

SepMamba: State-space models for speaker separation using Mamba

Thor Højhus Avenstrup*, Boldizsár Elek*, István László Mádi*, András Bence Schin*,
Morten Mørup*, Bjørn Sand Jensen* and Kenny Olsen*†

*Technical University of Denmark

†WS Audiology A/S

Abstract—Deep learning-based single-channel speaker separation has improved significantly in recent years in large part due to the introduction of the transformer-based attention mechanism. However, these improvements come with intense computational demands, precluding their use in many practical applications. As a computationally efficient alternative with similar modeling capabilities, Mamba was recently introduced. We propose SepMamba, a U-Net-based architecture composed of bidirectional Mamba layers. We find that our approach outperforms similarly-sized prominent models — including transformer-based models — on the WSJ0 2-speaker dataset while enjoying significant computational benefits in terms of multiply-accumulates, peak memory usage, and wall-clock time. We additionally report strong results for causal variants of SepMamba. Our approach provides a computationally favorable alternative to transformer-based architectures for deep speech separation.

Index Terms—speaker separation, deep learning, selective state-space models

I. INTRODUCTION

Speech separation is the problem of extracting separate source signals from a single mixture, also known as the cocktail party problem [1], and is a fundamental task in many modern audio processing systems, such as hearing aids or telecommunication devices — systems which are typically low-resource environments making computationally expensive methods impractical.

Recent deep learning models have greatly improved the quality of speaker separation systems, but at the cost of significant computational overhead. Prior approaches to deep learning-based speaker separation have until recently chiefly been based on using the short-time Fourier transform (STFT) to map input audio mixture into the frequency-domain, performing separation on the (complex) frequency representation, and then using the inverse STFT to construct the estimated sources [2]. These methods come with several drawbacks. Firstly, the STFT transforms the signal into the complex domain, where both magnitude and phase must be explicitly modelled. In practice, using the phase information is challenging, and without careful design models often end up reusing the input phase for the estimated sources, which sets an upper bound on the separation quality [3]. Secondly, for high-quality speech separation large frame lengths are typically used which heavily limits the usefulness of such models in cases where low latency is required.

Instead of modeling the mixture signal in the time-frequency representation, several recent approaches attempt to model the

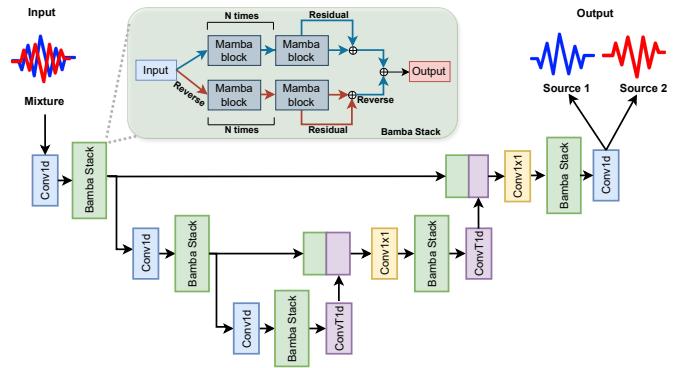


Fig. 1. SepMamba has 5 stages of processing with Bamba stacks. The downsampling and upsampling is handled by convolutional and matching transposed convolutional layers. The skip connections are projected into the required dimension with 1×1 convolutions. We double the dimension of the Mamba blocks after each downsampling by a factor of 2, and halve it after each upsampling (ensuring matching dimensions on the same level).

mixture directly in the time-domain. The first time-domain solution to demonstrate superior performance to STFT-based systems was the TasNet [3]. TasNet proposed an encoder-masker-decoder architecture where separation is performed in a learned linearly encoded basis, and decoded using another linear basis, reminiscent of STFT-based systems but using real-valued learnable bases. The masking network in the original TasNet was a deep long short-term memory (LSTM) network, which was later replaced in Conv-TasNet [4] using only convolutional layers throughout the whole model. Notably, Conv-TasNet outperformed all existing solutions (in terms of SI-SNR and SDR) on the WSJ0-2mix dataset in the non-causal setting and reached comparable results on the causal setting with a substantially smaller model size. Later SudoRM-RF [5], a convolution-based model, proposed a resource efficient architecture with improved performance over its predecessors.

