling factor of frequency bands becomes 2.

The learnable Sigmoid (LSigmoid) [17] is adopted as the last layer of the decoder to predict the mask, given by

$$\mathrm{LSigmoid}(x) = \beta / \left(1 + e^{-\alpha x}\right), \tag{3}$$

where β is a constant scalar set to 2.0, and $\alpha \in \mathbb{R}^F$ is trainable. Considering distinct patterns in high and low frequency bands of speech signals, LSigmoid can adaptively learn the function shapes for different bands.

C. Dual-Path Recurrent (DPR) Module

Since the emergence of DPRNN [11], dual-path dependent models have shown to be very promising for SE [10], [18], [19]. Specifically, time and frequency sequences are modeled sequentially to capture the time dependency in each band and the frequency dependency in each frame. In this work, we propose a DPR module as the backbone in Fig. 1(c), where the feature map $\boldsymbol{M} \in \mathbb{R}^{B\times C\times T\times F}$ is reshaped to $BT\times F\times C$ for frequency modeling and then reshaped to $BF\times T\times C$ for temporal modeling with B and C being the batch size and hidden dimension, respectively. To ensure the causality of our

model, we apply bidirectional GRU (Bi-GRU) for frequency modeling and unidirectional GRU for temporal modeling. The RNN nodes need to be decreased to reduce the computational complexity for real-time applications, resulting in the decay of modeling capacity. To address this issue, we utilize the convolutional gated linear unit (ConvGLU) modified from [20] as a channel mixer to fuse inter-channel information, which is composed of a linear layer, depthwise convolution [21] (DWConv) and Mish [22] activation. The DWConv can aggregate the nearest information and the gate mechanism allows for a fine-grained channel attention.

D. Phase Refinement

Instead of directly using the noisy phase $\boldsymbol{P}_x$ to synthesize the output speech [5], [23], we propose to utilize the Griffin-Lim Algorithm (GLA) [24], [25] as a post-processor to refine the phase spectrum. The GLA performs phase retrieval depending on the spectral consistency, which iteratively applies STFT and inverse STFT (iSTFT) to update the phase as

$$\boldsymbol{P}^{(k)} = \angle \mathrm{STFT}\left(\mathrm{iSTFT}\left(\boldsymbol{A}e^{i\boldsymbol{P}^{(k-1)}}\right)\right), \tag{4}$$

where k denotes the iteration number and $\boldsymbol{A}$ denotes the given magnitude. The GLA has been considered for blind source separation, e.g., in [26]. However, errors might be introduced at each iteration due to the possible speech distortion and residual noise in the enhanced magnitude spectrum, resulting in a limited phase accuracy [27]. Since advanced SE models can generate a more accurate magnitude, we propose to reconsider GLA for phase refinement as

$$\hat{\boldsymbol{P}}_y = \mathrm{GLA}\left(\hat{\boldsymbol{A}}_y, \boldsymbol{P}_x\right), \tag{5}$$

where $\hat{\boldsymbol{A}}_y$ denotes the estimated magnitude spectrum.

E. Noise Detector (ND)

The efficiency of existing SE models might decrease in the case of instantaneous noises, where the noise duration can be much shorter than the speech length. Turning on the SE module continuously would cause a large amount of unnecessary resource consumption. Therefore, as depicted in Fig. 1(d), we propose a noise detector as an optional front-end to detect frame-level noisy regions. We extract 64-dimensional Mel spectrograms to reduce the redundancy of the input features. The output binary values of the Sigmoid layer indicate whether each frame contains noise (0 for clean and 1 for noisy). Only noisy frames need to be sent to the subsequent SE back-end, such that the overall computational burden can be further saved, particularly in short noise cases.

F. Loss Function

In this work, we consider multiple loss functions to train the proposed LiSenNet. Similarly to [19], we apply the mean squared error (MSE) on the power-compressed magnitude spectrum and complex spectrum as

$$\mathcal{L}_{\mathrm{mag}} = \mathbb{E}\left[||(\boldsymbol{A}_y)^c - (\hat{\boldsymbol{A}}_y)^c||^2\right], \tag{6}$$

$$\mathcal{L}_{\mathrm{comp}} = \mathbb{E}\left[||\boldsymbol{Y}^c - \hat{\boldsymbol{Y}}^c||^2\right], \tag{7}$$

where $\boldsymbol{Y}^c = (\boldsymbol{A}_y)^c e^{i\boldsymbol{P}_y}$ and $\hat{\boldsymbol{Y}}^c = (\hat{\boldsymbol{A}}_y)^c e^{i\hat{\boldsymbol{P}}_y}$ denote the power-compressed complex spectrum of the clean and estimated speech, respectively.

