A Bio-Inspired Spiking Attentional Neural Network for Attentional Selection in the Listening Brain

Siqi Cai, *Member, IEEE*, Peiwen Li, and Haizhou Li, *Fellow, IEEE*

***Abstract*— Humans show a remarkable ability in solving the cocktail party problem. Decoding auditory attention from the brain signals is a major step toward the development of bionic ears emulating human capabilities. Electroencephalography (EEG)-based auditory attention detection (AAD) has attracted considerable interest recently. Despite much progress, the performance of traditional AAD decoders remains to be improved, especially in low-latency settings. State-of-the-art AAD decoders based on deep neural networks generally lack the intrinsic temporal coding ability in biological networks. In this study, we first propose a bio-inspired spiking attentional neural network, denoted as BSAnet, for decoding auditory attention. BSAnet is capable of exploiting the temporal dynamics of EEG signals using biologically plausible neurons and an attentional mechanism. Experiments on two publicly available datasets confirm the superior performance of BSAnet over other state-of-the-art systems across various evaluation conditions. Moreover, BSAnet imitates realistic brain-like information processing, through which we show the advantage of brain-inspired computational models.**

***Index Terms*— Auditory attention, brain–computer interface (BCI), cocktail party problem, electroencephalography (EEG), spiking feature representation.**

I. Introduction

HUMAN listening is characterized by an abundance of sounds that compete for our attention. Humans have the ability to attend to the sound of interest and follow it selectively over time, that is described as "cocktail party effect" [1]. Understanding how the brain solves the cocktail party problem can shed light on the biological process of *auditory attention*, and pave the way toward more effective auditory prostheses [2]. Auditory attention has been a long-lasting research topic in neuroscience, however, the neural underpinnings of this cognitive capacity remain unclear. Recent studies have predominantly focused on "selective neural speech tracking," [3], [4] that is to track the brain activity that attends to a speech stimulus while ignoring others [5], [6]. This neuroscientific insight lays the groundwork for decoding auditory attention directly from the brain, which is generally referred to as *auditory attention detection (AAD)*.

The study of AAD on a number of neuro-recording modalities, e.g., electrocorticography (ECoG) [5], magnetoencephalography (MEG) [7], and electroencephalography (EEG) [8], has attracted increasing interest. Among them, EEG-based AAD emerges as a promising solution for the cognitive control of hearing aids, i.e., neuro-steered hearing aids, as EEG is a noninvasive, wearable, widely accessible, and relatively low-cost technique [9], as summarized in [10]. The AAD algorithms traditionally adopt a stimulus-reconstruction approach [8]. Briefly, they first reconstruct the envelope corresponding to the attended speech from EEG signals and evaluate the correlation between the reconstructed envelope and the original single-speaker speech stimulus. The speech stimulus with a high correlation is identified as the attended one. Despite much progress [11], [12], [13], [14], [15], [16], [17], [18], the AAD studies suffer from two limitations.

1) The single-speaker speech stimulus may not be available in practice due to various reasons such as technological constraints, interference, sound overlapping, and recording circumstances. While it is possible to obtain single-speaker speech stimulus from the recorded mixture of sound sources using speech separation method [2], this introduces an additional overhead [10].
2) The stimulus-reconstruction approach typically requires a large decision window [19], [20], e.g., 10 s, to effectively decode auditory attention, e.g., at an accuracy of 75%–85%, in a two-speaker scenario. Such technique sees a notable accuracy drop with shorter decision windows [8], [21], [22], thus limiting the scope of real-time applications in brain–computer interfaces (BCIs) [23], [24].

A recent study on auditory spatial attention detection [19], [25], that is termed *ASAD*, does not involve the speech

Manuscript received 7 August 2022; revised 4 June 2023; accepted 4 August 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62271432; in part by the Internal Project of Shenzhen Research Institute of Big Data under Grant T00120220002; in part by the Guangdong Provincial Key Laboratory of Big Data Computing, The Chinese University of Hong Kong, Shenzhen, China, under Grant B10120210117-KP02; in part by the Human-Robot Collaborative AI for Advanced Manufacturing and Engineering under Grant A18A2b0046; in part by the Agency for Science, Technology and Research, Singapore; in part by the Deutsche Forschungsgemeinschaft (DFG) under Germany's Excellence Strategy (University Allowance, EXC 2077, University of Bremen, Germany). *(Corresponding author: Haizhou Li.)*

Siqi Cai is with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583.

Peiwen Li is with the Shien-Ming Wu School of Intelligent Engineering, South China University of Technology, Guangzhou, Guangdong 510460, China.

Haizhou Li is with the School of Data Science, Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), Shenzhen 518172, China, also with the Electrical and Computer Engineering, National University of Singapore, Singapore 117583, and also with the Machine Listening Laboratory, University of Bremen, 28359 Bremen, Germany (e-mail: haizhouli@cuhk.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2023.3303308.

Digital Object Identifier 10.1109/TNNLS.2023.3303308

stimulus during the decoding, thus improving the practical applicability of auditory attention for neuro-steered hearing applications. In terms of system implementation, the convolutional neural networks (CNNs) [26], [27], one of the popular models in artificial neural networks (ANNs), are among the most effective neural architectures for EEG-based ASAD that achieve the state-of-the-art performance [20], [28], [29], [30]. However, we argue that CNN is not the most suitable neural architecture for EEG signal processing.

Spiking neural networks (SNNs), a biologically inspired computational model [31], offer a competitive alternative to ANNs. First, increasing evidence indicates that time-asynchronous SNNs outperform layer-synchronous deep ANNs for the classification of nonstationary signals [32], [33] in terms of accuracy. Second, an SNN is potentially energy-efficient where neurons process discrete events in the spiking trains in an event-driven, time-asynchronous manner, that follows the neuronal communication principles observed in the human brain [6], [34], [35]. Furthermore, the low-frequency, spiky EEG signals can be effectively encoded into spiking trains as a way of feature representation. Therefore, SNNs represent an alternative solution to the EEG-based ASAD problem.

Neural circuits exhibit repetitive or recurrent temporal dynamics, which is thought to play a key role in many cognitive states in the human brain [36], [37]. Moreover, the brain recurrently uses available sensory information in the top-down and bottom-up processes of selective auditory attention [38], [39]. We are particularly interested in exploring whether a computational model with recurrently connected spiking neurons is advantageous in decoding such attention activities in the listening brain.

Attentive listening relies on the differentiated neural representation between attended and ignored speech stimuli [6]. Specifically, an essential aspect of selective cortical representation entails the linking over time of responses produced by the attended speech stimulus, whereas simultaneously separating it from others produced by other stimuli [3], [4], [40]. Inspired by how auditory attention shapes the internal representation of speech stimuli, we further integrate a temporal attention mechanism into the ASAD decoder. The mechanism is capable of assigning weights to each recurrent encoded temporal slice of EEG and aggregating the attentive temporal information to form a final representation.

In this article, we study a bio-inspired spiking attentional neural network, which is referred to as the *BSAnet* hereafter, that does not require the individual speech stimulus and performs ASAD on a short window, i.e., high time resolution, of EEG signals. The main contributions of this study include as follows.

1) We propose BSAnet, a novel modularized end-to-end pipeline. The neural architecture allows us to understand the effect of the intermediate latent representations, and the contributions of individual modules analytically.
2) BSAnet is composed of biologically realistic leaky integrate-and-fire (LIF) neurons and an event-driven neural encoder. It also consists of a temporal attention mechanism and a recurrent spiking layer that are biologically motivated. BSAnet is inherently well-suited for modeling brain signals with a complex temporal structure.
3) Through comprehensive experiments, for the first time, we show that a biologically realistic SNN achieves state-of-the-art performance in ASAD tasks.

The remainder of this article is organized as follows. In Section II, we present a BSAnet pipeline for decoding auditory spatial attention. Extensive experiments are conducted in Section III. The experimental results are reported in Section IV. We look into the results from different perspectives in Section V. Finally, Section VI concludes the study.

II. Bio-Inspired Spiking Attention Net

We now formulate a BSAnet for decoding selective attention in the listening brain. BSAnet consists of three main modules, namely the neural encoder, the spiking temporal attention, and the recurrent spiking layer. We next discuss each of the modules in detail.

A. Neuronal Function

The spiking neuron is the elementary unit in a SNN. The spiking neuron is defined by a neuronal function that processes input signals and produces output spiking trains that are relevant to the intended cognitive task. A spike train is typically a binary time series with the spiking time carrying the information. In the input layer, the spiking trains are ideally generated by event-driven sensors [41]. In this study, the EEG signals are seen as event-driven spiking trains. While the EEG signal is spiky, it takes continuous real values rather than binary ones. Therefore, the spiking neurons in the input layer also play the role to transform the input real-valued EEG time series into a binary spike train.

