U-Mamba-Net: A highly efficient Mamba-based U-net style network for noisy and reverberant speech separation

Shaoxiang Dang*, Tetsuya Matsumoto*, Yoshinori Takeuchi†, and Hiroaki Kudo*
* Graduate School of Informatics, Nagoya University, Nagoya, Japan
† School of Informatics, Daido University, Nagoya, Japan
E-mail: dang.shaoxiang.s0@s.mail.nagoya-u.ac.jp

***Abstract*—The topic of speech separation involves separating mixed speech with multiple overlapping speakers into several streams, with each stream containing speech from only one speaker. Many highly effective models have emerged and proliferated rapidly over time. However, the size and computational load of these models have also increased accordingly. This is a disaster for the community, as researchers need more time and computational resources to reproduce and compare existing models. In this paper, we propose U-mamba-net: a lightweight Mamba-based U-style model for speech separation in complex environments. Mamba is a state space sequence model that incorporates feature selection capabilities. U-style network is a fully convolutional neural network whose symmetric contracting and expansive paths are able to learn multi-resolution features. In our work, Mamba serves as a feature filter, alternating with U-Net. We test the proposed model on Libri2mix. The results show that U-Mamba-Net achieves improved performance with quite low computational cost.**

I. Introduction

Speech separation in complex environments is one of the most challenging branches of speech separation. Noise, reverberation, and other factors severely interfere with the desired signal, making it even more difficult to capture long-term dependencies, which has long been valued by the speech separation community [1].

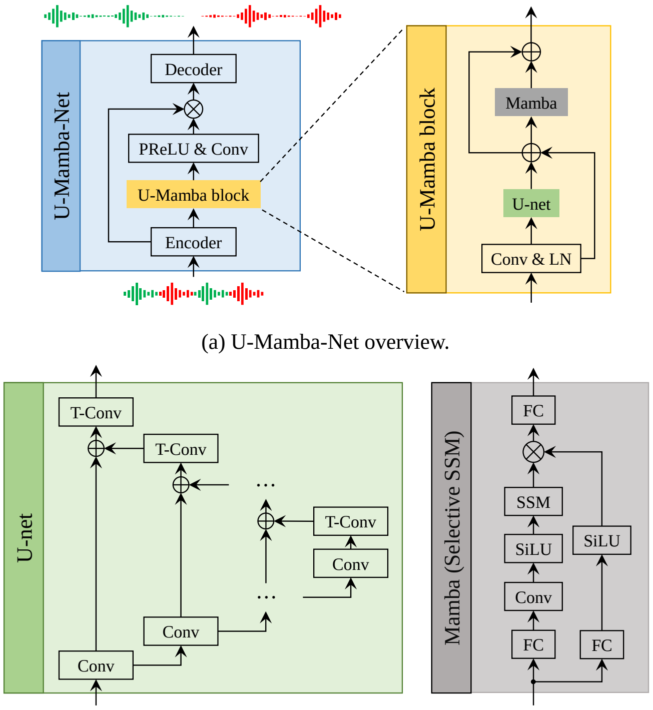
In the endeavor to capture long-term dependencies, recurrent neural networks (RNNs) are the first popular structure that demonstrates effectiveness on both time-frequency [2] and time-domain processing [3]. In time-domain separation, RNNs are typically used in conjunction with dual-path (DP) structure [3]. The DP structure assists the network in modeling global information, complementing the processing of local information. Later on, self-attention mechanism like Transformer [4], as an alternative to RNNs, that allows parallel computing is introduced into the DP structure [5]. As for drawbacks, the DP structure adds extra computational overhead, and the quadratic scaling of the self-attention mechanism in transformers also make training time-consuming and resource-intensive, which has been a growing consensus and concern in this community.

The cascaded multi-task learning (CMTL) methods advocates solving complex speech separation task structurally [1]. Concretely, it decomposes a complex task into simpler, sequential sub-tasks. Through progressive supervision, cascaded multi-task structure can achieve further improvements in performance. However, due to the stacking of sub-modules, CMTL cannot effectively control the size of the model. Worse still, there may be gradient conflicts between modules, which significantly limits the model's capabilities [6].

