When Audio Denoising Meets Spiking Neural Network

Xiang Hao
Department of Computing
The Hong Kong Polytechnic University
Hong Kong SAR, China
haoxiangsnr@gmail.com

Chenxiang Ma
Department of Computing
The Hong Kong Polytechnic University
Hong Kong SAR, China
chenxiang.ma@connect.polyu.hk

Qu Yang
College of Design and Engineering
National University of Singapore
Singapore
quyang@u.nus.edu

Kay Chen Tan
Department of Computing
The Hong Kong Polytechnic University
Hong Kong SAR, China
kaychen.tan@polyu.edu.hk

Jibin Wu
Department of Computing
The Hong Kong Polytechnic University
Hong Kong SAR, China
jibin.wu@polyu.edu.hk

***Abstract*—Audio denoising techniques are essential tools for enhancing audio quality. Spiking neural networks (SNNs) offer promising opportunities for audio denoising, as they leverage brain-inspired architectures and computational principles to efficiently process and analyze audio signals, enabling real-time denoising with improved accuracy and reduced computational overhead. This paper introduces Spiking-FullSubNet, a real-time audio denoising model based on SNN. Our proposed model incorporates a novel gated spiking neuron model (GSN) to effectively capture multi-scale temporal information, which is crucial for achieving high-fidelity audio denoising. Furthermore, we propose the integration of GSNs within an optimized FullSubNet neural architecture, enabling efficient processing of full-band and sub-band frequencies while significantly reducing computational overhead. Alongside the architectural advancements, we incorporate a metric discriminator-based loss function that selectively enhances the desired performance metrics without compromising others. Empirical evaluations show the superior performance of Spiking-FullSubNet, ranking it as the winner of Track 1 (Algorithmic) of the Intel Neuromorphic Deep Noise Suppression Challenge.**

***Index Terms*—speech denoising, spiking neural network, neuromorphic computing, audio signal processing**

I. Introduction

Spiking neural networks (SNNs) are emerging as an energy-efficient alternative to traditional artificial neural networks (ANNs) [1]. However, SNNs are primarily explored for classification tasks, there has been limited progress in applying them to regression tasks, hindering their broader application [2]–[4]. Audio denoising, a typical regression task, plays a crucial role in various applications, particularly on power-constrained edge devices such as headsets, hearing aids, and smartphones [5]. These devices require real-time processing capabilities while operating under restricted power budget to ensure a seamless user experience. Traditional methods driven by ANNs for audio denoising often struggle to meet these demands due to their heavy computational complexity [6]–[8].

In light of these challenges, SNNs present a promising solution for audio denoising [9]. By capitalizing on the event-driven computation and high-level sparsity of spiking events, the compute load could be significantly reduced [10], enabling real-time processing without compromising the power consumption limitation. Nevertheless, developing an SNN-based system that can deliver denoising performance comparable to conventional solutions is challenging. It remains an open question to determine suitable spiking neuron models, network architectures, and loss functions that can fully unleash the power of SNNs in audio denoising.

This paper presents a novel neuromorphic audio denoising system that is grounded on a comprehensive study of these perspectives. Firstly, we argue that existing spiking neuron models struggle to retain multi-scale temporal information, which is crucial for high-quality audio denoising. To overcome this challenge, we propose a novel gated spiking neuron model (GSN) that can dynamically control information storage, ensuring the preservation of critical historical information. Furthermore, we propose the integration of GSN within an enhanced FullSubNet [11] framework, which can handle different frequency bands with varying levels of detail, thereby improving computational efficiency. Additionally, we incorporate a metric discriminator-based loss function [12], [13] into our framework, which selectively improves targeted evaluation metrics without negatively impacting other performance measures. Our main contributions are threefold:

- We propose Spiking-FullSubNet, a novel real-time neuromorphic audio denoising system, by integrating the GSN model, FullSubNet architecture, and metric discriminator-based loss function. Our approach performs superior in capturing multi-scale temporal information, leading to accurate and efficient denoising results.

The first two authors contributed equally. Jibin Wu is the corresponding author. This work was supported by the Research Grants Council of the Hong Kong SAR (Grant No. PolyU11211521, PolyU15218622, PolyU15215623, and PolyU25216423), The Hong Kong Polytechnic University (Project IDs: P0039734, P0035379, P0043563, and P0046094), and the National Natural Science Foundation of China (Grant No. U21A20512, and 62306259).