A number of spiking neuronal functions have been studied that are inspired by how the human nervous system operates [42]. In general, the computational cost of a spiking neuron model grows with increasing biological properties. The LIF neuron is among the most common ones, which features strong biology support and effective computation [43]. It is devised to emulate the successive information propagation through the spiking process of a biological neuron, as shown in Fig. 1.

The membrane potential U_i^l of LIF neuron i at layer l can be formulated by

$$\tau_m \frac{dU_i^l}{dt} = -\left[U_i^l(t) - U_0\right] + RI_i^l(t) \tag{1}$$

where τ_m denotes the membrane time constant of the neuron. U_0 and R denote the resting potential and the membrane resistance of the spiking neuron, respectively. When U_i^l reaches the membrane threshold U_{th}, a spike is emitted. Then, $U_i^l(t)$ resets to the reset potential U_r. In general, we set U_0, U_r to zero, and R to unitary [44], [45]. $I_i^l(t)$ represents the time-dependent input current to the neuron i, which can be derived from the spike trains of all presynaptic neurons j connected

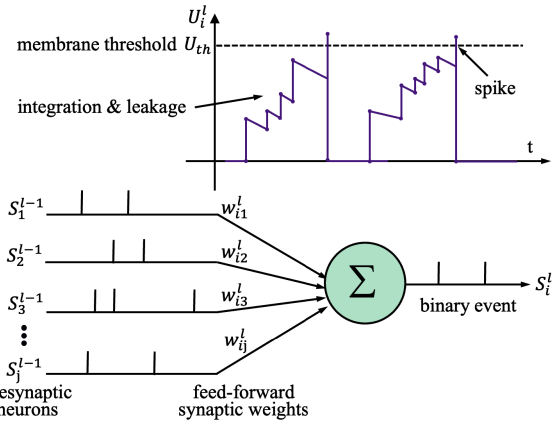

Fig. 1. Output potential of a LIF neuron is driven by multiple input spike trains S_j^{l-1}. Each spike in the input has a certain weight w_{ij}^l that represents the strength of the connection between the neurons. As the inputs are integrated, they contribute to changing the neuron's membrane potential U_i^l. An output spike is fired when the potential exceeds a threshold U_{th}. Once a neuron fires a spike, its membrane potential is reset to the reset potential.

to neuron i as follows:

$$\frac{dI_i^l}{dt} = \sum_j w_{ij}^l S_j^{l-1}(t) \tag{2}$$

$$S_j^{l-1}(t) = \sum_k \delta\left(t - t_j^k\right) \tag{3}$$

where w_{ij}^l represents the feed-forward synaptic weight from neuron j in layer $l-1$ to neuron i in layer l. Spikes emitted by the jth neuron in layer $l-1$ at a finite set of times t_j^k can be described as a spike train S_j^{l-1}, which is defined by (3).

In practice, the LIF model needs to be discretized for implementation. The square bracket is used to index variables that change in a discrete time context [46], where t denotes a continuous value in $f(t)$ or an integer in $f[t]$. As we discretize continuous time into successive time step Δt, the spikes can only take place at the multiples of Δt. Therefore, the linear differential equation for a LIF neuron can be approximated as follows:

$$I_i^l(t+\Delta t) = I_i^l[t+1] = \sum_j w_{ij}^l S_j^{l-1}[t+1] + b_i^l \tag{4}$$

$$U_i^l(t+\Delta t) = U_i^l[t+1] = \alpha U_i^l[t]\left(1 - S_j^l[t]\right) + I_i^l[t+1] \tag{5}$$

$$S_i^l(t+\Delta t) = S_i^l[t+1] = g\left(U_i^l[t+1]\right) \tag{6}$$

where

$$g(x) = \begin{cases} 1, & x \geqslant U_{\text{th}} \\ 0, & x < U_{\text{th}}. \end{cases} \tag{7}$$

Here, $\alpha = \exp(-\Delta t/\tau_m)$ denotes a leaking factor. $I_i^l[t+1]$ denotes the current contributions from presynaptic neurons to the neuron i at time step $t+1$. w_{ij} represents the strength of connection between the presynaptic neuron j and postsynaptic neuron i. b_i^l is the constant injecting current to the neuron i. $S_i^l \in \{0, 1\}$ represents the output of the ith neuron determined by the step function $g(\cdot)$, where $S_i^l = 1$ denotes a spike activity and $S_i^l = 0$ denotes otherwise.

Considering the discretization process in a LIF neuron, we may see a spiking layer as three cascaded sub-layers: the current from (4), the membrane potential from (5), and the spike from (6). Therefore, the output spike trains of the J^l neurons in the jth layer can be expressed by a spiking linear function $\eta(\cdot)$

$$\mathbf{S}^l = \eta\left(\mathbf{S}^{l-1}, \mathbf{W}^l, \mathbf{B}^l\right) \tag{8}$$

where $\eta(\cdot)$ reflects the collective behavior of a set of LIF neurons. A LIF neuron is defined in Fig. 1 and formulated in (1)–(7). In particular, $\mathbf{S}^l$ denotes the J^l output spike trains, $\mathbf{W}^l$ denotes the feed-forward synaptic weight matrix from the neurons at the $(l-1)$th layer to those at the jth layer, and $\mathbf{B}^l$ represents the constant injecting current to the neurons.

B. Neural Encoder

In a two-speaker scenario, AAD is a binary classification problem [19], [20], [25], [28], [29], [30] that takes a window of EEG data as input and makes a binary decision. An input EEG signal is represented as a sequence of shifting *decision windows*. Let $\mathbf{E} = [e_1, \ldots, e_c, \ldots, e_N] \in \mathbb{R}^{N\times T}$ denote an EEG window, where $e_c \in \mathbb{R}^{1\times T}$ is a time series of T samples from the cth EEG channel of a total of N channels.

As illustrated in Fig. 2, the proposed BSAnet pipeline takes $\mathbf{E}$ as the input and is optimized to decide if the listener has leftward or rightward auditory attention. For the SNN to take the real-valued EEG signals $\mathbf{E}$ as the time-dependent input currents, we design an input layer that converts the EEG signals to spike trains. We further design an input layer, that is called the neuron encoder, which is composed of N_s LIF neurons as shown in Fig. 2(a) and expressed as follows:

$$\mathbf{S}_e = \eta(\mathbf{E}, \mathbf{W}_e, \mathbf{B}_e) \tag{9}$$

where $\mathbf{S}_e \in \mathbb{R}^{N_s\times T}$ denotes the encoded EEG with N_s streams of output, each coming from one LIF neuron. $\mathbf{S}_e$ is therefore referred to as the spiking feature representation. Note that typically we set $N_s < N$. In this way, the neural encoder serves as a learnable feature extractor to reduce the volume of the input data. $\mathbf{W}_e \in \mathbb{R}^{N_s\times N_s}$ and $\mathbf{B}_e \in \mathbb{R}^{N_s}$ denotes the weights and the biases that are the trainable parameters of the LIF neurons. This neural encoding scheme is capable of converting real-valued EEG signals into spike trains with enough precision and high temporal resolution, as demonstrated in previous research [47].

C. Spiking Temporal Attention

Attentional modulation serves an important role in human cognitive processes, and the cocktail party effect is a typical example [1], [5]. Simply speaking, selective listening is a result of attentional filtering that separates relevant from irrelevant stimuli [4]. Recently, there is an increasing interest in adopting the neural attention mechanism in deep learning approaches [48]. The idea is to model attentional modulation by assigning differentiated weights to the incoming data points in the time series at run-time dynamically. Such an attention mechanism is expected to bias the allocation of available resources toward the most informative part of a signal.

