Biomedical Engineering Letters (2024) 14:981–991
https://doi.org/10.1007/s13534-024-00406-y

REVIEW ARTICLE

Snn and sound: a comprehensive review of spiking neural networks in sound

Suwhan Baek[1,2] · Jaewon Lee[3]

Received: 4 March 2024 / Revised: 8 May 2024 / Accepted: 24 June 2024 / Published online: 11 July 2024

Abstract

The rapid advancement of AI and machine learning has significantly enhanced sound and acoustic recognition technologies, moving beyond traditional models to more sophisticated neural network-based methods. Among these, Spiking Neural Networks (SNNs) are particularly noteworthy. SNNs mimic biological neurons and operate on principles similar to the human brain, using analog computing mechanisms. This capability allows for efficient sound processing with low power consumption and minimal latency, ideal for real-time applications in embedded systems. This paper reviews recent developments in SNNs for sound recognition, underscoring their potential to overcome the limitations of digital computing and suggesting directions for future research. The unique attributes of SNNs could lead to breakthroughs in mimicking human auditory processing more closely.

Keywords Spiking neural network · Sound · Neuromorphic engineering · Analog computing

1 Introduction

The evolution of modern technology explores new frontiers in artificial intelligence and machine learning, with special attention to sound and acoustic recognition. Research in sound and acoustic recognition aims to mimic the human auditory system, with a focus on precise processing and deep understanding of auditory data [1–4]. Traditional sound and acoustic recognition systems have primarily relied on classical machine learning and mathematical models [5, 6]. However, recent advancements in AI technology have led to the predominance of neural network-based research in this area [1–4, 7, 8]. These auditory-based deep learning systems are utilized in fields such as sound classification, localization, and voice recognition, with various software-based neural network models actively researched in each category.

This review paper focuses on Spiking Neural Network (SNN) [9, 10] for sound and acoustic recognition, while highlighting their potential to transcend digital computer architectures by leveraging analog computing structures. SNNs communicate and learn information in a similar manner to biological neurons by using sporadic electrical firings [11], otherwise known as ’spikes’. Unlike software-centric neural networks, SNNs are based on gate-level hardware, operating on analog computing principles rather than von-Neumann or digital computer architectures. This leads to characteristics like low power consumption and low latency, making SNNs particularly promising for embedded devices [10]. Moreover, SNNs’ capacity to naturally integrate temporal dynamics makes them exceptionally well-suited for handling time-series data like sound [12–15].

This paper will delve into the latest research trends in sound recognition using SNNs, reviewing key technical advancements and innovations. It will also discuss how this technology could be applied across various application areas and outline future research directions. This paper will first cover the characteristics of SNNs, then discuss recent SNN technologies and research related to sound and provide a summary of reported research findings, and lastly suggest

✉ Suwhan Baek
sw.baek@posco-inc.com

Jaewon Lee
sieonlee5@snu.ac.kr

[1] AI R &D Laboratory, Posco-Holdings, Cheongam-ro, Pohang-si, Gyeongsangbuk-do 37673, Korea

[2] Department of Computer Science, Kwangwoon University, Gwangun-ro, Nowon-gu, Seoul 01899, Republic of Korea

[3] Department of Psychology, Seoul National University, Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea

future research directions based on current technology levels and limitations.

2 Auditory SNN overview

2.1 SNN outlines

SNN [9] attempt to mimic the neural cells from the hardware level, unlike other software-based networks. This is in contrast to traditional software-based virtual neural networks, demonstrating exceptional performance in embedded devices and dynamic environments that must operate under harsh conditions. Such characteristics arise because SNN fundamentally follow the structure of analog computers, employing design units of analog signals to emulate the physical properties of brain neural networks through pure electrical signal-based operations. Recent studies have shown that SNN offer superior results in energy efficiency and processing speed for voice signal processing, especially in time-sensitive sound recognition tasks, compared to conventional deep learning methods. This advancement stems from SNNs' unique structure as 'analog computer'-based neural network systems, surpassing traditional neural networks in terms of power efficiency and latency in sound recognition.

Although SNN models initially focused on simulating simple spiking activities, recent research has shifted attention to their learning capabilities and real-time data processing abilities. These virtues shine in mobile and wearable technologies, and so recent studies tend to focus on how the technology may be applied in these environments [16–19].

The operation of a typical SNN can be mathematically divided into two main components: the behaviour of neurons and the transmission of spikes.

Behavior of Neurons: Each neuron changes its state according to the input signals. The state of a neuron is commonly represented by its voltage state, which varies over time. The voltage $V(t)$ of a neuron is expressed as a function of time t, and it generates a spike when it reaches a certain threshold voltage $V_{\text{threshold}}$ [20, 21]. The voltage state of a neuron can be modeled by the following differential equation:

$$\tau \frac{dV(t)}{dt} = -V(t) + I(t) \tag{1}$$

Here, τ is the time constant, and $I(t)$ is the input signal over time.

Transmission of Spikes: When a neuron generates a spike, this spike is transmitted to other neurons. The effect of the spike is regulated through weights. The transmission of a spike from neuron j to neuron i is determined by the weight W_{ij}. This can be expressed by the following equation:

$$I_i(t) = \sum_j W_{ij} S_j(t) \tag{2}$$

Where $I_i(t)$ is the total input signal reaching neuron i, W_{ij} is the connection strength from neuron j to i, and $S_j(t)$ is the spike time function of neuron j. In this way, SNNs can respond in a more natural manner to temporally varying inputs.