In this work, we shed light on U-net, which though was developed for biomedical image segmentation [7]. The U-net architecture consists entirely of convolutional neural layers, making it compact in size and low in computational complexity. Demucs [8] and SuDoRMRF [9] are successful attempts at using the U-Net architecture for music sound separation and audio source separation, respectively. To address its inability to learn global relationships, we add a Mamba module after each U-Net block. Mamba is a selective structured state space sequence model [10]. Concretely, Mamba features high-order polynomial projection operators (HiPPO) initialization and input-dependent structure while maintain linear computational complexity. HiPPO initialization assists in decomposing signals into basis functions [11]. We conduct experiments in the simulated noisy and reveberant version of open-source public dataset Libri2mix [12]. According to performance comparison of proposed model and previous models, we find that U-Mamba-Net is not only able to performance better cross various metrics, but also achieve so with low computational resource.

II. Proposed U-Mamba-Net

U-Mamba-Net consists of an encoder, U-Mamba blocks, and a decoder, similar to the prevailing structures used previously [13]. The encoder is a 1-dimensional convolutional layer that maps waveform to time-frequency like representation. U-Mamba blocks aim to learn features with high representational capability. A convolutional layer following U-Mamba blocks assists estimating masks for all sources. After applying the estimated masks to the representations of the mixed speech, the features of all sources are decoded by the 1-dimensional transposed convolutional layer, resulting in the final separated speeches, The overview is illustrated in Fig. 1.



(a) U-Mamba-Net overview.

(b) Illustration of U-net (left) and Mamba (right).

Fig. 1. Overview of U-Mamba-Net.

A. *U-Mamba blocks*

Given the feature $\boldsymbol{X} \in \mathbb{R}^{F\times T}$ of a noisy and reverberant mixed speech, the goal of the B stacked U-mamba blocks is to estimate robust representation $\boldsymbol{M} \in \mathbb{R}^{F\times T}$. Then, a convolutional layer helps to generate masks equal to the number of sources S. As the core part of our proposed model, one U-Mamba block mainly owns a U-net module and a Mamba module. The U-net module is composed of L successive downsampling and upsampling layers, with each pair of sub-layers at corresponding depths connected by residual connections. Subsequently, the output of the U-net is fed into the Mamba module. The U-Mamba block outputs the estimated representations with another residual connection to the output of the U-net in this module. The detailed statement is presented in Algorithm 1.

B. *Mamba*

Mamba is an extension of state space models (SSMs) [14]. SSMs formulates a mapping from $\boldsymbol{x}(t) \in \mathbb{R}^F$ to $\boldsymbol{y}(t) \in \mathbb{R}^F$ via hidden state space $\boldsymbol{h}(t) \in \mathbb{R}^N$:

$$\boldsymbol{h}^{'}(t) = \boldsymbol{A}\boldsymbol{h}(t) + \boldsymbol{B}\boldsymbol{x}(t) \tag{1}$$

$$\boldsymbol{y}(t) = \boldsymbol{C}\boldsymbol{h}(t) + \boldsymbol{D}\boldsymbol{x}(t) \tag{2}$$

where $\boldsymbol{A} \in \mathbb{R}^{N\times N}$, $\boldsymbol{B} \in \mathbb{R}^{N\times F}$, $\boldsymbol{C} \in \mathbb{R}^{F\times N}$, and $\boldsymbol{D} \in \mathbb{R}^{F\times F}$ are state matrices. N is number of state space dimension.

Algorithm 1 An algorithm for b-th U-Mamba block

```
Input: $\boldsymbol{M}_b$, $L$
Output: $\boldsymbol{M}_{b+1}$
  $\boldsymbol{D}^{(0)} \leftarrow$ PReLU(LN(BottleNeckConv($\boldsymbol{M}_b$)))
                                        ▷ Bottle Neck Convolution
  $l \leftarrow 1$
  while $l \leq L$ do                                   ▷ Downsampling
      $\boldsymbol{D}^{(l)} \leftarrow$ LN(Conv1D($\boldsymbol{D}^{(l-1)}$))
      $l \leftarrow l + 1$
  end while
  $l \leftarrow L$
  $\boldsymbol{U}^{(L)} \leftarrow \boldsymbol{D}^{(L)}$
  while $l \geq 1$ do                                   ▷ Upsampling
      $\boldsymbol{U}^{(l-1)} \leftarrow$ T-Conv1D($\boldsymbol{U}^{(l)}$) + $\boldsymbol{D}^{(l-1)}$
                                        ▷ Residual connection
      $l \leftarrow l - 1$
  end while
  $\boldsymbol{M}_b \leftarrow$ PReLU($\boldsymbol{U}^{(0)}$)
  $\boldsymbol{M}_{b+1} \leftarrow$ Mamba($\boldsymbol{M}_b$) + $\boldsymbol{M}_b$
                    ▷ Mamba module and residual connection
  return $\boldsymbol{M}_{b+1}$
```