Following [23], we use a discriminator D to further improve the perceptual speech quality, which is trained to predict the perceptual evaluation of speech quality (PESQ) [28], which is scaled to (0, 1). The discriminator encourages the model to generate the magnitude with a larger PESQ score, and the PESQ loss is given by

$$\mathcal{L}_{\mathrm{pesq}} = \mathbb{E}\left[||D((\boldsymbol{A}_y)^c, (\hat{\boldsymbol{A}}_y)^c) - 1||^2\right]. \tag{8}$$

If the noise detector is included, we choose the binary cross entropy (BCE) loss $\mathcal{L}_{\mathrm{bce}}$ as the training criterion. Therefore, the overall loss function for model training is given by

$$\mathcal{L} = \lambda_1\mathcal{L}_{\mathrm{mag}} + \lambda_2\mathcal{L}_{\mathrm{comp}} + \lambda_3\mathcal{L}_{\mathrm{pesq}} + \mathcal{L}_{\mathrm{bce}}, \tag{9}$$

where $\lambda_1, \lambda_2, \lambda_3$ are the empirical weights, which are respectively set to 0.9, 0.1 and 0.05 in this work.

III. EXPERIMENTS

A. Experimental Setup

We use two datasets to evaluate our proposed method. The first one is the widely-used SE benchmark Voice-Bank+DEMAND [29], which provides clean-noisy audio pairs. There are 11,572 utterances from 28 speakers for training and 872 utterances from 2 speakers for testing. Clean signals are mixed with noises at signal-to-noise ratios (SNRs) of $\{0, 5, 10, 15\}$ dB in the training set and $\{2.5, 7.5, 12.5, 17.5\}$ dB in the test set. All clips are resampled from 48 kHz to 16 kHz in experiments.

In addition, we create the WSJ0+ESC50 dataset for the evaluation of the noise detector, with clean speech utterances selected from Wall Street Journal (WSJ0) dataset [30] and noise samples sourced from ESC50 [31]. We randomly select a 4-second clip of each speech utterance, and the silent regions in noise recordings are removed using an energy threshold. Each 4-second clip contains 0.5 to 2 seconds of noise, with the SNR sampled uniformly between -10 and 5 dB. This dataset contains 14.2 hours for training and 0.7 hours for testing.

We perform STFT using a Hanning window with a length of 512 (32ms) and a hop size of 256 (16ms). The numbers of output channels for Conv blocks in the encoder, decoder, and noise detector are $\{4, 8, 12, 16\}$, $\{12, 8, 4, 1\}$, and $\{4, 8, 16, 32\}$, respectively. The number of hidden states is 24 for unidirectional GRU and 12 for Bi-GRU. The DPR module is repeated 2 times (i.e., N = 2 in Fig. 1(a)) in the network and the iterations for GLA are set to 2. Our model is trained using the AdamW [32] optimizer with 100 epochs. The L_2 norm for gradient clipping is set to 5.0. The learning rate starts from 5e-4 and reduces at a decay factor of 0.98 in each epoch.

B. Experimental Results

First, we show the performance of the proposed LiSenNet and other state-of-the-art (SOTA) methods in terms of the parameter amount (Para.), MACs, PESQ, and STOI [34] on

TABLE I
PERFORMANCE COMPARISION ON VOICEBANK+DEMAND DATASET.

Method	Para.	MACs	PESQ	STOI
Noisy	-	-	1.97	0.921
RNNoise [8]	0.06M	**0.04G**	2.33	0.922
CRN* [5]	17.58M	2.56G	2.54	0.936
DCCRN* [6]	3.67M	14.38G	2.78	0.940
FullSubNet* [7]	5.64M	30.87G	2.89	0.940
Fast FullSubNet* [33]	6.84M	4.14G	2.79	0.935
CCFNet+ (Lite) [9]	160k	390M	2.94	-
FSPEN [10]	79k	89M	2.97	**0.942**
LiSenNet	**37k**	56M	**3.07**	0.939

* We retrained these models to obtain results.

TABLE II
PERFORMANCE COMPARISION ON WSJ0+ESC50 DATASET.