To provide appropriate inputs for an SNN designed in this way, it is necessary to embed conventional time-series data into spike signals [22–24]. There are typically two methods used for signal encoding: Rate Coding and Temporal Coding. These two methods offer different ways of encoding aspects of the signal into spikes.

Rate Coding: In this method, the intensity of the input signal is converted into the firing frequency of neurons [22]. A more intense signal is converted into a series of frequent spikes, whereas a less intense signal would be transformed into a series of infrequent spikes. Given a time-series signal $x(t)$, rate coding to convert this into a spike sequence can be expressed as follows:

$$S(t) = f(x(t)) \tag{3}$$

Where $S(t)$ represents the occurrence of spikes at time t, and f is the function that converts the signal into spiking frequency.

Temporal Coding: In this method, the temporal pattern of the input signal is encoded into the timing in which the signal changes [22, 23]. Here, the timing of signal changes is more important than the signal's intensity. For a time-series signal $x(t)$, temporally coded spikes can have the following relationship:

$$S(t) = g(\Delta x(t)) \tag{4}$$

Where $\Delta x(t)$ is the change in the signal over time, and g is the function that converts this change into the timing of spikes.

Both encoding methods are used to convert important characteristics of time-series data into spikes. Rate Coding reflects the intensity information of the input signal, while Temporal Coding captures the temporal pattern and changes in the signal more effectively. These converted spikes are processed through the SNN.

2.2 SNN in sound domain

2.2.1 Sound localization

Sound localization is an important ability that allows humans and animals to identify the source of sounds in their environment and determine their direction [3]. This process utilizes various signal processing mechanisms such as the Time Difference of Arrival (TDOA), Interaural Level Difference (ILD), and the filtering effects caused by the shape of the ears citestrutt1907our,grumiaux2022survey,niu2019 deep,desai2022review. Research on sound localization using SNNs leverages the capability of these networks to process complex signals in real-time. For instance, SNNs can precisely calculate the TDOA of sound signals to estimate their direction, opening up potential applications in robotics, hearing aids, and environmental monitoring systems. Moreover, SNNs can be used to extract various features from sound signals and then use their findings to identify the source of the sounds. This can be particularly useful in distinguishing multiple sound sources in complex environments [25, 26].

Liu et al. [27] proposes a spiking neural network inspired by the auditory sensory neural pathway of actual mammals. In mammals, three parts of the brain stem are associated with locating a sound source: the medial superior olive (MSO), lateral superior olive (LSO), and the inferior colliculus (IC). Liu et al. [27] also uses three types of artificial neurons to get the TDOA, ILD, and the azimuth angle, respectively. Under the assumption that the azimuth angle is related to TDOA and ILD, Liu et al. [27] use the Bayes theorem to better the estimation of the azimuthal location. It should be noted that unlike previous studies that did not use the ILD information, the localization performance when using sound data with frequencies above 1KHz was greatly improved.

Wall et al. [28] also tries to mimic a mammalian auditory processing nuclei, namely the MSO. Wall et al. [28] introduces the Ben's Spiker Algorithm (BSA) in order to convert the sound signals into a biologically realistic spike train. These spike trains are then filtered through another layer to remove any erroneous spikes that might have resulted from noise. The resulting spike trains are finally sent to a train of neurons simulating the MSO using Jeffress' model [29]. This network produces biologically realistic spike trains while also proving that biologically inspired sound localization can compare favorably with classical techniques like cross-correlation SNNs.

On their next research, [30] takes inspiration from the LSO and another distinct part of the auditory brainstem nuclei called the medial nucleus of the trapezoid body (MNTB). The network uses two layers derived from each nuclei. The data from both 'ears' are filtered through a layer mimicking the MNTB, with the input from ipsilateral 'ear' relative to the sound source acting as an excitatory signal while the input from the contralateral 'ear' acts as an inhibitory signal. Both signals are then used in the layer mimicking the LSO when determining ILD.

Pan et al. [31] continue the trend of taking hints from the biological workings of sound localization. Although the study also takes hint from Jeffress' model [29] similar to Wall et al. [28], Pan et al. [31] introduces Multi-Tone Phase Coding (MTPC) before the Jeffress model layer for better simulating human hearing and also improving localization performance. MTPC breaks down a single sound into multiple tones, similar to how the human cochlear recognizes tones from a specific frequency. The model then uses the Jeffress model to estimate ITD information from different tones and then converts them into spikes. For Pan et al. [31] convert pure tones rather than a single complex sound to spikes, the model is computationally efficient and improves the sound localization accuracy.

Roozbehi et al. [32] advance the domain of sound localization by implementing a dynamic-structured reservoir SNN (rSNN). Although this network also incorporates the Jeffress model, the unique integration of Adaptive Resonance Theory (ART) within the rSNN framework enhances the adaptability and efficiency of sound localization. The ART-rSNN model optimizes the neural arrangement dynamically to amplify the detection of energy near sound sources, significantly enhancing localization precision. This innovative architecture allows for real-time adjustments based on the acoustic environment, offering a substantial improvement in computational efficiency and localization accuracy over traditional models.

Lastly, Haghighatshoar and Muir [33] extend the technological boundaries of sound source localization by innovating a low-power SNN method using a Hilbert Transform spike encoding scheme. Unlike the above-mentioned studies, which background heavily lay in biological findings, this study introduces a novel short-time Hilbert transform (STHT) that circumvents the need for complex band-pass filtering. This approach simplifies the auditory signal processing pipeline by directly obtaining a robust phase signal from wideband audio, which is then used to derive a new beamforming method. The result is a state-of-the-art localization accuracy that rivals traditional non-SNN methods but with significantly lower power consumption, making it ideal for integration into low-power IoT devices. The implementation on ultra-low-power SNN hardware demonstrates how signal processing and neural network design can be co-optimized for high efficiency, pushing forward the capabilities of neuromorphic computing in practical applications.

