ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10094718

TIME-DOMAIN SPEECH ENHANCEMENT ASSISTED BY MULTI-RESOLUTION FREQUENCY ENCODER AND DECODER

Hao Shi[1], *Masato Mimura*[1], *Longbiao Wang*[2], *Jianwu Dang*[2], *Tatsuya Kawahara*[1]

[1]Graduate School of Informatics, Kyoto University, Kyoto, Japan
[2]Tianjin Key Laboratory of Cognitive Computing and Application,
College of Intelligence and Computing, Tianjin University, Tianjin, China

ABSTRACT

Time-domain speech enhancement (SE) has recently been intensively investigated. Among recent works, DEMUCS [1] introduces multi-resolution STFT loss to enhance performance. However, some resolutions used for STFT contain non-stationary signals, and it is challenging to learn multi-resolution frequency losses simultaneously with only one output. For better use of multi-resolution frequency information, we supplement multiple spectrograms in different frame lengths into the time-domain encoders. They extract stationary frequency information in both narrowband and wideband. We also adopt multiple decoder outputs, each of which computes its corresponding resolution frequency loss. Experimental results show that (1) it is more effective to fuse stationary frequency features than non-stationary features in the encoder, and (2) the multiple outputs consistent with the frequency loss improve performance. Experiments on the Voice-Bank dataset show that the proposed method obtained a 0.14 PESQ improvement.

Index Terms— Speech enhancement, time domain, multi-resolution spectrograms.

1. INTRODUCTION

Speech enhancement (SE) has been extensively studied because noise often corrupts speech signals collected in real-world scenarios [2, 3, 4], which significantly degrades the performance of speech applications [5, 6, 7, 8, 9, 10, 11]. SE aims to recover speech components from noisy signals [12]. Deep learning-based SE [13, 14] methods have been shown to perform better than traditional methods [15, 16]. Supervised learning-based SE can be classified into frequency-domain [13], and time-domain [17] methods. Frequency-domain SE that uses only magnitude information has been mainly studied because it presumes that the human ear is less sensitive to phase information than magnitude information. This issue was supplemented and corrected by subsequent studies [18].

Recently, SE systems that process magnitude and phase information simultaneously achieves impressive performance [19, 20]. There are two approaches to handle phase information: enhancement in complex-domain [19, 20] and time-domain [1]. Complex-domain SE [20, 21] processes the Fourier transform's real and imaginary parts. Time-domain SE [1, 22] directly inputs the time-domain waveform and outputs enhanced features.

Among time-domain SE models, DEMUCS [1] has demonstrated state-of-the-art performance. It is based on the standard U-Net [23] structure and optimized by minimizing the L1 regression loss and supplemented by multi-resolution spectrogram domain losses [24]. DEMUCS exploits frequency-domain information through the spectrogram-domain loss, which significantly improves the stability of the model training. Furthermore, different from other time-domain models [25, 26, 17], DEMUCS introduces upsampling [27], and downsampling [27] processing based on sinc interpolation before the encoder and after the decoder, respectively, and the interpolation of the redundant information can alleviate information loss or distortion [28] caused by SE.

DEMUCS still has two drawbacks: (1) Some of the supplemented frequency-domain information contains non-stationary signals. Speech signals can be regarded as short-term stationery with an interval between 10ms and 30ms [29]. The Fourier transform presupposes that the signal is stationary [30]. However, in addition to 32ms, DEMUCS also adopts STFT of 64ms and 128ms. (2) It needs to learn the frequency loss of different resolutions simultaneously with one output. Multiple learning targets with a single output make training the neural network difficult.

In this study, we investigate the better use of multi-resolution frequency information from the perspective of the encoder and decoder, respectively. First, instead of using non-stationary frequency information in the output, we incorporate multi-resolution stationary frequency-domain information into the time-domain SE encoder layer by layer. The multi-resolution spectrograms are supplemented to provide frequency domain information. According to the length of framing time, spectrograms can be divided into wideband and narrowband [31]. These two kinds of spectrograms are much different and show a certain complementarity: Wideband (about 3ms length of framing time) spectrogram can capture the rapid amplitude changes [32] and clear speech formant information; Narrowband (about 20ms length of framing time) spectrogram has better spectral resolutions and captures harmonics information. Furthermore, the SE system [24] trained with multi-resolution information outputs significantly higher perceived quality in mean opinion score (MOS). Considering these, we incorporate frequency information of 8ms, 16ms, and 32ms into the model. Second, we propose using multiple time-domain decoders by downsampling, each corresponding to only one resolution of frequency-domain loss.

