The following publication X. Hao, C. Ma, Q. Yang, J. Wu and K. C. Tan, "Toward Ultralow-Power Neuromorphic Speech Enhancement With Spiking-FullSubNet," in IEEE Transactions on Neural Networks and Learning Systems, vol. 36, no. 9, pp. 17350-17364, Sept. 2025 is available at https://doi.org/10.1109/TNNLS.2025.3566021.

Towards Ultra-Low-Power Neuromorphic Speech Enhancement with Spiking-FullSubNet

Xiang Hao*, Chenxiang Ma*, Qu Yang, Jibin Wu, *Member, IEEE,* and Kay Chen Tan, *Fellow, IEEE*

***Abstract*—Speech enhancement is critical for improving speech intelligibility and quality in various audio devices. In recent years, deep learning-based methods have significantly improved speech enhancement performance, but they often come with a high computational cost, which is prohibitive for a large number of edge devices, such as headsets and hearing aids. This work proposes an ultra-low-power speech enhancement system based on the brain-inspired spiking neural network (SNN) called Spiking-FullSubNet. Spiking-FullSubNet follows a full-band and sub-band fusioned approach to effectively capture both global and local spectral information. To enhance the efficiency of computationally expensive sub-band modeling, we introduce a frequency partitioning method inspired by the sensitivity profile of the human peripheral auditory system. Furthermore, we introduce a novel spiking neuron model that can dynamically control the input information integration and forgetting, enhancing the multi-scale temporal processing capability of SNN, which is critical for speech denoising. Experiments conducted on the recent Intel Neuromorphic Deep Noise Suppression (N-DNS) Challenge dataset show that the Spiking-FullSubNet surpasses state-of-the-art methods by large margins in terms of both speech quality and energy efficiency metrics. Notably, our system won the championship of the Intel N-DNS Challenge (Algorithmic Track), opening up a myriad of opportunities for ultra-low-power speech enhancement at the edge. Our source code and model checkpoints are publicly available at github.com/haoxiangsnr/spiking-fullsubnet.**

***Index Terms*—Speech enhancement, spiking neural network, neuromorphic computing, neuromorphic speech processing**

I. Introduction

MICROPHONES inevitably pick up various interferences from the surrounding environments, such as ambient noise, which can drastically degrade the quality of perceived speech signals. This prevalent issue calls for speech enhancement (SE) techniques in real-world applications like headsets, hands-free communication, teleconferencing, and hearing aids [1], [2], etc. Furthermore, SE can also benefit a variety of downstream tasks, such as automatic speech recognition [3], speaker recognition/verification [4], and speech diarization [5]. It serves as a crucial front-end processing step to improve the robustness of these systems against signal degradation.

Early SE methods, such as spectral subtraction [6] and Wiener filtering [2], were developed to improve speech signals. However, these methods often face challenges in real-world conditions, particularly when dealing with low signal-to-noise ratio (SNR) scenarios [7], [8]. During the last decade, deep learning techniques, such as Long Short-Term Memory (LSTM) [9]–[11], Convolutional Neural Network (CNN) [12], [13], and Transformer [14], have improved the SE performance by leaps and bounds [7], [8]. While these deep learning-based SE models offer superior performance, their high computational cost and latency can be prohibitive for deployment on ubiquitous resource-constrained edge devices, such as headsets and hearing aids [9], [15].

Recently, brain-inspired Spiking Neural Networks (SNNs) have received increasing attention as energy-efficient alternatives to the prevailing deep learning models [16]–[23]. SNNs employ spike trains to encode and convey information, closely mimicking the operations of biological neural networks [16]. This spike-based communication leads to asynchronous, event-driven computation, where information processing is triggered solely by the arrival of incoming spikes. Furthermore, spiking neurons exhibit rich neuronal dynamics that are believed to be crucial for information processing in the brain [24]. This spike-based communication, coupled with the inherent temporal dynamics of spiking neurons, enables efficient and effective temporal signal processing using SNNs. Notably, when deployed on emerging neuromorphic chips, SNN models can yield orders of magnitude improvements in energy efficiency and reduced latency compared to conventional artificial neural networks (ANNs) used in deep learning [25]–[27].

The superior energy efficiency, low latency, and ability to effectively process temporal signals position SNNs as a compelling approach for performing SE on resource-constrained devices. This potential has been notably recognized in the recent Intel Neuromorphic Deep Noise Suppression (N-DNS) Challenge [28]. However, the development of high-performance SNNs for the SE task currently faces several key challenges. One primary challenge arises from the inherent complexity of speech signals, which exhibit temporal variations across multiple time scales. Current SNN models, commonly equipped with simplified spiking neurons like the Leaky Integrate-and-Fire (LIF) model, often face difficulties in effectively handling the high temporal complexity present in speech signals [29]. Moreover, achieving state-of-the-art (SOTA) performance on a par with or surpassing conventional

This work was supported by the National Natural Science Foundation of China (Grant No. 62306259), Research Grants Council of the Hong Kong SAR (Grant No. C5052-23G, PolyU25216423, PolyU15217424, PolyU15218622, PolyU15215623, and PolyU15229824), and The Hong Kong Polytechnic University (Project IDs: P0043563, P0046094, P0053699, P0052694, P0053758).

*Xiang Hao and Chenxiang Ma contributed equally to this work. Corresponding Author: Jibin Wu (jibin.wu@polyu.edu.hk).

Xiang Hao, Chenxiang Ma, Jibin Wu and Kay Chen Tan are with Department of Data Science and Artificial Intelligence, The Hong Kong Polytechnic University, Hong Kong SAR. Jibin Wu is also with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong SAR. Jibin Wu and Kay Chen Tan are also with Research Center of Data Science and Artificial Intelligence, The Hong Kong Polytechnic University, Hong Kong SAR.

Qu Yang is with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore.

ANN systems while ensuring real-time audio streaming, as required in practical applications, demands a holistic system design. Such a design necessitates not only enhanced spiking neuron models, but also the development of effective enhancement workflows and training techniques tailored to the unique characteristics of SNNs. To the best of our knowledge, no existing study has effectively addressed these challenges.

In this work, we introduce a real-time neuromorphic SE system that demonstrates competitive denoising performance while exhibiting significantly improved energy efficiency compared to ANN methods. First of all, drawing inspiration from recent advancements in deep-learning-based SE research works [10], [13], [14], [30], we propose a novel SNN-based SE model named *Spiking-FullSubNet*. This model effectively integrates both full-band and sub-band information, allowing it to capture both global and local spectral characteristics. Specifically, the full-band component receives inputs covering the entire frequency partition, enabling it to capture the global spectral structure of the speech signal. On the other hand, each sub-band component focuses exclusively on a specific frequency band, facilitating effective modeling of local spectral structures. Moreover, to enhance the computation efficiency of the sub-band components, we introduce a brain-inspired frequency band partitioning method. Specifically, motivated by the frequency sensitivity profile of the peripheral auditory system, we employ a finer granularity for low-frequency bands and a coarser granularity for high-frequency bands.

Furthermore, we propose a novel spiking neuron model called Gated Spiking Neuron (GSN) to enhance the temporal signal processing capability of spiking neurons. Unlike existing spiking neuron models where input information decays exponentially over time, the GSN model dynamically controls the integration of input information and the forgetting of historical information. This unique feature enables the GSN to identify and retain crucial temporal information that is essential for speech enhancement. In summary, our main contributions are threefold:

- We propose *Spiking-FullSubNet*, a novel real-time neuromorphic SE model that combines recent advancements in speech enhancement and neuromorphic computing. This cross-disciplinary approach significantly enhances both speech enhancement performance and energy efficiency.
- We propose a novel spiking neuron model called GSN. Unlike existing spiking neuron models that passively filter input information, the GSN model dynamically controls the input information integration and forgetting, thereby facilitating effective temporal information processing that is critical for speech enhancement.
- We conducted comprehensive experiments on the Intel N-DNS Challenge dataset to evaluate the proposed neuromorphic SE model. Our model not only demonstrates exceptional speech enhancement capabilities, but also showcases a remarkable improvement in energy efficiency, surpassing SOTA ANN models by almost three orders of magnitude.

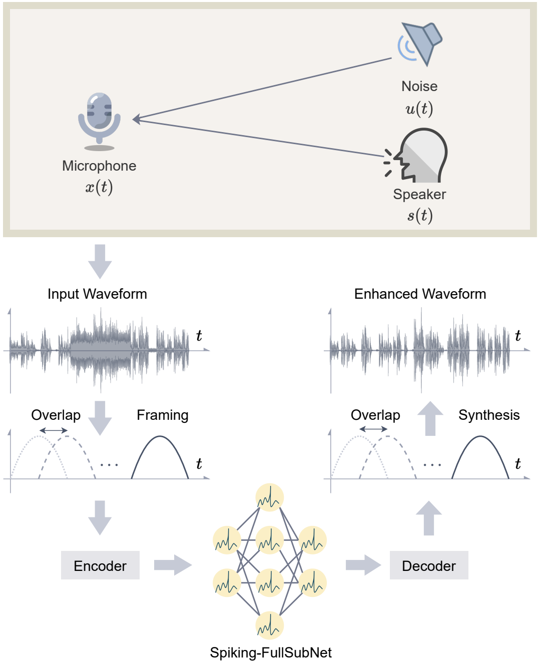


Fig. 1: Block diagram of the proposed real-time neuromorphic speech enhancement system.

II. Related Works

A. Spiking Neuron Model

Among various spiking neuron models introduced in neuroscience [31], [32], the Leaky Integrate-and-Fire (LIF) model [29] is the most commonly employed one for constructing large-scale neuromorphic computing systems. Despite their simplicity, LIF neurons have limitations in handling temporal signals with complex structures due to their oversimplified neuronal dynamics. To overcome this limitation, several new spiking neuron models have been introduced recently. For example, the parametric LIF (PLIF) model [33] incorporates learnable time constants, allowing historical information to decay at different rates. The adaptive LIF (ALIF) model [34] adjusts the firing threshold after each spike, effectively achieving firing rate adaptation. This endows spiking neurons with context-dependent processing capability. In a similar vein of research, the GLIF model [35] selectively regulates input currents integration, membrane decaying factors, and neuronal state resetting to enrich neuronal dynamics. These regulations are achieved through trainable parameters along the temporal dimension. While the GLIF model exhibits flexibility in temporal processing, it encounters difficulties when dealing with signals of varying time lengths. This issue is particularly relevant to the speech enhancement task addressed in this paper, which involves handling signals with variable durations.

In addition to incorporating adaptive state variables into spiking neurons to enhance their temporal processing capability, recent studies have introduced multi-compartment neuron models. These models aim to enrich neuronal dynamics by considering the complex morphology of biological neurons. For instance, the TC-LIF model [36] incorporates two compartments representing soma and dendrites, enabling the simulation of interactive dynamics between these distinct neuronal structures. Building upon this concept, PMSN model [37] incorporates multiple compartments for multi-scale temporal information processing. However, despite these advancements, existing spiking neuron models still face challenges in effectively capturing and retaining essential temporal information, limiting their ability to process signals with complex temporal structures.

