Two-Stage Speech Enhancement Network Based on Complex Spectrum Mapping and Voice Activity Detection

2024 IEEE International Conference on Software System and Information Processing (ICSSIP) | 979-8-3315-0613-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICSSIP63203.2024.11012578

Wang Chen
School of Communication and Information Engineering
Chongqing University of Posts and Telecommunications
Chongqing, China
1467598798@qq.com

Yi Zhou*
School of Communication and Information Engineering
Chongqing University of Posts and Telecommunications
Chongqing, China
zhouy@cqupt.edu.cn
*Corresponding author

Liwen Tan
School of Communication and Information Engineering
Chongqing University of Posts and Telecommunications
Chongqing, China
2395989874@qq.com

Yin Liu
School of Communication and Information Engineering
Chongqing University of Posts and Telecommunications
Chongqing, China
liuyin@cqupt.edu.cn

***Abstract*—In the multi-stage method, each stage model only focuses on one task, and the collaborative learning of multiple tasks often achieves better results than the single-stage model. Inspired by this, this paper proposes a two-stage speech enhancement model based on complex spectrum mapping and voice activity detection(VAD). In the first stage, the model directly predicts the real and imaginary parts of the clean speech. In the second stage, the model learns the frame-level information, and then applies it to the predicted complex spectrum of the first stage. In order to extract richer time-frequency information in the encoder, this paper adopts multi-scale gated convolution. At the same time, Conformer is used in the bottleneck layer to realize global and local attention calculation in the time dimension and frequency dimension. Through this design, the model proposed in this paper can effectively enhance the speech signal and accurately detect voice activity. Finally, the experimental results on the Voicebank+Demand dataset show that the model proposed in this paper is superior to other comparison models in terms of objective and subjective indicators. At the same time, in order to further verify the generalization ability of the model, the results on the Librispeech dataset show that the model proposed in this paper has good results under different noises and signal-to-noise ratios.**

***Keywords*—speech enhancement, voice activity detection, two stage, signal processing**

I. Introduction

In real-world communication scenarios, speech signals are inevitably subjected to various kinds of interferences. Some of these interferences arise from the speech acquisition environment, while others are due to issues such as compression and packet loss during the transmission of speech signals. The goal of speech enhancement is to remove these interferences as much as possible, while preserving the original speech information. Traditional speech enhancement methods typically rely on statistical signal processing principles, such as spectral subtraction [1], minimum mean square error estimation [2], Wiener filtering [3], and Kalman filtering [4]. However, these conventional methods often make the following assumptions: speech and noise are uncorrelated, and noise follows a Gaussian distribution. As a result, when dealing with non-stationary noise, these methods often achieve limited performance.

Recently, deep neural networks (DNNs) have greatly advanced speech enhancement algorithms due to their superior ability to handle non-stationary noises, outperforming traditional statistical signal processing methods. Deep learning-based speech enhancement approaches can be categorized into time-domain methods and time-frequency domain methods. The former directly takes noisy speech as input to a neural network and outputs the enhanced speech[5,6]. Time-frequency methods can be further classified into two types: mask-based methods and mapping-based methods. Mask-based methods learn a filter and apply it to the noisy spectrogram before performing an inverse Fourier transform to reconstruct the speech signal. Mapping-based methods, on the other hand, directly establish a mapping relationship between the noisy and clean speech.

