AN INVESTIGATION OF INCORPORATING MAMBA FOR SPEECH ENHANCEMENT

Rong Chao[‡§], *Wen-Huang Cheng*[§], *Moreno La Quatra*[†]
Sabato Marco Siniscalchi[+], *Chao-Han Huck Yang*[*], *Szu-Wei Fu*[*], *Yu Tsao*[‡]

[‡]Academia Sinica, [§]National Taiwan University, [†]Kore University of Enna,
[+]University of Palermo, [*]NVIDIA

ABSTRACT

This work aims to investigate the use of a recently proposed, attention-free, scalable state-space model (SSM), Mamba, for the speech enhancement (SE) task. In particular, we employ Mamba to deploy different regression-based SE models (SEMamba) with different configurations, namely basic, advanced, causal, and non-causal. Furthermore, loss functions either based on signal-level distances or metric-oriented are considered. Experimental evidence shows that SEMamba attains a competitive PESQ of 3.55 on the VoiceBank-DEMAND dataset with the advanced, non-causal configuration. A new state-of-the-art PESQ of 3.69 is also reported when SEMamba is combined with Perceptual Contrast Stretching (PCS). Compared against Transformed-based equivalent SE solutions, a noticeable FLOPs reduction up to $\sim$12% is observed with the advanced non-causal configurations. Finally, SEMamba can be used as a pre-processing step before automatic speech recognition (ASR), showing competitive performance against recent SE solutions.

Index Terms— consistency loss, Mamba, speech enhancement, state-space machine, SEMamba

1. INTRODUCTION

A speech enhancement (SE) process involves retrieving clean speech components from a distorted signal to generate an improved version with enhanced acoustic properties [1, 2]. In various speech applications, such as assistive hearing technologies [3, 4], speaker recognition [5, 6], and automatic speech recognition (ASR) [7, 8], SE serves as a critical front-end processor. Recently, SE has been framed as a regression task utilizing neural-network-based mapping functions to convert input noisy speech signals into cleaner outputs. Specifically, deep neural models are employed to implement the mapping function that enhances noisy speech. Several deep architectures, including deep denoising autoencoders [9], deep fully connected neural networks [10, 11], convolutional neural networks [12], long short-term memory [13], convolutional recurrent neural networks [14, 15], and more recently, Transformer and its variant Conformer, have been used to form the mapping function, resulting in notably promising SE performance [16, 17, 18, 19].

In addition to exploring advanced model architectures, several studies have aimed to develop effective objective functions to boost SE capability. Typically, a signal-level distance measure serves as the fundamental objective function during training, e.g., L1/L2 norm [20], SI-SDR [21], or multiple-resolution loss [22]. Objective functions specifically crafted and optimized for the downstream SE tasks have also been proposed. For instance, objective scores, such as perceptual evaluation of speech quality (PESQ) [23], or short-time objective intelligibility (STOI) [24], are learned using neural network models in [25, 26] and then employed to optimize the SE model. Similarly, in [27], subjective assessment results are learned through a quality estimator and utilized for training the SE model. Furthermore, a few studies have shown improved SE results by preprocessing the target speech prior to model training. For example, perceptual contrast stretching (PCS) [28] enhances the contrast of target features based on their perceptual importance, effectively boosting SE performance without a notable increase in computational cost during runtime.

Recently, a highly promising neural architecture, referred to as Mamba [29], has emerged, harnessing the state-space model with a novel selection mechanism. Mamba exhibits comparable or superior performance to state-of-the-art (SOTA) Transformer-based models across diverse tasks [30, 31] due to its intrinsic capability in modeling extremely long-range dependencies. Mamba stands out for its efficient use of computational resources, scaling linearly in sequence length compared to the quadratic complexity of Transformers, making it much more resource-efficient for long sequences.