Based on the continuous form above, its discrete form can be formulated by:

$$\boldsymbol{h}_t = \overline{\boldsymbol{A}}\boldsymbol{h}_{t-1} + \overline{\boldsymbol{B}}\boldsymbol{x}_t \tag{3}$$

$$\boldsymbol{y}_t = \overline{\boldsymbol{C}}\boldsymbol{h}_t + \overline{\boldsymbol{D}}\boldsymbol{x}_t \tag{4}$$

where $\overline{\boldsymbol{A}}$ and $\overline{\boldsymbol{B}}$ are discretized parameters, which are converted from $\boldsymbol{A}$, $\boldsymbol{B}$, and a step parameter Δ using bilinear method:

$$\overline{\boldsymbol{A}} = (\boldsymbol{I} - \frac{\Delta}{2}\boldsymbol{A})^{-1}(\boldsymbol{I} + \frac{\Delta}{2}\boldsymbol{A}) \tag{5}$$

$$\overline{\boldsymbol{B}} = (\boldsymbol{I} - \frac{\Delta}{2}\boldsymbol{A})^{-1}\boldsymbol{\Delta}\boldsymbol{B} \tag{6}$$

$$\overline{\boldsymbol{C}} = \boldsymbol{C} \tag{7}$$

$$\overline{\boldsymbol{D}} = \boldsymbol{D} \tag{8}$$

System in equations 3 and 4 can be rewritten as a convolutional form of structured kernel $\overline{\boldsymbol{K}}$ and input $\boldsymbol{x}$.

$$\boldsymbol{y} = \overline{\boldsymbol{K}} * \boldsymbol{x}, \overline{\boldsymbol{K}} = (\overline{\boldsymbol{CB}}, \overline{\boldsymbol{CAB}}, \cdots, \overline{\boldsymbol{CA}}^{\boldsymbol{T-1}}\overline{\boldsymbol{B}}) \tag{9}$$

Structured SSMs (S4) feature a HiPPO initialization instead of random initialization, which is confirmed to easier decompose representation into orthogonal polynomials.

Mamba improves S4 with a selective mechanism [10]. Concretely, it adds input-dependency into SSM matrices via parallel scan algorithm, Furthermore, state variables are kept in SRAM, which holds a fast and efficient GPU memory hierarchy. Mamba is illustrated in the right part of Fig. 1b.

TABLE I
REVERBERATION CONFIGURATION.

Room	L (m)	$\mathcal{U}$(5,10)
	W (m)	$\mathcal{U}$(5,10)
	H (m)	$\mathcal{U}$(3,4)
T60	T (s)	$\mathcal{U}$(0.2,0.6)
Receiver	L (m)	$\frac{L_{room}}{2} + \mathcal{U}$(-0.2,0.2)
	W (m)	$\frac{W_{room}}{2} + \mathcal{U}$(-0.2,0.2)
	H (m)	$\mathcal{U}$(0.9,1.8)
Sources	H (m)	$\mathcal{U}$(0.9,1.8)
	Dist. (m)	$\mathcal{U}$(0.66,2)
	θ	$\mathcal{U}$(0,2π)

TABLE II
HYPERPARAMETERS OF U-MAMBA-NET.