Due to the nonstructural nature of phase and its wrapping characteristics, previous studies have mostly predicted the magnitude spectrum and then coupled the predicted magnitude with the noisy signal's phase to reconstruct the speech signal. However, the lack of phase information inevitably causes distortion. Hu et al. [7] demonstrated that better utilization of phase information in speech signals can significantly improve the quality of enhanced speech. Recently, multi-stage methods have gained widespread use in speech enhancement tasks. Schroter et al. [8] proposed a real-time two-stage speech enhancement network, where the first stage operates in the Equivalent Rectangular Bandwidth(ERB) domain to enhance the speech envelope,and the second stage uses deep filtering to enhance the speech's periodic components.Earlier studies have also attempted to combine speech enhancement tasks with voice activity detection (VAD) tasks to create multi-task learning frameworks [9,10]. However, these studies mainly treat speech enhancement networks as preprocessing or auxiliary modules to improve the accuracy of VAD. Zhang et al. [11] integrated speech enhancement with VAD to create a multi-task learning framework, but they treated VAD as an auxiliary task and did not fully utilize the model's VAD predictions. Moreover, their

This work is supported in part by National Natural Science Foundation of China funded project(62301096).

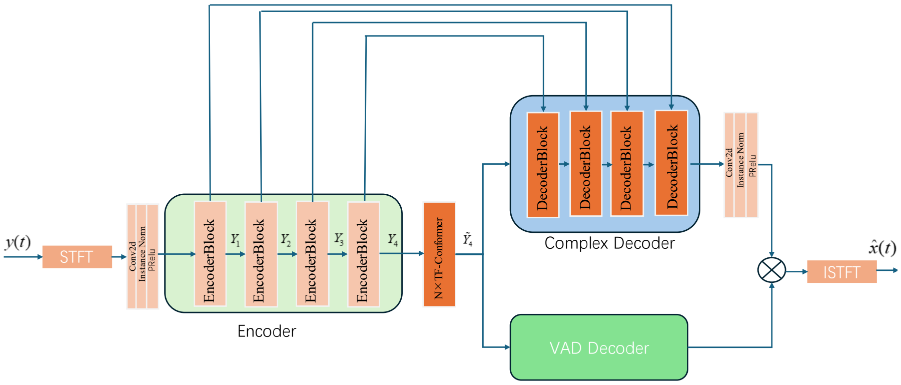

Figure. 1. System flowchart of CSM-VANet.

focus was mainly on magnitude information while ignoring phase information. However, speech signal reconstruction depends heavily on phase information, leading to suboptimal reconstruction results. In this paper, the first stage directly predicts the real and imaginary parts of the speech signal's spectrum, avoiding the direct prediction of the phase spectrum, which is often considered a challenging task.

The main contributions of this paper are summarized as follows:

1. We propose a two-stage network: Stage 1 predicts the complex spectrogram, and Stage 2 predicts VAD.
2. We apply Multi-Scale Gated Convolutional(MSGC) layers rather than conventional 2D convolutions. In the content branch, we use depthwise separable convolutions with different dilation rates and kernel sizes, which allows us to extract richer time-frequency information with fewer computational resources.
3. We employ a time-frequency Conformer module as the bottleneck layer, which can effectively model both the time and frequency dimensions of the features.

Since we propose a Complex Spectrum Mapping (CSM) framework that integrates VAD for enhanced audio processing, the overall network is named CSM-VANet.

The paper is organized as follows: Section II describes the proposed method.Section III details the datasets and experimental setup.Section IV presents experimental results, and Section V concludes the whole paper.

II. Methodology

A. Signal Model

Assume that the time-domain signals $x(t)$ 、 $n(t)$ 、 $y(t)$, represent the clean speech signal, noise signal, and noisy signal, respectively. this relationship in the time domain can be expressed as:

$$y(t) = x(t) + n(t) \tag{1}$$

Transforming equation (1) into the time-frequency domain :

$$Y(k,f) = X(k,f) + N(k,f) \tag{2}$$

where $X(k,f)$ 、 $N(k,f)$ 、 $Y(k,f)$ represent the complex spectra of the clean speech signal, noise signal, and noisy signal, respectively; f denotes the frequency index, and k denotes the time index. For ease of understanding, the time and frequency indices will be omitted in the subsequent equations.

In a speech enhancement network based on complex spectral mapping, a DNN is used to directly predict the real and imaginary parts of the clean speech spectrum. Then, the inverse short-time Fourier transform (ISTFT) is applied to reconstruct the time-domain signal.