Transformer-based models have become the new standard backbone in deep speech separation models [6]–[8] due to their strong modelling capacity and efficient use of parallelism on GPUs during training. SepFormer [6] was the first model to rely entirely on a transformer-based masker network in a TasNet-like structure and achieved state-of-the-art performance on the WSJ0-2/3mix dataset. Here, the masker network employs transformer blocks to capture both short-term and long-term dependencies. Later MossFormer [7] and MossFormer2 [9]

employed a unified attentive gating model, further improving performance on the WSJ0-2/3mix datasets.

A notable drawback of these models, and generally of models building on the transformer-based attention mechanism [10], is their quadratic complexity over the input sequence length, which necessitates processing samples in shorter chunks (e.g. SepFormer uses a chunk size of 250 samples). While this approach effectively handles dependencies within and across chunks due to the dual-path structure — first introduced by DP-RNN [11] —, it may still struggle with long-range dependencies if critical information spans multiple chunks. Furthermore, the process of segmenting the input during training, which is counter-intuitive for audio — an inherently continuous signal — leaves the question of whether sequence models that can model a full signal without segmentation could further improve performance and efficiency.

State-space models have emerged as a promising alternative to existing sequence modeling architectures. Recently, Mamba [12] has shown strong results in language modeling and the modeling of DNA within genomics where it beat or matched transformer-based architectures. Furthermore, Mamba has been explored in biomedical image segmentation tasks [13] as well, where it also outperformed CNN- and transformer-based segmentation networks. Notably, Mamba was also used for the unconditional generation of audio [12], using a Sashimi architecture [14], and was found to outperform its predecessor, the S4 [15] layer.

Mamba has previously been combined with U-Nets, e.g., in the imaging domain [16], [17] and U-Net inspired hierarchical structures with Mamba has also been successfully developed in the context of sequence modeling for sensor based data [18]. In the speech separation space, SP-Mamba [19] builds on the successful TF-GridNet [8] architecture by exchanging the bidirectional LSTM components with bidirectional Mamba blocks. While SP-Mamba reports strong results, it still employs the transformer-based attention mechanism, making it computationally resource-demanding.

The rise of state-space models — especially Mamba — as a general sequential model poses an opportunity to develop new, less resource-intensive and more computationally efficient solutions in the speech separation space as well, while achieving performance comparable to current state-of-the-art transformer-based models [6]–[8].

We propose SepMamba, the first Mamba-based architecture for speaker separation in the time domain that does not rely on expensive transformer-based attention mechanisms. The new architecture achieves competitive performance in both causal and non-causal settings by incorporating Mamba layers into a U-Net architecture to efficiently learn multi-scale structures in sound, enabling inexpensive learning of long-range dependencies. We provide a comprehensive overview of existing speaker separation results on WSJ0-2mix and report strong performance, matching or outperforming transformer-based architectures at a fraction of the computational costs. Additionally, SepMamba achieves much lower forward and backward pass wall-clock timings and a significantly lower

peak memory use than comparable models.*

II. METHODS

The Mamba layer is a sequence-to-sequence transformation mapping a 1-D input signal with discrete-time samples $x_t \in \mathbb{R}$ to a 1-D output signal $y_t \in \mathbb{R}$ through intermediate hidden states $h_t \in \mathbb{R}^D$, using the (discretized) state transition eqs. (1) and (2) with parameters $\mathbf{A}, \mathbf{B}, \mathbf{C}, \Delta$ and discretization rule eqs. (3) and (4),

$$\mathbf{h}_t = \bar{\mathbf{A}}\mathbf{h}_{t-1} + \bar{\mathbf{B}}x_t, \quad (1)$$

$$y_t = \mathbf{C}\mathbf{h}_t, \quad (2)$$

$$\bar{\mathbf{A}} = \exp(\Delta \mathbf{A}), \quad (3)$$

$$\bar{\mathbf{B}} = (\Delta \mathbf{A})^{-1} \exp(\Delta \mathbf{A} - \mathbf{I}) \cdot \Delta \mathbf{B}. \quad (4)$$

