A Neural-Inspired Architecture for EEG-Based Auditory Attention Detection

Siqi Cai, *Member, IEEE*, Peiwen Li, Enze Su, Qi Liu, *Member, IEEE*, and Longhan Xie, *Member, IEEE*

***Abstract*—Humans have the ability to focus on one of the sound sources in a noisy scene, which is critical for everyday communication. Auditory attention detection (AAD) seeks to detect selective attention from one's brain signals. For AAD to be useful in brain–computer interface applications, new approaches with low computational cost, high classification performance, and low latency are required to be developed. In this study, we proposed a novel neural-inspired architecture to mimic the neural computation and coding strategy in the brain for electroencephalography-based AAD. We validated our model through data visualization, and conducted experiments on two publicly available databases. For both KUL and DTU databases, it outperforms both linear and convolutional neural network (CNN) models with consistent improvements from 1 s to 5 s decision windows in terms of detection accuracy. Although the accuracy of the proposed neural-inspired model is inferior to the state-of-the-art spatio-spectral feature (SSF)-CNN model, the computational cost of our model is less than 1% of SSF-CNN's. Moreover, the neural-inspired decoder is more hardware friendly and energy-efficient due to its biological computing scheme. Overall, the proposed neural-inspired architecture realizes a fast, accurate, and low energy expenditure AAD, which is a big step forward towards practical neuro-steered hearing aids.**

***Index Terms*—Auditory attention, brain–computer interface (BCI), electroencephalography, neural-inspired architecture.**

I. Introduction

THE ability to selectively attend to speech in so-called "cocktail party" scenarios is critical for everyday communication [1]. However, millions of people around the world with hearing loss struggle to listening under such conditions [2]. Modern hearing devices have been developed to produce better performance by using noise suppression systems. Nevertheless, the selection of the attended speaker is still a fundamental problem in cocktail party environments. Recent research in neuroscience has demonstrated that the selective auditory attention can be decoded using recordings of brain activity, such as electrocorticography (ECoG) [3], magnetoencephalography (MEG) [4], [5] and electroencephalography (EEG) [6]–[10]. These findings open up new opportunities to develop a new generation hearing aid that extracts the attention-related information directly from the brain, and then enhances the target speaker, i.e., neuro-steered hearing aids.

Inspired by this new insight, many auditory attention detection (AAD) algorithms have been developed. Specifically, AAD tackles the challenge of detecting which speaker is attended by the subject in a multispeaker scenario. The most common approach to AAD, which is known as stimulus reconstruction, focuses on decoding which speech envelope corresponds to the attended speaker [6]. Neural activities are used to approximate the envelope of the speech heard by the subject, that is then compared with the original speech envelopes. The speaker with a higher correlation coefficient is determined as the attended speaker. Different variations of the stimulus reconstruction algorithm have been proposed to improve the AAD performance [8], [11]–[14]. However, the stimulus reconstruction based AAD decoders still suffer from the following limitations.

1) The temporal resolution of stimulus reconstruction approach for reliable AAD is on the order of ∼10 s, which is not practical for real-time BCI applications [15]–[17]. There is a tradeoff between real-time operation and performance of the attentional state estimates [18]. The major reason is the stochastic fluctuations and uncertainties in correlations between the reconstructed and the original speech envelopes when computing over smaller windows of length [19]. Humans can switch attention from one speaker to another one at a temporal resolution of ∼1 s [11]. It remains a challenge for stimulus reconstruction based AAD models to accurately detect auditory attention at such a high temporal resolution.

Manuscript received July 5, 2021; revised March 5, 2022 and May 12, 2022; accepted May 13, 2022. Date of publication June 1, 2022; date of current version July 14, 2022. This work was supported in part by the Guangdong Provincial Key Laboratory of Big Data Computing, The Chinese University of Hong Kong, Shenzhen under Grant B10120210117-KP02, in part by the Guangdong Provincial Key Laboratory of Human Digital Twin under Grant 2022B1212010004, and in part by the IAF, A*STAR, SOITEC, NXP and National University of Singapore under FD-fAbrICS: Joint Lab for FD-SOI Always-on Intelligent and Connected Systems under Grant I2001E0053. The work of Siqi Cai was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (University Allowance, EXC 2077, University of Bremen, Germany). The work of Longhan Xie was supported by the National Natural Science Foundation of China under Grant 52075177. This article was recommended by Associate Editor T. H. Falk. *(Siqi Cai and Peiwen Li contributed equally to this work.) (Corresponding authors: Qi Liu; Longhan Xie.)*

Siqi Cai and Qi Liu are with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore (e-mail: 2447379329@qq.com; drliuqi@scut.edu.cn).