B. Model structure

1. Overall structure

As shown in Figure 1,the proposed CSM-VANet consists of an input convolution, an encoder, a bottleneck layer, a complex decoder, a VAD decoder, and an output convolution. The input features are the noisy speech's magnitude spectrum, real part spectrum, and imaginary part spectrum. The input convolution maps the input feature channels to 24 and is composed of 2D convolutions, instance normalization, and parametric rectified linear unit (PReLU). The output convolution has the same structure as the input convolution, directly taking the output from the complex decoder as its input to predict the magnitude spectrum. The encoder is made up of four MSGC with down-sampling, where each down-sampling operation halves the frequency dimension to reduce computational cost. Correspondingly, the complex decoder consists of four MSGC with up-sampling. The bottleneck layer is responsible for modeling the temporal and frequency dimensions of the feature maps. The VAD decoder combines the outputs of the encoder and the bottleneck layer to predict frame-level VAD information. Skip connections are used between the encoder and complex decoder to mitigate the vanishing gradient problem.

2. EncoderBlock

As shown in Figure 2, the encoder's EncoderBlock consists of multi-scale gated convolutions, which include the gated convolution branch, the context branch, and down-sampling convolutions. Multi-scale gated convolutions have been successfully applied in the field of computer vision[12],so we use them instead of 2D convolution layers to extract features.

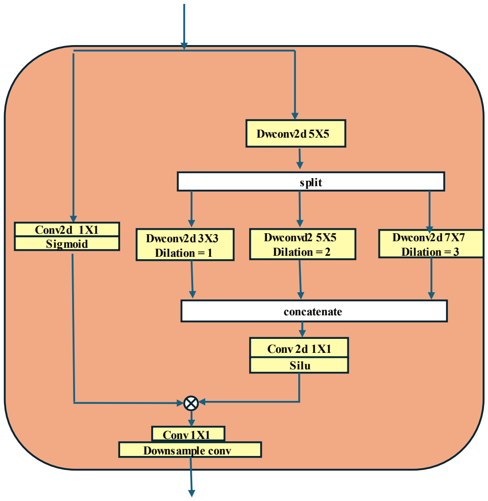

Figure. 2. Multi-Scale Gated Convolution in the Encoder Block

The context branch uses three depthwise convolution (DWConv) layers with different dilation rates to capture time-frequency information at small, medium, and large receptive fields. To further reduce computational complexity, the features are decomposed along the channel dimension before being input into the different scale convolutions. Finally, the outputs of the three scales are concatenated along the channel dimension and then multiplied pointwise with the gated output to obtain the multi-scale output. Afterward, a down-sampling convolution module is applied to halve the frequency dimension and reduce computation. The decoder block has a similar structure, but it uses transposed convolutions for up-sampling.

3. *Bottleneck layer*

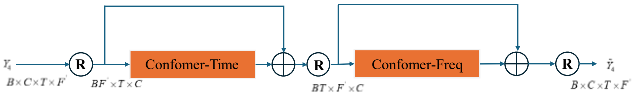

Figure. 3. Diagrams of the TF-Conformer.

As shown in Figure 3,the input feature map $Y_4 \in R^{B\times C\times T\times F^{'}}$ is reshaped into $Y_{\mathrm{T}} \in R^{BF^{'}\times T\times C}$ to capture the temporal dependencies for the first Conformer block. The output is added element-wise to the input to form a new feature map $Y_F \in R^{BT\times F^{'}\times C}$, which is then passed to the second Conformer block to capture the frequency dependencies. The final output is reshaped back to the input size. More details can be found in the CMGAN[13].

4. *VAD decoder*

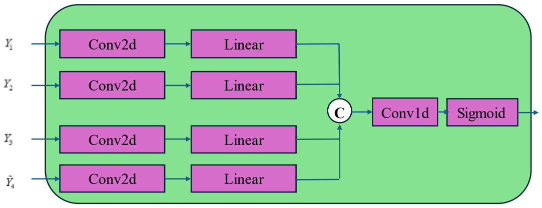

Figure. 4. The structure of the VAD Decoder.