B. Speech Enhancement

Speech enhancement aims to improve speech intelligibility and quality [2], [6], which needs to remove noise from noisy signals captured by microphones as shown in Fig. 1. Existing methods can be divided into time-domain and frequency-domain approaches. Time-domain methods [38], [39] directly estimate the clean speech signal, bypassing spectral analysis and waveform synthesis, while frequency-domain methods [7], [8], [10]–[12], [40] estimate the spectrogram of the clean speech and then convert it back to the time-domain signal. Between these two approaches, frequency-domain methods have received significant research attention, primarily due to the sparse nature of speech in the frequency domain. Specifically, they can be broadly categorized into spectral magnitude-only enhancement and complex spectrum enhancement. Spectral magnitude-only methods [7], [8] focus on estimating the magnitude of the clean speech spectrum, utilizing the noisy phase for reconstructing the time-domain signal. On the other hand, complex spectrum enhancement methods [10]–[12], [40] estimate both the real and imaginary parts of the complex spectrum, which have exhibited greater potential in enhancing speech quality by leveraging the full spectral information.

C. Sub-band Modeling in Speech Enhancement

In recent years, there has been a notable shift in research focus towards the utilization of sub-band modeling in both single-channel and multi-channel SE [10], [11], [13], [14], [30]. In contrast to traditional full-band modeling, sub-band modeling involves the separation of input audio into multiple frequency bands, which are then processed independently. Specifically, each sub-band model takes in a noisy sub-band signal along with its adjacent frequency bands, and then predicts the corresponding clean sub-band signal. This method leverages the distinct stationary characteristics of speech and noise. Speech signals are inherently non-stationary, exhibiting dynamic and variable properties over time. Conversely, many types of noise are relatively stationary, meaning their statistical properties remain more consistent and stable [41], [42]. In addition, sub-band modeling focuses on the local spectral pattern presented in the current and neighboring frequencies, which has been proven informative for discriminating between speech and other signals [43], [44]. Furthermore, the sub-band models are also effective in modeling the reverberation as the room's reverberation time (RT60) is frequency-dependent [45].

However, the sub-band modeling approach comes with a trade-off. While it leverages the distinct stationary characteristics of speech and noise, it can also result in the loss of the global spectral structure of the speech signal. This global spectral information is also crucial for effective SE. To address this issue, recent works have proposed a full-band and sub-band fusion modeling approach [10], [11], [13], [14], [30]. In these works, a full-band model and several sub-band models are combined, allowing them to complement each other and capture both the local and global spectral information [13]. Though these fusion-based methods have led to significant improvements, sub-band modeling can still be computationally costly, as it requires processing each frequency band separately. This poses challenges for real-time edge applications. In this work, we propose a novel approach that applies different granularity levels to various frequency sub-bands, significantly reducing the computational cost while maintaining the speech enhancement performance.

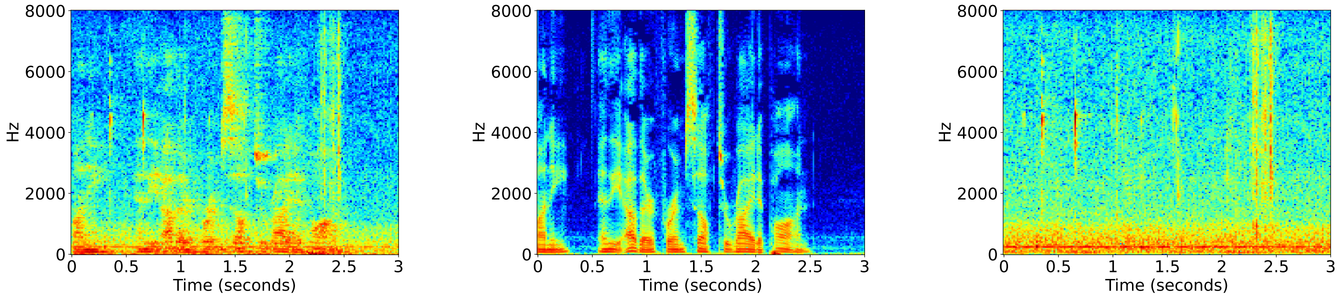
D. Neuromorphic Speech Processing

SNNs have recently emerged as a promising approach for power-efficient speech processing. Compared to traditional ANNs, SNNs can offer significant advantages in terms of reduced computational complexity. An early work [46] employs Izhikevich neurons [31] as feature extractors, which are trained using unsupervised spiking-timing-dependent plasticity (STDP) [47] to recognize spoken digits. Subsequent works [48], [49] introduce a SOM-SNN framework that incorporates a self-organizing map (SOM) for feature representation, followed by an SNN for pattern classification. Recent studies leverage deeper SNNs and more advanced learning rules to enhance classification performance. For instance, deep convolutional SNNs coupled with the tandem learning rule have achieved significantly improved performance on keyword spotting tasks [50], [51]. Recurrent networks of spiking neurons (RSNNs) are also exploited for speech recognition [52], [53], holding enhanced memory capacity and bringing improvements over the feedforward counterparts. These earlier studies focus on small vocabulary speech recognition tasks. More recently, Wu et al. [54] apply deep SNNs to large vocabulary continuous speech recognition tasks and demonstrate competitive accuracy compared to ANN-based systems. Notably, a recent benchmark study [55] compared the performance of neuromorphic keyword spotting systems to ANN-based systems that deployed on conventional computing hardware. The study evaluated metrics such as inference speed, energy cost per inference, and dynamic energy consumption. The findings from this benchmark study suggest that neuromorphic systems can significantly reduce the energy costs per inference while maintaining equivalent inference accuracy compared to their traditional ANN counterparts. However, there is a lack of study of neuromorphic technologies for the speech denoising task.

III. Background

A. Spiking Neuron Model

The most commonly used spiking neuron model is the leaky integrated-and-fire (LIF) neuron [29]. This model is favorable in terms of computational complexity and analytical tractability. The LIF neuron maintains an internal state known



(a) Spectrogram of Noisy Speech Signal

(b) Spectrogram of Clean Speech Signal

(c) Spectrogram of Noise Signal

Fig. 2: Time-frequency magnitude spectrogram of different signals. The speech enhancement methods aim at recovering a clean speech from a noisy observation by removing the unwanted noise.

as the membrane potential, which decays over time at a rate determined by the time constant τ. Meanwhile, the neuron integrates the input current. When the membrane potential surpasses the predefined threshold ϑ, an output spike is generated and transmitted to downstream neurons. This is then followed by a resetting process, where the membrane potential is reset to a specific value. Such neuronal dynamics can be described by the following discrete-time formulation:

$$\boldsymbol{i}^l(t) = \boldsymbol{W}^l_{mn}\boldsymbol{o}^{l-1}(t) + \boldsymbol{W}^l_{nn}\boldsymbol{o}^l(t-1) + \boldsymbol{b}^l \tag{1}$$

$$\boldsymbol{u}^l(t) = \lambda\boldsymbol{u}^l(t-1) + \boldsymbol{i}^l(t) \tag{2}$$

$$o^l_i(t) = \Theta(u^l_i(t) - \vartheta) = \begin{cases} 1 & u^l_i(t) >= \vartheta \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$$\boldsymbol{u}^l(t) = \boldsymbol{u}^l(t) - \vartheta\boldsymbol{o}^l(t) \tag{4}$$

where $\boldsymbol{W}^l_{mn}$ and $\boldsymbol{W}^l_{nn}$ denote the feed-forward and recurrent weight matrices at the l^{th} layer, respectively, $\lambda = \exp(-1/\tau)$ controls the decay rate of the membrane potential $\boldsymbol{u}^l(t)$, $\boldsymbol{i}^l(t)$ represents the input current, and $\boldsymbol{b}^l$ is the bias term.

B. Formulation of Speech Enhancement

In a typical acoustic environment, as depicted in Fig. 1, the input to a speech enhancement system is a time-varying waveform that undergoes continuous sampling and quantization. The signal captured by the microphone can be represented as

$$x(t) = s(t) + u(t) \quad t = 1, 2, \ldots, T, \tag{5}$$

where $s(t)$ represents the clean speech signal and $u(t)$ denotes the noise. Due to the complexity of directly enhancing the time-domain signal, which can be highly non-stationary, many speech enhancement methods focus on working in the frequency domain, where signals exhibit more stable properties [10], [12], [40], [56], [57]. To process the signal in the frequency domain, the noisy signal is converted into its Short-Time Fourier Transform (STFT) representation, which can be expressed as:

$$x(n, f) = s(n, f) + u(n, f), \tag{6}$$

where $n = 1, 2, \ldots, N$ denotes the time-frame index and $f = 1, 2, \ldots, F$ represents the frequency-bin index. $x(n, f)$ is the noisy spectrogram, with its time-frequency (TF) magnitude spectrogram shown in Fig. 2(a). The clean reference is $s(n, f)$, as depicted in Fig. 2(b). The noise reference is $u(n, f)$, as depicted in Fig. 2(c).”

The resulting noisy TF magnitude spectrogram is the input feature used by the model. In recent years, mask-based speech enhancement is a widely used technique in the frequency domain [58]–[60]. The core idea is to estimate a mask that will highlight the speech components and suppress the noise components. In these methods, the model predicts a time-frequency mask, $\hat{m}(n, f)$, that is applied to the noisy spectrogram to produce the enhanced speech signal. The enhanced spectrogram, $\hat{s}(n, f)$, is then computed as:

$$\hat{s}(n, f) = \hat{m}(n, f) \odot x(n, f). \tag{7}$$

During training, the model learns to predict the optimal mask that separates the speech signal from the noise. The loss function is typically used between the estimated enhanced speech $\hat{s}(n, f)$ and the clean speech $s(n, f)$, is given by $\mathcal{L}(\hat{s}, s)$. This objective encourages the model to estimate a mask that reduces noise while preserving the speech content.

IV. Method

In this section, we elaborate on the proposed Spiking-FullSubNet model. We start with an overview of the model. We then introduce a novel spiking neuron model called GSN, specifically designed for speech processing. Next, we present the approach of full-band and sub-band fusion for speech denoising that we have adopted in this work. Lastly, we introduce the model training details.