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00275

- We validate the effectiveness of our system on the Intel N-DNS Challenge dataset, demonstrating significant improvements over other spike-based baselines in denoising quality metrics. Additionally, our system achieves a 10× reduction in energy consumption compared to conventional ANN-based solutions while maintaining comparable performance, making it a promising solution for power-constrained edge devices.
- We will open-source our code and release model checkpoints to facilitate future explorations and promote innovative solutions in neuromorphic audio denoising.

II. Method

Figure 1 illustrates the proposed Spiking-FullSubNet system, which integrates a full-band model and multiple sub-band models to effectively process the audio signal. Within each model, the GSN is encapsulated to effectively capture multi-scale temporal information within the input audio signal. We will dive into details in the following sections.

A. Problem Formulation

In audio signal processing, the signal $x(t)$ captured by a microphone is typically composed of a desired source signal $s(t)$ and a mixture of stationary or non-stationary noises $u(t)$. Audio denoising aims to remove unwanted noises while keeping the source signal. To achieve this, this work first represents the input audio signal in the frequency domain via the Short-Time Fourier Transform (STFT) as follows:

$$X(n,f) = S(n,f) + U(n,f), \tag{1}$$

where $X(n,f)$, $S(n,f)$, and $U(n,f)$ correspond to the complex-valued time-frequency (T-F) bins at discrete time frame n and frequency bin f, with $n = \{1,\dots,N\}$ and $f = \{0,\cdots,F-1\}$. The variables N and F represent the total number of frames and frequency bins, respectively.

B. Gated Spiking Neuron (GSN)

The spiking neuron serves as the fundamental computing unit in an SNN. The frequently used Leaky Integrate-and-Fire (LIF) neuron model, however, struggles to achieve high performance in audio denoising [9]. This is mainly due to the fixed decay factor $\lambda \in \mathbb{R}$ used for every neuron, which restricts their ability to retain multi-scale temporal information that is critical for audio denoising. A recently proposed Parametric LIF (PLIF) [14] replaces the fixed λ with learnable ones, whose values are regulated via a sigmoid function $\sigma(\boldsymbol{\lambda}) \in \mathbb{R}^N$. However, it still falls short as the decay factor remains constant across different time steps. To overcome this limitation, we introduce a gating function to regulate the decay rate at each time step. This allows each neuron to dynamically adjust its membrane potential, strengthening its capability to process temporal tasks. The neuronal dynamics of GSN can be formally expressed as follows:

$$\boldsymbol{i}^l[t] = \boldsymbol{W}_{mn}\boldsymbol{o}^{l-1}[t] + \boldsymbol{W}_{nn}\boldsymbol{o}^l[t-1] + \boldsymbol{b} \tag{2}$$

$$\boldsymbol{\lambda}^l[t] = \sigma(\boldsymbol{W}_{mn}\boldsymbol{o}^{l-1}[t] + \boldsymbol{W}_{nn}\boldsymbol{o}^l[t-1] + \boldsymbol{b}) \tag{3}$$

$$\boldsymbol{u}^l[t] = \boldsymbol{\lambda}^l[t]\boldsymbol{u}^l[t-1] + (1-\boldsymbol{\lambda}^l[t])\boldsymbol{i}^l[t] \tag{4}$$

When the membrane potential surpasses a predefined threshold, an output spike is triggered, followed by a resetting process. To save parameters, we reuse the same weight matrices for calculating $\boldsymbol{\lambda}^l[t]$ as those used in Equation (2). As a result, our proposed GSN model has the same number of parameters as PLIF [14]. The Spiking-FullSubNet further encapsulates the proposed GSN model into an improved FullSubNet architecture that will be introduced in the following subsection.

C. Improved FullSubNet

FullSubNet [11] is a popular audio denoising model that synergistically combines a full-band model and a sub-band model. In FullSubNet, the full-band model capture global spectral information as well as cross-band dependencies, while the sub-band model independently processes each frequency band, focusing on local spectral patterns, reverberation characteristics, and signal stationarity. Experimental evidence supports the effective integration of these two complementary models within a single framework. However, FullSubNet's Achilles' heel lies in the computationally intensive sub-band component, which processes each band at the same frequency granularity. This approach contrasts with the human auditory system which is more sensitive to low-frequency sounds [15], [16]. To address this issue, we introduce a frequency partitioning technique, which applies different processing granularity across the frequency bands, mirroring the human auditory system. Specifically, frequency partitioning allows for tailored processing, with more deep filtering [17] applied to the low-frequency bands and fewer to high-frequency bands. This refinement to the FullSubNet model not only reduces computational demand but also maintains output audio quality, as confirmed in our experiments.