In this work, we are interested in comparing and contrasting Transformer- and Mamba-based SE solutions, and we thereby introduce a new Mamba-based SE solution, named SEMamba. We evaluate different SEMamba configuration: first, we deploy a basic SE configuration, termed SEMamba-basic. Next, we implement SEMamba with an advanced configuration, termed SEMamba-advanced. Both causal, and non-causal solutions are tested along with conventional distance-based and metric-oriented losses. We assess our

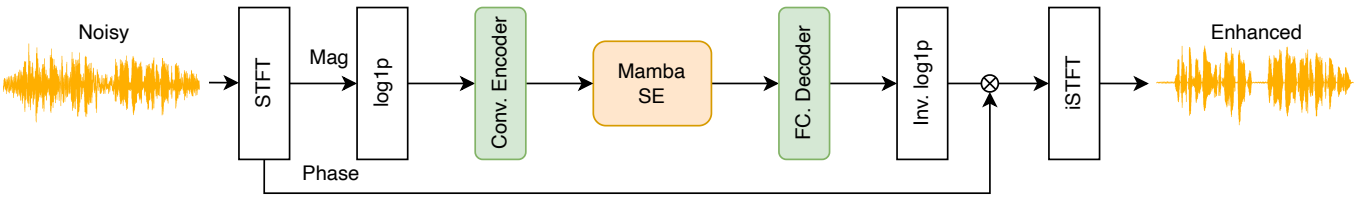

Fig. 1. Architecture of our basic Mamba-based Speech Enhancement (SE) model, SEMamba-basic.

SEMamba on the VoiceBank-DEMAND dataset [32], and contrast it with transformer-based counterparts. Experimental results demonstrate that (i) SEMamba attains a competitive PESQ of 3.55 on the VoiceBank-DEMAND dataset with the advanced, non-causal configuration, and (ii) a new state-of-the-art PESQ of 3.69 is delivered when combined with PCS. Compared with Transformed-based basic architecture, a reduction in the number of parameters up to $\sim$60% and $\sim$28% is also observed along with a floating-point operations per second (FLOPs) reduction up to $\sim$66% and $\sim$53% with causal and non-causal configurations, respectively. In the advanced configuration, a notable FLOPs reduction up to $\sim$12% is observed when replacing Transformers with Mamba, while keeping the similar amount of parameters. Finally, SEMamba can be used as a pre-processing step before ASR, showing competitive performance against recent SE solutions.

2. MAMBA: LINEAR-TIME SEQUENCE MODELING WITH SELECTIVE STATE SPACES

The structured state space model (SSM) [33] has demonstrated the ability to manage long-dependent sequences with low computation and memory requirements. It can serve as a replacement for either a CNN for efficient parallel training or an RNN for rapid autoregressive generation [34]. More recently, Mamba [29] has introduced notable advancements in discrete data modeling of SSM by introducing two key improvements. First, Mamba incorporates an input-dependent selection mechanism, enabling efficient information filtering by parameterizing the SSM module based on the input information. Second, Mamba introduces a hardware-aware algorithm that scales linearly with input sequence length, facilitating faster computation of the model recurrently with a scan. The Mamba architecture, which integrates SSM blocks with linear layers, is notably simpler than its Transformer-based counterparts and has demonstrated SOTA performance across various long-sequence patterns, including language and genomics. Therefore, noticeable computational efficiency during both training and inference phases has been attained.

Structured SSMs, as described in the Mamba method [29], operate by mapping an input $\boldsymbol{x}$ to an output $\boldsymbol{y}$ through a higher dimensional latent state $\boldsymbol{h}$, as follows: $h_n = \bar{\boldsymbol{A}} h_{n-1} + \bar{\boldsymbol{B}} x_n$, and $y_n = \boldsymbol{C} h_n$. $\bar{\boldsymbol{A}}$ and $\bar{\boldsymbol{B}}$ represent discretized state matrices. The discretization process converts continuous parameters $(\boldsymbol{\Delta}, \boldsymbol{A}, \boldsymbol{B})$ into discrete counterparts $(\bar{\boldsymbol{A}}, \bar{\boldsymbol{B}})$. Mamba integrates components from the H3 architecture [35] and a gated multilayer perceptron block into a stacked structure, thereby expanding the model's internal representation dimension and concentrating most parameters in linear projections. This design results in fewer parameters for the inner SSM and adopts the SiLU/Swish activation function along with standard normalization and residual connections. For further details, the interested reader is referred to [29].

3. MAMBA IN SPEECH ENHANCEMENT

Two SEMamba architectures are deployed: The first architecture integrates Mamba in a basic SE architecture (see Fig. 1). The second architecture leverages MP-SENet but uses the Mamba block in place of attention-based ones (see in Fig. 2).