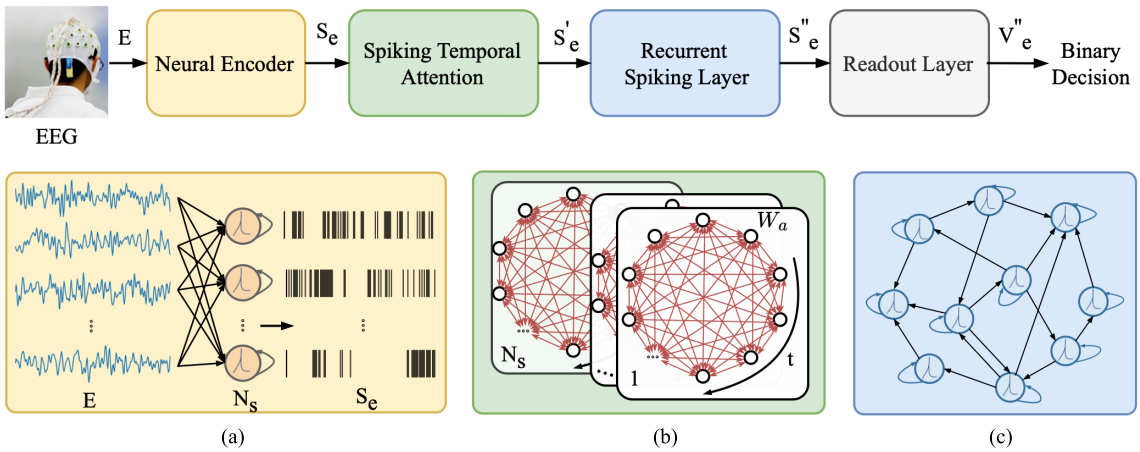

Fig. 2. Schematic of the proposed BSAnet mainly consists of three components: a neuron encoder, a spiking temporal attention module, and a recurrent spiking layer. Taking the EEG signal as input, the network is trained to detect auditory spatial attention via binary decisions. (a) Neural encoder. (b) Spiking temporal attention. (c) Recurrent spiking layer.

In view of the fact that neural responses to auditory stimuli are a temporal process [3], [4], [5], [6], [7], [8], a temporal attention mechanism that models the temporal progression of spike trains is desirable. Therefore, we design a spiking temporal attention mechanism that facilitates the modulation of spike trains by relating various EEG data points within a decision window. This mechanism can be implemented in three steps.

First, spike trains $\mathbf{S}_e$ are transformed into query $\mathbf{S}_q$, key $\mathbf{S}_k$, and value $\mathbf{S}_v$ via spiking linear projections

$$\begin{aligned}\mathbf{S}_q &= \eta(\mathbf{S}_e, \mathbf{W}_q, \mathbf{B}_q)\\ \mathbf{S}_k &= \eta(\mathbf{S}_e, \mathbf{W}_k, \mathbf{B}_k)\\ \mathbf{S}_v &= \eta(\mathbf{S}_e, \mathbf{W}_v, \mathbf{B}_v).\end{aligned} \tag{10}$$

Second, the relationship between the query and key is computed via a dot product

$$\mathbf{W}_a = \frac{\mathbf{S}_q^T \mathbf{S}_k}{d_s} \tag{11}$$

where $\mathbf{W}_a \in \mathbb{R}^{T\times T}$ is the temporal attention mask, as indicated in Fig. 2(b), which assigns different weights to the spiking trains $\mathbf{S}_v$ over time steps by $\mathbf{I}_a = \mathbf{S}_v\mathbf{W}_a$. The spike train $\mathbf{S}_a \in \mathbb{R}^{N_s\times T}$ can be obtained according to (5) and (6). d_s is a scale factor that keeps the weights, thus the number of spikes, within a reasonable range.

Finally, an attention modulated spiking feature representation $\mathbf{S}'_e \in \mathbb{R}^{N_s\times T}$ can be calculated as follows:

$$\mathbf{S}'_e = \eta(\mathbf{S}_a, \mathbf{W}_s, \mathbf{B}_s). \tag{12}$$

D. Recurrent Spiking Layer

Findings in psychoacoustic study and neuroscience suggest that the human auditory system processes incoming speech segment by segment, which facilitates temporal prediction [34], and cognition in a recurrent fashion [36], [37], [38], [39]. This inspires us to develop a recurrent spiking layer to decode the auditory activities manifested in the human brain.

While the spiking temporal attention answers where in time the model should look at, the recurrent spiking layer seeks to answer how the contextual history influences the cognitive prediction. As illustrated in Fig. 2(c), a recurrently connected layer is designed to process the modulated spike trains $\mathbf{S}'_e$. In this study, we adopt an LIF neuron driven by decaying synaptic currents generated by its synaptic afferent as a recurrent neuron [46], [49]

$$\frac{dI_i^l}{dt} = -\frac{I_i^l(t)}{\tau_s} + \sum_j w_{ij}^l S_j^{l-1}(t) + \sum_j v_{ij}^l S_j^l(t) \tag{13}$$

where τ_s is the synaptic time constant, v_{ij}^l denotes the recurrent weight from neuron j in layer l to neuron i in layer l.

Then, the discretization for the current of a recurrent neuron can be approximated as

$$I_i^l[t+1] = \beta I_i^l[t] + \zeta \tag{14}$$

with

$$\zeta = \sum_j w_{ij}^l S_j^{l-1}[t+1] + \sum_j v_{ij}^l S_j^l[t+1] + b_i^l \tag{15}$$

where $\beta = \exp(-\Delta t/\tau_s)$ is a leaking factor for the recurrent neuron.

Finally, we can obtain the output spiking feature representation from the recurrent layer, $\mathbf{S}''_e \in \mathbb{R}^{N_s\times T}$, by iterating (5), (6), and (14) over a number of time steps.

E. Optimization and Training

The neural architecture of BSAnet takes a window of EEG signal as input and outputs an attention decision via an SNN classifier, as shown in Fig. 2. BSAnet is trained in an end-to-end manner. The SNN classifier is a linear readout layer consisting of leaky integrator (LI) neurons that do not spike. It was suggested [46], [47] that the LI neuron provides a smooth learning curve, as it allows us to derive continuous error gradients from the continuous output. Therefore, we directly use the integrated membrane potential, V''_e, instead of binary spike trains for neural decoding in this study. The free aggregate membrane potential is averaged along time,

$\bar{V}''_e$, and the binary cross-entropy loss is used as the learning objective

$$\mathcal{L} = -\frac{1}{M}\sum_{m=1}^{M}\left[y_m \cdot \log P_m + (1-y_m)\cdot \log(1-P_m)\right] \quad (16)$$

with

$$P_m = \text{softmax}\left(\bar{V}''_e\right) \quad (17)$$

where y_m is the label of the mth decision window, P_m is its corresponding predicted probability, and M is the number of decision windows in a training batch.

To train a BSAnet model, we apply the backpropagation through time algorithm [46], with which the prediction error is back-propagated from the output to the input, including the previous input, to update the parameters by gradient descent. In view of the fact that (7) is nondifferentiable, the backward pass needs to be modified. We follow the derivative approximation method in [44] to have:

$$h(U_i^l) = \frac{1}{\sqrt{2\pi\rho}} e^{-\frac{(U_i^l - \upsilon_{\text{th}})^2}{2\rho}} \quad (18)$$

where ρ determines the curve steepness, i.e., the peak width. In this way, the network follows a step function as in (7) in the forward pass, while in the backward pass, it follows a Gaussian cumulative distribution function in (18).

During training, we use Kaiming uniform initialization [50] for the weights w_{ij}, the recurrent weights v_{ij} and bias b_i. The training algorithm of BSAnet is summarized in Algorithm 1.

Algorithm 1 Training Algorithm of BSAnet

Input: EEG signals $\mathbf{E}$, the class labels y_m corresponding to the EEG, and other model hyper-parameters

Output: The model parameters of BSAnet

1. Initialize the connection weights w_{ij}, the recurrent weights v_{ij}, and bias b_i using Kaiming uniform initialization.
2. Randomly initialize other model parameters in BSAnet.

Repeat

Forward Pass:

a. Encode EEG data $\mathbf{E}$ into spike trains $\mathbf{S}_e$ according to 9

b. Calculate the attention modulated spiking representation $\mathbf{S}'_e$ according to 10–12

c. Calculate the recurrent spiking layer and get the result $\mathbf{S}''_e$ according to 5, 6, and 14

d. Calculate the continuous free aggregate membrane potentials V''_e according to 4 and 5

e. Calculate the average free aggregate membrane potential $\bar{V}''_e$ at all time steps

f. Calculate the loss function according to 16

Backward Pass:

a. Update parameters by the back-propagation using the approximate derivative 18

until The iteration satisfies the predefined algorithm convergence condition

III. Experiments

A. Data Specifications

Experiments are carried out on two popular EEG datasets, which are hereafter referred to as KUL [51] and DTU [52] for short. In both datasets, 64 channels of signals from 64 electrodes were recorded following the international 10/20 protocol and by a BioSemi ActiveTwo system.

1) KUL Dataset [51]: The dataset was recorded from 16 individuals with normal hearing in an electromagnetically shielded and soundproof room, where the subjects listened to two competing speakers. They were instructed to pay attention to one speech stream and ignore another. The two speech streams were presented either 90° to the left or 90° to the right to the subjects. Each trial lasted 6 min, and each subject participated in eight trials. The EEG signals were sampled at 8192 Hz.