D. Loss Function Optimized with Black-Box Metrics

We employ a blend of loss functions for optimization. First, we use the Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) [18] loss function $\mathcal{L}_{\text{SISDR}}$ to ensure the time domain alignment consistency. Then, we incorporate loss function $\mathcal{L}_{\text{Freq}}$ on complex and magnitude spectrogram for frequency-level optimization. Finally, we include a MetricGAN+ [13] loss $\mathcal{L}_{\text{Gen}}$ to predict the Deep Noise Suppression Mean Opinion Score (DNSMOS) [19], a perceptual metric miming human auditory impressions. The final combined loss function is

$$\begin{aligned}\mathcal{L} = {} & \alpha(100 - \mathcal{L}_{\text{SISDR}}) + \beta\mathcal{L}_{\text{Gen}} \\ & + \underbrace{||\hat{S}(t,f)|^p - |S(t,f)|^p| + |\hat{S}(t,f) - S(t,f)|}_{\mathcal{L}_{\text{Freq}}},\end{aligned} \tag{5}$$

where α and β are hyperparameters that balance the SI-SDR loss, frequency loss, and generator loss. p is the ratio of dynamic range compression.

III. Experimental Settings

A. Datasets

We adopted the Intel N-DNS Challenge dataset [9] for model evaluation. Using the official synthesizer script, we

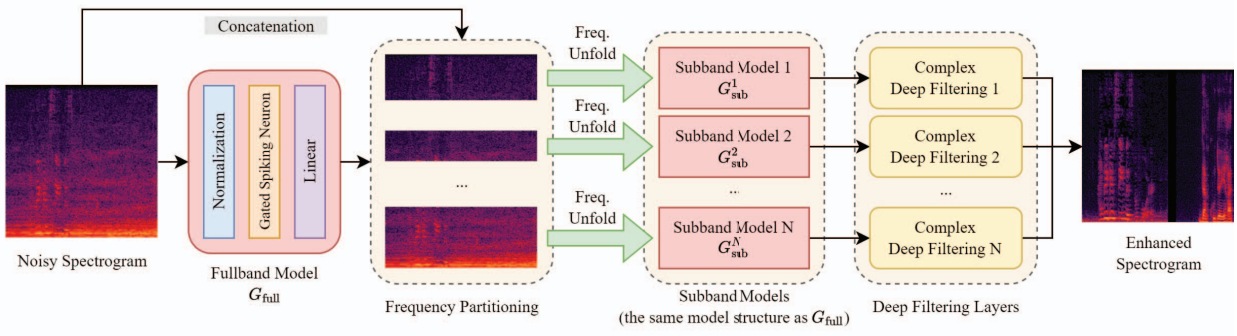

Fig. 1. Diagram of the proposed Spiking-FullSubNet architecture.

TABLE I

CONFIGURATION SETTINGS FOR SPIKING-FULLSUBNET MODELS OF DIFFERENT SIZES. THE TABLE DETAILS THE HIDDEN SIZE OF THE GSN NEURON FOR FULL-BAND ($\mathcal{H}_{\text{FULL}}$) AND SUBBAND ($\mathcal{H}_{\text{SUB}}$) MODELS, THE FREQUENCY CUTOFFS FOR DISCRETE FREQUENCY PARTITIONS (FREQ. CUTOFFS), THE ORDER OF DEEP FILTERING ($\mathcal{N}_{\text{DF}}$), AND THE NUMBER OF CENTER FREQUENCIES (# CENTER FREQS.).

Model Size	$\mathcal{H}_{\text{full}}$	$\mathcal{H}_{\text{sub}}$	Freq. Cutoffs	$\mathcal{N}_{\text{df}}$	# Ctr. Freqs.
Small (s)	240	160	$\{[0, 32), [32, 128), [128, 256]\}$	$[3, 1, 1]$	$[4, 32, 64]$
Middle (m)	320	224	$\{[0, 32), [32, 128), [128, 256]\}$	$[5, 3, 1]$	$[4, 32, 64]$
Large (l)	320	256	$\{[0, 32), [32, 128), [128, 192), [192, 256]\}$	$[5, 3, 1, 1]$	$[2, 4, 32, 64]$

synthesized a 495-hour subset, a 5-hour subset, and a 5-hour subset for training, validation, and testing, respectively. The audio samples, with a sampling rate of 16,000 Hz, were synthesized to maintain a consistent 30-second duration. For audio shorter than 30 seconds, concatenation with other speech signals from the same speaker was performed, with a 0.2-second silence interval inserted between clean speech utterances. The noisy audio was composed of randomly selected speech and noise data with Signal-to-Noise Ratios (SNRs) ranging from -5dB to 20dB. Loudness normalization was applied to each noisy audio sample to simulate agnostic input loudness levels from -35 to -15 decibels relative to full scale (dBFS). To evaluate the audio quality, we employed metrics specified by the Intel N-DNS Challenge, including SI-SDR [18], and Deep Noise Suppression Mean Opinion Score (DNSMOS) [19]. Computational resource usage was also measured, taking into account network latency, power consumption, Power Delay Product (PDP), and model size, as described in the Section VI of the official Intel DNS Challenge paper [9].