3.1. SEMamba-basic

SEMamba-basic is causal, i.e., the output at time t, y_t, is influenced only by preceding inputs x_n with $n \in \{0, \ldots, t\}$. As depicted in Fig. 1, the input noisy waveform is first transformed into its spectral representation using the Short-Time Fourier Transform (STFT). Next, the magnitude (Mag) component of the STFT is compressed using the $log1p$ function $(log1p(z) = log(1+z))$ [36] to regulate the dynamic range of the magnitude. The compressed spectral magnitude is then fed into the SE module, which comprises (i) a convolutional encoder having 4 convolutional layers, (ii) two uni-directional Mamba blocks, and (ii) a fully connected decoder layer. The enhanced magnitude is decompressed by an inverse of $log1p$ before being combined with the noisy phase in the iSTFT block to reconstruct the enhanced waveform. We also implement the basic SE architecture with Transformers, following the architecture used in [36]. On the other hand, SEMamba-basic employs the Mamba blocks to replace the Transformer blocks. Both basic models are trained using the mean absolute error of the magnitude as the loss function.

3.2. SEMamba-advanced

In the architecture shown in Fig. 2, the causality constraint is removed, and a more sophisticated SE structure is employed.

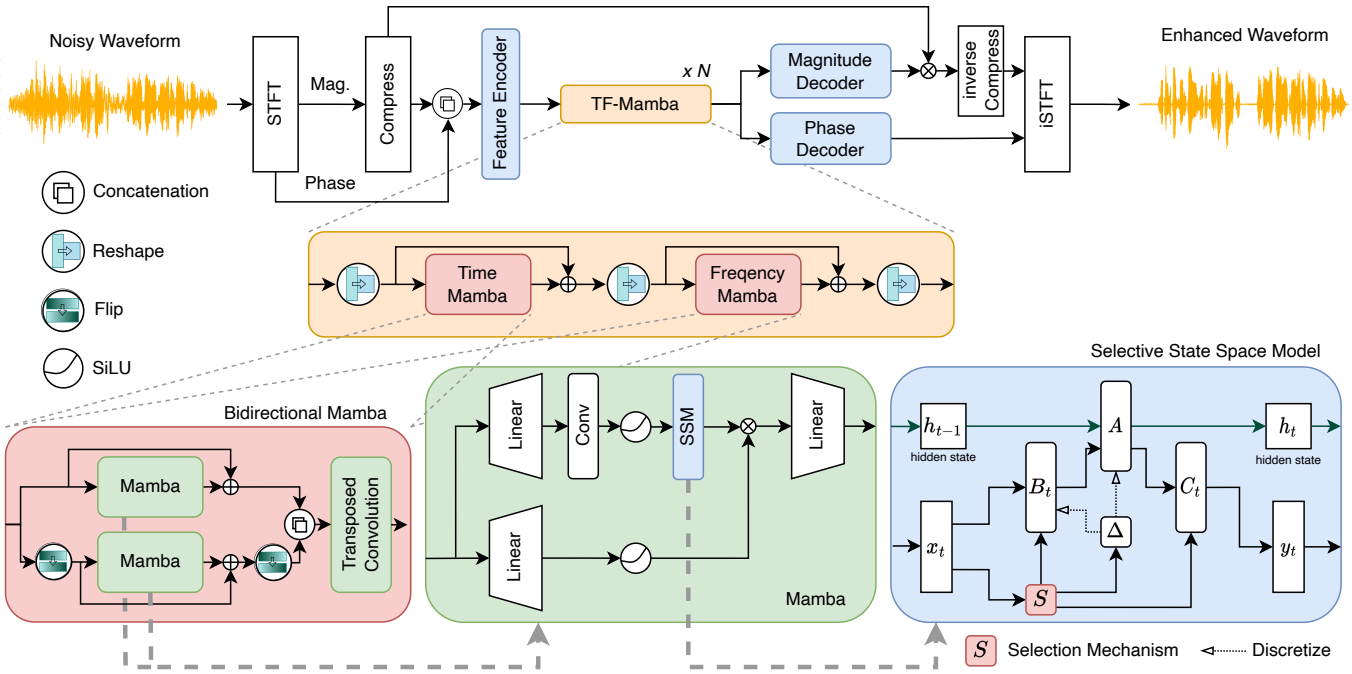

Fig. 2. Architecture of the proposed SEMamba-advanced with Time-Frequency (TF) and Selective-SSM mechanism.