A. Overview of Spiking-FullSubNet

As shown in Fig. 3, the Spiking-FullSubNet architecture is composed of a full-band model and multiple sub-band models, with the proposed GSN as the core component of each model, to effectively enhance noisy speech signals. Initially, noisy magnitude spectrogram frames are directly input into the full-band model, which is designed to capture global spectral patterns. The first GSN layer functions as an encoding layer, converting the real-valued input into spike trains [21], [22], [33], [54]. The output of the full-band model, combined with frequency bins from the noisy spectrogram, is then processed

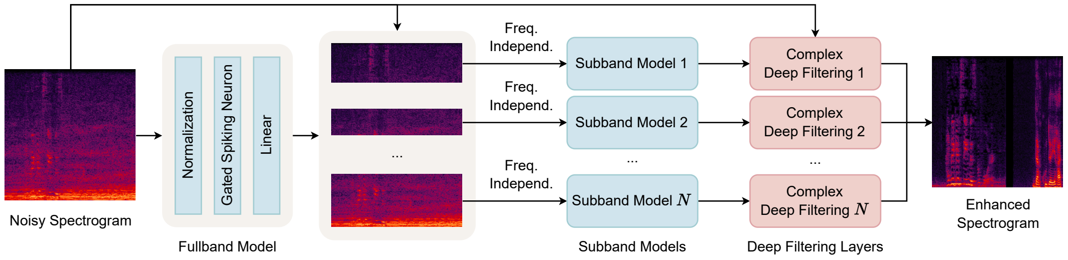


Fig. 3: Diagram of the proposed Spiking-FullSubNet architecture. The architecture integrates a full-band model and sub-band models, with gated spiking neurons serving as the core of each model, to effectively enhance noisy speech signals. The full-band model operates on the noisy magnitude spectrogram to capture global spectral patterns, while the sub-band components focus on specific frequency bands to effectively model local spectral information. By incorporating newly proposed GSNs into both the full-band and sub-band models, the temporal processing capability is greatly improved. Finally, deep filtering is employed as the training target to obtain the enhanced spectrogram.

by several sub-band models to capture local spectral characteristics. The outputs of the sub-band models are passed through linear layers, which generate both the real and imaginary parts of complex values. Finally, these complex values are utilized as an auditory mask for each time-frequency bin (i.e., multi-frame deep filtering (MFDF)) to effectively separates speech from noise. By applying the estimated mask to the noisy spectrogram, we obtain the enhanced speech signal.

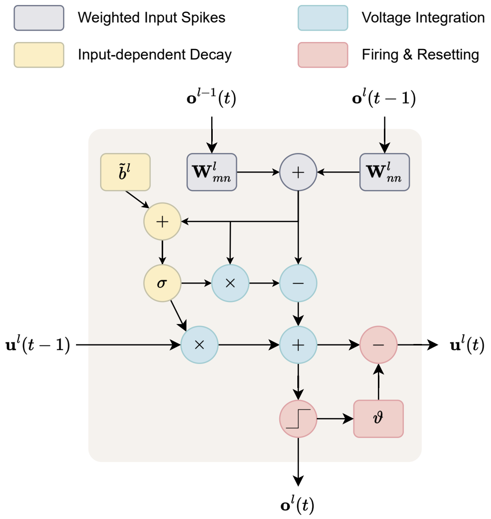


Fig. 4: Illustration of the proposed GSN model, which regulates the membrane decay rate at each time step based on the feedforward and recurrent inputs.

B. Gated Spiking Neuron

For speech enhancement, the objective is to remove unwanted noises from audio recordings. These noises can originate from various sources, such as background sounds, microphone interference, or transmission distortions, and their signal intensity varies over time. Due to its constant membrane decay rate, the commonly-used LIF model [29] cannot handle these dynamic changes. In these scenarios, the LIF neuron encounters two challenges. It might struggle to filter out noise during periods of high noise intensity effectively, or it may excessively filter informative audio signals during periods of low noise intensity. These challenges occur as the fixed decay factor either cannot provide sufficient attenuation of the noisy signal, resulting in inadequate enhancement or decays the neuron's potential too quickly, leading to a loss of desired audio contents.

A straightforward solution is to adopt different decay factors at each time step, allowing flexible adaptation to the changing noise levels in the input audio signals. However, this would introduce a huge number of parameters, especially for long time duration. Moreover, it would not work well for audio signals with a variable time duration, which are commonly encountered in SE tasks [15]. To address this issue, we propose a novel spiking neuron model called GSN that regulates the decay rate at each time step in an input-dependent manner. The input-dependent decay in our GSN is implemented by modeling the decay as a function of both feed-forward and recurrent input spikes, as shown in Fig. 4. This adaptive decay mechanism allows GSN to effectively handle noise variations. A sigmoid function $\sigma(\cdot)$ is applied to constrain the decay rate within the range of 0 to 1. The corresponding membrane potential dynamics are formally expressed as:

$$\boldsymbol{i}^l(t) = \boldsymbol{W}^l_{mn}\boldsymbol{o}^{l-1}(t) + \boldsymbol{W}^l_{nn}\boldsymbol{o}^l(t-1) + \boldsymbol{b}^l, \tag{8}$$

$$\boldsymbol{\lambda}^l(t) = \sigma(\boldsymbol{W}^l_{mn}\boldsymbol{o}^{l-1}(t) + \boldsymbol{W}^l_{nn}\boldsymbol{o}^l(t-1) + \widetilde{\boldsymbol{b}}^l), \tag{9}$$

$$\boldsymbol{u}^l(t) = \boldsymbol{\lambda}^l(t)\boldsymbol{u}^l(t-1) + (1-\boldsymbol{\lambda}^l(t))\boldsymbol{i}^l(t). \tag{10}$$

In order to save model parameters and reduce overall computation, we employ weight-sharing by using the same weight matrices, denoted as $\boldsymbol{W}^l_{mn}$ and $\boldsymbol{W}^l_{nn}$, in Eqs. (8) and (9). Notably, the GSN model can adaptively modulate the neuron's membrane potential along the temporal dimension while avoiding the need for a large number of parameters associated with the total time steps. Additionally, GSN bears a resemblance to the forget gate widely used as a critical component of the LSTM architecture [61]. However, this temporal gating mechanism has been underexplored in existing spiking neuron models.

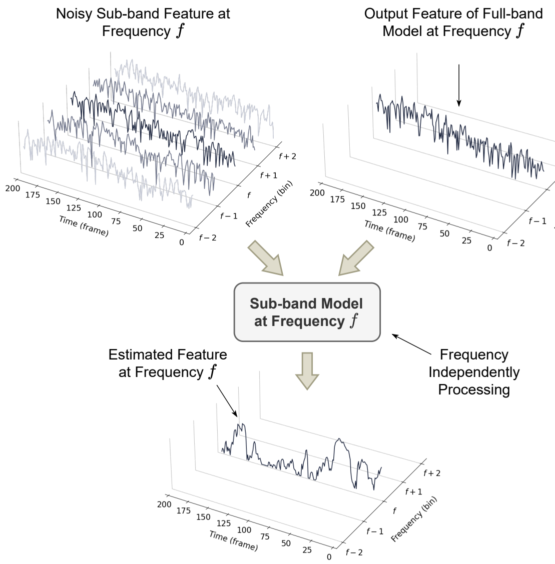

Noisy Sub-band Feature at Frequency f
Output Feature of Full-band Model at Frequency f
Sub-band Model at Frequency f
Estimated Feature at Frequency f
Frequency Independently Processing
Time (frame)
Frequency (bin)
200
175
150
125
100
75
50
25
0
$f-2$
$f-1$
f
$f+1$
$f+2$

Fig. 5: Illustration of the sub-band processing in Spiking-FullSubNet. The input is a vector $\mathbf{x}_f(n)$ comprising the magnitude bin of frequency f, its 4 neighboring frequency bins, and the corresponding bin in spectral embedding output from the full-band model.

C. Spiking-FullSubNet Architecture

Based on the GSN model introduced in Section IV-B, we further develop the Spiking-FullSubNet that can achieve high-performance real-time speech enhancement. Spiking-FullSubNet exploits full-band and sub-band modeling to capture both global (long-distance cross-band dependencies) and local (signal stationarity differences) spectral patterns, respectively. Notably, we propose to employ varying processing granularity for different frequency partitions, mimicking human auditory perception, to enhance the model's processing efficiency.

1) Full-Band Processing: The full-band model operates on magnitude spectral features extracted from the noisy speech signal. The input feature vector $\mathbf{x}(n)$ on audio frame n is given by

$$\mathbf{x}(n) = [|x(n,1)|, |x(n,2)|, \dots, |x(n,F)|]^\top \in \mathbb{R}^F, \tag{11}$$

where F denotes the total number of frequency bins, $x(n,f)$ represents the complex Fourier coefficient of the frame n and frequency bin f, and $|\cdot|$ denotes extracting the magnitude of the complex Fourier coefficient. The sequence of feature vectors across all frames is denoted by $\mathbf{X}$:

$$\mathbf{X} = [\mathbf{x}(1), \mathbf{x}(2), \dots, \mathbf{x}(T)] \in \mathbb{R}^{T\times F}, \tag{12}$$

where T is the total number of discrete time frames. We stack GSNs to process $\mathbf{X}$ by capturing both the global spectral content and the interactions between frequency bins, yielding a spectral embedding $\mathbf{E}$ of the same dimensions as $\mathbf{X}$, i.e., $\mathbf{E} \in \mathbb{R}^{T\times F}$.

2) Existing Sub-Band Processing Approach: The sub-band models process each frequency band independently, focusing on the stationarity differences between the speech and noise, local spectral patterns, and reverberation characteristics [10], [13]. As shown in Fig. 5, for a given time n at frequency f, the input to the sub-band model is a vector $\mathbf{x}_f(n)$ comprising the noisy magnitude bin at frequency f, its $2\times N$ neighboring frequency bins, and the corresponding bin in spectral embedding output by the full-band model:

$$\begin{aligned}\mathbf{x}_f(n) =& [\underbrace{|x(n,f-N)|, \dots, |x(n,f-1)|}_{\text{Lower neighboring frequencies}}, \\ & |x(n,f)|, \mathbf{E}(n,f), \\ & \underbrace{|x(n,f+1)|, \dots, |x(n,f+N)|}_{\text{Higher neighboring frequencies}}]^\top.\end{aligned} \tag{13}$$

This frequency-independent processing is inspired by the conventional signal processing-based noise reduction algorithms, such as noise density estimation and wiener filter [2]. It allows for a detailed analysis of the signal's local spectral pattern and stationarity. Experimental evidence supports the effective integration of a full-band model and a sub-band model within a single framework [10]. However, the computational challenge for the existing full-band and sub-band modeling methods lies in their sub-band part, which processes each sub-band at the same granularity. This approach contrasts with the human auditory system, which is more sensitive to low-frequency sound and less sensitive to high-frequency sound [62]. Motivated by human auditory perception, we introduce a frequency partitioning technique that applies different processing granularity across the frequency bands to address this issue.