B. Implementation Details

We employ the magnitude feature in the frequency domain for both input and output. The STFT is configured with a window length of 512 and a hop length of 128. We utilize AdamW as the optimizer with a learning rate of 1×10^{-3} and set the gradient norm clipping to 10. In the loss function $\mathcal{L}$, the weights are set to $\{\alpha = 0.001, \gamma_2 = 0.05, p = 0.5\}$. We developed three variants of the Spiking-FullSubNet model with different model sizes. They vary in the following aspects: the hidden unit sizes for both the full-band and sub-band model, the granularity of frequency partitioning, the order of deep filtering, and the number of central frequencies. Detailed settings for each variant are presented in Table I. The Spiking-FullSubNet is trained using the backpropagation through time (BPTT), where the gradient of the nondifferentiable spike firing function is replaced with a surrogate one.

IV. RESULTS

We compare the proposed Spiking-FullSubNet against established baselines, including the Microsoft NsNet2 [20], Intel DNS network [9], and the SDNN network [9], which are summarized in Table II. In this table, the noisy row shows the evaluting metrics of the unprocessed noisy audio. The Microsoft NsNet2 serves as the benchmark for the Microsoft DNS 2022 and represents an ANN-based approach, as does the Intel DNS network, which is a proprietary network utilized in Intel's production environments. The latter employs a causal architecture incorporating LSTM and 2D convolution layers. The SDNN network, on the other hand, utilizes a sigma-delta method and is the official baseline for the Intel N-DNS Challenge. The last three rows of Table II present the performance of the proposed Spiking-FullSubNet under different parameter configurations. A key distinction among these configurations lies in the subband processing granularity, with a comprehensive exposition provided in Table I, which outlines the differences in subband processing granularity among the models, providing insight into the underlying factors contributing to the observed performance enhancements.

We evaluated the models using several audio quality metrics, such as DNSMOS scores, SISNR, and the improvement in SI-SNR (SI-SNRi). All networks under study employ STFT encoding and ISTFT decoding, ensuring lossless transformation. This uniformity in encoding and decoding allows for direct comparison of relative performance differences across models in terms of SI-SNR and SI-SNRi, as shown in the last three rows of Table II.

Our results in Table II demonstrate that the SNN-based models, including the SDNN baseline and the proposed

TABLE II
EVALUATION METRICS COMPARISON.

Entry	SI-SNR (dB)	SI-SNRi		DNSMOS			Latency		Power proxy (M-Ops/s)	PDP proxy (M-Ops)	Param count ($\times 10^3$)	Model size (KB)
		data (dB)	enc+dec (dB)	OVR	SIG	BAK	enc+dec (ms)	total (ms)				
Noisy	7.37	-	-	2.44	3.16	2.69	-	-	-	-	-	-
Microsoft NsNet2	11.89	4.26	4.26	2.95	3.27	3.94	0.024	20.024	136.13	2.72	2,681	10,500
Intel DNS Network	12.71	5.09	5.09	3.09	3.35	4.08	0.036	32.036	-	-	1,901	3,802
SDNN baseline	11.85	4.48	4.48	2.69	3.21	3.45	0.036	32.036	14.54	0.44	525	465
Spiking-FullSubNet (Small)	13.89	6.52	6.52	2.97	3.28	3.93	0.03	32.03	29.24	0.94	521	2,084
Spiking-FullSubNet (Middle)	14.71	7.34	7.34	3.05	3.35	3.97	0.03	32.03	53.60	1.72	953	3,816
Spiking-FullSubNet (Large)	14.80	7.43	7.43	3.03	3.33	3.96	0.03	32.03	74.10	2.37	1,289	5,156

Spiking-FullSubNet, are significantly more efficient than the ANN-based solutions, such as the NsNet2 baseline and Intel DNS Network. This efficiency, measured by an order of magnitude, underscores the potential of SNN-based methods for ubiquitous audio denoising tasks. Further examination of the proposed Spiking-FullSubNet, especially the small-sized model variant, reveals a compelling balance between computational and network metrics comparable to that of the SDNN baseline. Moreover, the Spiking-FullSubNet exhibits superior performance in audio quality metrics, significantly surpassing the baseline models. It is worth mentioning that the proposed Spiking-FullSubNet ranked as the top entry for Track 1 (Algorithmic) of the Intel Neuromorphic Deep Noise Suppression Challenge.