The research trajectory in sound localization using SNNs demonstrates a significant evolution, moving from initial biologically-inspired models to integrating diverse computational strategies that enhance efficiency and adaptability. It must be noted that most papers aim to advance practical usage of computational sound localization while also trying to better understand the inner workings of the biological human sound locating system. As the research of SNNs on sound localization takes insights from neuro-physiology and also tries to share its knowledge to further other fields related to this task, the potential for cross-disciplinary innovation grows significantly, promising better technological advancements in this area.

Table. 1 reports the summary results of the main research results reported above. Since each paper evaluated the performance and developed the model in different datasets and experimental environments, it is impossible to evaluate the performance with one quantitative indicator, but the performance factor technology has been developed in each performance feature area. Table. 2 reports the results once again, focusing on the papers that can compare quantitative performance. Each paper summarized the data as similar as possible and the test results that were conducted in the environment.

2.2.2 Sound classification

The integration of AI in sound or voice-related applications has become essential in most embedded devices [61–63]. Given the necessary technological requirements for deployment on embedded systems, the low latency and low power consumption advantages of SNNs are particularly appealing. Against this backdrop, a significant amount of research on SNN-based sound, typically speech, classification-based applications is being conducted.

Tavanaei and Maida [40] develop a foundational approach by integrating Spike Timing Dependent Plasticity (STDP) with backpropagation, creating a hybrid model known as BP-STDP. This model aligns with the principles of biological neural networks while enhancing computational efficiency-a theme that recurs in subsequent SNN research, particularly in adapting SNNs to more efficiently emulate functionalities like those of rectified linear units (ReLUs) in traditional ANNs.

Dong et al. [41] extend this effort towards unsupervised learning in SNNs by employing a convolutional architecture and STDP for speech recognition, much like Tavanaei and Maida [40]. Dong et al. [41] focus on energy efficiency and the biological feasibility of their network, highlighting the unsupervised feature extraction capability which directly addresses the challenge of high power consumption noted in conventional ANNs. This reflects a shared emphasis on low power consumption that is crucial for embedded systems.

Martinelli et al. [43] further this exploration within the specific context of Voice Activity Detection (VAD). Similar to the previous studies, this research focuses on the power efficiency of SNNs but also tackles the challenge of effective training algorithms for SNNs. By adapting recurrent network training methodologies to SNNs, they manage to achieve state-of-the-art VAD performance while maintaining the low-energy consumption characteristic of SNNs, bridging a common gap in performance between artificial neural networks and SNNs.

Table 1 Performance comparison and major contributions of SNN-based sound localization models

Paper title	Dataset description	Highlights	Performance	Notes
[27]	Directional audio data in realistic acoustic settings	High accuracy at high frequencies	MAE: 1.02 °, Classification Accuracy: 100%	–
[28]	Stereo mics on a robot, low-frequency tones	Cross-correlation: minimal error	Cross-correlation: Error ± 3.3° and for SNN: Error ± 3.8°	Tested in dynamic, noisy environments
[30]	Data derived from cat's HRTF	High accuracy in high-frequency sound localization	Generalization accuracy across different test frequencies: 75.76% at 15 kHz Accuracy within ± 10°: 96.15%	Effective in high-frequency sounds
[31]	Directional audio dataset with varied environmental conditions	Maintains high accuracy in noisy environments	Mean Error: ± 1.5 °	–
Roozbehi et al. [32]	Real environmental sounds from omnidirectional microphones	Dynamic SNN structure enhances localization	MAE: 3.4 °, MDE: 0.38 m, SD: 0.322 m	Uses ART for dynamic adaptation in noisy conditions
Haghighatshoar & Muir [33]	Circular microphone arrays with home audio devices	Hilbert Transform spike encoding scheme for low-power SNNs	MAE: 1.08 ° at SNR 10 dB on noisy narrowband signals	Implemented in ultra-low-power SNN hardware, Hilbert beamforming technique

Table 2 Comparison of different systems for sound localization and classification

System	Front/Backend spiking	#mics	Azimuth range	Resolution	Data source, sound type, accuracy
[34]	yes/no	2	−90–90	10°	HRTF, Pure tones, 74.56% (±10°)
[30]	yes/yes	2	−60–60	10°	HRTF, Pure tones, 95.38% (±10°)
[35]	yes/no	2	−45–45	15°	Microphone data, Speech, 72.50% (±15°)
[27]	yes/no	2	−45–45	10°	HRTF, Speech, 90.00% (±10°)
[36]	yes/no	4	−180–180	15°	HRTF, Speech, sounds, 4°–8° MAE
[37]	yes/no	2	−90–90	15°	Microphone data, Speech, 91.00% (±15°)
[38]	yes/no	2	−90–90	30°	Microphone data, Speech, 80.00% (±30°)
[39][a]	no/no	2	−90–90	5°	HRTF, Speech, 99.70% (±5°)
[39][b]	no/no	2	−180–180	5°	HRTF, Speech, 100.00% (±5°)
[31]	yes/yes	2	−90–90	5°	Microphone data, Speech, 75.97% (±5°) or 3.91° MAE
[31]	yes/yes	4	−180–180	5°	Microphone data, Speech, 100.00% (±5°) or 1.02° MAE
[32]	yes/yes	2	variable ranges	5°	Microphone data, various sounds, 3.4° MAE
[33]	yes/yes	2	−90–90	5°	Microphone data, Speech, 0.29° MAE at 20 dB SNR

[a] No head movement

[b] With head movement

Amin [42] introduces another innovative approach through the Adaptive Threshold Module (ATM), which dynamically adjusts neuron thresholds to enhance feature extraction. This adaptation is a direct response to optimizing the processing of input spike trains, a fundamental challenge across SNN applications aimed at achieving real-time processing capabilities. This model echoes the prior emphasis on improving computational efficiency and reducing energy use.