Specifically, MP-SENet [18], which leverages both magnitude and phase spectra, serves as the architectural backbone. A spectral representation is obtained from the input noisy waveform using the STFT. The magnitude component is first compressed and then stacked together with the phase components. These components are processed by a feature encoder, featuring a dilated DenseNet core flanked by two convolutional layers. The output of the feature encoder undergoes multiple transformations via the Time-Frequency Mamba block, which is executed N times (N=4 in this study) to enhance its spectral properties. The output of the Time-Frequency Mamba block is then routed to two separate decoders for magnitude and phase processing, respectively. Each decoder comprises a dilated DenseNet followed by a deconvolutional block, and it ends with an output layer that employs a 2D-convolution.

The loss function used to train SEMamba-advanced leverages a linear combination of losses, including PESQ-based GAN discriminator, time, magnitude, complex, and phase losses, as in MP-SENet.

3.3. SEMamba-advanced & additional designs

3.3.1. From uni- to bi-directional Mamba

We explore the possible benefit of modifying the Mamba structure from a uni-directional to a bi-directional configuration. This modification involves adapting the module to process input sequences both in their original and reversed forms. The inputs are processed in parallel using Mamba modules, after which the outputs are concatenated. The output is then fed to a $Conv1D$ layer:

$$\mathbf{y} = Conv1D(M_{uni}(\mathbf{x}) \oplus flip(M_{uni}(flip(\mathbf{x})))), \quad (1)$$

where $\mathbf{x}$, $\mathbf{y}$, $M_{uni}()$, $flip()$, $Conv1D()$, and $\oplus$, respectively, denote input, output, uni-directional Mamba, flipping operation, 1-D convolution, and concatenation.

3.3.2. Consistency loss (CL)

Stability during the training process can be improved by leveraging the consistency loss proposed in [37]. The consistency loss aims to minimize the gap between the complex spectrum obtained directly from the model output (comprising amplitude and phase) and the complex spectrum derived after applying inverse Short-Time Fourier Transform (iSTFT) and then re-applying STFT to the resulting waveform. This gap arises because the complex spectrum predicted directly by the model may fall outside the STFT domain. Once processed through iSTFT and then subjected to STFT, the spectrum re-enters the STFT domain, thereby closing this gap. This mechanism ensures that our enhancements in the complex domain translate effectively in the time-frequency domain.

Table 1. Comparison of basic SE architectures leveraging either Mamba or Transformer blocks.

	Causal	PESQ	STOI	FLOPs	Param.
noisy-speech	–	1.97	0.92	–	–
Transf. [36]	Yes	2.76	0.94	2.26G	9.05M
Mamba	Yes	2.76	0.94	**0.76**G	**3.60M**
Transf. [36]	No	2.84	0.94	2.26G	9.05M
Mamba	No	**2.85**	0.94	**1.06G**	**6.49M**

Table 2. Comparison of Mamba and Transformer within an advanced SE architecture (excluding CL and PCS).

	PESQ	STOI	FLOPs	Parameters
noisy-speech	1.97	0.92	–	–
Conformer	3.50	0.96	74.29G	2.05M
Mamba (Uni)	3.29	0.95	**53.09G**	**1.41M**
Mamba (Bi)	**3.52**	0.96	65.46G	2.26M

3.3.3. Perceptual contrast stretching (PCS)

PCS is a spectral processing technique that aims to improve the perceptual quality of speech signals. It leverages empirical observations of varying sensitivity levels in the human auditory system. PCS exploits this phenomenon by stretching the magnitude spectrum of the signal based on each frequency band's perceived importance. In this paper, we introduce PCS as an auxiliary step following the enhancement process, aiming to refine the perceptual quality of the speech signal [28].

3.4. Dataset

We utilized the VoiceBank-DEMAND dataset [32] for our study. This dataset comprises noisy speech recordings generated by mixing clean speech from the VoiceBank collection [38] with noise from the DEMAND dataset [39] and is widely used as one of the SE benchmark dataset. It encompasses 30 distinct speakers, with 28 speakers allocated for training and 2 for testing. Clean samples were mixed with noise samples at four signal-to-noise ratios (SNRs) during training ([0, 5, 10, 15] dB) and testing ([2.5, 7.5, 12.5, 17.5] dB). The training dataset comprised a total of 11,572 utterances, while the testing set included 824 utterances.