2) DTU Dataset [52], [53]: The EEG signals were taken from 18 subjects of normal hearing when the subjects attended to one target speaker in a competing acoustic environment. The two speech stimuli are narrated by one male and a female speaker. They were presented either 60° to the left or 60° to the right of the listening subject. Each trial last around 50 s, and each subject performed 60 trials. The arriving direction of the speech stimulus and the gender of the speaker were randomized across trials. The EEG signals were sampled at 512 Hz.

B. Data Preprocessing

The EEG data are first rereferenced to the average response of all channels. Previous studies suggest that nonlinear ASAD could benefit from a broader bandwidth [10], [13], [20], [30]. Therefore, all EEG data are bandpass filtered between 1 and 32 Hz, and downsampled to 128 Hz. It is worth noting that the proposed BSAnet is a data-driven approach that neither involves any manual removal of artifacts nor feature engineering. We believe the end-to-end learning offers a unique benefit to the implementation of neuro-steered hearing aids, as well as more generally BCI systems, where systems are required to adapt to a new working environment with minimum human supervision.

For rapid tracking of auditory attention, a short decision window with a low-latency response is preferred [19]. Humans can switch attention from one speaker to another within 2 s [14]. Therefore, we study four decision window sizes in this work, i.e., 0.25, 0.5, 1.0, and 2.0 s.

C. Network Configuration

The performance of the proposed model is evaluated using fivefold cross-validation (CV) in this study [54]. In accordance with previous studies [20], [28], [29], and [30], the ASAD accuracy is defined as the percentage of correctly classified decision windows on the test set. The average performance of the fivefold validation process is reported as the final results. All models in this study are implemented with PyTorch.

Taking 1-s decision window as an example, BSAnet configuration is described as follows. As the input to the end-to-end

TABLE I
PARAMETERS SET IN OUR EXPERIMENT

Network parameter	Description	Value
U_{th}	Membrane threshold	0.5mV
U_0	Resting potential	0
U_r	Reset potential	0
α	Decay factor	0.25
β	Decay factor	0.25
ρ	Derivative approximation parameters	0.15

model, 1-s EEG signals are denoted as $\mathbf{E} \in \mathbb{R}^{64\times128}$ with 64-channel and 128 samples. The neuron encoder converts the real-valued EEG inputs into spike trains $\mathbf{S}_e \in \mathbb{R}^{10\times128}$. Then, the spiking temporal attention mechanism modulates the spike trains $\mathbf{S}_e$ into $\mathbf{S}'_e \in \mathbb{R}^{10\times128}$. Specifically, the dimension of the weight matrices $\mathbf{W}_q$, $\mathbf{W}_k$, and $\mathbf{W}_v$ is 10×10. The scale factor d_s is 10. Then the recurrent spiking layer, which consists of a single layer of ten recurrent units, further modulates the spiking feature representation and outputs $\mathbf{S}''_e \in \mathbb{R}^{10\times128}$. Finally, a readout layer, which consists of two LI neurons, is used for making binary decisions. As the threshold-dependent batch normalization (tdBN) technique [55] effectively alleviates the problem of gradient vanishing or explosion in the SNN, a tdBN layer is added after each spiking linear layer, except the spiking temporal attention layer, i.e., Fig. 2(b). We set the batch size M to 32.

During training, the adaptive moment estimation (Adam) optimizer is implemented at the learning rate of 10^{-3}. Most of the model parameters are empirically chosen to be consistent with the biological findings [44], [45], [46], as shown in Table I. Additional hyper-parameters are chosen through a fivefold CV grid search.

In an ablation study, we also implement a reduced version of BSAnet by skipping the spiking temporal attention module, which is referred to as the *BSnet* hereafter.

IV. RESULTS

A. BSnet versus CNN Decoder

The BSnet and the CNN decoder [20] adopt different neuronal functions, i.e., spiking neuron versus rectified linear unit (ReLU); they also represent different neural architectures, i.e., recurrent spiking layer versus convolutional layer. By comparing the two networks, we hope to answer the question of whether recurrently connected spiking neurons are advantageous in decoding auditory attention.

The CNN module is composed of one convolution layer with the kernel size of 64×17, one average pooling layer, and two fully connected layers (Input: 5, hidden: 5, output: 2). It employs the ReLU activate function and the cross-entropy loss. We follow the CNN implementation that is available at [20], and fine-tune the hyperparameters in the same way as BSnet on both datasets.

As shown in Fig. 3, the CNN model attains the mean ASAD accuracy of 63.3% (standard deviation (SD): 5.96%) in the DTU dataset and 84.1% (SD: 10.16%) in the KUL dataset with 1-s decision window, respectively. BSnet, the reduced version of BSAnet, consistently outperforms CNN model by a large margin in both datasets. Specifically, BSnet achieves

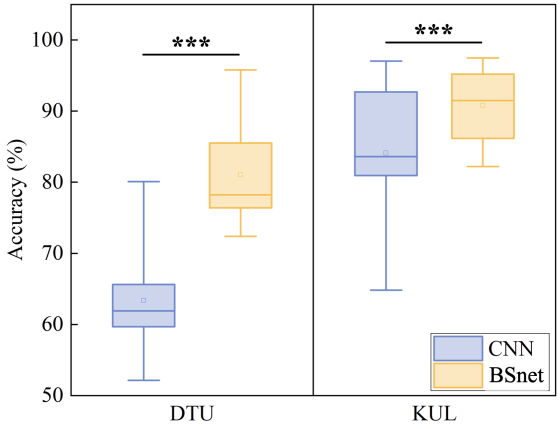

Fig. 3. ASAD accuracy of CNN and BSnet among all subjects in the DTU and KUL datasets with 1-s decision window. Statistically significant difference: $^{***}p < 0.001$.

an average improvement of 17.7% (mean: 81.0%, SD: 7.09%) or an error reduction of 48.2%, i.e., from 36.7% to 19.0%, on the DTU dataset, 6.7% (mean: 90.8%, SD: 5.05%) or an error reduction of 42.1%, i.e., from 15.9% to 9.2%, on the KUL dataset, respectively. Moreover, it is noteworthy that the number of parameters of the BSnet is markedly lower than that of the CNN model. Specifically, the CNN model approximately consists of 5500 parameters [20], whereas the proposed BSnet just has around 900 parameters. Given that the dataset size of EEG signals is usually limited, the high-representative and light-weighted BSnet provides a promising solution to EEG-based BCIs.

Statistical analyses were performed using IBM SPSS statistics software. Descriptive statistics were used for means and SDs. A level of significance of 0.05 was used in this study. Specifically, the average accuracy obtained by the BSnet significantly outperforms the CNN model (paired t-test: $p < 0.001$) in both datasets. These results support our hypothesis that BSnet effectively learns the temporal patterns of EEG signals and generates discriminative features for the ASAD tasks in a better way than CNN-based decoder.

B. Size of Decision Window

We now evaluate BSAnet with four decision windows, from 0.25 to 2 s, and report in Fig. 4.

On the KUL dataset, BSAnet obtains a relatively high ASAD accuracy of 95.2% (SD: 3.08) for 2 s and 93.7% (SD: 4.02) for 1 s of window. These results are consistent with the previous findings that decoding accuracy positively correlates with the decision window size [10], [19], [20], [25], [29], [30]. The ASAD accuracy degrades to 90.3% (SD: 5.19) for a window of 0.5 s. It is worth noting that BSAnet obtains an acceptable ASAD performance (mean: 84.7%, SD: 6.46) despite a very narrow window of 0.25 s or 250 ms.

On the DTU dataset, BSAnet attains an average accuracy of 75.3% (SD: 7.32) for a window of 0.25 s, 79.8% (SD: 6.99) for 0.5 s, 83.1% (SD: 6.75) for 1 s, and 85.6% (SD: 6.47) for 2 s. We observe that the performance on the KUL dataset is generally better than that on the DTU dataset.