In the following sections, we will introduce related work in Section 2. We will introduce the proposed method in Section 3. In Section 4, the experimental settings and results will be introduced. The conclusion will be introduced in Section 5.

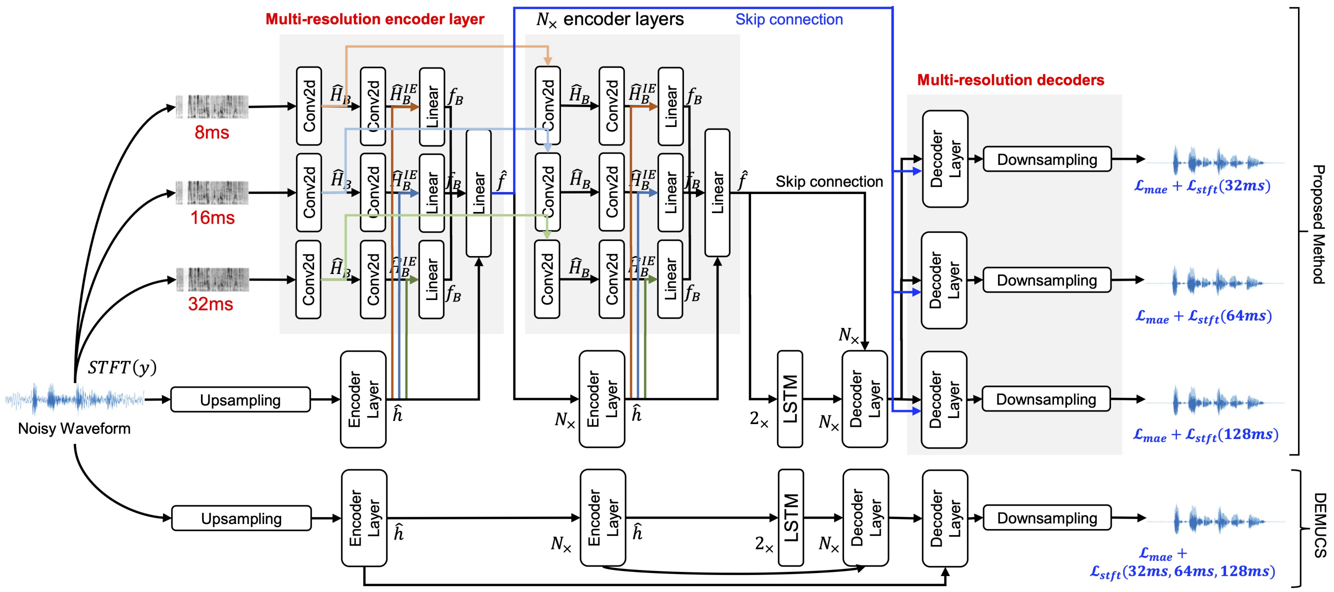
Multi-resolution encoder layer
$N_\times$ encoder layers
Skip connection
Multi-resolution decoders
Proposed Method
DEMUCS
8ms
16ms
32ms
$STFT(y)$
Noisy Waveform
Upsampling
Conv2d
Linear
$\hat{H}_B$
$\hat{H}_B^{IE}$
f_B
$\hat{f}$
$\hat{h}$
Encoder Layer
$N_\times$
$2_\times$
LSTM
Decoder Layer
Skip connection
Downsampling
$\mathcal{L}_{mae} + \mathcal{L}_{stft}(32ms)$
$\mathcal{L}_{mae} + \mathcal{L}_{stft}(64ms)$
$\mathcal{L}_{mae} + \mathcal{L}_{stft}(128ms)$
$\mathcal{L}_{mae} + \mathcal{L}_{stft}(32ms, 64ms, 128ms)$

Fig. 1. Flowchart of the proposed method and DEMUCS. The proposed multi-resolution encoder and decoders are highlighted in the flowchart.