3.5. Experimental Setup

Following [40], all recordings were downsampled from 48 kHz to 16 kHz. The evaluation metrics used to assess the SE performance include: (i) Wide-band PESQ [23], (ii) Prediction of the signal distortion (CSIG), (iii) Prediction of the background intrusiveness (CBAK), (iv) Prediction of the overall speech quality (COVL), and (v) STOI [24]. For all metrics, higher values indicate better enhancement outcomes.

Table 3. The results of SEMamba and several well-known SE solutions on the VoiceBank-DEMAND dataset.

Model	PESQ	CSIG	CBAK	COVL	STOI
noisy-speech	1.97	0.92	3.34	2.63	–
SEGAN [40]	2.16	3.48	2.94	2.80	–
MetricGAN+ [26]	3.15	0.93	4.14	3.64	–
DPT [41]	3.33	0.96	4.58	4.00	–
CMGAN [17]	3.41	4.63	3.94	4.12	**0.96**
MP-SENet [18]	3.50	4.73	3.95	4.22	**0.96**
S4DSE [42]	2.55	3.94	3.00	3.23	0.93
S4ND-UNet [43]	3.15	4.52	3.62	3.85	–
Spiking-S4 [44]	3.39	**4.92**	2.64	4.31	–
SEMamba (-CL)	3.52	4.75	**3.98**	4.26	**0.96**
SEMamba	3.55	4.77	3.95	4.29	**0.96**
SEMamba (+PCS)	**3.69**	4.79	3.63	**4.37**	**0.96**

In our experimental setup, we systematically evaluate various aspects to comprehensively assess the performance of the proposed SEMamba system[1]. First, we assess SEMamba-basic. Next, we assess Mamba with the more advanced model architecture. Finally, we examine the effectiveness of SEMamba-advanced with CL and PCS and evaluate its impact as pre-processing step in an ASR scenario as well.

4. EXPERIMENTS

4.1. Experimental Results

Table 1 reports the experimental results with the basic SE architecture leveraging either the Mamba or Transformer block, as presented in Section II-A. Additionally, a non-causal configuration is also implemented as a support comparison. The results show that Mamba delivers comparable or superior performance to the Transformer in both causal and non-causal configurations while utilizing fewer FLOPs and parameters. Notably, the Mamba block reduces computational demands, needing 66.37% fewer FLOPs and 60.22% fewer parameters than the Transformer in causal configurations, and 53.09% fewer FLOPs with 28.28% fewer parameters in non-causal setups.

In the second set of experiments, we compare Mamba- and Transformer-based advanced SE architectures. The results are listed in Table 2. The original MP-SENet uses a Conformer as the core component. In our experiments, Conformer and Transformer yield similar results. All three SE systems presented in Table 2 are non-causal due to dilated DenseNet in feature encoder. Mamba (Bi) utilizes a bi-directional architecture, whereas Mamba (Uni) employs a uni-directional architecture. Comparing the results in Tables 1 and 2, we observe that integrating Mamba into an

[1]The detailed results and source codes of the model will be available for public access at `https://github.com/RoyChao19477/SEMamba`.

advanced model architecture notably improves performance. Additionally, Table 2 reveals that Mamba not only achieves superior performance but also requires 11.88% fewer FLOPs compared to Conformer.

Third, we contrast the proposed SEMamba with several well-known SE methods on the VoiceBank-DEMAND dataset. The results are listed in Table 3. For comparison, we include three SE methods based on the state-space model: S4DSE, S4ND-UNet, and Spiking-S4 [42, 43, 44]. The table also reports the results obtained with and without consistency loss (-CL), as well as with (+PCS) and without PCS, facilitating a detailed performance analysis. From Table 3, we first note that SEMamba (-CL) yields performance comparable to MP-SENet. The major distinction between SEMamba (-CL) and MP-SENet is that SEMamba (-CL) utilizes Mamba as the core model, whereas MP-SENet employs Conformer. The results again confirm that Mamba yields similar evaluation scores. Notably, SEMamba reaches a high PESQ score of 3.55 on the VoiceBank-DEMAND dataset when consistency loss (CL) is included. Finally, by applying PCS, SEMamba's attains a PESQ score of 3.69, achieving SOTA results on this dataset.

