SPIKING STRUCTURED STATE SPACE MODEL FOR MONAURAL SPEECH ENHANCEMENT

Yu Du*, Tsinghua University Department of Precision Instrument,Beijing, China.
Xu Liu†,Yansong Chua‡, China Nanhu Academyof Electronics and Information Technology,Jiaxing, China.

*Thanks to the National Natural Science Foundation of China (6183600462236009).
†Co frst authorThanks to the STI 2030 Major Proiects2021ZD0200300.
‡Corresponding authorThanks to the STI 2030 Major Projects2021ZD0200300.

Abstract

Speech enhancement seeks to extract clean speech from noisy signals. Traditional deep learning methods face two challenges: efficiently using information in long speech sequences and high computational costs. To address these, we introduce the Spiking Structured State Space Model (Spiking-S4). This approach merges the energy efficiency of Spiking Neural Networks (SNN) with the long-range sequence modeling capabilities of Structured State Space Models (S4), offering a compelling solution. Evaluation on the DNS Challenge and VoiceBank+Demand Datasets confirms that Spiking-S4 rivals existing Artificial Neural Network (ANN) methods but with fewer computational resources, as evidenced by reduced parameters and Floating Point Operations (FLOPs).

Method

As shown in Fig. 1, the noisy speech signal is first transformed into the time-frequency domain using the Short-Time Fourier Transform (STFT) layer. Then, the magnitude is fed into the linear encoder to generate the input u of shape $K \times L$ for the spiking S4 layers, where K is the length of the input sequence and L is the independent S4 kernel number of each sequence element. Next it is passed to N spiking S4 layers and a linear decoder, which produces a magnitude mask $\widehat{M}$. This mask is then multiplied with the original magnitude to obtain the denoised magnitude. Finally, the denoised magnitude and the phase information are combined and converted back to the time domain using Inverse Short-Time Fourier Transform (ISTFT) layer.

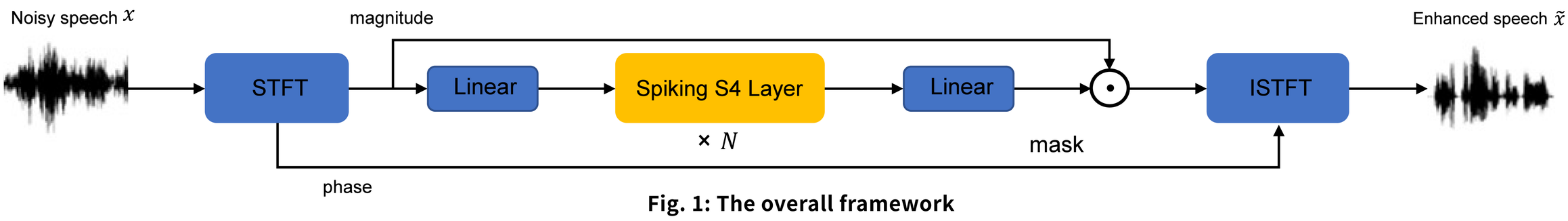

Fig. 1: The overall framework

As shown in Fig. 2, each step of the encoded feature is first passed to L independent S4 kernels with the hidden size of H . Then it is passed to an emission layer, and a LIF node which collects input signals gradually, accumulating them over a duration of time, and generates a spike when the membrane potential reaches a predefined threshold. Finally, the spikes are fed into a linear decoder to be restored back to the real domain.