A. SepMamba Architecture

Our proposed SepMamba architecture operates on raw audio waveform, in contrast to STFT-domain models such as TF-GridNet [8] or SP-Mamba [19], and is based on the U-Net [20] architecture — composed of five stages of down/up-sampling with a bidirectional Mamba (Bamba) block at each stage. Each Bamba block additively combines the outputs of a stack of Mamba blocks with those of a separate stack of Mamba blocks run on a reversed copy of the input,

$$\text{Bamba}(x) = \text{Mamba}_1(x) + \text{flip}(\text{Mamba}_2(\text{flip}(x))), \quad (5)$$

where in each stage of processing we use the same number of Mamba blocks per Bamba stack, but double the dimensionality of the Mamba layers (the channel dimension) with every downsampling, and halve it with every upsampling. We use standard convolutions for downsampling and matching transposed convolutions during upsampling. Skip connections between features with different channel dimensions are projected to match the target dimensionality using 1×1 convolutional layers. We use ReLU activations throughout the whole network. For causal variants we match the number of Mamba blocks per stage, but without reversing the inputs of either branch. The architecture is illustrated in fig. 1, and model configurations can be found in table I.

Besides the small number of convolutions used for down/up-sampling, our model is the first instance of a speaker separation network relying only on Mamba layers to learn temporal dependencies, while other methods that have also started to incorporate Mamba blocks — such as SP-Mamba, which replaces the bidirectional LSTM module in TF-GridNet with Mamba layers — still rely on transformer-based layers for the bulk of their computation.

B. Experimental Setup

The models were trained and evaluated on the Wall Street Journal 0 (WSJ0) 2-speaker setup [21] (WSJ0-2mix) using the dynamic mixing (DM) data augmentation technique [22], which creates new mixtures from randomly sampled speaker

*The implementation of SepMamba is available at: <https://github.com/andrasschin/SepMamba>.

utterances on-the-fly. The sources are mixed by uniformly sampling an SNR value from the interval $[-2.5, 2.5]$. Additionally, speed perturbation is employed on the sources, changing the speed of the audio to between 95% and 105% of the original. During training, gradient clipping of 5.0 is used to ensure stable convergence of the model parameters.

The model parameters are inferred by minimizing the negative Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) [23] and use utterance-level Permutation Invariant Training [24] (uPIT). During training, we threshold the loss at -30 as in WaveSplit [25]. The AdamW [26] optimizer is used with an initial learning rate of 15×10^{-5} , weight decay of 0.1 and $\beta = (0.9, 0.999)$. Training loss was monitored for convergence at the initial constant learning rate, upon which an exponential decay learning rate schedule is introduced with a gamma value of 0.98 to 0.99, depending on the model. Training typically takes 6–7 days on an NVIDIA A100 GPU. A batch size of 1 was used for all training runs, as we found larger batch sizes either produced significantly worse results or diverged entirely, a phenomenon we also observed while attempting to replicate other models such as SepFormer and SuDoRM-RF when trained using uPIT.

The SI-SDR improvement (SI-SDR_i) is reported on the held-out test data, which measures the difference between the SI-SDR after processing and the SI-SDR before processing. We additionally report the Signal-Distortion Ratio improvement (SDR_i) and Scale-Invariant Signal-to-Noise Ratio improvement (SI-SNR_i) to facilitate comparisons in table II.

We also report compute requirements in terms of giga-multiply-accumulates (GMAC) as a hardware-independent measure of computational intensity, as well as wall-clock timings specific to A100 GPUs and peak memory use.

III. RESULTS AND DISCUSSION

Table II compares the results achieved on the WSJ0-2mix dataset by SepMamba and other prominent architectures. We highlight that SepMamba (M) outperforms the transformer-based SepFormer and MossFormer (M) architectures with a substantially reduced compute and memory footprint. In a similar manner, SepMamba (M) also outperforms the related SP-Mamba model with significantly lower computational and memory requirements, demonstrating the advantages of a fully Mamba-based architecture. Lastly, our smaller model, SepMamba (S) outperforms prior models at a similar parameter count.

Our models in the causal setting achieved SI-SNR_i scores of 21.4 and 19.2. These results outperform current state-of-the-art causal models, such as UX-NET (SI-SNR_i 13.6) [28] and Causal Deep Casa (SI-SNR_i 15.2) [29].