Encoder / Decoder	channel / window / hop	F / - / -	128 / 41 / 20
U-net	Input / output channel	-	128
Mamba	Input / output channel	-	128
U-net	Down / upsampling depths	L	4
U-Mamba	Number of repeat core	B	16
Sources	-	S	2

III. EXPERIMENTS

A. Datasets

We use open-source mixture dataset Libri2mix to conduct experiments [12]. Source audios are from Librispeech, ambient noises are sampled from WHAM! [15]. As for simulation of indoor surrounding, we exploit Pyacousticroom toolkit to generate reverberant version of clean source [16]. The simulated mixture is produced by adding each reverberant source and ambient noise together in time domain. The configuration of indoor reverebration is shown in Table I. T60 represents the reverberation intensity, indicating the time required for the sound energy to decay by 60 dB. As a result, our training, validation, and test datasets owns 13900, 3000, 3000 samples, respectively. All the samples are in 8 kHz.

B. Network

The basic hyperparameters that we use are exhibited in Table II. For the ablation studies, we indicate different parameter while keeping the same alphabet symbols. Additionally, we evaluate several different upsampling methods. Besides T-Conv1D, we also investigate the nearest neighbor (NN) and linear upsampling techniques.

C. Training phase

We train the proposed model on GeForce 4070Ti Super with a batch size of 4 and initial learning rate of 0.00015. Each sample is randomly cut into 3-s long. Notably, we report the computational load based on the input utterance with this length. We set the maximum epoch at 120. As for objective function, we adopt permutation-invariant scale-invariant single-to-noise ratio (SI-SNR) [13], shown as follows:

$$\mathcal{L} = -\max_{\pi \in \mathcal{P}} \frac{1}{I} \sum_{i} \text{SI-SNR}(\hat{s}_{\pi(i)}, s_i) \tag{10}$$

where π is the best permutation mapping set that allow overall SI-SNR to achieve maximum [17].

TABLE III
MAIN RESULTS.

Methods	CMTL	SI-SNRi	SDRi	SIRi	#Param (M)	GMACs
TasNet [22]	✗	5.70	5.05	10.84	23.2 M	27.8
SuDoRM-RF [9]	✗	2.90	3.36	6.14	2.6 M	3.6
SuDoRM-RF+ [9]	✗	5.33	6.05	11.02	2.7 M	3.0
Conv-TasNet [13]	✗	6.88	7.17	14.32	6.3 M	18.7
DPRNN [3]	✗	7.59	7.88	15.16	3.7 M	23.9
DPRNN	✓	8.08	8.62	16.39	5.6 M	40.2
U-Mamba-Net	✗	**8.50**	**8.62**	**17.67**	4.4 M	2.5

D. Evaluation phase

SI-SNR improvement (SI-SNRi) is an extension of SI-SNR, it checks how much information get excluded from mixture, we use SI-SNR improvement (SI-SNRi). We use signal-to-interference ratio improvement (SIRi) for separation effectiveness [18]. At the perceptual aspect, we use short-time objective intelligibility (STOI) [18] and perceptual evaluation of speech quality (PESQ) [19]. For denoising performance, we use predicted rating of speech distortion (CSIG), predicted rating of background distortion (CBAK), and predicted rating of overall quality (COVL) [20]. Additionally, we present the model size and computational load in terms of Giga Multiply-Add Operations per Second (GMACs) [21].

IV. EXPERIMENTAL RESULTS

A. Main results

We compare the results of U-Mamba-Net with several previous models in Table III. In the implementation of CMTL using DPRNN, we adopt an Enhancement Priority Pipeline (EPP). We can observe that, firstly, the proposed U-mamba-net achieves the best results cross major metrics. It is 0.92 dB better than DPRNN in SI-SNRi. Compared to CMTL, which uses DPRNN as sub-modules, U-mamba-net is still 0.42dB higher. Moreover, U-Mamba-Net model size is 20% smaller than DPRNN (CMTL). Most importantly, its computational efficiency (measured in GMACs) is only one-sixteenth of that of DPRNN (CMTL) (2.5 vs. 40.2) and one-ninth of that of DRRNN (2.5 vs. 23.9). This strongly demonstrates the proposed model's high efficiency and effectiveness.

B. Performance of perceptual quality and denoising effectiveness

TABLE IV
PERCEPTUAL AND DENOISING PERFORMANCE.