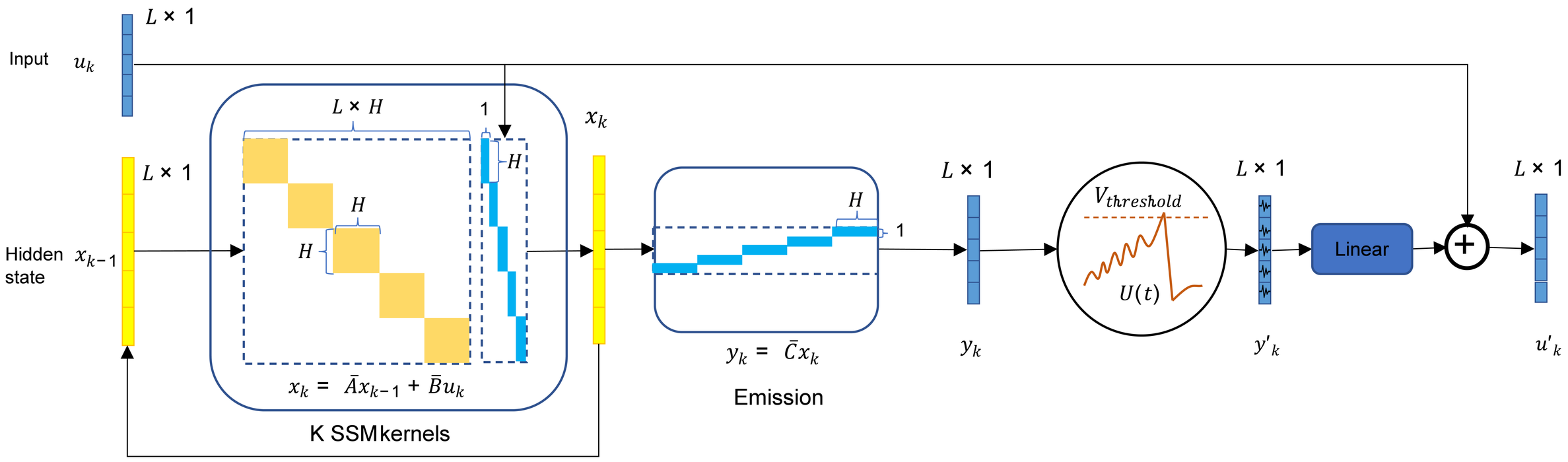

Fig. 2: The spiking S4 layer

Results

Table 1: Results on DNS Challenge 2023 validation set and test set.The ANN and SNN-based models are separated by a horizontal line.

Method	SISNR	PESQ	STOI	DNSMOS		
Validation set						
Wave-U-Net [23]	13.70	1.80	0.88	2.91	3.15	3.52
S4 [8]	14.82	1.99	0.89	2.93	3.23	3.91
FRCRN [2]	**15.51**	2.50	**0.92**	**3.09**	**3.41**	**3.96**
Sigma-Delta [17]	11.7	1.69	0.86	2.67	3.17	3.44
Spiking-S4	14.42	**2.73**	0.89	2.85	3.21	3.74
Test set						
Wave-U-Net [23]	13.90	1.85	1.02	3.01	3.25	3.65
S4 [8]	15.01	2.76	0.89	2.93	3.24	3.89
FRCRN [2]	**15.67**	2.52	**0.92**	**3.08**	**3.41**	**3.95**
Sigma-Delta [17]	11.21	2.43	0.86	2.68	3.14	3.51
Spiking-S4	14.58	**2.75**	0.89	2.85	3.21	3.74

Table 2: Results on Voice-Bank+Demand dataset.The ANN and SNN-based models are separated by a horizontal line.

Method	WB-PESQ	CSIG	CBAK	COVL
Wave-U-Net [23]	3.25	4.20	3.61	3.30
GaGNet [24]	2.94	4.26	3.45	3.59
MetricGAN+ [25]	3.15	4.14	3.16	3.64
PERL-AE [26]	3.17	4.43	3.53	3.83
FRCRN [2]	3.21	4.23	3.64	3.37
S4 [8]	3.38	**4.93**	2.63	4.30
Sigma-Delta [17]	3.20	4.89	2.59	4.15
Spiking-S4	**3.39**	4.92	**2.64**	**4.31**

Table 3: Comparision of parameters and FLOPs.

Method	Parameter	FLOPs
Wave-U-Net [23]	70.1M	3.36×10^{10}
GaGNet [24]	5.9M	8.13×10^{9}
Sigma-Delta [17]	0.53M	1.97×10^{9}
FRCRN [2]	14.0M	1.13×10^{12}
S4[8]	0.79M	2.48×10^{9}
Spiking-S4	**0.53M**	1.50×10^{9}

References

[17] Jonathan Timcheck, Sumit Bam Shrestha, Daniel Ben Dayan Rubin, Adam Kupryjanow, Garrick Orchard, Lukasz Pindor, Timothy Shea, and Mike Davies, “The intel neuromorphic dns challenge,” Neuromorphic Computing and Engineering, vol.3, no. 3, pp. 034005, 2023.
[23] Craig Macartney and Tillman Weyde, “Improved speech enhancement with the wave-u-net,” arXiv preprint arXiv:1811.11307, 2018.
[24] Andong Li, Chengshi Zheng, Lu Zhang, and Xiaodong Li, “Glance and gaze: A collaborative learning framework for single-channel speech enhancement,” Applied Acoustics, vol.187, pp. 108499, 2022.
[25] Szu-Wei Fu, Cheng Yu, Tsun-An Hsieh, Peter Plantinga, Mirco Ravanelli, Xugang Lu, and Yu Tsao, “Metricgan+: An improved version of metricgan for speech enhancement,” arXiv preprint arXiv:2104.03538, 2021.
[26] Saurabh Kataria, Jes′us Villalba, and Najim Dehak, “Perceptual loss based speech denoising with an ensemble of audio pattern recognition and self-supervised models,” in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 7118–7122.
[2] Shengkui Zhao, Bin Ma, Karn N Watcharasupat, and WoonSeng Gan, “Frcrn: Boosting feature representation using frequency recurrence for monaural speech enhancement,” in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022, pp.9281–9285.
[8] Albert Gu, Karan Goel, and Christopher R′e, “Efficiently modeling long sequences with structured state spaces,” arXiv preprint arXiv:2111.00396, 2021.