As shown in Figure 4, let $Y_i \in R^{B\times C_i\times T\times F_i}$ represent the output of the i-th EncoderBlock and $\tilde{Y}_4 \in R^{B\times C_4\times T\times F_4}$ represent the output of the bottleneck layer. The VAD decoder first passes the input through convolution and linear layers to map the channel dimension and frequency dimension to 1, respectively. The output of the linear layers is then concatenated along the channel dimension to form the feature map, which is passed through a 1D convolution layer to map the channels to 1. Finally, a sigmoid activation function is applied to output frame-level VAD information.

C. *Multi-Target Loss*

This paper employs multiple loss functions to enhance the performance of speech enhancement models across various domains. Specifically, we use time-domain loss, magnitude loss, complex loss, and VAD loss:

$$L_{Time} = E_{x,\hat{x}}\left[\| x - \hat{x} \|_1\right], \tag{3}$$

$$L_{Mag} = E_{X_m,\hat{X}_m}[\| X_m - \hat{X}_m \|^2], \tag{4}$$

$$L_{com} = E_{X_r,\hat{X}_r}[\| X_r - \hat{X}_r \|^2] + E_{X_i,\hat{X}_i}[\| X_i - \hat{X}_i \|^2], \tag{5}$$

$$L_{vad} = L_{BCE}(v,\hat{v}) \cdot \tag{6}$$

In these functions, x、X_m、X_r、X_i、v represent clean speech, magnitude spectrum, real part spectrum, imaginary part spectrum, and VAD label, respectively. The symbols with ^ on top of the letters indicate the model's outputs.The final loss function, L_{total} is a linear combination of these individual loss functions:

$$L_{total} = \alpha_1 L_{Time} + \alpha_2 L_{Mag} + \alpha_3 L_{com} + \alpha_4 L_{vad} \tag{7}$$

where α_1、α_2、α_3 and α_4 are empirically set to 0.2、0.8、0.2 and 0.1 respectively.

III. Experiment

A. *Datasets*

To evaluate the model proposed in this paper, we used two datasets: the VoiceBank + DEMAND dataset[14] and the LibriSpeech dataset[15]. VoiceBank + DEMAND includes paired noisy and clean speech clips sampled at 48 kHz. The clean speech clips are sourced from the Voice Bank corpus, comprising 11,572 speech clips from 28 different speakers for training and 872 clips from two unseen speakers for testing. The clean clips are mixed with 10 types of noise with SNR levels of 0 dB, 5 dB, 10 dB, and 15 dB in the training set, and 2.5 dB, 7.5 dB, 12.5 dB, and 17.5 dB in the test set.

To validate the noise reduction performance of CSM-VANet under various noise types and SNR levels, we synthesize noisy speech based on the LibriSpeech dataset. The train-clean and test-clean subsets are used as the training and testing corpora, respectively. The test set includes three noise types—PCAFETER, PSTATION, and NRIVER—from the DEMAND database, with the remaining noise types used for training.For the training set, SNRs are randomly selected within the range of -10 dB to 15 dB to create noisy speech. For the test set, clean speech from the LibriSpeech test set and unused noise types are combined to create noisy speech at SNR levels of -5 dB, 0 dB, 5 dB, 10 dB, and 15 dB. Finally, 200,000 audio clips,

TABLE I. Comparison and ablation results on VoiceBank+DEMAND dataset. "–" indicates that the results are not provided in the original paper.