2. DEMUCS

The time domain SE directly inputs a noisy speech waveform y and outputs enhanced waveform $\hat{x}$:

$$\hat{x} = \mathcal{N}(y) \tag{1}$$

The mean absolute error between $\hat{x}$ and the original signal x is the common loss function for training time-domain SE:

$$\mathcal{L}_{mae} = \frac{1}{T}||\hat{x} - x||_F^1, \tag{2}$$

where T is the time points in the waveform.

DEMUCS is one of the time-domain SE shown in Fig.1. It is based on the standard U-Net structure. It contains five encoder layers, two Long Short-term Memory (LSTM) layers, and five decoder layers. During training, in addition to the time-domain loss in Eq. (2), DEMUCS also introduces the following two frequency-domain losses:

$$\begin{aligned} \mathcal{L}_{stft} &= \mathcal{L}_{sc} + \mathcal{L}_{mag} \\ \mathcal{L}_{sc} &= \frac{|||STFT(\hat{x})| - |STFT(x)|||_F^1}{|STFT(x)|} \\ \mathcal{L}_{mag} &= \frac{1}{T}||\log|STFT(\hat{x})| - \log|STFT(x)|||_F^1 \end{aligned} \tag{3}$$

The final multi-resolution loss function of DEMUCS is:

$$\mathcal{L}_{demucs} = \alpha\mathcal{L}_{mae} + (1-\alpha)\sum_{r=1}^{R}(\mathcal{L}_{stft}(r)) \tag{4}$$

where R represents the multi-resolution number. In the conventional standard DEMUCS, $R = 3$ and its STFT points are {32ms, 64ms, 128ms}, the hop size are {3.125ms, 7.5ms, 15ms}, and the window (Hanning window) length are {15ms, 37.5ms, 75ms}.

3. PROPOSED METHOD

Although the STFT loss introduced in DEMUCS shows a significant improvement, it still has two problems: framing lengths {64ms, 128ms} are non-stationary for speech signals, and single output for multiple resolution STFT information may increase the learning burden of the neural network due to the mismatch. Fig. 1 shows a flowchart of the proposed method. We address the above two issues from the following two aspects.

3.1. Fusing Frequency Information in Encoder

Instead of introducing STFT information in the loss calculation, multi-resolution stationary frequency information is incorporated into the encoder layer by layer. Time-domain branch is processed:

$$\hat{h} = GLU(Conv1d((ReLU(Conv1d(h))))) \tag{5}$$

where h represents the time information from the output of the previous encoder layer or the original time-domain input feature. $\hat{h}$ is a hidden representation obtained by convolutional processing in Eq. (5).

Three different window-size spectrograms are adopted to provide stationary multi-resolution frequency information. Spectrograms can be divided into wideband and narrowband spectrograms according to the number of STFT points with certain information complementarity. Taking into account the short-term stability of the speech signal, we choose {8ms, 16ms, and 32ms} with {4ms, 8ms, 16ms} hop size and {8ms, 16ms, 32ms} window length as frequency input features.

Frequency information in each encoder layer is processed as follows:

$$\hat{H_B} = ELU(BatchNorm2d(Conv2d(H_B))), \tag{6}$$

H_B represents the $B-$th frequency information from the output of the previous encoder layer or the original frequency feature, B is among {32ms, 16ms or 8ms}. $\hat{H}_B$ is a hidden representation obtained by convolutional processing in Eq. (6). $\hat{H}_B$ is adopted as the B-th frequency input to the next encoder layer. Furthermore, $\hat{H}_B$ is used as auxiliary information to improve the time-domain SE branch.

In order to extract a more suitable feature representation from frequency information, the $\hat{H}_B$ is processed by two more convolutional processing:

$$\begin{aligned} H_B^{IE} &= ELU(BatchNorm2d(Conv2d(\hat{H}_B))), \\ \hat{H}_B^{IE} &= ELU(BatchNorm2d(Conv2d(H_B^{IE}))), \end{aligned} \tag{7}$$

where $\hat{H}_B^{IE}$ is the extracted information from $\hat{H}_B$.

Finally, the frequency information is incorporated into the time-domain branch as follows:

$$\begin{aligned} f_B &= Linear(Concat(\hat{h}, \hat{H}_B^{IE})) \\ \hat{f} &= Linear(ReLU(Linear(\hat{h} + f_8 + f_{16} + f_{32})) \end{aligned} \tag{8}$$

$\hat{f}$ and $\hat{H}_B$ are the time and frequency domain outputs of the encoder layer, respectively. $\hat{f}$ is also adopted as skip connection information and is input into the corresponding decoder layer. We refer to the model with multi-resolution frequency encoder as **DEMUCS-MRE**. The loss function is the same with Eq. (4).