Wall-clock timings: Figure 2 (Left) shows the average forward pass wall-clock time on an A100 GPU. It is calculated by measuring the number of forward passes each model can complete in 10 seconds and averaging. Both SepMamba (S) and SepMamba (M) have significantly lower forward pass time than other models while outperforming most of them in terms of SI-SDR_i.

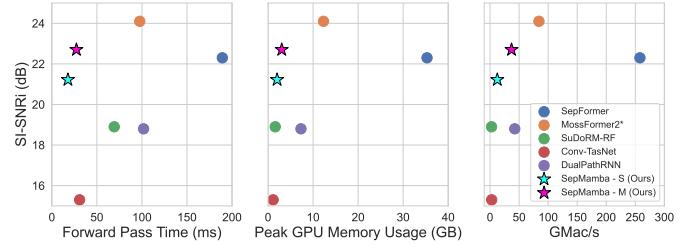


Fig. 2. **(Left)** Average forward pass time on an NVIDIA A100 GPU for 4 seconds of audio samples at 8 kHz. **(Middle)** Peak GPU memory usage during the backpropagation of a 4 seconds sample at 8kHz on an NVIDIA A100 GPU. **(Right)** Multiply-accumulate (MAC) operations per seconds. *For MossFormer2 SI-SDR_i is listed instead of SI-SNR_i.

Memory usage: Figure 2 (Middle) shows the peak memory usage by each model during backpropagation on a four second audio sample at 8 kHz. Both SepMamba (S) and SepMamba (M) have memory use comparable to that of the Conv-TasNet and SudoRM-RF, while being notably more efficient than MossFormer 2, and significantly more efficient compared to SepFormer.

Compute intensity: Figure 2 (Right) shows the number of giga-multiply-accumulates (GMAC) performed by each model during a forward pass per second of audio sampled at 8kHz compared with SI-SNR_i on WSJ0-2mix. SepMamba (M) outperforms SepFormer with a fraction of the computational needs. While MossFormer2 achieves higher SI-SDR_i, it requires more than twice the GMAC/s. Similarly, the smaller SepMamba (S) has a strong performance with a relatively low computational overhead that matches previous efficient models.

While in most cases SepMamba outperforms other methods at a lower compute and memory footprint, several other methods achieve higher performance as a function of parameter count. In the case where parameter count is the limiting factor on a system, we note that there is plenty of opportunity for optimizing the parameter-efficiency of SepMamba, such as parameter-sharing (e.g. between branches in Bamba stacks) or adjusting the kernel size of the convolutions to trade-off compute and parameter efficiency.

In our analysis we considered GMAC/s and wall-clock timings on A100 GPUs as the two primary metrics for compute performance, but real-world performance depends heavily on the characteristics of the system in question — indeed the Mamba layer is a very memory-bound operation with a low arithmetic intensity [12], and so its runtime on GPUs becomes dominated by transfers between local and global GPU memory, whereas the attention mechanism still incurs quadratic compute cost, even when using the highly optimized FlashAttention [30] implementation. This implies that we may expect an even greater performance advantage for Mamba-based architectures on systems where less parallelism and compute is available.

A key advantage of Mamba is its compute and memory efficiency, especially over the transformer-based attention mechanism. Another advantage is that it features a substantially smaller state when operating in a recurrent, real-time context, as only the current hidden state h_t is required to perform

	Size	Dim	# Blocks	Kernel size	Stride	# Params	Causal*	GMAC/s	SI-SNRi
SepMamba	S	64	8	16	2	7.2M	✓	12.46	19.2
SepMamba	S	64	8	16	2	7.2M	✗	12.46	21.2
SepMamba	M	128	6	16	2	22M	✓	37.0	21.4
SepMamba	M	128	6	16	2	22M	✗	37.0	22.7

TABLE I
PARAMETERIZATIONS FOR THE HIGHLIGHTED MODELS. **DIM** REFERS TO THE DIMENSIONS OF THE MAMBA BLOCKS IN THE FIRST AND THE LAST STAGE. **# BLOCKS** REFER TO THE TOTAL NUMBER OF MAMBA BLOCKS PER STAGE.