3) Sub-Band Processing Based on Frequency Partitioning: To account for the varying perceptual importance of different frequency bins, the input magnitude spectral is divided into K non-overlapping frequency partitions:

$$\mathcal{P} = \{\mathcal{P}_1, \mathcal{P}_2, \dots, \mathcal{P}_K\} \quad \text{where} \quad \sum_{k=1}^{K} |\mathcal{P}_k| = F, \tag{14}$$

where $|\cdot|$ is the size of the frequency partition set. Each frequency partition $\mathcal{P}_k$ is defined by a set of contiguous discrete frequency bins, thereby covering the entire frequency spectrum. The lowest frequency partition $\mathcal{P}_1$ includes the indices of the discrete frequency bins $\{1, \dots, f_{c_1}\}$, where f_{c_1} is the cutoff frequency for this first partition. Each subsequent partition $\mathcal{P}_k$ for $k = 2, \dots, K-1$ starts from the cutoff frequency of the previous partition $f_{c_{k-1}} + 1$ and extends up to its own cutoff frequency f_{c_k}. The final partition $\mathcal{P}_K$ contains frequencies from $f_{c_{K-1}} + 1$ to the upper limit F. This partitioning can be formally expressed as

$$\begin{aligned}&\mathcal{P}_1 = \{1, \dots, f_{c_1}\}, \\ &\mathcal{P}_k = \{f_{c_{k-1}} + 1, \dots, f_{c_k}\} \quad \text{for} \quad k = 2, \dots, K-1, \\ &\mathcal{P}_K = \{f_{c_{K-1}} + 1, \dots, F\}.\end{aligned} \tag{15}$$

For each frequency partition $\mathcal{P}_k$, the Spiking-FullSubNet model adjusts the processing granularity by setting a grouping parameter g_k that dictates how many discrete frequency bins to be processed jointly in the sub-band model. Specifically, for each frequency partition $\mathcal{P}_k$, we process groups of $g_k + 1$ adjacent discrete frequency bins, along with a context window of N bins on either side, and the corresponding embedding vector output by the full-band model:

$$\begin{aligned}\mathbf{x}_f^k(n) =&[\underbrace{|x(n,f-N)|,\ldots,|x(n,f-1)|}_{\text{Lower neighboring frequencies}},\\&|x(n,f)|,|x(n,f+1)|,\ldots,|x(n,f+g_k)|,\\&\mathbf{E}(n,f),\mathbf{E}(n,f+1),\ldots,\mathbf{E}(n,f+g_k),\\&\underbrace{|x(n,f+g_k+1)|,\ldots,|x(n,f+g_k+N)|}_{\text{Higher neighboring frequencies}}]^\top,\end{aligned} \tag{16}$$

where $\mathbf{x}_f^k(n)$ is the grouped feature vector for the k^{th} frequency partition at time frame n, and f is the starting frequency bin of the group within the partition $\mathcal{P}_k$. The value of f ranges from the lower bound of $\mathcal{P}_k$ to an upper bound, ensuring the group of g_k+1 bins is within the partition. The grouping parameter g_k allows for a flexible adjustment of the sub-band resolution and computational complexity within each sub-band. For lower frequency partitions, where finer sub-band resolution is often more important for speech perception, g_k may be set to a smaller value. Conversely, for higher frequency intervals, g_k may be larger, reflecting the reduced perceptual importance of spectral details and allowing for reduced computational cost. This frequency partition processing strategy enables the Spiking-FullSubNet model to efficiently handle the spectral information with varying resolution across the frequency spectrum, which is more aligned with the non-uniform frequency resolution of human hearing.

D. Learning Target

Conventional speech enhancement methods work within the Short-Time Fourier Transform (STFT) domain [7], [8]. The typical methodology involves the estimation of time-frequency (TF) masks, e.g., Ideal Binary Mask (IBM) [58], Ideal Ratio Mask (IRM) [58], and Complex Ideal Ratio Mask (cIRM) [63] via neural networks, which are then applied element-wise to the complex STFT of the noisy speech mixture to extract the enhanced signal. However, the performance of this methodology will degrade if the frequency resolution gets too low [64]. Recently, multi-frame deep filtering (MFDF) [64] in the frequency domain has been proposed, where a filter is applied to multiple adjacent TF bins, enabling recovery of degraded signals like notch-filters or time-frame zeroing.

The proposed Spiking-FullSubNet estimates a complex-valued MFDF filter for each TF bin within the STFT domain. We set the filter length (order) d_k for input within the k^{th} frequency partition and define the noisy multi-frame vector of the $\mathbf{x}_f^k(n)$ for the k^{th} frequency partition as

$$\overline{\mathbf{x}}_f^k(n) = \begin{bmatrix} x(n,f) & x(n-1,f) & \ldots & x(n-d_k,f)\\ x(n,f+1) & x(n-1,f+1) & \ldots & x(n-d_k,f+1)\\ \vdots & \vdots & \ddots & \vdots\\ x(n,f+g_k) & x(n-1,f+g_k) & \ldots & x(n-d_k,f+g_k)\end{bmatrix}, \tag{17}$$

which includes only the streaming historical frames, thereby avoiding the introduction of future latency. For the noisy multi-frame input $\mathbf{x}_f^k(n)$, the output complex-valued MFDF filter is

$$\overline{\mathbf{w}}_f^k(n) = \begin{bmatrix} w_0(f) & w_1(f) & \ldots & w_{d_k}(f)\\ w_0(f+1) & w_1(f+1) & \ldots & w_{d_k}(f+1)\\ \vdots & \vdots & \ddots & \vdots\\ w_0(f+g_k) & w_1(f+g_k) & \ldots & w_{d_k}(f+g_k)\end{bmatrix}. \tag{18}$$

The application of the MFDF is then expressed as

$$\begin{aligned}\hat{\mathbf{s}}_f^k(n) &= \sum_j \left(\overline{\mathbf{w}}_f^k(n)\odot\overline{\mathbf{x}}_f^k(n)\right)_{ij}\\ &= [\hat{s}(n,f),\hat{s}(n,f+1),\ldots\hat{s}(n,f+g_k)]^\top,\end{aligned} \tag{19}$$

where $\odot$ denotes the complex-valued hadamard product, $\hat{\mathbf{s}}_f^k(n)$ is the enhanced counterpart, and $\sum_j(\cdot)_{ij}$ means the sum by rows. Our Spiking-FullSubNet model accommodates variable deep filtering orders across different frequency partitions. Lower frequencies, which exhibit stronger temporal correlations, are processed with a higher deep filtering order, allowing the network to capture more complex temporal structures. In contrast, higher frequencies characterized by weaker temporal correlations are processed with a lower deep filtering order. This approach promotes computational efficiency without sacrificing the resolution of critical spectral details, thus maintaining the integrity of the speech signal.

E. Loss Function and SNN training

Inspired by Braun *et al.* [65], we use a linear combination of magnitude loss $\mathcal{L}_{\text{Mag.}}$ and complex loss $\mathcal{L}_{\text{RI}}$ in the TF-domain:

$$\begin{aligned}\mathcal{L}_{\text{TF}} =&\alpha\,\mathcal{L}_{\text{Mag.}} + (1-\alpha)\,\mathcal{L}_{\text{RI}},\\ \mathcal{L}_{\text{Mag.}} =&\mathbb{E}_{s,\hat{s}}\left[\|s(n,f)-\hat{s}(n,f)\|^2\right],\\ \mathcal{L}_{\text{RI}} =&\mathbb{E}_{s_r,\hat{s}_r}\left[\|s_r(n,f)-\hat{s}_r(n,f)\|^2\right]\\ &+\mathbb{E}_{s_i,\hat{s}_i}\left[\|s_i(n,f)-\hat{s}_i(n,f)\|^2\right],\end{aligned} \tag{20}$$

where α is a hyperparameter to balance the contribution from two losses. Moreover, an additional penalization on the waveform features is proven to help improve the speech quality [66]. In our model, we use the Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) [67] loss function, which is computed directly in the time domain and forces the model to learn how to precisely estimate the magnitude and the phase of the target speech signals:

$$\mathcal{L}_{\text{SI-SDR}} = -10\log_{10}\left(\frac{\|\frac{\hat{\mathbf{s}}^T\mathbf{s}}{\|\mathbf{s}\|^2}\mathbf{s}\|^2}{\|\frac{\hat{\mathbf{s}}^T\mathbf{s}}{\|\mathbf{s}\|^2}\mathbf{s}-\hat{\mathbf{s}}\|^2}\right). \tag{21}$$

The final loss function is formulated as follows:

$$\mathcal{L} = \gamma_1\,\mathcal{L}_{\text{TF}} + \gamma_2\,(100-\mathcal{L}_{\text{SI-SDR}}), \tag{22}$$

where γ_1 and γ_2 are the weights of the corresponding losses.

With the established loss function, the Spiking-FullSubNet model can thus be trained end-to-end using backpropagation through time (BPTT) [68]. However, BPTT cannot be directly used since the gradient of the spike generation function is zero almost everywhere except for the firing threshold, where it is infinity. To address this issue, we adopt a surrogate gradient to circumvent the non-differentiability issue [69], [70], formulated by

$$\frac{\partial o_i^l(t)}{\partial u_i^l(t)} = \max\left(0,\ 1 - |u_i^l(t) - \vartheta|\right). \tag{23}$$

V. Experimental Setup

In this section, we evaluate the proposed Spiking-FullSubNet on the publicly available speech enhancement dataset. Our study focuses on both the audio quality and energy efficiency metrics that are critical for edge devices.

A. Intel N-DNS Challenge Dataset for Speech Enhancement

We utilize the publicly available Intel N-DNS Challenge dataset [28] for the speech enhancement performance evaluation. This comprehensive dataset contains a wide array of human speech audio samples in multiple languages, including English, German, French, Spanish, and Russian, as well as several noise categories. The official Intel N-DNS Challenge repository provides a synthesizer script that generates clean (ground truth), noise (additive), and noisy (ground truth plus noise) audio segments for both training and validation. For testing, the challenge has officially released a large test set for a fair performance comparison. With this official synthesizer script, we synthesize two subsets: a 495-hour subset for training and a 5-hour subset for validation. All audio samples, synthesized at a sampling rate of 16 kHz, are kept at a consistent duration of 30 seconds. For audios shorter than 30 seconds, we concatenate them with other speech signals from the same speaker, inserting a 0.2-second silence interval between clean speech utterances. The noisy audio is simulated using randomly selected speech and noise data, with SNRs ranging from -5 to 20 dB. We apply loudness normalization to each noisy audio sample to simulate agnostic input loudness levels from -35 to -15 decibels relative to full scale (dBFS). For audio quality evaluation, we employ the metrics specified by the Intel N-DNS Challenge, which include SI-SDR [67], and Deep Noise Suppression Mean Opinion Score (DNSMOS) [71]. We also employ the power consumption metrics that will be introduced in Section V-B.