3.2. Multiple Decoders Consistent with the Learning Targets

In addition to the non-stationary loss issue, reducing the mismatch between the multi-resolution frequency losses and single network output is essential. We propose to use multiple outputs to alleviate the problem that the multiple learning targets are set for the single output.

In the proposed method, each output only calculates one resolution STFT loss. In this paper, the decoder depth is five. So there will be three output layers in parallel after the fourth decoder layer. For the decoder layer, we use the same structure as DEMUCS:

$$\hat{d} = ReLU(ConvTranspose1d(GLU(Conv1d(d)))), \tag{9}$$

where d is the output of the previous decoder layer or the output of the LSTM layers. $\hat{d}$ is the processed output. The last layer of the decoder does not use the ReLU activation function.

Different decoder outputs are expected to perform better on their corresponding resolution frequency-domain information. Averaging multiple waveforms, especially with complementary information, can improve the enhancement performance [12]. Thus, the final enhanced waveform is an average of three different outputs. We refer to the model that takes multiple time-domain outputs as **DEMUCS-MRD**. The loss function is the same with Eq. (4).

4. EXPERIMENTAL SETTINGS AND ANALYSIS

All neural networks were implemented with PyTorch. We used the causal DEMUCS, which can be used for streaming operations. The detailed neural network settings can be found in this URL[1].

We used a public dataset synthesized from the Voice Bank corpus [33]. The dataset can be accessed from this URL[2]. All speech data were sampled at 16 kHz.

4.1. Evaluation Metrics

We used several composite measures for evaluation. They are obtained by linearly combining existing objective measures. In this paper, we used multiple linear regression analysis to form the following composite measures: C_{sig} for a five-point scale of signal distortion (SIG) [34]; C_{bak} for a five-point scale of background intrusiveness (BAK) [34]; C_{ovl} for the overall quality (OVL, [1=bad, 2=poor, 3=fair, 4=good, 5=excellent]) [34]. The three composite measures are obtained from log-likelihood ratio (LLR) [34], the perceptual evaluation of speech quality (PESQ) [35], segmental SNR (segSNR) [34], and weighted-slope spectral (WSS) [36] distance. We also adopted the Short-Time Objective Intelligibility (STOI) [37]. For all metrics, higher values indicate better performance.

Table 1. Comparison of different resolution loss in DEMUCS.

System	STOI (%, ↑)	PESQ (↑)
DEMUCS-8ms,16ms,32ms	94.2	2.79
DEMUCS-32ms	94.6	2.92
DEMUCS-32ms,64ms,128ms (Conventional)	94.8	2.93

4.2. Effect of Different STFT losses in DEMUCS

First, we tested different STFT losses for DEMUCS. We refer to standard "DEMUCS" as "DEMUCS-32ms,64ms,128ms". "DEMUCS-8ms,16ms,32ms" is compared to see the effect of stationary STFT losses. Unexpectedly, Table 1 shows that its performance degrades compared with the standard "DEMUCS-32ms,64ms,128ms". This may be because although 64ms and 128ms are non-stationary signals for speech processing, they may benefit noise components. Therefore, we did not choose to change the resolutions of STFT losses in the time domain enhancement branch. Additionally, we compared models using multi-resolution versus single-resolution "DEMUCS-32ms" in Table 1. Their performances were almost the same, suggesting that improving the performance with a single network output is difficult.

4.3. Effect of Fusing Frequency Information in Encoder

Table 2 shows the results of different SE systems and the proposed method. With the proposed method of "DEMUCS-MRE", there is a 0.1 PESQ improvement when the redundant frequency-domain information is added to the time-domain encoder. CSIG, CBAK, and COVL improvements show that the proposed "DEMUCS-MRE" could maintain more speech signals, suppress more noise, and improve overall quality.

Furthermore, we also tried the same resolutions as the DEMUCS (32ms, 64ms, and 128ms STFT points) in DEMUCS-MRE. The results in Table 3 show that the performance of this system is slightly degraded compared to stationary "DEMUCS-MRE (8ms,16ms,32ms)". This shows that processing stationary signals in the frequency domain is more effective. Nevertheless, fusing the non-stationary frequency information in the encoder can also significantly improve the model performance. It is often pointed out that the frequency-domain SE systems have more stable enhancement performance than time-domain SE systems [45] because the instability of the phase information makes the time-domain waveform less stable than the frequency-domain magnitude of the spectrogram.