	SI-SNRi	SI-SDRi	SDRi	# Params	GMAC/s	Fw. pass (ms)	Mem. Usage (GB)
Conv-TasNet [4]	15.3	—	15.6	5.1M	2.82	30.79	1.13
DualPathRNN [11]	18.8	—	19.0	2.6M	42.52	101.83	7.31
SudoRM-RF [5]	18.9	—	—	2.6M	2.58	69.23	1.60
SepFormer [6]	20.4	—	—	26M	257.94	189.25	35.30
SepFormer [6] + DM	22.3	—	—	26M	257.94	189.25	35.30
MossFormer [7] (S)	—	20.9	—	10.8M	— [†]	— [†]	—
MossFormer [7] (M) + DM	—	22.5	—	25.3M	— [†]	— [†]	—
MossFormer [7] (L) + DM	—	22.8	—	42.1M	70.4	72.71	9.57
MossFormer2 [9] + DM	—	24.1	—	55.7M	84.2	97.60	12.30
TF-GridNet [8] (S)	—	20.6	—	8.2M	19.2	—	—
TF-GridNet [8] (M)	—	22.2	—	8.4M	36.2	—	—
TF-GridNet [8] (L)	—	23.4	23.5	14.4M	231.1	—	—
SP-Mamba [19]	22.5 [‡]	—	—	6.14M	119.35	148.11	14.40
SepMamba (S) + DM (ours)	21.2	21.2	21.4	7.2M	12.46	17.84	2.00
SepMamba (M) + DM (ours)	22.7	22.7	22.9	22M	37.0	27.25	3.04

TABLE II

PERFORMANCE COMPARISON ON THE WSJ0-2MIX DATASET. GMAC/S IS REPORTED FOR A FORWARD PASS OVER 1 SECOND OF 8 KHZ AUDIO AND CALCULATED USING PTFLOPS[§] [27]. FORWARD PASS TIME IS THE AVERAGE OF 4 SECONDS OF AUDIO SAMPLES ON 8 KHZ. MEMORY USAGE IS THE PEAK MEMORY USAGE DURING BACKPROPAGATION OF 4 SECONDS OF AUDIO SAMPLED AT 8 KHZ. ALL CALCULATIONS ARE IN FP32.

inference at the subsequent timestep, compared to the attention mechanism which requires storing states proportional to the entire length of the input sequence that it is attending to.

A. Architectural Considerations

In this section, we discuss how some of the architectural decisions influenced our training runs. Decreasing the stride to two is beneficial, as the performance gained outweighs the slightly longer forward and backward pass.

We also decided to construct our Bamba stacks with several blocks per branch in the stack because pilot studies found that recombining the forward and reversed inputs after every block leads to worse performance.

For the causal U-Net we additionally experimented with other activation functions as well, namely SiLU [31], Mish [32] and PReLU [33], but we did not see significant improvements in the pilot runs to justify further experimentation.

*Refers to the causal setting of the Mamba blocks, not the convolutions.

[†]Based on the descriptions in the public repository of MossFormer [7], we were not able to obtain the source code for the smaller models.

[‡]Result taken from <https://github.com/JasperLee/SPMamba>.

[§]Mossformer GMAC/s calculation used FlopCounterMode from PyTorch with the assumption MAC = FLOP/2. Mamba layer GMAC/s calculation is based on <https://github.com/state-spaces/mamba/issues/110>.

IV. CONCLUSION

In this work we proposed SepMamba, a highly efficient U-net architecture for speech separation based on Mamba layers, and demonstrated strong performance on WSJ0-2mix using a fraction of the compute and memory budget of comparable methods.

We provided an extensive performance per compute review of recent methods, and find that SepMamba broadly outperforms competing methods, especially at lower compute budgets.

SepMamba forms a promising efficiency-focused alternative to transformer-based models, suitable for use in lower-resource systems. Taken together with our strong causal results, we suggest that SepMamba may be a strong candidate for real-world, low-power and low-latency deployments.