Bensimon et al. [12] and Xiang et al. [44] both push the boundaries of traditional SNN applications by incorporating novel elements-SCTN and PCSNN models respectively-that integrate with sensory inputs or photonic components. These studies not only emphasize low power consumption but also explore novel hardware integrations to boost performance and efficiency, a common theme aimed at bridging the gap between laboratory research and real-world application.

Lastly, Yang and Chang [45] synthesize these themes into a practical application by developing an ultra-low-power speech recognition accelerator that utilizes an RSNN. Their work not only consolidates the advancements in reducing power consumption and computational complexity but also demonstrates effective integration of these strategies into a device suitable for edge computing.

Table. 3 summarizes the main sound classification models summarized above. Each paper is a paper that improved the model based on a new methodology in the field of research. Since the papers were tested with different datasets and environments, there was a problem with quantitative performance comparison. To solve this problem, Table. 4 and Table. 5 collect the results of the tested papers using the RWCP and TIDIGITS datasets during the performance reporting of the papers, respectively, and report them as quantitative figures.

The integration of SNN technology into voice classification systems is still in its early stages, with ongoing research focused on overcoming challenges such as effective training algorithms, spike encoding methods, and hardware integration. As these challenges are addressed, SNNs are expected to become increasingly viable for a wide range of voice processing applications, offering a path toward more efficient, responsive, and power-aware computing solutions.

2.2.3 Multimodal SNN

Recently, sound processing SNNs aim to add visual processing to create multimodal models. The visual and auditory cortices in the human brain are functionally and structurally connected, enabling the integration of sensory information [64, 65]. Given that SNNs aim to replicate the brain's processing structure and capabilities, the progression towards audio-visual multimodal SNNs seems like a natural development.

Rathi and Roy [66] introduces an unsupervised spiking neural network that integrates auditory and visual inputs using distinct unimodal networks linked through cross-modal connections trained via spike-timing-dependent plasticity (STDP). This approach harnesses the inherent correlations between sensory modalities, enhancing robustness and improving classification accuracy in noisy environments. The focus on unsupervised learning and organic integration of sensory data through cross-modal connections is particularly beneficial in environments where labeled data is scarce,

Table 3 Performance comparison and major contributions of SNN-based Sound classification models

Paper title	Dataset	Model	Highlights	Performacne	Structure
Tavanaei andMaida [40]	Aurora dataset	Multi-layer SNN	Uses unsupervised learning for feature extraction, applied in a HMM for spoken digit recognition	Up to 96.32%	Employs LIF neurons, Mel-scaled pooling, and probabilistic STDP for feature learning
Dong et al. [41]	TIDIGITS, TIMIT	Convolutional SNN with STDP	High accuracy on TIDIGITS (97.5%) and TIMIT (93.8%) datasets	97.5% on TIDIGITS, 93.8% on TIMIT	Utilizes STDP for unsupervised learning, achieving performance comparable to ANNs; employs local weight sharing and time-to-first-spike coding
Amin [42]	TIDIGITS, RWCP, Poisson spike trains	ATM based SNN	Features adaptive threshold for real-time spike train feature extraction	High on TIDIGITS 97.64% and RWCP 99.50% datasets	Employs an adaptive algorithm based on the internal spiking neuron threshold level for real-time feature extraction
Bensimonet al. [12]	RWCP database	SCTN-based SNN	Efficient hardware implementation with direct interfacing of acoustic sensors, avoiding costly DACs	98.73%	Implements a biologically inspired design for sound signal processing, leveraging low power SCTN
Martinelliet al. [43]	QUT-NOISE-TIMIT	Recurrent SNN trained with backpropagation	Employs backpropagation training on discrete-time SNNs, using surrogate gradients to allow error backpropagation through a non-differentiable spike function	DCF%: 2.4% to 26.5% across SNRs from +15dB to -10dB	Uses pruning to reduce network connections by 85% without performance loss, demonstrating energy efficiency and competitive performance
Xiang et al. [44]	TIDIGITS, TI20-Word, FSDD	PCSNN	Realizes speech recognition with up to 93.75% accuracy using a photonic convolutional spiking neural network	93.75% on TIDIGITS, 92.19% on TI 20-Word, 84.00% on FSDD	Combining a convolutional spiking neural network for unsupervised feature extraction and a photonic spiking neural network for classification using time-based supervised learning
Yang andChang [45]	TIMIT dataset	Low time-step RSNN	Implemented on a hardware accelerator consuming 71.2 μW, achieving 28.41 TOPS/W and 1903.11 GOPS/mm^2 in energy and area efficiency, respectively	PER of 22.6%	Uses parallel time steps, merged spike techniques, and is optimized for low power consumption

Table 4 Classification accuracy comparison with RWCP dataset

Author	Model	Accuracy (%)
Yu et al. [46]	RNN	95.35
Lai et al. [47]	LSTM	98.40
Pan et al. [48]	BAE-tempotron	95.03
Dennis et al. [49]	MFCC-HMM	47.30
Yao et al. [50]	DKP-SNN	99.10
Dennis et al. [49]	LSF-SNN	98.50
Xiao et al. [51]	LTF-SNN	97.50
Wu et al. [52]	SOM-SNN	99.60
Zhang et al. [53]	SFNN-IP (3 hidden layers)	99.50
Bensimon et al. [12]	SCTN-SNN	98.73
Amin [42]	ATM and 2-phases classifier	98.73

marking a significant shift towards exploiting natural sensory interactions without supervised training.