Method	Para.	MACs	RTF	PESQ
Noisy	-	-	-	1.86
CRN [5]	17.58M	2.56G	0.042	2.80
DCCRN [6]	3.67M	14.38G	0.184	3.05
FullSubNet [7]	5.64M	30.87G	0.386	2.99
Fast FullSubNet [33]	6.84M	4.14G	0.100	2.75
LiSenNet	**37k**	56M	0.028	**3.08**
LiSenNet+ND	55k	**15M∼70M**	**0.020**	3.00

the VoiceBank+DEMAND dataset in Table I. It is clear that LiSenNet achieves competitive performance with a tiny model size and low computational burden. Compared to RNNoise which has a comparable resource consumption, our model shows an obvious superiority in the SE performance. In comparison with FSPEN, our LiSenNet improves the PESQ score by 0.1 with only 47% parameters and 63% MACs.

Results on the WSJ0+ESC50 dataset are presented in Table II. The RTF is measured with Intel(R) Xeon(R) E5-2620 v4 CPU. LiSenNet obtains the best performance with a much lower complexity in comparison with SOTA methods. At the expense of a slight performance decay, which is caused by the imprecision of noise detection and the information loss of clean regions, the inclusion of noise detector (ND) can further reduce the computational complexity. But this (of LiSenNet+ND) depends on the noise proportion, ranging from 15M to 70M in MACs, which however is still much lower than other methods. In the extreme case (no noise in speech), the proposed method only requires 15M MACs/second. Fig. 2 shows the RTF curve in terms of the noise proportion, demonstrating the effectiveness of the noise detector as a front-end, especially when noise proportion in speech is low.

Finally, we carry out an ablation study to analyze the impact of different components of the proposed LiSenNet in Table III. The ablation of input features and loss function validates that PD helps the magnitude estimation and $\mathcal{L}_{\text{pesq}}$ is beneficial for the perceptual quality. Replacing the sub-band Conv with a normal convolution layer (w/o sub-band Conv), removing the ConvGLU module (w/o ConvGLU), or freezing the learnable parameters in LSigmoid (w/o LSigmoid) will

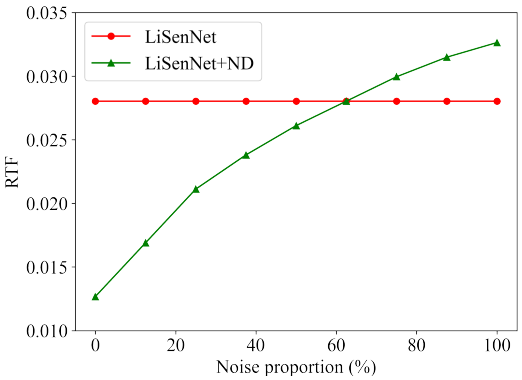


Fig. 2. The RTF under different noise proportion conditions.

TABLE III
ABLATION STUDY ON VOICEBANK+DEMAND DATASET.

Method	Para.	MACs	PESQ	STOI
LiSenNet	37k	56M	**3.07**	0.939
w/o PD	37k	56M	2.99	0.936
w/o $\mathcal{L}_{\text{pesq}}$	37k	56M	2.95	0.937
w/o sub-band Conv	30k	54M	3.01	0.938
w/o ConvGLU	31k	48M	2.98	0.937
w/o LSigmoid	37k	56M	3.05	0.938
DPR (N = 0)	**15k**	**22M**	2.68	0.913
DPR (N = 1)	26k	39M	2.97	0.935
DPR (N = 2)	37k	56M	3.07	0.939
DPR (N = 3)	48k	72M	3.06	**0.940**
GLA (Noisy Phase)	37k	56M	3.02	0.939
GLA (k = 1)	37k	56M	3.06	0.939
GLA (k = 2)	37k	56M	3.07	0.939
GLA (k = 3)	37k	56M	3.07	0.939

lead to a decrease in performance, verifying the necessity of these modules. The number of the DPR module can make a trade-off between complexity and performance, and N = 2 seems the best choice. Furthermore, the GLA is effective for refining the phase spectrum, but 2 iterations (k = 2) are enough for the convergence of performance.

IV. CONCLUSION

In this work, we proposed a lightweight network called LiSenNet for efficient real-time SE. The sub-band DS-Conv and US-Conv were proposed to maintain the resolution of low-frequency bands, and the dual-path recurrent module was used to model the intra-frame, inter-frame and inter-channel patterns. The denoised magnitude spectrum was estimated using the noisy magnitude spectrum and phase difference, and the GLA was adopted to refine the phase spectrum. The noise detector can be flexibly inserted to further reduce the computational burden depending on the noise proportion. The proposed LiSenNet showed a superiority in both the SE performance and complexity, which is thus more appropriate for the deployment onto low-resource edge devices.

REFERENCES

[1] Z. Chen, S. Watanabe, H. Erdogan, and J. R. Hershey, "Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks," in *Proc. Interspeech*, pp. 3274–3278, 2015.