REFERENCES

- [1] E. Colin Cherry, "Some experiments on the recognition of speech, with one and with two ears," *Journal of the Acoustical Society of America*, vol. 25, pp. 975–979, 1953.
- [2] Yusuf Isik, Jonathan Le Roux, Zhuo Chen, Shinji Watanabe, and John R. Hershey, "Single-channel multi-speaker separation using deep clustering," 2016.
- [3] Yi Luo and Nima Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," 2018.
- [4] Yi Luo and Nima 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, Aug. 2019.
- [5] Efthymios Tzinis, Zhepei Wang, and Paris Smaragdis, "Sudo rm -rf: Efficient networks for universal audio source separation," in *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2020, pp. 1–6.
- [6] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong, "Attention is all you need in speech separation," 2021.
- [7] Shengkui Zhao and Bin Ma, "Mossformer: Pushing the performance limit of monaural speech separation using gated single-head transformer with convolution-augmented joint self-attentions," 2023.
- [8] Zhong-Qiu Wang, Samuele Cornell, Shukjae Choi, Younglo Lee, Byeong-Yeol Kim, and Shinji Watanabe, "Tf-gridnet: Integrating full- and sub-band modeling for speech separation," .
- [9] Shengkui Zhao, Yukun Ma, Chongjia Ni, Chong Zhang, Hao Wang, Trung Hieu Nguyen, Kun Zhou, Jiaqi Yip, Dianwen Ng, and Bin Ma, "Mossformer2: Combining transformer and rnn-free recurrent network for enhanced time-domain monaural speech separation," 2023.
- [10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," 2023.
- [11] Yi Luo, Zhuo Chen, and Takuya Yoshioka, "Dual-path rnn: Efficient long sequence modeling for time-domain single-channel speech separation," .
- [12] Albert Gu and Tri Dao, "Mamba: Linear-time sequence modeling with selective state spaces," 2023.
- [13] Jun Ma, Feifei Li, and Bo Wang, "U-mamba: Enhancing long-range dependency for biomedical image segmentation," 2024.
- [14] Karan Goel, Albert Gu, Chris Donahue, and Christopher Ré, "It's raw! audio generation with state-space models," 2022.
- [15] Albert Gu, Karan Goel, and Christopher Ré, "Efficiently modeling long sequences with structured state spaces," .
- [16] Ziyang Wang, Jian-Qing Zheng, Yichi Zhang, Ge Cui, and Lei Li, "Mamba-unet: Unet-like pure visual mamba for medical image segmentation," 2024.
- [17] Weibin Liao, Yinghao Zhu, Xinyuan Wang, Chengwei Pan, Yasha Wang, and Liantao Ma, "Lightm-unet: Mamba assists in lightweight unet for medical image segmentation," 2024.
- [18] Raunaq Bhirangi, Chenyu Wang, Venkatesh Pattabiraman, Carmel Majidi, Abhinav Gupta, Tess Hellebrekers, and Lerrel Pinto, "Hierarchical state space models for continuous sequence-to-sequence modeling," .
- [19] Kai Li and Guo Chen, "Spmamba: State-space model is all you need in speech separation," 2024.
- [20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015.
- [21] John R. Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," 2015.
- [22] Neil Zeghidour and David Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," 2020.
- [23] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey, "Sdr - half-baked or well done?," 2018.
- [24] Morten Kolbæk, Dong Yu, Zheng-Hua Tan, and Jesper Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE Transactions on Audio, Speech and Language Processing*, 2017.
- [25] Neil Zeghidour and David Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," 2020.
- [26] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," 2019.
- [27] Vladislav Sovrasov, "ptflops: a flops counting tool for neural networks in pytorch framework," 2018-2023.
- [28] Kashyap Patel, Anton Kovalyov, and Issa Panahi, "Ux-net: Filter-and-process-based improved u-net for real-time time-domain audio separation," 2022.
- [29] Yuzhou Liu and DeLiang Wang, "Causal deep casa for monaural talker-independent speaker separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2109–2118, 2020.
- [30] Tri Dao, "Flashattention-2: Faster attention with better parallelism and work partitioning," .
- [31] Stefan Elfwing, Eiji Uchibe, and Kenji Doya, "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning," 2017.
- [32] Diganta Misra, "Mish: A self regularized non-monotonic activation function," 2020.
- [33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," 2015.