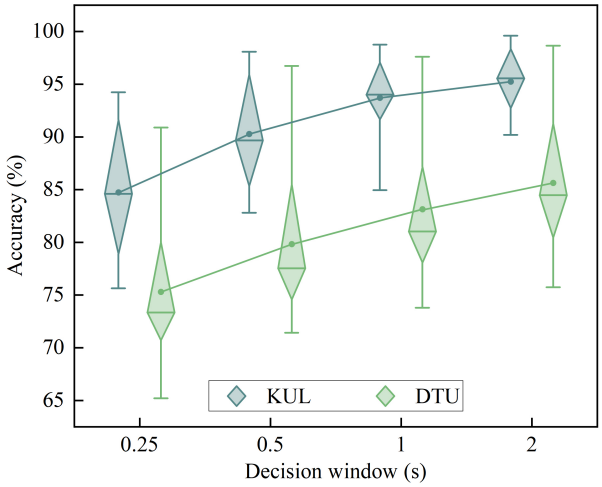

Fig. 4. ASAD performance of BSAnet for four different decision window sizes across all subjects in both DTU and KUL datasets.

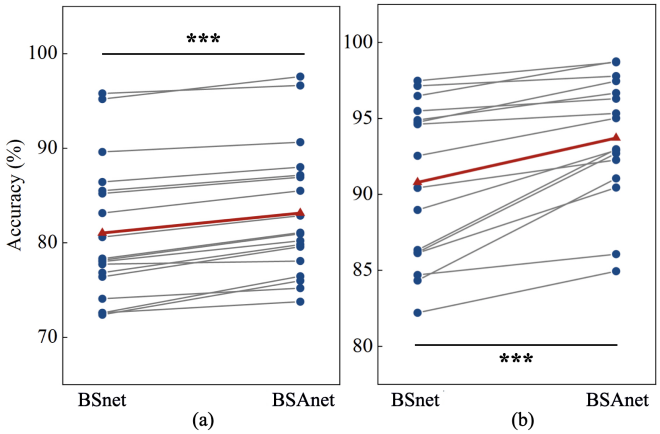

Fig. 5. ASAD accuracy of the BSnet and BSAnet with 1-s decision window on two datasets. (a) DTU dataset. (b) KUL dataset. Blue dots: individual results. Gray lines: same subjects. Red triangles: mean accuracies of all subjects. Statistically significant difference: $^{***}p < 0.001$.

This could be affected by the psychological and physiological characteristics of the individuals [56], [57], the acoustic content and the environment, and the physical layout of the experiments [51], [52].

As a narrow EEG window allows for decision at a high time resolution, BSAnet is therefore suitable for low-latency and real-time tracking of auditory attention in BCI, for instance, neuro-steered hearing aids.

C. Effect of Spiking Temporal Attention

To validate the effectiveness of the spiking temporal attention, i.e., the self-attention mechanism, we conduct ablation analysis on both datasets. As shown in Fig. 5, the ASAD accuracy of the BSnet and BSAnet are reported across all subjects with 1-s decision window.

On the KUL dataset, it is observed that the spiking temporal attention improves the decoding accuracy from 90.8% (SD: 5.05%) of BSnet to 93.7% (SD: 4.02%) of BSAnet for 1-s decision window. With the BSnet, half of the subjects achieved an average ASAD accuracy higher than 90%, while with BSAnet, 13 of 16 subjects did. Moreover, ASAD accuracy is significantly different between the BSnet and BSAnet (paired t-test: $p < 0.001$).

On the DTU dataset, we observe similar results. BSAnet achieves a relatively high ASAD accuracy (mean: 83.1%, SD: 6.75%), which significantly outperforms BSnet with an average improvement of 2.1% (paired t-test: $p < 0.001$).

The fact that BSAnet significantly outperforms BSnet on two publicly available datasets confirms that the proposed spiking temporal attention contributes to the performance gain by assigning differentiated weights dynamically to the data points in the spike trains during run-time inference.

V. Discussions

We hypothesize that it is advantageous to model the temporal dynamics of neural responses in the listening brain for AAD. To validate our hypothesis, we compare the ASAD performance and computational cost of the proposed BSAnet with other competing models in the literature. We also visualize the spiking feature representations at different stages of the BSAnet pipeline to understand how the processing modules operate.

A. Comparative Study

As summarized in Table II, we compare our models with several baselines on DTU [52] and KUL [51] datasets.

On the KUL dataset, we start by comparing our models with common spatial pattern (CSP)-based [19] and Riemannian geometry classifier (RGC)-based [25] ASAD models, which were reported with competitive results in low-latency settings. It is clear that both BSnet and BSAnet outperform CSP and RGC by a large margin. Specifically, BSnet gains an average accuracy increase of 10.4% and 10.2% across all decision window sizes over CSP-based and RGC-based models. BSAnet further improves the average accuracy by 13.0% and 12.8% over CSP and RGC, respectively.

These results, in line with the previous studies [10], [12], and [13], suggest that nonlinear approaches are advantageous for rapid and reliable decoding auditory attention (spatial) attention, especially with short decision windows. We then compare our models with other state-of-the-art nonlinear approaches. BSnet improves the accuracy with an increase of 6.1% and 3.4% across all decision window sizes over the CNN-based models [20], [29]. In addition, BSAnet further enhances the accuracy and significantly outperforms previous nonlinear models [20], [29], [30] ($p < 0.001$ for [20], $p = 0.002$, and [29], $p = 0.023$ for [30]), respectively.

Similar to the findings on the KUL dataset, the proposed BSnet significantly outperforms other nonlinear models on the DTU dataset across all decision window sizes ($p < 0.001$ for [20], $p = 0.002$, and [29], $p = 0.006$ for [30]). The proposed BSAnet further improves over CNN [20], CNN with channel mask (CNN-CM) [29], and spatiotemporal attention network (STAnet) [30] by a large margin, i.e., 18.7%, 13.8%, and 9.7% across all window sizes, respectively.

In summary, BSAnet outperforms all state-of-the-art models. We argue that the contributions come from the modeling of the temporal dynamics of EEG signals and more effective feature representation by the neural encoder.

TABLE II

ASAD PERFORMANCE COMPARISON OF DIFFERENT MODELS ON BOTH DATASETS FOR FOUR DIFFERENT DECISION WINDOW SIZES. NOTE THAT THE PROPOSED BSANET SIGNIFICANTLY OUTPERFORMS OTHER MODELS IN TERMS OF DECODING PERFORMANCE. CSP = COMMON SPATIAL PATTERN, RGC = RIEMANNIAN GEOMETRY CLASSIFIER

Dataset	Model	Decision window (second)			
		0.25	0.5	1	2
KUL [51]	CSP (Geirnaert et al. [19])	74.9%	77.6%	79.1%	80.3%
	RGC (Geirnaert et al. [25])	73.5%	78.1%	79.4%	81.6%
	CNN (Vandecappelle et al. [20])	78.4 ± 10.06%	80.6 ± 10.37%	84.1 ± 10.16%	86.2 ± 9.53%
	CNN-CM (Su et al. [29])	80.9 ± 8.29%	84.3 ± 8.56%	86.5 ± 7.99%	88.3 ± 7.89%
	STAnet (Su et al. [30])	84.4 ± 9.67%	87.2 ± 9.77%	90.1 ± 8.95%	91.4 ± 8.33%
	BSnet (This work)	82.2 ± 7.26%	87.4 ± 6.03%	90.8 ± 5.05%	93.1 ± 4.54%
	BSAnet (This work)	**84.7 ± 6.46%**	**90.3 ± 5.19%**	**93.7 ± 4.02%**	**95.2 ± 3.08%**
DTU [52]	CNN (Vandecappelle et al. [20])	58.9 ± 5.86%	61.7 ± 6.68%	63.3 ± 5.96%	65.2 ± 5.49%
	CNN-CM (Su et al. [29])	64.1 ± 6.78%	67.2 ± 7.74%	67.9 ± 7.41 %	69.5 ± 7.03%
	STAnet (Su et al. [30])	68.5 ± 6.92%	70.8 ± 8.04%	71.9 ± 8.94%	73.7 ± 9.59%
	BSnet (This work)	73.5 ± 7.42%	78.1 ± 7.11%	81.0 ± 7.09%	83.7 ± 6.78%
	BSAnet (This work)	**75.3 ± 7.32%**	**79.8± 6.99%**	**83.1 ± 6.75%**	**85.6 ± 6.47%**