Methods	CMTL	STOI	PESQ	CSIG	CBAK	COVL
SuDoRM-RF+	✗	65.92	1.52	2.37	1.61	1.84
Conv-TasNet	✗	69.61	1.60	2.58	1.76	2.00
DPRNN	✗	71.86	1.67	2.65	1.83	2.07
DPRNN	✓	73.19	**1.75**	**2.75**	**2.16**	**2.18**
U-Mamba-Net	✗	**73.99**	1.70	2.56	1.89	2.05

In this section, we compare the U-Mamba-Net with previous models in terms of perceptual quality and denoising performance in Table IV. In terms of the STOI metric, U-Mamba-Net consistently maintains its advantage, just as the results and

TABLE V
ABLATION STUDIES.

F	R	L	Upsampling	SI-SNR	SI-SNRi	SDRi	SIRi	STOI (%)	PESQ	CSIG	CBAK	COVL	#Param (M)	GMACs
64	16	4	T-Conv1D	1.29	7.12	7.59	14.64	70.59	1.60	2.52	1.75	1.97	1.3 M	0.7
128	12	4	T-Conv1D	2.31	8.14	8.39	16.31	73.28	1.68	2.53	1.85	2.02	3.3 M	1.9
128	16	4	T-Conv1D	2.67	8.50	8.62	17.67	73.97	1.70	2.53	1.89	2.04	4.4 M	2.5
128	20	4	T-Conv1D	2.76	8.59	8.76	17.46	74.25	1.71	2.56	1.90	2.06	5.5 M	3.1
128	16	8	T-Conv1D	2.59	8.42	8.52	17.39	73.72	1.69	2.50	1.87	2.01	4.6 M	2.5
128	16	4	NN	2.70	8.52	8.66	17.44	73.99	1.70	2.56	1.89	2.05	4.4 M	2.5
128	16	4	Linear	2.70	8.53	8.70	17.54	74.06	1.71	2.55	1.89	2.05	4.4 M	2.5
192	16	4	T-Conv1D	3.02	8.85	8.90	18.15	74.87	1.74	2.64	1.93	2.11	9.7 M	5.3

conclusions in the previous section. It also shows superiority in PESQ, CSIG, CBAK, and COVL when compared to all models except DPRNN (CMTL). However, DPRNN (CMTL) outperforms U-Mamba-Net in these metrics, particularly in CBAK. This indicates that a single-task model's performance on specific sub-task is not as effective as that of a CMTL architecture incorporating additional processing modules. We believe the reason is that DPRNN (CMTL) can leverage additional information like noise-free mixture as intermediate label to supervised model in a step-by-step manner. This suggests that combining specialized modules in a multi-task framework can offer superior performance by addressing various aspects of the problem more comprehensively.

C. Ablation studies

To better understand the role of each parameter in the model, we design several ablation experiments. Their results are displayed in Table V. Firstly, the impact of feature dimensions on the model is most noticeable. As the feature dimension F increases from 64 to 128 to 192, the separation performance significantly improves like in SI-SNRi from 7.12 dB to 8.50 dB to 8.85 dB, but the model size also increases accordingly. Increasing the number of U-mamba-Net blocks R also enhances the model's performance, although the improvement is not as pronounced. With 20 blocks, the model reaches an SI-SNRi score of 8.59 dB. Increasing the depth L of the model has a negative effect on the test set. We speculate that for separation tasks, excessively low resolution may not be beneficial. Using NN and linear upsampling method though has a subtle improvement on the test set, we observe a decreased performance in the validation set during training. Therefore, upsampling methods are less critical than any other hyperparameters else.

D. Visualization

This section visualizes an example of separation results using DPRNN and U-Mamba-Net in Fig 2. First, from the spectrograms of the results, both DPRNN and U-Mamba-Net exhibit effective denoising and dereverberation, as evidenced by the clarity of the spectrograms of separations. Regarding the separation results, DPRNN exhibits more erroneous separations, where information originally belonging to one source is allocated to another source. This issue is highlighted by the red boxes in the figure. We believe this problem arises from the model's less robust modeling of long-term dependencies. U-Mamba-Net also has some weaknesses. For instance, as shown in the white boxes, the generated speech spectrogram lines are not as clear as those produced by DPRNN. Even worse, when we listen to the separated speech, its perceptual quality remains relatively low, which is consistent with the STOI score indicated in the previous sections.