B. Power Consumption Measurement

The power consumption of a neuromorphic system is roughly proportional to the total number of computational primitives used, including synaptic operations (SynOPs) and neuron operations (NeuronOPs) [26], [28]. Based on the power estimation conducted on the Intel Loihi architecture, which indicates that the energy consumed by one NeuronOP is approximately equivalent to that of about 10×SynOPs [28], we derive the power consumption proxy P_{proxy} using the following formula:

$$P_{\text{proxy}} = \text{SynOPs} + 10 \times \text{NeuronOPs}, \tag{24}$$

$$\text{SynOPs} = \sum_{l=1}^{L-1} \sum_{i=1}^{\mathcal{N}^l} \mathcal{R}_i^l (\mathcal{N}^{l+1} + \mathcal{N}^l), \tag{25}$$

$$\text{NeuronOPs} = \sum_{l=1}^{L} \mathcal{N}^l, \tag{26}$$

where $\mathcal{R}_i^l$ denotes the firing rate of neuron i in layer l, $\mathcal{N}^l$ represents the number of neurons in layer l, and L is the total number of layers in the network. Eq. (25) is formulated based on the recurrent neural network architecture, encompassing both feedforward ($\mathcal{N}^{l+1}$) and recurrent outputs ($\mathcal{N}^l$).

We also utilize the power delay product (PDP) metric [26], [28], which consolidates latency and power consumption into a single metric, to assess and compare different systems with distinct focuses on speed and power efficiency. The PDP proxy $\text{PDP}_{\text{proxy}}$ is defined by

$$\text{PDP}_{\text{proxy}} = P_{\text{proxy}} \times \text{Latency}. \tag{27}$$

For the computational costs of Multiply-ACcumulate (MAC) and ACcumulate (AC) operations employed for SynOPs and NeuronOPs respectively, we refer to the findings presented in [72]. These findings indicate that with $45nm$ CMOS technology, one floating-point MAC operation consumes $4.6\ pJ$ of energy while one AC operation consumes $0.9\ pJ$.

It is worth noting that we exclude the encoding and decoding processes from our power consumption measurement, based on the assumption that neuromorphic power dominates in realistic applications. This is consistent with the power measurement guidelines in the Intel N-DNS Challenge [28]. Additionally, the sigmoid function used in GSN can be efficiently implemented on neuromorphic hardware through a lookup table, which requires negligible computational resources. Thus, its contribution is omitted from the power consumption measurement.

C. Implementation Details

All speech audio data are processed at a sampling rate of 16 kHz. The STFT is set up with a window length of 32 ms (512 samples) and a hop length of 8 ms (128 samples), utilizing a Hanning window and comprising 512 FFT frequency bins. The proposed Spiking-FullSubNet takes the magnitude spectrogram as its input. We utilize the AdamW optimizer [73] with a learning rate of 1×10^{-3} and set the gradient norm clipping to 10. In the loss function $\mathcal{L}$, the weights of different terms are set to $\alpha = 0.5, \gamma_1 = 0.5, \gamma_2 = 0.001$. We set the number of neighboring frequency bins to 15. To further improve energy efficiency, we add the total number of synaptic operations into the loss function to penalize excessive firing. We develop different variants of the Spiking-FullSubNet with varying model sizes, detailed in Table III. These variants differ in aspects, including the granularity of frequency partitioning and the order of deep filtering.

We divide all discrete frequency bins based on the study of human auditory perception [74], [74] to see the effects of frequency partitioning configuration on denoising performance and model efficiency. The fundamental frequency, a crucial characteristic of speech signals, extends up to 1 kHz. Therefore, we group the lowest 32 discrete frequency bins, corresponding to $0 \sim 1$ kHz, applying the smallest grouping size within this range. Then, given that the human voice's most significant content lies between $1 \sim 4$ kHz, a range particularly sensitive for the human auditory system and rich in harmonic structures, we group the following 96 discrete frequency bins, equivalent to $1 \sim 4$ kHz, with uniform processing granularity in this range. Lastly, we group the remaining high-frequency bins, 128 discrete frequency bins corresponding to $4 \sim 8$ kHz, as a separate group. These high-frequency bins have a more minor impact on human auditory perception, allowing for coarser granularity in processing. In summary, we establish three frequency partitions: $0 \sim 1$ kHz, $1 \sim 4$ kHz, and $4 \sim 8$ kHz, each tailored to optimize computational efficiency and performance based on human auditory characteristics. Additional implementation details can be found in our open-source code repository.

VI. Results and Discussion

In this section, we first evaluate the speech enhancement performance of the proposed Spiking-FullSubNet. Subsequently, we conduct ablation studies to analyze the impact of different components within the Spiking-FullSubNet as well as their configurations. Finally, we analyze the temporal information processing capability and energy efficiency of the proposed GSN neuron model.

A. Superior Speech Denoising Capability

To ensure a comprehensive evaluation of the speech enhancement performance, this analysis employs a range of metrics concerning audio quality and power consumption. As summarized in Table I, we compare the proposed Spiking-FullSubNet with SOTA real-time ANN baselines and top-ranking SNN models from the Intel N-DNS Challenge. For a fair comparison, all models utilize the same STFT encoding and inverse Short-Time Fourier Transform (iSTFT) decoding. Additionally, in order to provide a clear understanding of the characteristics of the noisy input audio, we also present the results for unprocessed noisy audio, indicated in the row labeled "Noisy (Unprocessed)".

Microsoft NsNet2 [75] is the official baseline benchmark for the Microsoft DNS Challenge 2022, which consists of a series of LSTM layers optimized using a perception-based loss function. *Intel DNS Network* [28] is a proprietary production-level model used in Intel's product environments, which is a streaming, real-time model that integrates LSTM and CNN. It is trained on a larger-scale dataset with data augmentation techniques. As claimed by the Intel N-DNS Challenge organizer [28]: "The network was trained using proprietary datasets and augmentation techniques, and as such we view its audio quality results as upper-bound aspirational targets for challenge submissions." *SDNN network* [28] utilizes a sigma-delta method and is the official SNN baseline for the Intel N-DNS Challenge. From these results, we find our Spiking-FullSubNet model performs considerably better than the above models, demonstrating remarkable effectiveness.

We further compare our method against SOTA real-time ANN models. Among the models under consideration, *DC-CRN* [12] adopts a complex-valued convolutional neural network and ranked 1st in the Microsoft DNS Challenge. Both *FullSubNet* [10] and its variant *Fast FullSubNet* [11] utilize full-band and sub-band modeling techniques similar to our approach. Fast FullSubNet differs from FullSubNet by strategically using mel-scale inputs, which reduces the number of sub-band features. However, our method distinguishes itself from these two models by employing a brain-inspired sub-band partitioning method and leveraging spiking neural networks, setting it apart from the aforementioned models. Furthermore, we include the recently proposed SOTA *CMGAN* [40], a complex-valued multi-scale generative adversarial network, in our comparison. To ensure a fair comparison, we retrain the SOTA models on the Intel N-DNS Challenge dataset with a comparable number of parameters. Our modifications only involve adjusting the number of hidden units in RNNs or the output channels in CNNs, without altering their core architectures. In comparison to these SOTA ANN models, it is evident that our proposed model, Spiking-FullSubNet, outperforms them all in terms of both audio quality and energy efficiency. Notably, when compared with the best-performing ANN model *CMGAN*, the energy consumption of our Spiking-FullSubNet model is almost three orders of magnitude smaller (1.47 m vs. 1.48 μ). This significant energy efficiency is attributable to the cheaper and sparser operations inherent to the SNN model.

In addition, we also compare our method with top-ranking systems [28] in the Intel N-DNS Challenge, as shown in the table. While the repositories for these methods are anonymous, preventing a direct comparison of specific implementations, it's important to note that all submissions to the Intel N-DNS Challenge underwent rigorous verification by the organizers. This inspection ensures the authenticity of their model's real-time capabilities, power consumption, and overall performance. Our Spiking-FullSubNet model stands out from these competitors and has won the championship for the algorithmic track of the Intel N-DNS Challenge. Remarkably, it achieves the highest DNSMOS overall score and SI-SNR, suggesting superior denoising capability. Additionally, the power proxy and PDP proxy are 51.30 M-Ops/s and 1.64 M-Ops, respectively, showing a good balance between audio enhancement performance and computational efficiency.

Besides this, we compare our approach with recently proposed real-time SNN speech enhancement systems (i.e., Spiking S4 [76] and DPSNN [77]). Spiking S4 combines the energy efficiency of SNNs with the long-range sequence modelling capabilities of Structured State Space Models (S4) with LIF neurons. DPSNN is a two-phase time-domain streaming SNN framework that incorporates a Spiking Convolutional Neural Networks and a Spiking Recurrent Neural Networks, with

[1] https://github.com/IntelLabs/IntelNeuromorphicDNSChallenge
[2] https://github.com/huyanxin/DeepComplexCRN
[3] https://github.com/Audio-WestlakeU/FullSubNet
[4] https://github.com/ruizhecao96/CMGAN

TABLE I: Comparison of speech enhancement performance among various baseline models on the Intel N-DNS Challenge dataset. The evaluation metrics include SI-SNR, SI-SNRi, DNSMOS, latency, and computational efficiency. The table comprises the official baselines from the Intel N-DNS Challenge, state-of-the-art real-time ANN models, top-performing SNN models from the competition, and state-of-the-art Spiking-SNN systems. Results are presented with the year of publication or the corresponding challenge rankings.

Entry	**Year / (Rank)**	**SI-SNR (↑) (dB)**	**SI-SNRi (↑)**		**DNSMOS (↑)**			**Latency (↓)**		**Power Proxy (↓) (Ops/s)**	**PDP Proxy (↓) (Ops)**	**Energy Cost (↓) (J)**	**Param Count ($\times 10^3$)**
			data (dB)	**enc+dec (dB)**	**OVR**	**SIG**	**BAK**	**enc+dec (ms)**	**total (ms)**				
Noisy (Unprocessed)	-	7.37	-	-	2.43	3.16	2.69	-	-	-	-	-	-
Official Intel N-DNS Challenge baseline systems. The results are directly quoted from the Intel N-DNS official repository[1].													
Microsoft NsNet2 [75]	2023	11.63	4.26	4.26	2.95	3.26	3.93	0.02	20.02	136.13 M	2.72 M	12.51 μ	2681
Intel DNS Network [28]	2023	12.51	5.14	5.14	3.08	3.34	4.07	0.02	32.02	-	-	-	1901
SDNN Network [28]	2023	12.26	4.89	4.89	2.70	3.20	3.45	0.02	32.02	14.52 M	0.46 M	0.41 μ	525
State-of-the-art real-time ANN baselines, which are adapted from their official code repositories and trained on the Intel N-DNS Challenge dataset.													
DCCRN[2] [12]	2021	12.46	5.09	5.09	2.85	3.11	3.77	0.02	20.02	5.07 G	0.10 G	0.46 m	1247
FullSubNet[3] [10]	2022	13.55	6.18	6.18	2.93	3.22	3.84	0.03	32.03	3.65 G	0.12 G	0.55 m	1141
Fast FullSubNet[3] [11]	2023	13.88	6.51	6.51	2.91	3.24	3.82	0.03	32.03	0.49 G	0.02 G	0.09 m	1141
CMGAN[4] [40]	2024	14.01	6.64	6.64	2.81	3.23	3.84	0.02	20.02	15.94 G	0.32 G	1.47 m	1413
Intel N-DNS Challenge top-ranking systems. The results are directly quoted from the Intel N-DNS official repository[1].													
CTDNN LAVADL[1]	rank 2	13.52	6.59	6.59	2.97	3.32	3.86	0.00	32.00	61.37 M	1.96 M	1.76 μ	905
Sparsity SDNN[1]	rank 3	12.16	4.80	4.80	2.70	3.19	3.46	0.01	32.01	9.32 M	0.30 M	0.27 μ	344
PSNN[1]	rank 4	12.32	4.96	4.96	2.68	2.91	3.96	0.00	32.00	57.24 M	1.83 M	1.65 μ	724
State-of-the-art real-time SNN systems. The results are directly quoted from their papers.													
Spiking-S4 [76]	2024	14.58	7.21	7.21	2.85	3.21	3.74	-	-	-	-	-	530
DPSNN [77]	2024	14.70	7.34	7.34	2.90	3.27	3.77	0.0	10.00	87.47 M	0.87 M	0.78μ	1320
Spiking-FullSubNet	**winner**	15.20	7.83	7.83	3.03	3.35	3.94	0.02	32.02	51.30 M	1.64 M	1.48 μ	965

PLIF and ALIF neurons at the core, respectively. We note that Spiking-FullSubNet significantly outperforms Spiking S4 and DPSNN, even though DPSNN has a much larger number of parameters.