V. CONCLUSION

This paper introduces the Spiking-FullSubNet, a groundbreaking SNN-based system tailored for real-time audio denoising tasks. The Spiking-FullSubNet incorporates a novel GSN neuron model capable of capturing multi-scale temporal information, as well as an improved FullSubNet neural architecture that mimics human auditory perception. Our experimental results demonstrate significant improvements in computational efficiency and denoising performance. By leveraging the algorithmic advancements of Spiking-FullSubNet, our work presents a promising solution for a wide range of devices equipped with auditory interfaces.

REFERENCES

[1] K. Roy, A. Jaiswal, and P. Panda, "Towards spike-based machine intelligence with neuromorphic computing," *Nature*, vol. 575, no. 7784, pp. 607–617, 2019.

[2] J. K. Eshraghian, M. Ward, E. O. Neftci, X. Wang, G. Lenz, G. Dwivedi, M. Bennamoun, D. S. Jeong, and W. D. Lu, "Training spiking neural networks using lessons from deep learning," *Proceedings of the IEEE*, 2023.

[3] J. Wu, E. Yılmaz, M. Zhang, H. Li, and K. C. Tan, "Deep spiking neural networks for large vocabulary automatic speech recognition," *Frontiers in neuroscience*, vol. 14, p. 199, 2020.

[4] J. Wu, Y. Chua, M. Zhang, H. Li, and K. C. Tan, "A spiking neural network framework for robust sound classification," *Frontiers in neuroscience*, vol. 12, p. 836, 2018.

[5] 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, pp. 1702–1726, 2017.

[6] I. Fedorov, M. Stamenovic, C. R. Jensen, L.-C. Yang, A. Mandell, Y. Gan, M. Mattina, and P. N. Whatmough, "Tinylstms: Efficient neural speech enhancement for hearing aids," in *Interspeech*, 2020.

[7] H.-S. Choi, S. Park, J. H. Lee, H. Heo, D. Jeon, and K. Lee, "Real-time denoising and dereverberation wtih tiny recurrent u-net," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5789–5793, 2021.

[8] K. Tan and D. Wang, "Towards model compression for deep learning based speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1785–1794, 2021.

[9] J. Timcheck, S. B. Shrestha, D. B. D. Rubin, A. Kupryjanow, G. Orchard, L. Pindor, T. Shea, and M. Davies, "The Intel neuromorphic DNS challenge," *Neuromorphic Computing and Engineering*, vol. 3, no. 3, p. 034005, Aug. 2023, publisher: IOP Publishing.

[10] M. Davies, N. Srinivasa, T.-H. Lin, G. Chinya, Y. Cao, S. H. Choday, G. Dimou, P. Joshi, N. Imam, S. Jain *et al.*, "Loihi: A neuromorphic manycore processor with on-chip learning," *Ieee Micro*, vol. 38, no. 1, pp. 82–99, 2018.

[11] X. Hao, X. Su, R. Horaud, and X. Li, "Fullsubnet: A Full-Band and Sub-Band Fusion Model for Real-Time Single-Channel Speech Enhancement," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2021, pp. 6633–6637, iSSN: 2379-190X.

[12] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2813–2821, 2016.

[13] 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 *International Conference on Machine Learning (ICML)*, 2019.

[14] W. Fang, Z. Yu, Y. Chen, T. Masquelier, T. Huang, and Y. Tian, "Incorporating learnable membrane time constant to enhance learning of spiking neural networks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 2661–2671.

[15] W. A. Yost, R. R. Fay, and A. N. Popper, *Auditory perception of sound sources*. Springer Science & Business Media, 2007, vol. 29.

[16] S. A. Shamma and C. Micheyl, "Behind the scenes of auditory perception," *Current Opinion in Neurobiology*, vol. 20, no. 3, pp. 361–366, 2010, sensory systems.

[17] H. Schroter, A. N. Escalante-B, T. Rosenkranz, and A. Maier, "Deepfilternet: A Low Complexity Speech Enhancement Framework for Full-Band Audio Based On Deep Filtering," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, pp. 7407–7411.

[18] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr – half-baked or well done?" *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 626–630, 2018.

[19] C. K. A. Reddy, V. Gopal, and R. Cutler, "Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6493–6497, 2020.

[20] Y. Xia, S. Braun, C. K. A. Reddy, H. Dubey, R. Cutler, and I. Tashev, "Weighted speech distortion losses for neural-network-based real-time speech enhancement," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 871–875.