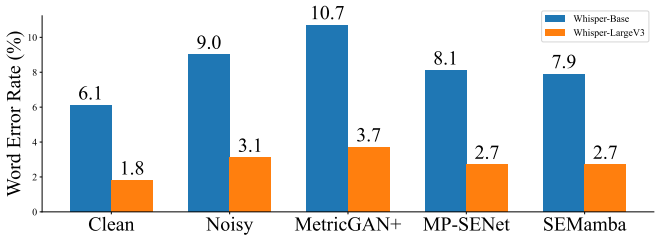

Fig. 3. Comparative analysis of WERs for SEMamba and related models on the VoiceBank-DEMAND dataset with Whisper ASR [45].

Finally, we have also tested ASR results using Whisper-Base and Whisper-LargeV3 on the test set of VoiceBank-DEMAND when the data are pre-processed by SEMamba-advanced[2]. For comparison, we tested related methods, MetricGAN+ and MP-SENet, as shown in Table 3 and both trained on VoiceBank-DEMAND. The results, presented in Fig. 3, show that SEMamba outperforms other methods in almost all conditions, achieving notable relative word error rate (WER) reductions of 12.22% (from 9.0% to 7.9%) on Whisper-Base and 12.90% (from 3.1% to 2.7%) on Whisper-LargeV3.

[2]For the ASR performance evaluation, the SEMamba configuration we referred to is SEMamba-Advanced (PESQ=3.55). We conducted the ASR performance tests directly on this result without retraining. The testing script is at: `https://github.com/RoyChao19477/SEMamba`.

5. CONCLUSION

This study explored Mamba, a novel SSM incorporating a selective mechanism, for tackling the SE task. Comprehensive investigations involving basic and advanced SE neural schemes and employing traditional signal-level and metric-oriented objective functions were carried out. Experimental results showed that Mamba holds noticeable promise in advancing SE performance. In fact, on the VoiceBank-DEMAND dataset, SEMamba with PCS achieved a SOTA PESQ score of 3.69. Furthermore, SEMamba used as a pre-processing step before ASR with advanced technologies allows, overall, to achieve lower WERs. In future research, we aim to explore Mamba's potential in other speech generation tasks.

6. REFERENCES

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, Inc., USA, 2nd edition, 2013.

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

[3] D. Wang, "Deep learning reinvents the hearing aid," *IEEE Spectrum*, vol. 54, no. 3, pp. 32–37, 2017.

[4] P. C. Loizou, "Speech processing in vocoder-centric cochlear implants," in *Cochlear and Brainstem Implants*, vol. 64, pp. 109–143. Karger Publishers, 2006.

[5] H. Taherian, Z.-Q. Wang, J. Chang, and D. Wang, "Robust speaker recognition based on single-channel and multi-channel speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1293–1302, 2020.

[6] D. Michelsanti and Z.-H. Tan, "Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification," in *Proc. INTERSPEECH*, 2017.

[7] T. Ochiai, K. Iwamoto, M. Delcroix, R. Ikeshita, H. Sato, S. Araki, and S. Katagiri, "Rethinking processing distortions: Disentangling the impact of speech enhancement errors on speech recognition performance," *arXiv preprint arXiv:2404.14860*, 2024.

[8] C.-C. Lee, Y. Tsao, H.-M. Wang, and C.-S. Chen, "D4AM: A general denoising framework for downstream acoustic models," in *Proc. ICLR*, 2023.

[9] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. INTERSPEECH*, 2013.

[10] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.

[11] D. Liu, P. Smaragdis, and M. Kim, "Experiments on deep learning for speech denoising," in *Proc. INTERSPEECH*, 2014.

[12] S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 26, no. 9, pp. 1570–1584, 2018.

[13] 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*, 2015.

[14] A. Li, M. Yuan, C. Zheng, and X. Li, "Speech enhancement using progressive learning-based convolutional recurrent neural network," *Applied Acoustics*, vol. 166, pp. 107347, 2020.

[15] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement.," in *Proc. INTERSPEECH*, 2018.

[16] Y. Koizumi, K. Yatabe, M. Delcroix, Y. Masuyama, and D. Takeuchi, "Speech enhancement using self-adaptation and multi-head self-attention," in *Proc. ICASSP*, 2020.