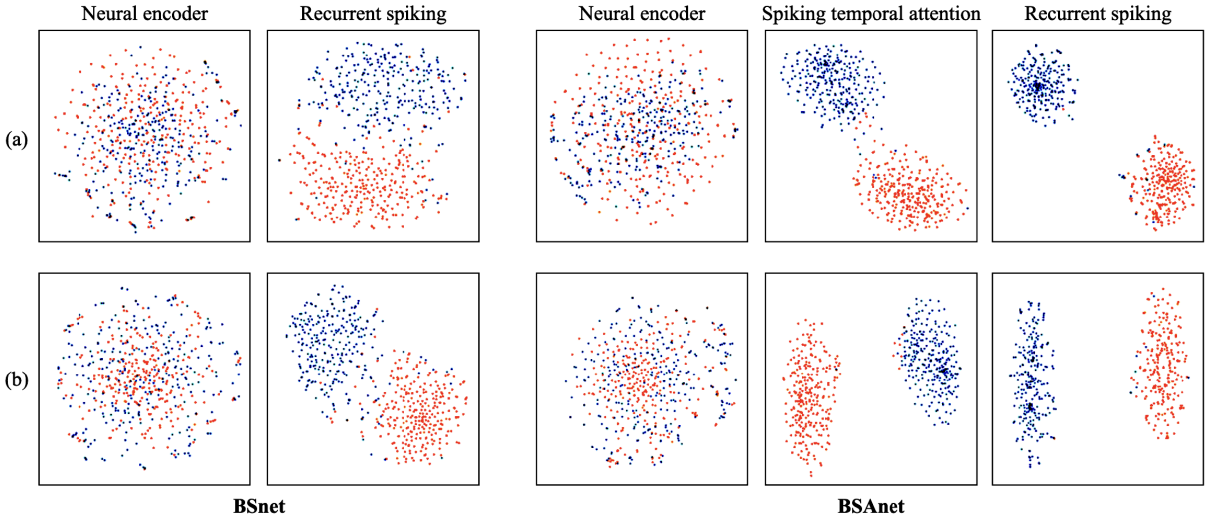

Fig. 6. T-SNE visualization of the feature representations of the proposed BSnet and BSAnet from two randomly selected subjects (a) and (b) in the KUL dataset. The colors denote the actual auditory attention, i.e., leftward or rightward.

B. Visualization of Feature Representation

Representation learning is commonly used in machine learning to automatically discover feature representations that are useful for downstream tasks. Typically the live activations in a neural architecture that respond to the input signals can be used as the feature representation of the input. The ability to interpret such representations, usually via visualization, allows us to gain insight into the research problem and build intuitions about how a neural architecture works. For example, a hierarchical 2-D feature representation learned for visual object recognition by a CNN at different layers captures different levels of abstract features ranging from edges and gradients to shapes and styles [58].

In the domain of auditory and speech perception, the temporal patterns in neural responses offer valuable insights into brain activities [3], [4], [5], [6], [7], [8]. Furthermore, it has been observed that attention-related neural activity in the human brain entrains rhythmically to the rhythmic stimulation present in the acoustic environment [59]. However, EEG signals are noisy because they reflect many concurrent brain activities, also being corrupted by electronic noises during signal acquisition. Effectively characterizing these temporal patterns remains a significant challenge.

It is found that binary spike trains, which capture spike time intervals, serve as salient features in representing information [60], [61]. As a departure from the ANN of analog neurons where a fixed dimension embedding is used as the feature representation, in this article, we advocate for representing the input in different processed forms of spike trains along the processing pipeline of the neural architecture, e.g., $\mathbf{S}_e$, $\mathbf{S}'_e$ as shown in Fig. 2. In other words, we consider the intermediate processed forms of spike trains as the feature representations of the input EEG signals.

We apply the t-stochastic neighbor embedding (t-SNE) visualization [62] to understand how our networks perform at different stages of the pipeline. As shown in Fig. 6, the feature representations of two randomly selected subjects from the KUL dataset are visualized for the BSnet and BSAnet. The leftward and rightward auditory attention samples are colored differently, which allows us to observe the distribution of the feature representations. It is apparent that the output feature distribution of the recurrent spiking layer in BSAnet is more

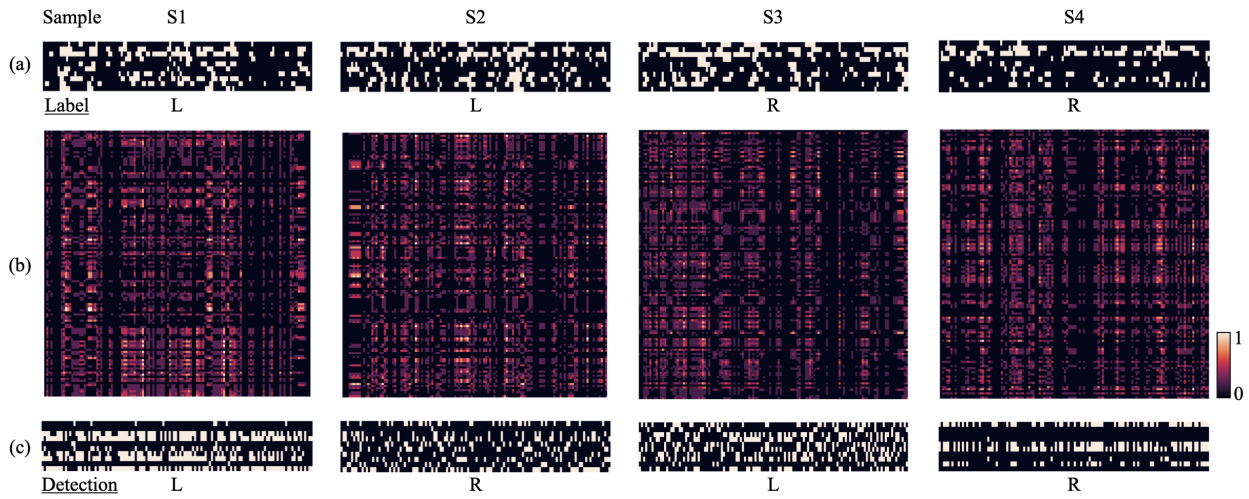

Fig. 7. Bio-plausible visualization of BSAnet for one randomly selected subject. (a) Spiking feature representations of four 1-s EEG windows, i.e., ten spike trains of 128 data points each, as the output of the neural encoder denoted as S1–S4. The "L/R" label denotes actual leftward or rightward auditory attention. The spiking events in the spiking trains are represented in white. (b) Attention masks, $\mathbf{W}_a$, are generated by the spiking temporal attention module for the leftward/rightward EEG samples. The color of the cells represents the weights with lighter colors corresponding to larger weights. (c) Spiking feature representations as the output of the spiking temporal attention module. The "L/R" indicates the detected leftward or rightward auditory attention by the BSAnet model.

discriminative than that of the BSnet. This further confirms the effectiveness of the spiking temporal attention.

We further visualize the feature representations of one randomly selected subject in Fig. 7. In Fig. 7(a), we illustrate the spike trains after the neural encoder in a 2-D display, $\mathbf{S}_e \in \mathbb{R}^{N_s \times T}$, where the y-axis represents the spiking neurons, and the x-axis represents the time index [63]. The biologically motivated temporal attention mechanism is expected to attend to salient data points in the EEG window relevant to the ASAD task. As illustrated in Fig. 7(b), the attention mask $\mathbf{W}_a$ assigns differentiated weights on the fly to the input EEG data points over time. Consistent with biological observations [57], leftward/rightward auditory stimuli are encoded into different spiking patterns by BSAnet.

We further visualize the optimized spiking representation of EEG signals, i.e., $\mathbf{S}'_e$, to understand the binary decision of BSAnet, as shown in Fig. 7(c).

Notably, we observe distinct and spatially clear spiking patterns for S1 and S4, indicating positive detection. Conversely, S2 and S3 exhibit noisy patterns, which correspond to negative detection. These results suggest a potential relationship between spiking patterns and AAD outcomes. Moreover, these findings highlight the vital role of temporal regularity in selective listening [35]. Additionally, the precise timing of individual neuronal spikes emerges as a crucial factor in facilitating prompt sensory responses in the human brain [6], [34], [35].

Overall, it is encouraging to see that BSAnet shows the temporal coding ability present in biological networks, which mimics the human auditory attention mechanism [3], [4], [5].

C. Computational Cost

Apart from accuracy, energy consumption is also a critical factor of consideration in system deployment. We further compare the proposed bio-inspired decoders, BSnet and BSAnet, with the CNN model [20] in terms of computational cost for various decision windows.

TABLE III
COMPUTATIONAL COST (PJ) COMPARISON OF DIFFERENT ASAD MODELS FOR DIFFERENT DECISION WINDOW SIZES

Model	Decision window (second)			
	0.25	0.5	1	2
CNN [20]	4.0+E5	1.2+E6	2.8+E6	6.0+E6
BSnet	9.6+E4	1.9+E5	3.8+E5	7.6+E5
BSAnet	1.0+E5	2.1+E5	4.5+E5	9.6+E5

We use the total number of floating point operations as the proxy of the calculation of computational cost, that is formulated in [64] based on the standard 45 nm CMOS process. As summarized in Table III, the computational cost of both BSnet and BSAnet is much lower than that of the CNN implementation across all window sizes. BSnet and BSAnet achieve an average computational cost reduction of 83.5%, and 81.4% over the basic CNN model, respectively. For 1-s decision window, BSnet and BSAnet achieve an improvement of 6.7% and 9.6% in terms of the ASAD performance, and a reduction of 86.4% and 83.9% in terms of the computational cost over the CNN model.