[1] https://github.com/hshi-speech/icassp2023/tree/main

[2] https://datashare.ed.ac.uk/handle/10283/1942

Table 2. Results of different SE systems and the proposed method. "Causal" indicates that the model can be used for streaming operations.

System	segSNR (↑)	CSIG (↑)	CBAK (↑)	COVL (↑)	STOI (%, ↑)	PESQ (↑)	Causal
Noisy	1.68	3.35	2.44	2.63	91.5	1.97	✘
SEGAN [38]	7.73	3.48	2.94	2.80	-	2.16	✘
SEGAN-D [39]	8.72	3.46	3.11	3.50	93.3	2.39	✘
Wave U-Net [40]	9.97	3.52	3.24	2.96	-	2.40	✘
MMSE-GAN [41]	-	3.80	3.12	3.14	93.0	2.53	✘
MetricGAN [42]	-	3.99	3.18	3.42	-	2.86	✘
S-DCCRN [43]	-	4.03	2.97	3.43	94.0	2.84	✔
DeepMMSE [44]	-	4.28	3.46	3.64	94.0	2.95	✘
PHASEN [19]	**10.18**	4.21	**3.55**	3.62	-	2.99	✘
DEMUCS [1]	8.74	4.22	3.25	3.52	94.8	2.93	✔
DEMUCS-MRE (proposed)	8.95	4.38	**3.52**	3.73	**95.1**	3.03	✔
DEMUCS-MRD (proposed)	**9.07**	4.33	3.49	3.68	94.6	2.98	✔
DEMUCS-MRE-MRD (proposed)	8.73	**4.40**	**3.52**	**3.77**	**95.1**	**3.07**	✔

Table 3. The enhancement performance of DEMUCS-MRE with non-stationary frequency information (32ms, 64ms, 128ms STFT points).

System	STOI (%, ↑)	PESQ (↑)
DEMUCS-MRE (8ms,16ms,32ms)	95.1	3.03
DEMUCS-MRE (32ms,64ms,128ms)	94.7	3.00

Table 4. Comparison different time-domain outputs in "DEMUCS-MRD".

System	32ms output	64ms output	128ms output
STOI (%, ↑)	94.6	94.6	94.6
PESQ (↑)	2.99	2.98	2.96

Incorporating frequency information into time-domain information can improve the stability of time-domain information.

4.4. Effect of Multiple Decoders

Table 2 shows that the "DEMUCS-MRD" can provide 0.05 PESQ improvement. We averaged multiple outputs to get the final enhanced waveform. Table 4 shows results of different time-domain outputs in "DEMUCS-MRD". The output of all different resolutions shows some improvement, which suggests that adopting multiple outputs can alleviate the problem of learning mismatch with a single output. The results of the different outputs were almost the same, especially for 32ms and 64ms. This may be because the model already has multi-resolution frequency domain information through the shared 1-th to 4-th decoder layers. Furthermore, "DEMUCS-MRD" achieved the best performance for segSNR, which means better performance at the segment level.

4.5. Effect of Improving the Model with Both Encoder and Decoder

The combination of the proposed MRE and MRD can provide further improvement, which is shown in the last row of Table 2. "DEMUCS-MRE-MRD" has stronger speech signal retention and signal overall quality recovery ability. The PESQ improvement from the baseline "DEMUCS" is 0.14.

5. CONCLUSIONS

In this paper, we proposed using multi-resolution encoders and decoders to solve the drawbacks of DEMUCS from both the encoder and decoder perspectives. We first added multi-resolution stationary frequency information to the time-domain enhancement layer by layer to solve the non-stationary STFT loss issue. The experimental results show that the stationary frequency information can significantly improve performance. Moreover, we adopted multiple time-domain outputs to alleviate the problem of learning mismatch with a single output. The results show that it can ensure the model has multi-resolution frequency information while improving the performance of all resolutions. Furthermore, the proposed MRE and MRD can be used jointly to achieve further improvement. In the future, we will fuse the multiple enhanced waveforms into one signal with a neural network instead of averaging.