[17] S. Abdulatif, R. Cao, and B. Yang, "CMGAN: Conformer-based metric-gan for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2477–2493, 2024.

[18] Y.-X. Lu, Y. Ai, and Z.-H. Ling, "MP-SENet: A Speech Enhancement Model with Parallel Denoising of Magnitude and Phase Spectra," in *Proc. INTERSPEECH*, 2023.

[19] Y.-X. Lu, Y. Ai, and Z.-H. Ling, "Explicit estimation of magnitude and phase spectra in parallel for high-quality speech enhancement," *arXiv preprint arXiv:2308.08926*, 2023.

[20] J. Qi, J. Du, S. M. Siniscalchi, X. Ma, and C.-H. Lee, "On mean absolute error for deep neural network based vector-to-vector regression," *IEEE Signal Processing Letters*, vol. 27, pp. 1485–1489, 2020.

[21] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR-half-baked or well done?," in *Proc. ICASSP*, 2019.

[22] A. Défossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," in *Proc. INTERSPEECH*, 2020.

[23] 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*, 2001.

[24] 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.

[25] 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.

[26] S.-W. Fu, C. Yu, T.-A. Hsieh, P. Plantinga, M. Ravanelli, X. Lu, and Y. Tsao, "MetricGAN+: An improved version of metricgan for speech enhancement," in *Proc. INTERSPEECH*, 2020.

[27] K. M. Nayem and D. Williamson, "Attention-based speech enhancement using human quality perception modelling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 250–260, 2023.

[28] R. Chao, C. Yu, S.-W. Fu, X. Lu, and Y. Tsao, "Perceptual contrast stretching on target feature for speech enhancement," in *Proc. INTERSPEECH*, 2022.

[29] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[30] X. Jiang, C. Han, and N. Mesgarani, "Dual-path mamba: Short and long-term bidirectional selective structured state space models for speech separation," *arXiv preprint arXiv:2403.18257*, 2024.

[31] K. Li and G. Chen, "Spmamba: State-space model is all you need in speech separation," *arXiv preprint arXiv:2404.02063*, 2024.

[32] 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*, 2016.

[33] A Gu, K Goel, and C Re, "Efficiently modeling long sequences with structured state spaces," in *Proc. ICLR*, 2021.

[34] A. Gu, K. Goel, A. Gupta, and C. Ré, "On the parameterization and initialization of diagonal state space models," in *Proc. NeurIPS*, 2022.

[35] D. Y. Fu, T. Dao, K. K. Saab, A. W. Thomas, A. Rudra, and C. Ré, "Hungry hungry hippos: Towards language modeling with state space models," *arXiv preprint arXiv:2212.14052*, 2022.

[36] S.-W. Fu, C.-F. Liao, T.-A. Hsieh, et al., "Boosting objective scores of a speech enhancement model by metricgan post-processing," in *Proc. APSIPA*, 2020.

[37] V. Zadorozhnyy, Q. Ye, and K. Koishida, "SCP-GAN: Self-correcting discriminator optimization for training consistency preserving metric gan on speech enhancement tasks," *arXiv:2210.14474*, 2022.

[38] 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.

[39] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings," in *Proc. Meetings on Acoustics*, 2013.

[40] S. Pascual, A. Bonafonte, and J. Serrà, "Speech enhancement generative adversarial network," in *Proc. INTERSPEECH*, 2017.

[41] F. Dang, H. Chen, and P. Zhang, "DPT-FSNet: Dual-path transformer based full-band and sub-band fusion network for speech enhancement," in *Proc. ICASSP*, 2022.

[42] L. Sun, S. Yuan, A. Gong, L. Ye, and E. S. Chng, "Dual-branch modeling based on state-space model for speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1457–1467, 2024.

[43] P.-J. Ku, C.-H. Huck Yang, S. M. Siniscalchi, and C.-H. Lee, "A multi-dimensional deep structured state space approach to speech enhancement using small-footprint models," in *Proc. INTERSPEECH*, 2023.

[44] Y. Du, X. Liu, and Y. Chua, "Spiking structured state space model for monaural speech enhancement," in *Proc. ICASSP*, 2024.

[45] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28492–28518.