The SNNs, i.e., BSnet and BSAnet, provide rapid and accurate decoding of the auditory spatial attention owing to event-driven computation [65], where computation occurs only when a neuron fires. In addition, a spiking neuron only requires an accumulate operation for each input spike, while a standard artificial neuron requires a multiply-and-accumulate operation for each input.

Overall, the proposed bio-inspired decoders not only allow for bio-realistic information processing, but also offer tremendous energy-saving benefits for devices with limited resources, such as neuro-steered hearing aids.

D. Future Study

One challenge in the implementation of neuro-steered hearing aids is the complex and real-life acoustic environment, i.e., several spatially separated speakers of different relevance.

In this study, we have only studied a two-speaker scenarios [52], [53]. Extending this study to multispeaker acoustic environments could be an interesting direction for future research.

The precise neurophysiological mechanisms underpinning the cocktail party effect are still not well understood. While the proposed model draws inspiration from biological studies, the neural architecture and functions are primarily based on existing findings, leaving room for further exploration to achieve full biological plausibility. Moreover, neuronal oscillation also plays a significant functional role in speech processing, which has not been thoroughly explored yet. There is a need for further research and development to gain better insight and to consider these aspects in the study.

VI. Conclusion

We have proposed a novel neural architecture for the detection of auditory attention. The end-to-end BSAnet effectively exploits the temporal information of EEG signals, extracts discriminative EEG features, that leads to overall performance gain. BSAnet shows the state-of-the-art performance on publicly available datasets. Moreover, the computational cost of the model is much lower than that of the ANN-based decoders. We also introduce the idea of using intermediate spiking trains as temporal feature representations that suggests a new direction for EEG pattern classification. Overall, we develop a rapid, accurate AAD decoder, that is suitable for low-latency and real-time tracking of auditory attention. The proposed BSAnet is a leap forward toward achieving energy-efficient and bio-realistic neuro-steered hearing aids, which can also be extended in the development of other EEG-based BCI applications.

References

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *J. Acoust. Soc. Amer.*, vol. 25, no. 5, pp. 975–979, Sep. 1953.

[2] S. Van Eyndhoven, T. Francart, and A. Bertrand, "EEG-informed attended speaker extraction from recorded speech mixtures with application in neuro-steered hearing prostheses," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 5, pp. 1045–1056, May 2017.

[3] E. M. Z. Golumbic et al., "Mechanisms underlying selective neuronal tracking of attended speech at a 'cocktail party,"' *Neuron*, vol. 77, no. 5, pp. 980–991, 2013.

[4] S. Tune, M. Alavash, L. Fiedler, and J. Obleser, "Neural attentional-filter mechanisms of listening success in middle-aged and older individuals," *Nature Commun.*, vol. 12, no. 1, pp. 1–14, Jul. 2021.

[5] N. Mesgarani and E. F. Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature*, vol. 485, no. 7397, pp. 233–236, May 2012.

[6] L. Meyer, "The neural oscillations of speech processing and language comprehension: State of the art and emerging mechanisms," *Eur. J. Neurosci.*, vol. 48, no. 7, pp. 2609–2621, Oct. 2018.

[7] N. Ding and J. Z. Simon, "Emergence of neural encoding of auditory objects while listening to competing speakers," *Proc. Nat. Acad. Sci. USA*, vol. 109, no. 29, pp. 11854–11859, Jul. 2012.

[8] J. A. O'Sullivan et al., "Attentional selection in a cocktail party environment can be decoded from single-trial EEG," *Cerebral Cortex*, vol. 25, no. 7, pp. 1697–1706, Jul. 2015.

[9] O.-Y. Kwon, M.-H. Lee, C. Guan, and S.-W. Lee, "Subject-independent brain–computer interfaces based on deep convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 10, pp. 3839–3852, Oct. 2020.

[10] S. Geirnaert et al., "Electroencephalography-based auditory attention decoding: Toward neurosteered hearing devices," *IEEE Signal Process. Mag.*, vol. 38, no. 4, pp. 89–102, Jul. 2021.

[11] A. de Cheveigné, D. D. E. Wong, G. M. D. Liberto, J. Hjortkjær, M. Slaney, and E. Lalor, "Decoding the auditory brain with canonical component analysis," *NeuroImage*, vol. 172, pp. 206–216, May 2018.

[12] G. Ciccarelli et al., "Comparison of two-talker attention decoding from EEG with nonlinear neural networks and linear methods," *Sci. Rep.*, vol. 9, no. 1, pp. 1–10, Aug. 2019.

[13] T. Taillez, B. Kollmeier, and B. T. Meyer, "Machine learning for decoding listeners' attention from electroencephalography evoked by continuous speech," *Eur. J. Neurosci.*, vol. 51, no. 5, pp. 1234–1241, Mar. 2020.

[14] S. Cai, E. Su, Y. Song, L. Xie, and H. Li, "Low latency auditory attention detection with common spatial pattern analysis of EEG signals," in *Proc. Interspeech*, Oct. 2020, pp. 2772–2776.

[15] S. Cai, P. Li, E. Su, and L. Xie, "Auditory attention detection via cross-modal attention," *Frontiers Neurosci.*, vol. 15, Jul. 2021, Art. no. 652058.

[16] S. Cai, E. Su, L. Xie, and H. Li, "EEG-based auditory attention detection via frequency and channel neural attention," *IEEE Trans. Human-Mach. Syst.*, vol. 52, no. 2, pp. 256–266, Apr. 2022.

[17] S. Geirnaert, T. Francart, and A. Bertrand, "Unsupervised self-adaptive auditory attention decoding," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 10, pp. 3955–3966, Oct. 2021.

[18] M. Hosseini, L. Celotti, and É. Plourde, "End-to-end brain-driven speech enhancement in multi-talker conditions," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 1718–1733, 2022.

[19] S. Geirnaert, T. Francart, and A. Bertrand, "Fast EEG-based decoding of the directional focus of auditory attention using common spatial patterns," *IEEE Trans. Biomed. Eng.*, vol. 68, no. 5, pp. 1557–1568, May 2021.

[20] S. Vandecappelle, L. Deckers, N. Das, A. H. Ansari, A. Bertrand, and T. Francart, "EEG-based detection of the locus of auditory attention with convolutional neural networks," *eLife*, vol. 10, Apr. 2021, Art. no. e56481.

[21] W. Biesmans, N. Das, T. Francart, and A. Bertrand, "Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 5, pp. 402–412, May 2017.

[22] S. Geirnaert, T. Francart, and A. Bertrand, "An interpretable performance metric for auditory attention decoding algorithms in a context of neuro-steered gain control," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 1, pp. 307–317, Jan. 2020.

[23] F. Fahimi, S. Dosen, K. K. Ang, N. Mrachacz-Kersting, and C. Guan, "Generative adversarial networks-based data augmentation for brain–computer interface," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 9, pp. 4039–4051, Sep. 2021.

[24] S. K. Khare and V. Bajaj, "Time–frequency representation and convolutional neural network-based emotion recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 7, pp. 2901–2909, Jul. 2021.

[25] S. Geirnaert, T. Francart, and A. Bertrand, "Riemannian geometry-based decoding of the directional focus of auditory attention using EEG," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 1115–1119.

[26] Z. Gao et al., "EEG-based spatio-temporal convolutional neural network for driver fatigue evaluation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 9, pp. 2755–2763, Sep. 2019.

[27] J.-S. Bang, M.-H. Lee, S. Fazli, C. Guan, and S.-W. Lee, "Spatio-spectral feature representation for motor imagery classification using convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 7, pp. 3038–3049, Jul. 2022.

[28] S. Cai, P. Sun, T. Schultz, and H. Li, "Low-latency auditory spatial attention detection based on spectro-spatial features from EEG," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 5812–5815.

[29] E. Su, S. Cai, P. Li, L. Xie, and H. Li, "Auditory attention detection with EEG channel attention," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 5804–5807.

[30] E. Su, S. Cai, L. Xie, H. Li, and T. Schultz, "STANet: A spatiotemporal attention network for decoding auditory spatial attention from EEG," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 7, pp. 2233–2242, Jul. 2022.