In contrast, Liu et al. [67] builds upon the concept of multimodal integration in SNNs but adopts a different approach by using a recurrent SNN for auditory data and a convolutional SNN for visual data, each optimized for temporal and spatial features, respectively. These networks are then fused through an attention-based cross-modal sub-network that dynamically evaluates and adjusts the weights assigned to each sensory input, optimizing integration based on data reliability. This architecture not only demonstrates the benefits of multimodal sensory processing but also significantly enhances performance over unimodal auditory SNNs by about 7%p, illustrating the effectiveness of attention mechanisms in complex sensory environments. This contrast with [66]'s method underscores the diversity in strategies for integrating multimodal data in neural networks, each offering unique advantages depending on the application context.

Guo et al. [68] introduces the Spiking Multi-Model Transformer (SMMT), which advances multimodal SNNs by integrating auditory and visual data using a Spiking Cross-Attention (SCA) mechanism. This model refines Liu et al. [67]'s attention-based method by employing a unified transformer framework that processes both modalities simultaneously, simplifying the architecture and potentially increasing integration efficiency. This approach demonstrates improved performance on diverse datasets, showcasing a significant evolution in SNNs for handling complex multimodal data.

Table. 6 summarizes the multimodal SNN model, which is combined with a different domain than the previously reported sound domain.

3 Discussion

SNN demonstrate potential in bridging the gap between current artificial intelligence systems and biological neural processing for auditory signal processing. The temporal

Table 5 Classification accuracy comparison with TIDIGITS dataset

Author	Model	Accuracy (%)
Pan et al. [48]	BAE-MPDAL	97.4
Zhang et al. [54]	Liquid state machine	92.3
Tavanaei and Maida [40]	SNN-SVM	91.0
Tavanaei and Maida [55]	Spiking CNN-HMM	96.0
Abdollahi and Liu [56]	AER silicon cochlea-SVM	95.6
Wu et al. [52]	SOM-SNN	97.4
Anumula et al. [57]	MFCC and GRU RNN	97.75
Neil et al. [58]	AER Silicon Cochlea & Deep RNN	96.10
Zhang and Li [59]	SpiKL-IP (With IP, Reservoir size 540)	94.38
Zhang et al. [60]	Spiking MLPs	94.38
Zhang et al. [53]	SFNN-IP (3 hidden layers)	97.7
Amin [42]	ATM and 2-phases classifier	97.64

Table 6 Performance of SNN-based multimodal learning Models

Paper Title	Model	Multimodal types	Datasets	Performance
Rathi and Roy [66]	SNN+STDP	Audio, Image	MNIST, TI46	98%
Liu et al. [67]	SNN+Attention	Audio, Video	MNIST-DVS + N-TIDIGITS	Audio+Video: 98.95% (Best Performance)
Guo et al. [68]	SNN+Transformer	Audio, Image	CIFAR10-AV, UrbanSound8K-AV, MNIST-DVS+N-TIDIGITS	CIFAR10-AV: 98.01% UrbanSound8K-AV: 96.85% MNIST-DVS+N-TIDIGITS: 99.82%

dynamics of acoustic signals align well with the characteristics of SNN, marking a significant step towards mimicking human neural activity. This not only suggests the possibility of advancing AI but also underscores the importance of biological fidelity in future technology development. The architecture of SNN is designed to emulate the operating principles of the human nervous system, offering a path to overcome the limitations of traditional software-based neural networks. By replicating how human neurons encode and process stimuli, SNN are expected to increase biological accuracy in AI systems. However, this poses the challenge of achieving the computational efficiency and speed of the human brain, which is not easily attainable within current digital computing paradigms.

The significance of this study lies in its pioneering role in extending the application scope of SNNs from the visual domain to various sensory data processing tasks, including audition. The study introduces diverse methods for encoding non-visual external stimuli, such as sound, into spikes. These advancements are expected to further broaden the potential of SNNs in sensory information processing, paving the way for novel applications and insights into biological neural processing. Encoding external stimuli into spikes presents a novel approach to AI development, potentially redefining the boundaries of neural network technology by providing a solution to the challenge of accurate stimulus embedding.

SNN research in the sound domain is currently making remarkable progress as we reported in this. In sound localization, studies by Liu et al. [27], Wall et al. [8, 28], Pan et al. [31], and Haghighatshoar and Muir [33] have demonstrated that SNN can accurately estimate sound source locations even in complex auditory environments. Moreover, in speech recognition, research by Dong et al. [41], Wu et al. [52, 69], and Xiang et al. [44] has confirmed that SNN can achieve performance comparable to existing deep learning models.

Furthermore, recent studies on multimodal information processing using SNNs have gained significant attention. [66] proposed an unsupervised learning-based multimodal SNN that integrates visual and auditory information, while [67] introduced a supervised learning-based multimodal SNN with an attention mechanism. These approaches demonstrate similarities to human multimodal information processing by flexibly combining and modeling interactions among different sensory inputs. Notably, cross-modal connections at the spike signal level differentiate these models from traditional fusion methods that operate on abstracted features. These biologically plausible multimodal SNN models can contribute to a more sophisticated understanding and emulation of human sensory information processing. Moreover, they have the potential to model and interpret human perceptual confusion phenomena arising from interactions between senses. For instance, research could explore using SNN models to explain illusions caused by mismatches between visual and auditory information, such as the McGurk effect [70].