6. ACKNOWLEDGE

This work was supported by JST, the establishment of university fellowships towards the creation of science technology innovation, Grant Number JPMJFS2123.

7. REFERENCES

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

[2] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An Overview of Noise-Robust Automatic Speech Recognition," *IEEE/ACM TASLP*, vol. 22, no. 4, pp. 745–777, 2014.

[3] C. Chen, Y. Hu, W. Weng, and E. S. Chng, "Metric-oriented speech enhancement using diffusion probabilistic model," *arXiv preprint arXiv:2302.11989*, 2023.

[4] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, “Speech Enhancement with LSTM Recurrent Neural Networks and its Application to Noise-Robust ASR,” in *Latent Variable Analysis and Signal Separation*, 2015, pp. 91–99.

[5] C. Chen, Y. Hu, N. Hou, X. Qi, H. Zou, and E. S. Chng, “Self-critical sequence training for automatic speech recognition,” in *Proc. ICASSP*. IEEE, 2022, pp. 3688–3692.

[6] A. Maas, Q. V. Le, T. M. O’Neil, O. Vinyals, P. Nguyen, and A. Y. Ng, “Recurrent Neural Networks for Noise Reduction in Robust ASR,” in *Proc. INTERSPEECH*, 2012.

[7] A. Narayanan and D. Wang, “Ideal ratio mask estimation using deep neural networks for robust speech recognition,” in *Proc. ICASSP*, 2013, pp. 7092–7096.

[8] M. Mimura, S. Sakai, and T. Kawahara, “Exploring deep neural networks and deep autoencoders in reverberant speech recognition,” in *Proc. HSCMA*, 2014, pp. 197–201.

[9] C. Chen, N. Hou, Y. Hu, S. Shirol, and E. S. Chng, “Noise-robust speech recognition with 10 minutes unparalleled in-domain data,” in *Proc. ICASSP*, 2022, pp. 4298–4302.

[10] C. Chen, Y. Hu, Q. Zhang, H. Zou, B. Zhu, and E. S. Chng, “Leveraging modality-specific representations for audio-visual speech recognition via reinforcement learning,” *arXiv preprint arXiv:2212.05301*, 2022.

[11] H. Shi, L. Wang, S. Li, C. Fan, J. Dang, and T. Kawahara, “Spectrograms Fusion-based End-to-end Robust Automatic Speech Recognition,” in *Proc. APSIPA ASC*, 2021, pp. 438–442.

[12] H. Shi, L. Wang, M. Ge, S. Li, and J. Dang, “Spectrograms Fusion with Minimum Difference Masks Estimation for Monaural Speech Dereverberation,” in *Proc. ICASSP*, 2020, pp. 7544–7548.

[13] Z.-Q. Wang, P. Wang, and D. Wang, “Complex Spectral Mapping for Single- and Multi-Channel Speech Enhancement and Robust ASR,” *IEEE/ACM TASLP*, vol. 28, pp. 1778–1787, 2020.

[14] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “A Regression Approach to Speech Enhancement Based on Deep Neural Networks,” *IEEE/ACM TASLP*, vol. 23, no. 1, pp. 7–19, 2015.

[15] S. Kamath and P. Loizou, “A multi-band spectral subtraction method for enhancing speech corrupted by colored noise,” in *Proc. ICASSP*, vol. 4, 2002, pp. IV–4164–IV–4164.

[16] Y. Ephraim and H. Van Trees, “A signal subspace approach for speech enhancement,” *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 251–266, 1995.

[17] A. Pandey and D. Wang, “Tcnn: Temporal convolutional neural network for real-time speech enhancement in the time domain,” in *Proc. ICASSP*, 2019, pp. 6875–6879.

[18] M. Schroeder, “Models of hearing,” *Proceedings of the IEEE*, vol. 63, no. 9, pp. 1332–1350, 1975.

[19] D. Yin, C. Luo, Z. Xiong, and W. Zeng, “PHASEN: A Phase-and-Harmonics-Aware Speech Enhancement Network,” *Proc. AAAI*, vol. 34, no. 05, pp. 9458–9465, 2020.

[20] K. Tan and D. Wang, “Complex Spectral Mapping with a Convolutional Recurrent Network for Monaural Speech Enhancement,” in *Proc. ICASSP*, 2019, pp. 6865–6869.