Model	#Param.	PESQ	STOI	CSIG	CBAK	COVL
Noisy		1.97	0.91	3.35	2.44	2.63
SEGAN	43.2M	2.16	0.92	3.48	2.94	2.80
METRICGAN[19]	1.86 M	2.86	–	3.99	3.18	3.42
PHASE[20]	–	2.99	–	4.21	3.55	3.62
TSTNN[21]	0.92M	2.96	0.95	4.17	3.53	3.49
DEMUCS	128M	3.07	0.95	4.31	3.40	3.63
METRICGAN+[22]	–	3.15	–	4.14	3.16	3.64
SE-CONFORMER[23]	–	3.13	0.95	4.45	3.55	3.82
GaGNet[24]	5.94 M	2.94	0.95	4.26	3.45	3.59
DeepFilter2	1.78	2.81	0.94	4.14	3.31	3.46
Vsanet	3.1	2.98		4.21	3.51	3.60
CSM-VANet	8.81M	**3.31**	**0.96**	**4.62**	**3.82**	**4.06**
w/o msgc	8.79M	3.24	0.95	4.50	3.61	3.81
w/o vad decoder	4.67M	3.22	0.94	4.48	3.58	3.76
w/o time loss	8.81M	3.29	0.95	4.57	3.79	4.01
w/o mag loss	8.81M	3.18	0.94	4.33	3.61	3.64

each 2 seconds long, were simulated for training, generating 50 audio clips for each noise type and SNR level, resulting in a total of 750 test data clips .

B. Evaluation Metrics

To evaluate the quality of the enhanced speech, we use a comprehensive set of evaluation metrics: PESQ [16], STOI, CSIG, CBAK, and COVL. PESQ is used to assess speech quality, with scores ranging from [-1, 4.5]. STOI evaluates speech intelligibility, with values ranging from [0, 1]. The three composite measures—CSIG, CBAK, and COVL—predict the mean opinion score (MOS) for signal distortion, background noise intrusion, and overall quality, respectively, with values ranging from [1, 5]. For all of these metrics, higher values indicate better performance.

C. Experimental setup

In the experiments, all audio clips are resample to 16 kHz, and during training, the audio is divided into 2-second segments. Input features are extracted from the waveform using STFT, with the FFT size, Hanning window size, and hop size set to 400, 400, and 100, respectively. The model is trained using the AdamW optimizer [17] for 100 epochs, with the initial learning rate set to 0.0005, halved every 30 epochs.Given the effectiveness of power compression in speech enhancement[18], we apply power compression to the magnitude while keeping the phase unchanged. The optimal compression coefficient is set to 0.3.The number of channels in the encoders is {24, 48, 96, 192}. The convolution kernel size and stride are set to (3, 3) and (1, 2), respectively. The number of conformer blocks (N), the batch size (B), and the number of channels (C) in the input convolution are set to 4, 4, and 24, respectively. In the VAD decoder, the kernel size of the 1D convolution is set to 3, with padding of 1 to ensure consistent length.

IV. Results and Analysis

A. Comparison with previous advanced baseline

As shown in Table I, we can observe that the CSM-VANet model performs exceptionally well in the speech enhancement task. With 8.81M parameters, CSM-VANet achieves PESQ, STOI, CSIG, CBAK, and COVL scores of 3.31, 0.96, 4.62, 3.82, and 4.06, respectively. These scores are the highest among all models, demonstrating the strong performance of CSM-VANet in speech enhancement.

B. Ablation study

In the ablation experiments, we examine the impact of removing different components on the performance of CSM-VANet. For instance, "w/o msgc" represents replacing the multi-scale gated convolution with a regular gated convolution. The PESQ score is 3.24, and the STOI is 0.95. Although slightly lower than CSM-VANet, it still maintains high performance. "w/o vad decoder" represents removing the VAD decoder, thus performing single-stage speech enhancement only. In this case, the PESQ score is 3.22, and STOI is 0.94, demonstrating that the VAD decoder also contributes significantly to model performance.

To further validate the effectiveness of the multi-objective loss function, we conducte an ablation study on the loss functions used. "w/o time loss" refers to removing the time-domain loss function. The model's PESQ score is 3.29, and STOI is 0.95, which is slightly lower than CSM-VANet but still higher than the other ablation models. "w/o mag loss" refers to removing the magnitude loss function, resulting in a PESQ score of 3.18 and STOI of 0.94.