[31] J. L. Lobo, J. Del Ser, A. Bifet, and N. Kasabov, "Spiking neural networks and online learning: An overview and perspectives," *Neural Netw.*, vol. 121, pp. 88–100, Jan. 2020.

[32] L. Deng et al., "Rethinking the performance comparison between SNNS and ANNS," *Neural Netw.*, vol. 121, pp. 294–307, Jan. 2020.

[33] D. Borra, S. Fantozzi, and E. Magosso, "Interpretable and lightweight convolutional neural network for EEG decoding: Application to movement execution and imagination," *Neural Netw.*, vol. 129, pp. 55–74, Sep. 2020.

[34] A.-L. Giraud and D. Poeppel, "Cortical oscillations and speech processing: Emerging computational principles and operations," *Nature Neurosci.*, vol. 15, no. 4, pp. 511–517, Apr. 2012.
[35] A. Zalta, S. Petkoski, and B. Morillon, "Natural rhythms of periodic temporal attention," *Nature Commun.*, vol. 11, no. 1, pp. 1–12, Feb. 2020.
[36] R. J. Douglas and K. A. C. Martin, "Recurrent neuronal circuits in the neocortex," *Current Biol.*, vol. 17, no. 13, pp. R496–R500, Jul. 2007.
[37] I. Diez and J. Sepulcre, "Neurogenetic profiles delineate large-scale connectivity dynamics of the human brain," *Nature Commun.*, vol. 9, no. 1, p. 3876, Sep. 2018.
[38] K. Friston, "A theory of cortical responses," *Phil. Trans. Roy. Soc. B, Biol. Sci.*, vol. 360, no. 1456, pp. 815–836, Apr. 2005.
[39] M. Pefkou, L. H. Arnal, L. Fontolan, and A.-L. Giraud, "Θ-band and β-band neural activity reflects independent syllable tracking and comprehension of time-compressed speech," *J. Neurosci.*, vol. 37, no. 33, pp. 7930–7938, Aug. 2017.
[40] S. A. Shamma, M. Elhilali, and C. Micheyl, "Temporal coherence and attention in auditory scene analysis," *Trends Neurosci.*, vol. 34, no. 3, pp. 114–123, Mar. 2011.
[41] S.-C. Liu, A. van Schaik, B. A. Minch, and T. Delbruck, "Asynchronous binaural spatial audition sensor with $2\times64\times4$ channel output," *IEEE Trans. Biomed. Circuits Syst.*, vol. 8, no. 4, pp. 453–464, Aug. 2014.
[42] S. S. Radhakrishnan, A. Sebastian, A. Oberoi, S. Das, and S. Das, "A biomimetic neural encoder for spiking neural network," *Nature Commun.*, vol. 12, no. 1, p. 2143, Apr. 2021.
[43] A. N. Burkitt, "A review of the Integrate-and-fire neuron model: I. Homogeneous synaptic input," *Biol. Cybern.*, vol. 95, no. 1, pp. 1–19, Jul. 2006.
[44] Y. Wu, L. Deng, G. Li, J. Zhu, and L. Shi, "Spatio-temporal backpropagation for training high-performance spiking neural networks," *Frontiers Neurosci.*, vol. 12, p. 331, May 2018.
[45] N. Perez-Nieves, V. C. H. Leung, P. L. Dragotti, and D. F. M. Goodman, "Neural heterogeneity promotes robust learning," *Nature Commun.*, vol. 12, no. 1, pp. 1–9, Oct. 2021.
[46] B. Cramer, Y. Stradmann, J. Schemmel, and F. Zenke, "The Heidelberg spiking data sets for the systematic evaluation of spiking neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 7, pp. 2744–2757, Jul. 2022.
[47] 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 Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 1, pp. 446–460, Jan. 2023.
[48] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[49] G. Bellec, D. Salaj, A. Subramoney, R. Legenstein, and W. Maass, "Long short-term memory and learning-to-learn in networks of spiking neurons," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 795–805.
[50] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.
[51] N. Das, T. Francart, and A. Bertrand, *Auditory Attention Detection Dataset KULeuven*, document Version 1.1.0, Aug. 2020, doi: 10.5281/zenodo.3997352.
[52] S. A. Fuglsang, D. D. Wong, and J. Hjortkjær, "EEG and audio dataset for auditory attention decoding," Mar. 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1199011
[53] S. A. Fuglsang, T. Dau, and J. Hjortkjær, "Noise-robust cortical tracking of attended speech in real-world acoustic scenes," *NeuroImage*, vol. 156, pp. 435–444, Aug. 2017.
[54] P. Refaeilzadeh, L. Tang, H. Liu, L. Angeles, and C. D. Scientist, "Cross-validation," *Encyclopedia Database Syst.*, vol. 5, pp. 532–538, Jan. 2020.
[55] H. Zheng, Y. Wu, L. Deng, Y. Hu, and G. Li, "Going deeper with directly-trained larger spiking neural networks," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 12, May 2021, pp. 11062–11070.
[56] I. Choi, L. Wang, H. Bharadwaj, and B. Shinn-Cunningham, "Individual differences in attentional modulation of cortical responses correlate with selective attention performance," *Hearing Res.*, vol. 314, pp. 10–19, Aug. 2014.
[57] Y. Deng, I. Choi, and B. Shinn-Cunningham, "Topographic specificity of alpha power during auditory spatial attention," *NeuroImage*, vol. 207, Feb. 2020, Art. no. 116360.
[58] B. Zhou, D. Bau, A. Oliva, and A. Torralba, "Interpreting deep visual representations via network dissection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2131–2145, Sep. 2019.
[59] I. C. Fiebelkorn and S. Kastner, "A rhythmic theory of attention," *Trends Cognit. Sci.*, vol. 23, no. 2, pp. 87–101, Feb. 2019.
[60] L.-V. Andreou, M. Kashino, and M. Chait, "The role of temporal regularity in auditory segregation," *Hearing Res.*, vol. 280, nos. 1–2, pp. 228–235, Oct. 2011.
[61] N. Ding and J. Z. Simon, "Adaptive temporal encoding leads to a background-insensitive cortical representation of speech," *J. Neurosci.*, vol. 33, no. 13, pp. 5728–5735, Mar. 2013.
[62] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.
[63] A. Bednar, F. M. Boland, and E. C. Lalor, "Different spatio-temporal electroencephalography features drive the successful decoding of binaural and monaural cues for sound localization," *Eur. J. Neurosci.*, vol. 45, no. 5, pp. 679–689, Mar. 2017.
[64] M. Horowitz, "Computing's energy problem (and what we can do about it)," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2014, pp. 10–14.
[65] B. Han, A. Sengupta, and K. Roy, "On the energy benefits of spiking deep neural networks: A case study," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 971–976.

Siqi Cai (Member, IEEE) received the Ph.D. degree from the Department of Shien-Ming Wu School of Intelligent Engineering, South China University of Technology, Guangzhou, China, in 2020.

She is currently a Research Fellow at the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. Her research interests include brain-computer interface, biosignal processing, and robotics.

Dr. Cai has served as the Local Arrangement Chair of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialog (SIGDial) 2021 and the Workshop Chair of the 47th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2022.

Peiwen Li received the B.E. degree from the School of Mechanical and Automotive Engineering, South China University of Technology, Guangzhou, China, in 2019, and the M.Sc. degree from the Department of Shien-Ming Wu School of Intelligent Engineering, South China University of Technology, in 2023.

Haizhou Li (Fellow, IEEE) is a Presidential Chair Professor and an Executive Dean at the School of Data Science, The Chinese University of Hong Kong, Shenzhen, China. He is also with the Department of Electrical and Computer Engineering, National University of Singapore (NUS), Singapore. His research interests include automatic speech recognition, natural language processing, and neuromorphic computing.

Prof. Li is a fellow of the Academy of Engineering Singapore. He was a recipient of the President's Technology Award 2013. He was named a Nokia Visiting Professor in 2009, ISCA Fellow in 2018, and Bremen Excellence Chair Professor in 2019. He was the General Chair of the 50th Annual Meeting of the Association for Computational Linguistics (ACL) 2012, INTERSPEECH 2014, IEEE ASRU 2019, and the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2022. He was the Editor-in-Chief of the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING from 2015 to 2018, the President of the International Speech Communication Association from 2015 to 2017, the President of Asia Pacific Signal and Information Processing Association from 2015 to 2016, the President of the Asian Federation of Natural Language Processing from 2017 to 2018, and the Vice President of the Conferences of IEEE Signal Processing Society from 2024 to 2026.