Finally, in Figure 6, we provide a detailed visual comparison of the spectrograms for a single test sample, showcasing the noisy input signal, enhanced signal, and the clean reference signal. The noisy speech input has an SNR of -5 dB and contains both stationary and non-stationary noise components, presenting a significant challenge for speech enhancement systems. It can be seen that the Spiking-FullSubNet model not only preserves essential speech features but also significantly reduces the background noise. In contrast, the Intel N-DNS official baseline model, referred to as the "SDNN network", struggles to handle the various noise components, as highlighted by the residual artifacts and noise components within the white translucent rectangles. This comparison underscores the effectiveness of our model in addressing the challenges posed by diverse noise conditions and illustrates its potential for practical implementations in real-world speech enhancement scenarios.

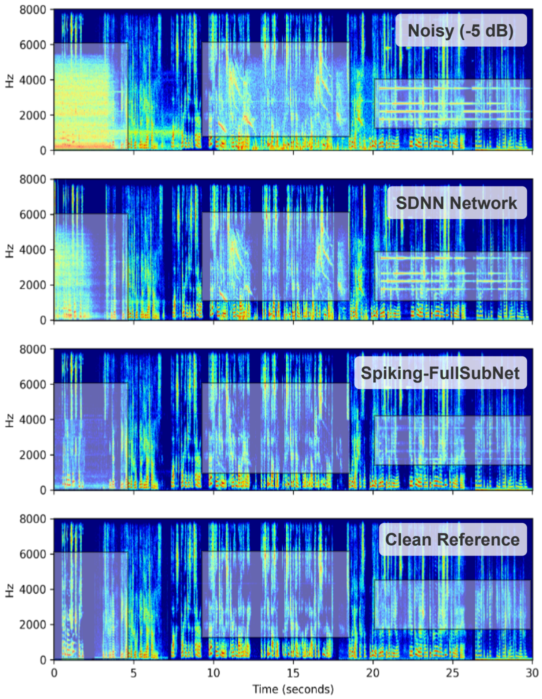


Fig. 6: Comparison of time-frequency magnitude spectrograms for an example from the Intel N-DNS Challenge test set.

B. Ablation Studies for Core Components of Spiking-FullSubNet

To further understand the effectiveness of different components within the proposed Spiking-FullSubNet, we conduct ablation studies by replacing the proposed GSN model with other commonly used spiking neuron models, removing the sub-band modeling, and removing deep filtering components from Spiking-FullSubNet. As shown in Table II, our model equipped with the newly proposed GSN model demonstrates superior performance with an SI-SNR of 15.20 dB and DNS-

TABLE II: Ablation Studies for Core Components of Spiking-FullSubNet. "w/o Sub-band models" means the sub-band GSN models are removed from the Spiking-FullSubNet. To fairly compare the performance, "w/o Sub-band models" has more layers and parameters than that of the Spiking-FullSubNet. "w/o Deep Filtering" means the deep filtering is fully removed from the Spiking-FullSubNet.

Entry	**# Para.** ($\times 10^3$)	**SI-SNR** (dB)	**DNSMOS**		
			OVR	SIG	BAK
Noisy	-	7.37	2.43	3.16	2.69
Spiking-FullSubNet	965	15.20	3.03	3.35	3.94
Replace the GSN model with other spiking neuron models.					
w/ LIF neuron	948	9.72	2.73	3.26	3.40
w/ PLIF neuron	950	11.37	2.82	3.25	3.65
w/ ALIF neuron	952	11.43	2.85	3.26	3.68
w/o Sub-band models	991	13.97	2.91	3.26	3.84
w/o Deep Filtering	918	15.10	3.01	3.28	3.94

MOS scores of 3.03 (OVR), 3.35 (SIG), and 3.94 (BAK). When the GSN is replaced with the LIF neuron, there is a noticeable drop in performance, with the SI-SNR falling to 9.72 dB. This decline is also visible in the DNSMOS metrics, which decreases across all categories, and most notably in the overall rating (OVR), which drops to 2.73. The use of the PLIF neuron, while yielding better results than LIF with an SI-SNR of 11.37 dB, still underperforms our GSN model by a large margin. This decrease is consistent across the DNSMOS metrics. The performance of ALIF is better than that of PLIF, but the margin is limited. These results indicate that the proposed GSN model is favorably advantageous in the speech enhancement task, outperforming existing spiking neuron models.

In addition, we remove the sub-band models from Spiking-FullSubNet and only keep the prior full band model to verify the improvement brought by the sub-band models. To make a fair comparison, we further increase the network layers and parameter counts of the remaining full band model. From the table, we can notice that removing the sub-band components from Spiking-FullSubNet results in decreased performance, with the SI-SNR dropping to 13.97 dB and DNSMOS scores decreasing across all categories. Our result indicates that the sub-band component can effectively learn to focus on complementary cues to the full-band component, leading to better speech enhancement performance. This observation aligns with findings from other full-band and sub-band modeling studies [10], [13].

Finally, we observe that removing the deep filtering component from the Spiking-FullSubNet will also result in lower performance. The SI-SNR drops to 15.10 dB, and DNSMOS scores decrease across all categories. This suggests that the deep filtering component enhances the temporal information processing capability of the Spiking-FullSubNet, further contributing to its improved speech enhancement performance.

C. Ablation Studies for Different Network Configurations of Spiking-FullSubNet

In this section, we delve deeper into the impact of different network component configurations within Spiking-FullSubNet and present the results in Table III. Specifically, we investigate the effects of varying the grouping parameter set ($\{g_k\}$) and the deep filtering order set ($\{d_k\}$) across all frequency partitions. To clearly demonstrate how these configurations influence power consumption, we remove the SynOPs regularization term during network training. In the following, we discuss how different configurations of $\{g_k\}$ and $\{d_k\}$ affect network performance.

1) Grouping Parameters $\{g_k\}$: Processing each discrete frequency independently, by setting the grouping parameter to 1 for all partitions (i.e., $\{1, 1, 1\}$, similar to the original FullSubNet [10]), results in notably high performance. However, this approach incurs a very high computational cost of 1900.47 M-Ops/s. By employing our proposed frequency partitioning method, which divides discrete frequencies according to human auditory perception characteristics and assigns different processing granularities to each partition, we significantly reduce computational cost while maintaining performance close to that achieved without frequency partitioning.

To be specific, we first investigate the impact of varying processing granularity for low frequencies ($0 \sim 1$ kHz), ranging from fine to coarse ($2 \rightarrow 4 \rightarrow 8 \rightarrow 16 \rightarrow 32$). As expected, computational cost decreases as processing granularity becomes coarser, as the model processes more discrete frequencies simultaneously, reducing the required computational operations. In addition, we notice that initially increasing granularity has minimal impact on performance. However, beyond a certain point, performance degrades. Specifically, setting the grouping parameter to 4 reduces the computational cost to 6.39% of the baseline setting (i.e., $\{1, 1, 1\}$) while maintaining similar performance. Setting the grouping parameter to 8 slightly decreases performance but reduces the computational cost to 4.94% of the baseline. However, beyond this point, performance significantly degrades with grouping parameters set to 16 and 32. This degradation is likely due to overly coarse granularity hindering the model's ability to differentiate temporal stationarity between speech and noise signals in the frequency domain, which is crucial for effective sub-band processing.

Then, we investigate the impact of varying processing granularity for the mid-frequency partition ($1 \sim 4$ kHz) while keeping the granularity for the low and high-frequency partitions constant. We observe a significant performance drop when increasing the granularity from 32 to 96 in the mid-frequency partition. This is likely because the mid-frequency partition contains many harmonic structures and exhibits distinct differences in the stationarity of speech and noise, similar to the low-frequency partition. Fine granularity is necessary to capture these details. Finally, we explore increasing processing granularity only for the high-frequency partition. While performance still declines, the decrease is less pronounced compared to the drop observed when coarsening the mid-frequency processing granularity. This can be explained by

TABLE III: The impact of different network components configurations within the Spiking-FullSubNet. $\{g_k\}$ is the grouping parameter set. The value in $\{g_k\}$ means how many discrete frequencies are processed jointly in each frequency partition. $\{d_k\}$ is the deep filtering order set for all frequency partitions.

Grouping Parameters $\{g_k\}$			Filter Orders $\{d_k\}$			Param Count ($\times 10^3$)	Power Proxy (M-Ops/s)	PDP Proxy (M-Ops)	SI-SNR (dB)	DNSMOS		
0∼1 KHz	1∼4 KHz	4∼8 KHz	0∼1 KHz	1∼4 KHz	4∼ 8 KHz					OVR	SIG	BAK
Unprocessed noisy speech						-	-	-	7.37	2.44	3.16	2.69
1	1	1	3	1	1	832	1900.47	63.75	15.26	3.05	3.36	3.96
2	32	64				918	175.90	5.63	15.27	3.04	3.35	3.97
4	32	64				922	121.37	3.89	15.24	3.04	3.35	3.96
8	32	64				929	93.83	3.01	15.16	3.02	3.34	3.93
16	32	64				944	80.52	2.58	14.86	2.97	3.29	3.88
32	32	64				972	75.45	2.41	13.73	2.87	3.26	3.74
8	96	64				987	82.34	2.64	14.74	2.98	3.32	3.88
8	32	128				987	88.28	2.83	14.97	2.99	3.33	3.88
8	32	64	1	1	1	922	93.21	2.98	15.10	3.01	3.28	3.94
			5	1	1	936	96.08	3.08	15.18	3.02	3.35	3.93
			7	1	1	943	97.46	3.12	15.11	3.00	3.21	3.87
			5	3	1	965	99.21	3.17	15.20	3.03	3.35	3.94
			5	5	1	994	103.72	3.32	15.11	3.03	3.33	3.96
			5	3	3	1023	105.44	3.37	15.02	2.97	3.21	3.77

the fact that high-frequency components may possess more redundant information for human auditory perception, allowing for coarser granularity without a significant loss in model performance.