C. Results on librispeech

To further evaluate the generalization ability of CSM-VANet, we conducte experiments using three noise types that did not appear in the training set. We computed the PESQ and STOI scores for CSM-VANet on the LibriSpeech dataset under different types of noise and various signal-to-noise ratios (SNRs). Table 2 shows the PESQ and STOI scores for CSM-VANet and other networks on the LibriSpeech dataset at different SNRs.

From Table II, it can be seen that, compared to noisy speech, the proposed method improves PESQ scores by 62.3%, 93.1 %, 124.6 %, 84.3%, and 53.2% at SNRs of -5 dB, 0 dB, 5 dB, 10 dB, and 15 dB, respectively. The STOI scores improve by 12.1%, 5.9%, 3.3%, 2.1%, and 1%, respectively. Compared to the two-stage model Deepfilter2, both PESQ and STOI scores show significant improvements, indicating that the proposed model has stronger generalization ability and better speech enhancement performance.

TABLE II. Results of different noise and signal-to-noise ratio on the librispeech dataset.

SNR	Method	PCAFETER		PSTATION		NRIVER		Averge	
		PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI
-5dB	Noisy	1.06	0.74	1.06	0.71	1.16	0.76	1.09	0.74
	Deepfilter2	1.46	0.79	1.66	0.79	1.70	0.81	1.61	0.8
	CSM-VANet	1.62	0.84	1.72	0.81	1.97	0.85	1.77	0.83
0 dB	Noisy	1.16	0.85	1.17	0.85	1.14	0.84	1.16	0.85
	Deepfilter2	2.03	0.88	2.01	0.87	1.99	0.82	2.01	0.86
	CSM-VANet	2.16	0.90	2.33	0.91	2.23	0.88	2.24	0.9
5 dB	Noisy	1.24	0.90	1.23	0.91	1.31	0.92	1.26	0.91
	Deepfilter2	2.57	0.91	2.62	0.92	2.63	0.91	2.61	0.91
	CSM-VANet	2.83	0.93	2.83	0.94	2.83	0.94	2.83	0.94
10 dB	Noisy	1.74	0.94	1.77	0.94	1.84	0.95	1.78	0.94
	Deepfilter2	2.94	0.96	2.87	0.92	2.91	0.93	2.91	0.94
	CSM-VANet	3.24	0.97	3.28	0.96	3.32	0.95	3.28	0.96
15 dB	Noisy	2.30	0.95	2.37	0.96	2.33	0.96	2.33	0.96
	Deepfilter2	3.21	0.96	3.25	0.97	3.44	0.97	3.3	0.97
	CSM-VANet	3.57	0.96	3.63	0.97	3.50	0.98	3.57	0.97

V. CONCLUSIONS

This paper proposes a two-stage speech enhancement model based on complex spectrum mapping and VAD, achieving significant performance improvements. In the first stage, the model predicts the real and imaginary parts of clean speech, while the second stage incorporates frame-level speech activity detection. Multi-scale gated convolution and a Conformer structure are used to extract richer time-frequency information and perform global and local attention in both time and frequency dimensions. Experimental results on the Voicebank and LibriSpeech datasets demonstrate good noise reduction and generalization. Future work aims to reduce the model's computational and parameter requirements for real-time applications.

REFERENCES

[1] Boll S. Suppression of acoustic noise in speech using spectral subtraction[J]. IEEE Transactions on acoustics, speech, and signal processing, 1979, 27(2): 113-120.

[2] Ephraim Y, Malah D. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator[J]. IEEE Transactions on acoustics, speech, and signal processing, 1984, 32(6): 1109-1121.

[3] Lim J, Oppenheim A. All-pole modeling of degraded speech[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1978, 26(3): 197-210.