Despite their potential, SNN are often overlooked in favor of models that rely on large numbers of parameters and high computing resources, as SNN currently lag behind in terms of scalability, performance, and ease of development. However, once developed, SNN consume significantly less power, making them environmentally friendly and capable of achieving sufficient performance with fewer parameters. These characteristic features demonstrate that SNN is suitable for solving energy problems required for future AI technologies and for the development of local embedded AI for solving security and speed problems, etc., and can make a significant contribution to the sustainable development of AI. The development of AI technology is at a crossroads. While the pursuit of ever-larger models with massive computational requirements has yielded impressive results, it has also raised concerns about energy consumption and sustainability. SNN offer an alternative path forward, one that prioritizes efficiency and environmental responsibility without sacrificing performance. By focusing on energy-efficient, locally executable models like SNN, we can create a more sustainable and accessible AI ecosystem. This will require ongoing research and collaboration across disciplines, as well as a willingness to challenge the prevailing paradigms of AI development.

However, challenges remain before SNN can be widely applied in practical applications. Key issues include improving SNN model learning efficiency, optimizing spike encoding methods, and implementing low-power neuromorphic hardware. Overcoming these challenges will position SNN as the next-generation auditory information processing technology, surpassing current limitations.

4 Conclusion

This review examined the latest research trends and achievements of SNN in auditory signal processing applications. By emulating the information processing methods of biological neural networks, SNN possess the potential to overcome the limitations of existing AI technologies. Notably, SNN have demonstrated excellent performance in sound localization and speech recognition and have shown the possibility of more sophisticated modeling of human auditory information processing mechanisms. However, for SNN technology to reach the application stage, various technical challenges must be addressed, including model learning efficiency, spike encoding, and hardware implementation. Overcoming these challenges requires interdisciplinary collaborative research among experts from neuroscience, semiconductor engineering, and algorithms.

By expanding the horizons of SNN research and breaking down technological barriers, we can implement artificial intelligence systems that approximate human auditory cognition capabilities. This goes beyond mere technological advancement; it is also an intellectual challenge that elevates our understanding of human intelligence. We anticipate that SNN research will open new frontiers in artificial intelligence technology across various domains, including the auditory field.

Acknowledgements We would like to thank Peter Flach for his help in the pursuit and completion of this study.

Funding This study was funded by the MOTIE (Ministry of Trade, Industry, and Energy) in Korea, under the Fostering Global Talents for Innovative Growth Program (P0017308) supervised by the Korea Institute for Advancement of Technology (KIAT).

Declarations

Conflict of interest Author A declares that he/she has no Conflict of interest.

Ethical approval This article does not contain any studies with human participants performed by any of the authors.