[21] 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, pp. 2472–2476.

[22] C. Chen, N. Hou, D. Ma, and E. S. Chng, “Time domain speech enhancement with attentive multi-scale approach,” in *Proc. APSIPA ASC*, 2021, pp. 679–683.

[23] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” in *Proc. MICCAI*, 2015, pp. 234–241.

[24] R. Yamamoto, E. Song, and J.-M. Kim, “Parallel Wavegan: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram,” in *Proc. ICASSP*, 2020, pp. 6199–6203.

[25] A. Pandey and D. Wang, “A new framework for cnn-based speech enhancement in the time domain,” *IEEE/ACM TASLP*, vol. 27, no. 7, pp. 1179–1188, 2019.

[26] S. Pascual, J. Serrà, and A. Bonafonte, “Time-domain speech enhancement using generative adversarial networks,” *Speech Communication*, vol. 114, pp. 10–21, 2019.

[27] J. Smith and P. Gossett, “A flexible sampling-rate conversion method,” in *Proc. ICASSP*, vol. 9, 1984, pp. 112–115.

[28] K. Iwamoto, T. Ochiai, M. Delcroix, R. Ikeshita, H. Sato, S. Araki, and S. Katagiri, “How Bad Are Artifacts?: Analyzing the Impact of Speech Enhancement Errors on ASR,” 2022.

[29] Q. Zhu and A. Alwan, “On the use of variable frame rate analysis in speech recognition,” in *Proc. ICASSP*, vol. 3, 2000, pp. 1783–1786.

[30] D. Griffin and J. Lim, “Signal estimation from modified short-time Fourier transform,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[31] S. Cheung and J. Lim, “Combined multi-resolution (wideband/narrowband) spectrogram,” in *Proc. ICASSP*, 1991, pp. 457–460 vol.1.

[32] A. V. Oppenheim, “Speech spectrograms using the fast fourier transform,” *IEEE Spectrum*, vol. 7, no. 8, pp. 57–62, 1970.

[33] C. Veaux, J. Yamagishi, and S. King, “The voice bank corpus: Design, collection and data analysis of a large regional accent speech database,” in *Proc. O-COCOSDA/CASLRE*, 2013, pp. 1–4.

[34] Y. Hu and P. C. Loizou, “Evaluation of objective quality measures for speech enhancement,” *IEEE TASLP*, vol. 16, no. 1, pp. 229–238, 2008.

[35] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs,” in *Proc. ICASSP*, vol. 2, 2001, pp. 749–752.

[36] D. Klatt, “Prediction of perceived phonetic distance from critical-band spectra: A first step,” in *Proc. ICASSP*, vol. 7, 1982, pp. 1278–1281.

[37] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time–frequency weighted noisy speech,” *IEEE TASLP*, vol. 19, no. 7, pp. 2125–2136, 2011.

[38] S. Pascual, A. Bonafonte, and J. Serrà, “SEGAN: Speech Enhancement Generative Adversarial Network,” in *Proc. Interspeech*, 2017, pp. 3642–3646.

[39] H. Phan, I. V. McLoughlin, L. Pham, O. Y. Chén, P. Koch, M. De Vos, and A. Mertins, “Improving GANs for Speech Enhancement,” *IEEE Signal Processing Letters*, vol. 27, pp. 1700–1704, 2020.

[40] C. Macartney and T. Weyde, “Improved Speech Enhancement with the Wave-U-Net,” *arXiv*, vol. abs/1811.11307, 2018.

[41] M. H. Soni, N. Shah, and H. A. Patil, “Time-frequency masking-based speech enhancement using generative adversarial network,” in *Proc. ICASSP*, 2018, pp. 5039–5043.

[42] 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*, 2019, pp. 2031–2041.

[43] 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 *Proc. ICASSP*, 2022, pp. 7767–7771.

[44] Q. Zhang, A. Nicolson, M. Wang, K. K. Paliwal, and C. Wang, “DeepMMSE: A Deep Learning Approach to MMSE-Based Noise Power Spectral Density Estimation,” *IEEE/ACM TASLP*, vol. 28, pp. 1404–1415, 2020.

[45] Y. Zhao and D. Wang, “Noisy-Reverberant Speech Enhancement Using DenseUNet with Time-Frequency Attention,” in *Proc. Interspeech*, 2020, pp. 3261–3265.