[4] Paliwal K, Basu A. A speech enhancement method based on Kalman filtering[C]//ICASSP'87. IEEE International Conference on Acoustics, Speech, and Signal Processing. IEEE, 1987, 12: 177-180.

[5] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," in Proc. Interspeech, 2017, pp. 3642–3646.

[6] Alexandre D´efossez, Gabriel Synnaeve, and Yossi Adi, "Real Time Speech Enhancement in the Waveform Domain," inProc.Interspeech2020,2020,pp.3291–3295.

[7] Yanxin Hu, Yun Liu, Shubo Lv, Mengtao Xing, Shimin Zhang, Yihui Fu, Jian Wu, Bihong Zhang, and Lei Xie, "DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement," in Proc. Interspeech 2020, 2020, pp. 2472–2476.

[8] Schröter H, Maier A, Escalante-B A N, et al. Deepfilternet2: Towards real-time speech enhancement on embedded devices for full-band audio[C]//2022 International Workshop on Acoustic Signal Enhancement (IWAENC). IEEE, 2022: 1-5.

[9] Qing Wang, Jun Du, Xiao Bao, Zi-Rui Wang, Li-Rong Dai, and Chin-Hui Lee, "A universal VAD based on jointly trained deep neural networks," in Proc. Interspeech 2015, 2015, pp. 2282–2286.

[10] Xu Tan and Xiao-Lei Zhang, "Speech enhancement aided end-to-end multi-task learning for voice activity detection," in ICASSP 2021 - 2021 IEEE International ConferenceonAcoustics,SpeechandSignalProcessing (ICASSP), 2021, pp. 6823–6827.

[11] Zhang Y, Zou H, Zhu J. Vsanet: Real-Time Speech Enhancement Based on Voice Activity Detection and Causal Spatial Attention[C]//2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2023: 1-8.

[12] Li S, Wang Z, Liu Z, et al. Moganet: Multi-order gated aggregation network[C]//The Twelfth International Conference on Learning Representations. 2023.

[13] Abdulatif S, Cao R, Yang B. Cmgan: Conformer-based metric-gan for monaural speech enhancement[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024.

[14] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noiserobust text-to-speech." in Proc. SSW, 2016, pp. 146–152.

[15] Panayotov V, Chen G, Povey D, et al. Librispeech: an asr corpus based on public domain audio books[C]//2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2015: 5206-5210.

[16] Hu Y, Loizou P C. Evaluation of objective quality measures for speech enhancement[J]. IEEE Transactions on audio, speech, and language processing, 2007, 16(1): 229-238.

[17] Loshchilov I. Decoupled weight decay regularization[J]. arXiv preprint arXiv:1711.05101, 2017.

[18] Li A, Zheng C, Peng R, et al. On the importance of power compression and phase estimation in monaural speech dereverberation[J]. JASA express letters, 2021, 1(1).

[19] Fu S W, Liao C F, Tsao Y, et al. Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement[C]//International Conference on Machine Learning. PmLR, 2019: 2031-2041.

[20] D. Yin, C. Luo, Z. Xiong, and W. Zeng, "PHASEN: A phaseand-harmonics-aware speech enhancement network," in Proc. AAAI, 2020, vol. 34, pp. 9458–9465

[21] Wang K, He B, Zhu W P. TSTNN: Two-stage transformer based neural network for speech enhancement in the time domain[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 7098-7102.

[22] Fu S W, Yu C, Hsieh T A, et al. Metricgan+: An improved version of metricgan for speech enhancement[J]. arXiv preprint arXiv:2104.03538, 2021.

[23] E. Kim and H. Seo, "Se-Conformer: Time-Domain Speech Enhancement using Conformer," in Proc. Interspeech, 2021.

[24] Li A, Zheng C, Zhang L, et al. Glance and gaze: A collaborative learning framework for single-channel speech enhancement[J]. Applied Acoustics, 2022, 187: 108499.