References

1. Francl A, McDermott JH. Deep neural network models of sound localization reveal how perception is adapted to real-world environments. Nature Human Behav. 2022;6(1):111–33.
2. Zhang X, Sun H, Wang S, Xu J. A new regional localization method for indoor sound source based on convolutional neural networks. IEEE Access. 2018;6:72073–82.
3. Yalta N, Nakadai K, Ogata T. Sound source localization using deep learning models. J Robot Mechatron. 2017;29(1):37–48.
4. Pak J, Shin JW. Sound localization based on phase difference enhancement using deep neural networks. IEEE/ACM Trans Audio Speech Lang Process. 2019;27(8):1335–45.
5. Asano F, Asoh H, Matsui T. Sound source localization and separation in near field. IEICE Trans Fundam Electron Commun Comput Sci. 2000;83(11):2286–94.
6. Laufer-Goldshtein B, Talmon R, Gannot S. Semi-supervised sound source localization based on manifold regularization. IEEE/ACM Trans Audio Speech Lang Process. 2016;24(8):1393–407.
7. Grumiaux P-A, Kitić S, Girin L, Guérin A. A survey of sound source localization with deep learning methods. J Acoust Soc of Am. 2022;152(1):107–51.
8. Vera-Diaz JM, Pizarro D, Macias-Guarasa J. Towards end-to-end acoustic localization using deep learning: from audio signals to source position coordinates. Sensors. 2018;18(10):3418.
9. Ghosh-Dastidar S, Adeli H. Spiking neural networks. Int J Neural Syst. 2009;19(04):295–308.
10. Tavanaei A, Ghodrati M, Kheradpisheh SR, Masquelier T, Maida A. Deep learning in spiking neural networks. Neural Netw. 2019;111:47–63.
11. Zhang W, Gao B, Tang J, Yao P, Yu S, Chang M-F, Yoo H-J, Qian H, Wu H. Neuro-inspired computing chips. Nature electron. 2020;3(7):371–82.
12. Bensimon M, Greenberg S, Haiut M. Using a low-power spiking continuous time neuron (sctn) for sound signal processing. Sensors. 2021;21(4):1065.
13. Deng B, Fan Y, Wang J, Yang S. Auditory perception architecture with spiking neural network and implementation on fpga. Neural Netw. 2023;165:31–42.
14. Cai S, Li P, Li H. A bio-inspired spiking attentional neural network for attentional selection in the listening brain. IEEE Trans Neural Netw Learn Syst. 2023.
15. Yan F, Liu W, Dong F, Hirota K. A quantum-inspired online spiking neural network for time-series predictions. Nonlinear Dyn. 2023;1–13
16. Shan H, Feng L, Zhang Y, Yang L, Zhu Z. Compact seizure detection based on spiking neural network and support vector machine for efficient neuromorphic implementation. Biomed Signal Process Control. 2023;86:105268.
17. Li Y, Yin R, Kim Y, Panda P. Efficient human activity recognition with spatio-temporal spiking neural networks. Front Neurosci. 2023;17:1233037.
18. Xiaoxue L, Xiaofan Z, Xin Y, Dan L, He W, Bowen Z, Bohan Z, Di Z, Liqun W. Review of medical data analysis based on spiking neural networks. Procedia Comput Sci. 2023;221:1527–38.
19. Yan Z, Zhou J, Wong W-F. Energy efficient ECG classification with spiking neural network. Biomed Signal Process Control. 2021;63:102170.
20. Ahmed F, Yusob B, Hamed H.N.A. Computing with spiking neuron networks: a review. Int J Adv Soft Comput Appl. 2014; 6(1)
21. Yamazaki K, Vo-Ho V-K, Bulsara D, Le N. Spiking neural networks and their applications: a review. Brain Sci. 2022;12(7):863.
22. Auge D, Hille J, Mueller E, Knoll A. A survey of encoding techniques for signal processing in spiking neural networks. Neural Process Lett. 2021;53(6):4693–710.
23. Petro B, Kasabov N, Kiss RM. Selection and optimization of temporal spike encoding methods for spiking neural networks. IEEE Trans Neural Netw Learn Syst. 2019;31(2):358–70.
24. Yu Q, Tang H, Tan KC, Yu H. A brain-inspired spiking neural network model with temporal encoding and learning. Neurocomputing. 2014;138:3–13.
25. Cerezuela-Escudero E, Jimenez-Fernandez A, Paz-Vicente R, Dominguez-Morales JP, Dominguez-Morales MJ, Linares-Barranco A. Sound recognition system using spiking and mlp neural networks. In: Artificial Neural Networks and Machine Learning–ICANN 2016: 25th International Conference on Artificial Neural Networks, Barcelona, Spain, September 6-9, 2016, Proceedings, Part II 25, 2016; 363–371 . Springer
26. Khatami F, Escabí MA. Spiking network optimized for word recognition in noise predicts auditory system hierarchy. PLOS Comput Biol. 2020;16(6):1007558.
27. Liu J, Perez-Gonzalez D, Rees A, Erwin H, Wermter S. A biologically inspired spiking neural network model of the auditory midbrain for sound source localisation. Neurocomputing. 2010;74(1–3):129–39.
28. Wall JA, McGinnity TM, Maguire LP. A comparison of sound localisation techniques using cross-correlation and spiking neural networks for mobile robotics. In: The 2011 International Joint Conference on Neural Networks, 2011;pp. 1981–1987 . IEEE
29. Jeffress LA. A place theory of sound localization. J Comp Physiol Psychol. 1948;41(1):35.
30. Wall JA, McDaid LJ, Maguire LP, McGinnity TM. Spiking neural network model of sound localization using the interaural intensity difference. IEEE Transactions Neural Netw Learn Syst. 2012;23(4):574–86.
31. Pan Z, Zhang M, Wu J, Wang J, Li H. Multi-tone phase coding of interaural time difference for sound source localization with spiking neural networks. IEEE/ACM Trans Audio Speech Lang Process. 2021;29:2656–70.