[2] N. Modhave, Y. Karuna, and S. Tonde, "Design of matrix Wiener filter for noise reduction and speech enhancement in hearing aids," in *Proc. RTEICT*, pp. 843–847, 2016.

[3] K. Tan, X. Zhang, and D. Wang, "Real-time speech enhancement using an efficient convolutional recurrent network for dual-microphone mobile phones in close-talk scenarios," in *Proc. ICASSP*, pp. 5751–5755, 2019.

[4] J. Zhang, R. Tao, J. Du, and L.-R. Dai, "SDW-SWF: Speech distortion weighted single-channel Wiener filter for noise reduction," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 3176–3189, 2023.

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

[6] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, *et al.*, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Proc. Interspeech*, pp. 2472–2476, 2020.

[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 *Proc. ICASSP*, pp. 6633–6637, 2021.

[8] J.-M. Valin, "A hybrid dsp/deep learning approach to real-time full-band speech enhancement," in *Proc. MMSP*, pp. 1–5, 2018.

[9] F. Dang, H. Chen, Q. Hu, P. Zhang, and Y. Yan, "First coarse, fine afterward: A lightweight two-stage complex approach for monaural speech enhancement," *Speech Communication*, vol. 146, pp. 32–44, 2023.

[10] L. Yang, W. Liu, R. Meng, G. Lee, S. Baek, and H.-G. Moon, "FSPEN: an ultra-lightweight network for real time speech enahncment," in *Proc. ICASSP*, pp. 10671–10675, 2024.

[11] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-Path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. ICASSP*, pp. 46–50, 2020.

[12] S. Braun and I. Tashev, "A consolidated view of loss functions for supervised deep learning-based speech enhancement," in *Proc. TSP*, pp. 72–76, 2021.

[13] P. Mowlaee and R. Saeidi, "Time-frequency constraints for phase estimation in single-channel speech enhancement," in *Proc. IWAENC*, pp. 337–341, 2014.

[14] M. Krawczyk and T. Gerkmann, "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1931–1940, 2014.

[15] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech.," *The Journal of the Acoustical Society of America*, vol. 87, pp. 1738–1752, 1990.

[16] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, *et al.*, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. CVPR*, pp. 1874–1883, 2016.

[17] S.-W. Fu, C. Yu, T.-A. Hsieh, P. Plantinga, M. Ravanelli, X. Lu, *et al.*, "MetricGAN+: An improved version of metricgan for speech enhancement," in *Proc. Interspeech*, pp. 201–205, 2021.

[18] X. Le, H. Chen, K. Chen, and J. Lu, "DPCRN: Dual-path convolution recurrent network for single channel speech enhancement," in *Proc. Interspeech*, pp. 2811–2815, 2021.

[19] R. Cao, S. Abdulatif, and B. Yang, "CMGAN: Conformer-based metric gan for speech enhancement," in *Proc. Interspeech*, pp. 936–940, 2022.

[20] D. Shi, "TransNeXt: Robust foveal visual perception for vision transformers," *arXiv preprint arXiv:2311.17132*, 2023.

[21] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. CVPR*, pp. 1800–1807, 2017.

[22] D. Misra, "Mish: A self regularized non-monotonic neural activation function," *arXiv preprint arXiv:1908.08681*, 2019.

[23] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *Proc. ICML*, vol. 97, pp. 2031–2041, 2019.

[24] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," in *Proc. ICASSP*, vol. 8, pp. 804–807, 1983.

[25] N. Perraudin, P. Balazs, and P. L. Søndergaard, "A fast Griffin-Lim algorithm," in *Proc. WASPAA*, pp. 1–4, 2013.

[26] N. Sturmel and L. Daudet, "Iterative phase reconstruction of Wiener filtered signals," in *Proc. ICASSP*, pp. 101–104, 2012.

[27] F. Mayer, D. S. Williamson, P. Mowlaee, and D. Wang, "Impact of phase estimation on single-channel speech separation based on time-frequency masking," *J. Acoust. Soc. Am.*, vol. 141, no. 6, pp. 4668–4679, 2017.

[28] 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 *Proc. ICASSP*, vol. 2, pp. 749–752, 2001.

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

[30] J. S. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Complete." [Online]. Available: https://catalog.ldc.upenn.edu/LDC93S6A.

[31] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. ACM MM*, pp. 1015–1018, 2015.

[32] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[33] X. Hao and X. Li, "Fast FullSubNet: Accelerate full-band and sub-band fusion model for single-channel speech enhancement," *arXiv preprint arXiv: 2212.09019*, 2023.

[34] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.