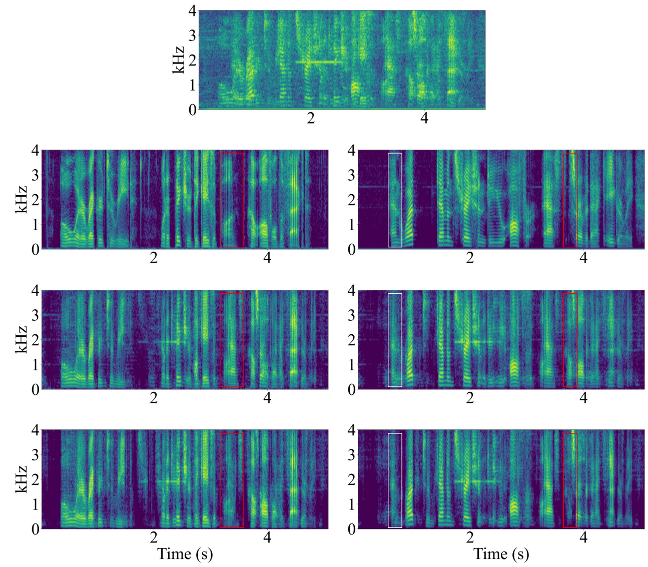


Fig. 2. Spectrogram of separation results. The sole spectrogram in the first row is noise and reverberant mixture. The following two spectrograms in the second row are ground truths. Third row are displaying the spectrograms of two separated results by DPRNN model. The last two are estimation of U-Mamba-Net. The red boxes highlight the places where DPRNN makes wrong separation but U-Mamba does not. The white box outlines the place where U-Mamba-Net performs worse. Because the fundamental frequencies and harmonics of U-Mamba-Net are not as clear as those of the DPRNN.

V. Conclusions

In this work, we proposed a lightweight U-Mamba-Net for noisy and reverberant speech separation. U-Mamba-Net not only demonstrates impressive separation capabilities but also maintains low computational load. In terms of SNR category metrics, it surpasses many previous models, demonstrating its overall robust separation capability. However, it still shows significant gaps in denoising and other specific metrics compared to CMTL.

Acknowledgment

This work is supported by a fellowship of the Nagoya University CIBoG WISE program from MEXT.

References

[1] M. Maciejewski, G. Wichern, E. McQuinn, and J. Le Roux, "Whamr!: Noisy and reverberant single-channel speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 696–700.

[2] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[3] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: Efficient long sequence modeling for time-domain single-channel speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 46–50.

[4] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[5] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 21–25.

[6] S. Dang, T. Matsumoto, Y. Takeuchi, and H. Kudo, "A separation priority pipeline for single-channel speech separation in noisy environments," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 12 511–12 515.

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, Springer, 2015, pp. 234–241.

[8] A. Défossez, N. Usunier, L. Bottou, and F. Bach, *Demucs: Deep extractor for music sources with extra unlabeled data remixed*, 2019. arXiv: 1909.01174 [cs.SD]. [Online]. Available: https://arxiv.org/abs/1909.01174.

[9] E. Tzinis, Z. Wang, and P. Smaragdis, "Sudo rm -rf: Efficient networks for universal audio source separation," in *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, IEEE, Sep. 2020.

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

[11] A. Gu, I. Johnson, A. Timalsina, A. Rudra, and C. Ré, "How to train your hippo: State space models with generalized orthogonal basis projections," *arXiv preprint arXiv:2206.12037*, 2022.

[12] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Librimix: An open-source dataset for generalizable speech separation," *arXiv preprint arXiv:2005.11262*, 2020.

[13] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[14] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," *arXiv preprint arXiv:2111.00396*, 2021.

[15] G. Wichern, J. Antognini, M. Flynn, *et al.*, "WHAM!: Extending speech separation to noisy environments," in *20th Annual Conference of the International Speech Communication Association, Interspeech 2019, Graz, Austria, September 15-19, 2019*, ISCA, 2019, pp. 1368–1372.

[16] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 351–355.

[17] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 626–630.

[18] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[19] 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.

[20] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2007.

[21] V. Sovrasov. "Ptflops: A flops counting tool for neural networks in pytorch framework." (2024), [Online]. Available: https://github.com/sovrasov/flops-counter.pytorch.

[22] Y. Luo and N. Mesgarani, "TasNet: Time-domain audio separation network for real-time, single-channel speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 696–700.