32. Roozbehi Z, Narayanan A, Mohaghegh M, Saeedinia SA. Dynamic-structured reservoir spiking neural network in sound localization. IEEE Access .2024.
33. Haghighatshoar S, Muir DR. Low-power snn-based audio source localisation using a hilbert transform spike encoding scheme. arXiv preprint arXiv:2402.11748 2024.
34. Xiao F, Weibei D. A biologically plausible spiking model for interaural level difference processing auditory pathway in human brain. In: 2016 international joint conference on neural networks (IJCNN), 2016;pp. 5029–5036 . IEEE
35. Voutsas K, Adamy J. A biologically inspired spiking neural network for sound source lateralization. IEEE Trans Neural Netw. 2007;18(6):1785–99.
36. Goodman DF, Brette R. Spike-timing-based computation in sound localization. PLoS Comput Biol. 2010;6(11):1000993.
37. Dávila-Chacón J, Heinrich S, Liu J, Wermter S. Biomimetic binaural sound source localisation with ego-noise cancellation. In: Artificial Neural Networks and Machine Learning–ICANN 2012: 22nd International Conference on Artificial Neural Networks, Lausanne, Switzerland, September 11-14, 2012, Proceedings, Part I 22, 2012; 239–246 . Springer
38. Anumula J, Ceolini E, He Z, Huber A, Liu SC. An event-driven probabilistic model of sound source localization using cochlea spikes. In: 2018 IEEE international symposium on circuits and systems (ISCAS), 2018;pp 1–5 . IEEE
39. Ma N, May T, Brown GJ. Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments. IEEE/ACM Trans Audio Speech Lang Process. 2017;25(12):2444–53.
40. Tavanaei A, Maida A. Bio-inspired multi-layer spiking neural network extracts discriminative features from speech signals. In: Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou, China, November 14–18, 2017, Proceedings, Part VI 24, 2017; 899–908 . Springer
41. Dong M, Huang X, Xu B. Unsupervised speech recognition through spike-timing-dependent plasticity in a convolutional spiking neural network. PloS one. 2018;13(11):0204596.
42. Amin HH. Automated adaptive threshold-based feature extraction and learning for spiking neural networks. IEEE Access. 2021;9:97366–83.
43. Martinelli F, Dellaferrera G, Mainar P, Cernak M. Spiking neural networks trained with backpropagation for low power neuromorphic implementation of voice activity detection. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020;pp 8544–8548 . IEEE
44. Xiang S, Zhang T, Han Y, Guo X, Zhang Y, Shi Y, Hao Y. Neuromorphic speech recognition with photonic convolutional spiking neural networks. IEEE Journal of Selected Topics in Quantum Electronics 29(6: Photonic Signal Processing), 2023;1–7
45. Yang C-C, Chang T-S. A 71.2- uw speech recognition accelerator with recurrent spiking neural network. IEEE transactions on circuits and systems I: Regular Papers.2024
46. Yu Y, Si X, Hu C, Zhang J. A review of recurrent neural networks: LSTM cells and network architectures. Neural Comput. 2019;31(7):1235–70. https://doi.org/10.1162/neco_a_01199.
47. Lai J, Chen B, Tan T, Tong S, Yu K. Phone-aware lstm-rnn for voice conversion. In: 2016 IEEE 13th international conference on signal processing (ICSP), 2016;pp. 177–182 . IEEE
48. Pan Z, Chua Y, Wu J, Zhang M, Li H, Ambikairajah E. An efficient and perceptually motivated auditory neural encoding and decoding algorithm for spiking neural networks. Front Neurosci. 2020;13:1420.
49. Dennis J, Yu Q, Tang H, Tran H.D, Li H. Temporal coding of local spectrogram features for robust sound recognition. In: 2013 IEEE international conference on acoustics, speech and signal processing, 2013;pp. 803–807. IEEE
50. Yao Y, Yu Q, Wang L, Dang J. A spiking neural network with distributed keypoint encoding for robust sound recognition. In: 2019 international joint conference on neural networks (IJCNN), 2019;pp. 1–8. IEEE
51. Xiao R, Tang H, Gu P, Xu X. Spike-based encoding and learning of spectrum features for robust sound recognition. Neurocomputing. 2018;313:65–73.
52. Wu J, Chua Y, Zhang M, Li H, Tan KC. A spiking neural network framework for robust sound classification. Front Neurosci. 2018;12:836.
53. Zhang A, Zhou H, Li X, Zhu W. Fast and robust learning in spiking feed-forward neural networks based on intrinsic plasticity mechanism. Neurocomputing. 2019;365:102–12.
54. Zhang Y, Li P, Jin Y, Choe Y. A digital liquid state machine with biologically inspired learning and its application to speech recognition. IEEE Trans Neural Netw Learn Syst. 2015;26(11):2635–49.
55. Tavanaei A, Maida AS. A spiking network that learns to extract spike signatures from speech signals. Neurocomputing. 2017;240:191–9.
56. Abdollahi M, Liu S-C. Speaker-independent isolated digit recognition using an aer silicon cochlea. In: 2011 IEEE biomedical circuits and systems conference (BioCAS), 2011;pp. 269–272 . IEEE
57. Anumula J, Neil D, Delbruck T, Liu S-C. Feature representations for neuromorphic audio spike streams. Front Neurosci. 2018;12:308889.
58. Neil D, Pfeiffer M, Liu S-C. Learning to be efficient: algorithms for training low-latency, low-compute deep spiking neural networks. In: Proceedings of the 31st Annual ACM Symposium on Applied Computing, 2016; 293–298
59. Zhang W, Li P. Information-theoretic intrinsic plasticity for online unsupervised learning in spiking neural networks. Front Neurosci. 2019;13:420224.
60. Zhang S, Zhang A, Ma Y, Zhu W. Intrinsic plasticity based inference acceleration for spiking multi-layer perceptron. IEEE Access. 2019;7:73685–93.
61. Park J, Boo Y, Choi I, Shin S, Sung W. Fully neural network based speech recognition on mobile and embedded devices. Adv Neural Inf Process Syst. 2018;31
62. Li S-A, Liu Y-Y, Chen Y-C, Feng H-M, Shen P-K, Wu Y-C. Voice interaction recognition design in real-life scenario mobile robot applications. Appl Sci. 2023;13(5):3359.
63. Price M, Glass J, Chandrakasan AP. A low-power speech recognizer and voice activity detector using deep neural networks. IEEE J Solid-State Circuits. 2017;53(1):66–75.
64. Calvert GA. Crossmodal processing in the human brain: insights from functional neuroimaging studies. Cerebral cortex. 2001;11(12):1110–23.
65. Eckert MA, Kamdar NV, Chang CE, Beckmann CF, Greicius MD, Menon V. A cross-modal system linking primary auditory and visual cortices: Evidence from intrinsic fmri connectivity analysis. Human brain mapp. 2008;29(7):848–57.
66. Rathi N, Roy K. Stdp based unsupervised multimodal learning with cross-modal processing in spiking neural networks. IEEE Trans Emerg Topics Comput Intell. 2018;5(1):143–53.
67. Liu Q, Xing D, Feng L, Tang H, Pan G. Event-based multimodal spiking neural network with attention mechanism. In: ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP), 2022;pp. 8922–8926. IEEE
68. Guo L, Gao Z, Qu J, Zheng S, Jiang R, Lu Y, Qiao H. Transformer-based spiking neural networks for multimodal audiovisual classification. IEEE Transactions on Cognitive and Developmental Systems. 2023

69. Wu J, Yılmaz E, Zhang M, Li H, Tan KC. Deep spiking neural networks for large vocabulary automatic speech recognition. Front Neurosci. 2020;14:199.
70. Tiippana K. What is the mcgurk effect? Front Psychol. 2014;5:91962.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.