In summary, varying the grouping parameters within different frequency partitions can substantially reduce the computational burden for sub-band processing, demonstrating the effectiveness of our approach. However, while computationally efficient, excessively coarse grouping parameters tend to decrease performance. To achieve a balance between performance and computational cost, we will use $\{8, 32, 64\}$ as our default configuration for subsequent ablation studies.

2) Deep Filter Orders $\{d_k\}$*:* We further investigate the impact of deep filtering on Spiking-FullSubNet by varying the order of the deep filtering for each frequency partition. As shown in Table III, we begin by setting the order to $\{1, 1, 1\}$, meaning deep filtering is not utilized as a training target. When increasing the deep filtering order for the $0 \sim 1$ kHz partition gradually improves performance as the order increases from 1 to 3 and then to 5. However, further increasing the order to 7 results in a performance decline. This suggests that while higher-order deep filtering is beneficial in the low-frequency partition, an excessively high order introduces too much redundant historical information. This prevents effective utilization of short-time correlations within the speech signal, which is the fundamental rationale behind the deep filtering [64]. Similarly, for the mid-frequency partition, increasing the filter order continues to be beneficial. However, when the order reaches 5, a slight decrease in SI-SNR is observed, i.e., from 15.20 to 15.11. However, in the high-frequency partition, applying deep filtering negatively impacts performance. Even a small order of 3 results in a noticeable performance drop. This may be because short-term correlations are less prevalent at high frequencies.

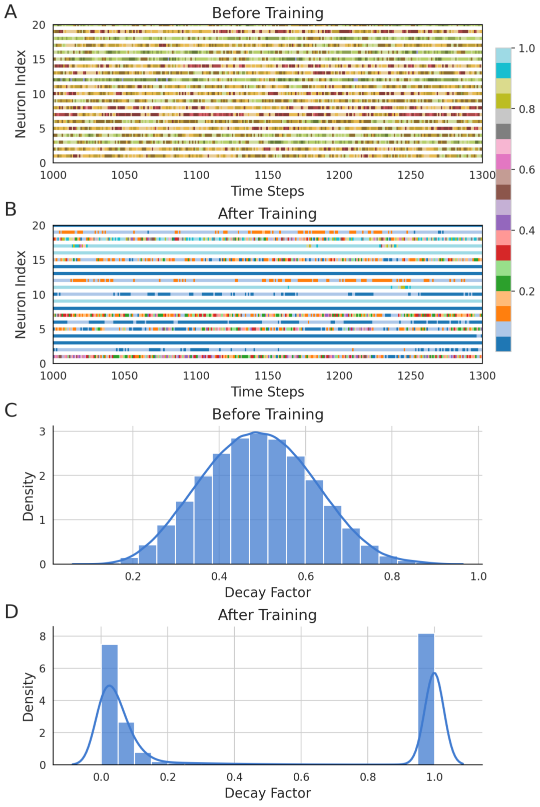


Fig. 7: Analyze the multi-scale temporal information processing capability of GSN. **(A)** and **(B)** depicts the evolution of decay factors over time, before and after training, respectively. **(C)** and **(D)** present the distributions of decay factors across all neurons and time steps, again comparing the states before and after the training process.

D. Enhanced multi-scale Temporal Information Processing

Unlike existing spiking neuron models, our GSN model can dynamically adjust decay factors over time in response to inputs. This dynamic adjustment of decay factors facilitates multi-scale temporal information processing, which is critical for achieving high performance in speech denoising tasks. To demonstrate this capability, we visualize the evolution of decay

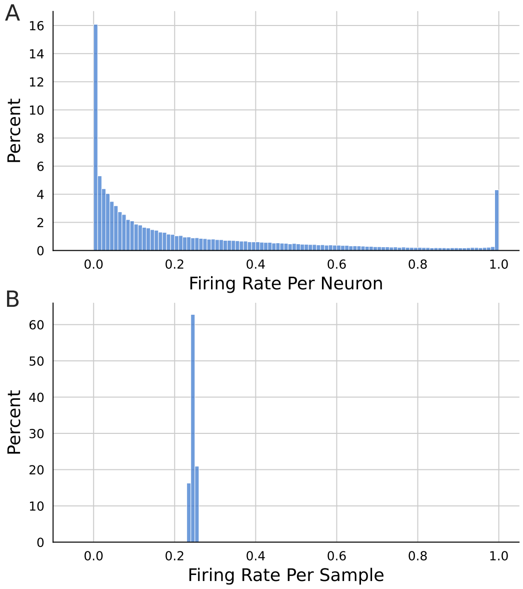


Fig. 8: (**A**) Distributions of per neuron's average firing rate, and (**B**) Distributions of per sample's average firing rate.

factors over time, as shown in Figs. 7**A** and 7**B**. Initially, the decay factors are relatively uniform across different time steps. However, after training, the decay factors show strong temporal variations, aligning with the temporal dynamics of the input signals. This suggests that the GSN model is able to adaptively modulate its temporal processing characteristics to better match the characteristics of the input speech signals. We further investigate the distributions of decay factors across all neurons and time steps. As shown in Figs. 7**C** and 7**D**, the initial decay factors closely follow a Gaussian distribution, with a mean value of approximately 0.5. However, after training, the decay factors exhibit a bimodal distribution, focusing on both slow (decay value close to 1) and rapid decay (decay value close to 0). This distribution shift allows the GSN model to effectively process multiple sound sources that have different temporal dynamics, which is a key requirement for successful speech denoising. The dynamic and adaptive nature of the decay factors in the GSN model is a significant advancement over existing spiking neuron models, enabling more effective multi-scale temporal processing for speech denoising and potentially other time-series signal processing tasks.

E. High Energy Efficiency

To assess the energy efficiency of our proposed Spiking-FullSubNet architecture, we consider the sparsity of output spikes, which is directly related to the computational cost. Specifically, we evaluate the firing rates of the neurons within the trained Spiking-FullSubNet on the test set of the Intel N-DNS dataset [28]. As depicted in Fig. 8**A**, the results show a high degree of sparsity in the neuronal activity. Over 16% of neurons remain inactive throughout the testing phase, and approximately 60% of neurons exhibit a firing rate of less than 0.2. This level of sparsity in neuronal activity is beneficial for achieving high energy efficiency on event-driven neuromorphic hardware implementations. Next, we calculate the firing rate for each sample by averaging the firing rates of all neurons in response to a given sample. Contrary to the wide range of firing rates observed per neuron, the firing rates of different samples are predominantly clustered around 0.25. This suggests that each input sample demands a relatively consistent and minimal number of firing spikes. The combination of high neuronal sparsity and consistent sample-level firing rates indicates that our Spiking-FullSubNet model is well-suited for deployment on energy-constrained neuromorphic hardware, as it can achieve significant computational and energy savings without compromising performance.

VII. Conclusion

This paper introduces Spiking-FullSubNet, an SNN-based model architecture designed for real-time, ultra-low-power speech enhancement tasks. Spiking-FullSubNet leverages a full-band and sub-band fusion approach to effectively capture global and local spectral features crucial for speech enhancement. It incorporates two novel features: 1) a GSN spiking neuron model capable of capturing multi-scale temporal information, and 2) an efficient sub-band frequency partitioning approach that mimics human auditory perception. Our experimental results demonstrate superior speech enhancement performance with a nearly three orders of magnitude reduction in power consumption compared to SOTA ANN models. By leveraging the algorithmic advancements of Spiking-FullSubNet, this work presents a promising denoising solution for a wide range of audio devices. In future work, we will focus on implementing the proposed model on neuromorphic computing chips, such as Intel Loihi [26] and Tianjic [27], to fully realize the potential of neuromorphic speech enhancement technologies.

References

[1] T. Van den Bogaert, S. Doclo, J. Wouters, and M. Moonen, "Speech enhancement with multichannel wiener filter techniques in multimicrophone binaural hearing aids," *The Journal of the Acoustical Society of America*, vol. 125, no. 1, pp. 360–371, 2009.

[2] S. Gannot, E. Vincent, S. M. Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, pp. 692–730, 2017.

[3] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas, S. Gannot, and B. Raj, "The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, oct 2013, pp. 1–4, iSSN: 1947-1629.

[4] 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.

[5] R. Ahmad, S. Zubair, and H. Alquhayz, "Speech enhancement for multimodal speaker diarization system," *IEEE Access*, vol. 8, pp. 126 671–126 680, 2020.

[6] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2007.

[7] 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, 2013.

[8] ——, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 23, no. 1, pp. 7–19, 2014.

[9] 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.

[10] 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.

[11] X. Hao and X. Li, "Fast FullSubNet: Accelerate full-band and sub-band fusion model for single-channel speech enhancement," mar 2023, arXiv:2212.09019 [eess].

[12] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Proc. INTERSPEECH 2020*, 2020, pp. 2472–2476.

[13] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "Tf-gridnet: Integrating full- and sub-band modeling for speech separationgridnet," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3221–3236, 2023.

[14] F. Xiong, W. Chen, P. Wang, X. Li, and J. Feng, "Spectro-temporal subnet for real-time monaural speech denoising and dereverberation," in *Proc. Interspeech 2022*, 2022, pp. 931–935.

[15] 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.

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

[17] X. Chen, Q. Yang, J. Wu, H. Li, and K. C. Tan, "A hybrid neural coding approach for pattern recognition with spiking neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–15, 2023.

[18] J. Wu, C. Xu, X. Han, D. Zhou, M. Zhang, H. Li, and K. C. Tan, "Progressive tandem learning for pattern recognition with deep spiking neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7824–7840, 2022.

[19] Z. Yan, J. Zhou, and W.-F. Wong, "Cq^{+}+ training: Minimizing accuracy loss in conversion from convolutional neural networks to spiking neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 11 600–11 611, 2023.

[20] L. Zhu, S. Dong, J. Li, T. Huang, and Y. Tian, "Ultra-high temporal resolution visual reconstruction from a fovea-like spike camera via spiking neuron model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 1233–1249, 2023.

[21] Q. Yu, C. Ma, S. Song, G. Zhang, J. Dang, and K. C. Tan, "Constructing accurate and efficient deep spiking neural networks with double-threshold and augmented schemes," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 4, pp. 1714–1726, 2021.

[22] M. Yao, G. Zhao, H. Zhang, Y. Hu, L. Deng, Y. Tian, B. Xu, and G. Li, "Attention spiking neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 9393–9410, 2023.

[23] Y. Hu, Q. Zheng, X. Jiang, and G. Pan, "Fast-snn: Fast spiking neural network by converting quantized ann," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 12, pp. 14 546–14 562, 2023.

[24] G. D. Birkhoff, *Dynamical systems*. American Mathematical Soc., 1927, vol. 9.

[25] F. Akopyan, J. Sawada, A. Cassidy, R. Alvarez-Icaza, J. Arthur, P. Merolla, N. Imam, Y. Nakamura, P. Datta, G.-J. Nam *et al.*, "Truenorth: Design and tool flow of a 65 mw 1 million neuron programmable neurosynaptic chip," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 34, no. 10, pp. 1537–1557, 2015.

[26] 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.

[27] J. Pei, L. Deng, S. Song, M. Zhao, Y. Zhang, S. Wu, G. Wang, Z. Zou, Z. Wu, W. He *et al.*, "Towards artificial general intelligence with hybrid tianjic chip architecture," *Nature*, vol. 572, no. 7767, pp. 106–111, 2019.

[28] 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.

[29] W. Gerstner and W. M. Kistler, *Spiking neuron models: Single neurons, populations, plasticity*. Cambridge University Press, 2002.

[30] Z. Chen and P. Zhang, "Lightweight full-band and sub-band fusion network for real time speech enhancement," in *Proc. Interspeech 2022*, 2022, pp. 921–925.

[31] E. M. Izhikevich, "Simple model of spiking neurons," *IEEE Transactions on Neural Networks*, vol. 14, no. 6, pp. 1569–1572, 2003.

[32] A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve," *The Journal of physiology*, vol. 117, no. 4, p. 500, 1952.

[33] 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.

[34] B. Yin, F. Corradi, and S. M. Bohté, "Accurate and efficient time-domain classification with adaptive spiking recurrent neural networks," *Nature Machine Intelligence*, vol. 3, no. 10, pp. 905–913, 2021.

[35] X. Yao, F. Li, Z. Mo, and J. Cheng, "Glif: A unified gated leaky integrate-and-fire neuron for spiking neural networks," *Advances in Neural Information Processing Systems*, vol. 35, pp. 32 160–32 171, 2022.

[36] S. Zhang, Q. Yang, C. Ma, J. Wu, H. Li, and K. C. Tan, "Tc-lif: A two-compartment spiking neuron model for long-term sequential modelling," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 15, 2024, pp. 16 838–16 847.

[37] X. Chen, J. Wu, C. Ma, Y. Yan, and K. C. Tan, "A parallel multi-compartment spiking neuron for multi-scale sequential modeling," 2024. [Online]. Available: https://openreview.net/forum?id=FlH6VB5sJN

[38] A. Pandey and D. Wang, "A new framework for cnn-based speech enhancement in the time domain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 7, pp. 1179–1188, 2019.

[39] ——, "Tcnn: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6875–6879.

[40] 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.

[41] T. Gerkmann and R. C. Hendriks, "Unbiased mmse-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, 2011.

[42] X. Li, L. Girin, S. Gannot, and R. Horaud, "Non-stationary noise power spectral density estimation based on regional statistics," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 181–185.

[43] E. J. Diethorn, "Subband noise reduction methods for speech enhancement," *Audio signal processing for next-generation multimedia communication systems*, pp. 91–115, 2004.

[44] L. Lin, W. H. Holmes, and E. Ambikairajah, "Subband noise estimation for speech enhancement using a perceptual wiener filter," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, vol. 1. IEEE, 2003, pp. I–80.

[45] H. Kuttruff, *Room acoustics*. Crc Press, 2016.

[46] A. Tavanaei and A. S. Maida, "A spiking network that learns to extract spike signatures from speech signals," *Neurocomputing*, vol. 240, pp. 191–199, 2017.

[47] N. Caporale and Y. Dan, "Spike timing–dependent plasticity: a hebbian learning rule," *Annu. Rev. Neurosci.*, vol. 31, pp. 25–46, 2008.

[48] J. Wu, Y. Chua, and H. Li, "A biologically plausible speech recognition framework based on spiking neural networks," in *2018 international joint conference on neural networks (IJCNN)*. IEEE, 2018, pp. 1–8.

[49] 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.

[50] J. Wu, Y. Chua, M. Zhang, G. Li, H. Li, and K. C. Tan, "A tandem learning rule for effective training and rapid inference of deep spiking neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 1, pp. 446–460, 2023.

[51] Q. Yang, Q. Liu, and H. Li, "Deep residual spiking neural network for keyword spotting in low-resource settings," *Interspeech 2022*, pp. 3023–3027, 2022.

[52] Y. Zhang, P. Li, Y. Jin, and Y. Choe, "A digital liquid state machine with biologically inspired learning and its application to speech recognition," *IEEE transactions on neural networks and learning systems*, vol. 26, no. 11, pp. 2635–2649, 2015.

[53] G. Bellec, D. Salaj, A. Subramoney, R. Legenstein, and W. Maass, "Long short-term memory and learning-to-learn in networks of spiking

neurons," *Advances in neural information processing systems*, vol. 31, 2018.

[54] 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.

[55] P. Blouw, X. Choo, E. Hunsberger, and C. Eliasmith, "Benchmarking keyword spotting efficiency on neuromorphic hardware," in *Proceedings of the 7th annual neuro-inspired computational elements workshop*, 2019, pp. 1–8.

[56] 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.

[57] Q. Zhang, A. Nicolson, M. Wang, K. K. Paliwal, and C. Wang, "DeepMMSE: A deep learning approach to mmse-based noise power spectral density estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1404–1415, 2020.

[58] 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.

[59] C. Zheng, H. Zhang, W. Liu, X. Luo, A. Li, X. Li, and B. C. J. Moore, "Sixty years of frequency-domain monaural speech enhancement: from traditional to deep learning methods," *Trends in Hearing*, vol. 27, p. 23312165231209913, jan 2023.

[60] T. Ochiai, M. Delcroix, T. Nakatani, and S. Araki, "Mask-based neural beamforming for moving speakers with self-attention-based tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 835–848, 2023.

[61] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

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

[63] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, 2016.

[64] 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.

[65] S. Braun and I. J. Tashev, "A consolidated view of loss functions for supervised deep learning-based speech enhancement," *2021 44th International Conference on Telecommunications and Signal Processing (TSP)*, pp. 72–76, 2020.

[66] S. Abdulatif, K. Armanious, J. T. Sajeev, K. Guirguis, and B. Yang, "Investigating cross-domain losses for speech enhancement," *2021 29th European Signal Processing Conference (EUSIPCO)*, pp. 411–415, 2020.

[67] 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.

[68] P. J. Werbos, "Backpropagation through time: what it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.

[69] Y. Wu, L. Deng, G. Li, J. Zhu, and L. Shi, "Spatio-temporal backpropagation for training high-performance spiking neural networks," *Frontiers in neuroscience*, vol. 12, p. 331, 2018.

[70] E. O. Neftci, H. Mostafa, and F. Zenke, "Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks," *IEEE Signal Processing Magazine*, vol. 36, no. 6, pp. 51–63, 2019.

[71] 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.

[72] S. Han, J. Pool, J. Tran, and W. Dally, "Learning both weights and connections for efficient neural network," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[73] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2017.

[74] R. T. Sataloff, "The human voice," *Scientific American*, vol. 267, no. 6, pp. 108–115, 1992.

[75] H. Dubey, V. Gopal, R. Cutler, A. Aazami, S. Matusevych, S. Braun, S. E. Eskimez, M. Thakker, T. Yoshioka, H. Gamper, and R. Aichner, "Icassp 2022 deep noise suppression challenge," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 9271–9275.

[76] Y. Du, X. Liu, and Y. Chua, "Spiking structured state space model for monaural speech enhancement," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 766–770.

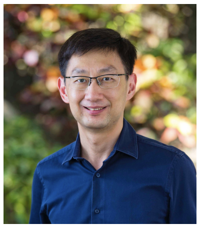
[77] T. Sun and S. Bohté, "Dpsnn: spiking neural network for low-latency streaming speech enhancement," *Neuromorphic Computing and Engineering*, vol. 4, no. 4, p. 044008, dec 2024.

Xiang Hao received his Master's degree in computer engineering from Inner Mongolia University, China, in 2021. He is currently working toward a Ph.D. degree in the Department of Data Science and Artificial Intelligence, The Hong Kong Polytechnic University. His research interests include speech enhancement and hearing aid technologies. He was the recipient of the 2024 IEEE Conference on Artificial Intelligence Best Paper Award.

Chenxiang Ma received his master's degree in computer science from Tianjin University, Tianjin, China, in 2022. He is currently pursuing a PhD degree with the Department of Data Science and Artificial Intelligence, The Hong Kong Polytechnic University. His current research interests include learning algorithms in spiking neural networks and efficient deep learning.

Qu Yang received a B.Eng. and an M.Sc. degree in electrical engineering from the National University of Singapore, Singapore in 2016 and 2019, respectively. She is currently a Ph.D. candidate in electrical engineering at the National University of Singapore. Her research focuses on neuromorphic computing and speech processing.

Jibin Wu (Member, IEEE) received the B.E. and Ph.D degree in Electrical Engineering from National University of Singapore, Singapore in 2016 and 2020, respectively. Dr. Wu is currently an Assistant Professor in the Department of Data Science and Artificial Intelligence, The Hong Kong Polytechnic University. His research interests broadly include brain-inspired artificial intelligence, neuromorphic computing, computational audition, speech processing, and machine learning. Dr. Wu has published over 40 papers in prestigious conferences and journals in artificial intelligence and speech processing, including NeurIPS, ICLR, AAAI, TPAMI, TNNLS, TASLP, and IEEE JSTSP. He is currently serving as the Associate Editors for IEEE Transactions on Neural Networks and Learning Systems and IEEE Transactions on Cognitive and Developmental Systems.



Kay Chen Tan (Fellow, IEEE) received the B.Eng. degree (with First-Class Hons.) and the Ph.D. degree from the University of Glasgow, Glasgow, U.K., in 1994 and 1997, respectively. He is currently the Head and Chair Professor of Computational Intelligence with the Department of Data Science and Artificial Intelligence, The Hong Kong Polytechnic University, Hong Kong. Prof. Tan was the Editor-in-Chief of IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION from 2015 to 2020, and IEEE Computational Intelligence Magazine from 2010 to 2013, and currently serves as an Editorial Board member of 10+ journals. He served as the Vice-President (Publications) of the IEEE Computational Intelligence Society, USA, from 2021 to 2024, and currently serves as the Chief Co-Editor of Springer Book Series on Machine Learning: Foundations, Methodologies, and Applications.