Peiwen Li, Enze Su, and Longhan Xie are with the Shien-Ming Wu School of Intelligent Engineering, South China University of Technology, Guangzhou, Guangdong 510460, China (e-mail: lintean@qq.com; enzesu@hotmail.com; melhxie@scut.edu.can).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/THMS.2022.3176212.

Digital Object Identifier 10.1109/THMS.2022.3176212

2) Most stimulus reconstruction approaches assume the availability of clean speech streams to perform AAD, which limits their applicabilities in the real world [20]. Although some studies have integrated the speech separation/extraction algorithms [21]–[23] to obtain the demixing speech envelopes into AAD systems, it significantly increases the computational complexity and adds lots of overheads [20]. Moreover, the demixing process often has negative effects on decoding accuracy or even makes the AAD system fail in real-life situations.

Motivated by the findings that locus of the auditory attention is neurally encoded [24]–[29], we hypothesize that we can decode the spatial location of the attended speaker from brain activities. This paradigm can realize AAD without access to the speech stimulus envelopes, which makes it be a highly competitive candidate for neuro-steered hearing aids [20]. Hence, it has the potential to achieve advanced performance at low-latency settings. Vandecappelle *et al.* [30] have firstly decoded the location of auditory attention based on EEG signals using a convolutional neural network (CNN), which achieves a relatively high AAD accuracy of 80.8% for 1 s decision window. Unfortunately, this approach does not benefit from feature representation as they directly take the $C \times T$-dimensional matrix, where C and T, respectively, denote the total number of EEG channels and time duration, as the input. Considering that the topographic specificity of alpha power (8 – 13 Hz) indicates the direction of auditory attention to speech [24], [25], [27], [28], we developed a novel spatio-spectral feature (SSF) representation method that retains task-related information in a pilot study [31] and outperforms the up-to-date models in AAD tasks [20], [30], [32].

Moreover, feature extraction is of particular importance for BCIs due to the low signal-to-noise ratio of EEG signals. In this study, we employed the neural model of Izhikevich [33] to learn and extract discriminative features for EEG-based AAD tasks, in which the EEG signals are processed biologically realistically. This prompts us to look for spiking neural networks (SNNs) with more biologically plausible spiking neurons. SNNs, as the third generation of artificial neural networks (ANNs) that more closely mimic biological neural functionality by processing information with sparse and asynchronous binary spikes (namely, events) over multiple time-steps [34]. ANNs succeed in tackling complex cognitive tasks, in which the neurons receive, process, and transmit analog information [35]. They, however, ignore the fact that mammalian brains process binary spike-based information using biological neurons. Contrary to traditional ANNs, SNNs carry information only when an action potential, that is, an intrinsic neuron along with its membrane electrical charge reaches a predefined threshold. Hence, the neuron fires spikes to carry information for subsequent neurons, which, in turn, decrease or increase their potentials in response to these input spikes.

Additionally, SNNs have shown favorable properties on neuromorphic hardware implementation, including low power consumption, massive parallelism, and on-chip learning, etc [34]. This makes them coincide with on-going interest toward real-world smart applications conditioned by limited hardware resources, such as mobile and wearable devices. Considering that EEG is essentially dynamic, and nonlinear time series signals, SNNs are designed by temporal coding approaches in isolating temporal characteristics of brain activities during different states, and also offer the prospect of event-driven hardware operation including the inherently biologically plausibility [36].

To this end, we developed a neural-inspired approach for EEG-based AAD task in this study, which is referred to as NI-AAD hereafter. The proposed NI-AAD method can detect the auditory spatial attention based on EEG alone, without the need of clean speech envelopes. To the best of our knowledge, this research is the first application of spiking neuron model to the EEG classification problem for AAD. The remainder of this article is organized as follows. In Section II, we formulate the proposed AAD pipeline, followed by the data processing and experimental setup in Section III. In Section IV, we report the experimental results and discuss the findings. Finally, Section V concludes this article.

II. Neural-Inspired AAD

Considering to formulate the EEG-based AAD as a binary classification problem in a two-speaker scenario [14], [30], [37], the proposed NI-AAD model consists of a spatio-spectral EEG feature representation, a spiking encoder for EEG feature extraction, and an SNN decoder for classification, as shown in Fig. 1. The advanced feature representation and spiking encoder are expected to extract spatial and spectral discriminative characteristics of raw EEG signals, and improve the AAD performance. Finally, an SNN decoder serves as a binary back-end classifier for decision making.

A. Spatio-Spectral Feature Representation

Previous studies have demonstrated that the topographic distribution of alpha power changed with the spatial focus of auditory attention [27], [28]. To improve the decoding performance, it makes sense to preserve the spatial and spectral information of EEG signals in feature representation from an AAD pipeline. As shown in Fig. 1, we employed the SSF representation method to extract the topographic specificity of alpha power from original EEG data [31].

First, a fast Fourier transform (FFT) is employed to calculate the power spectrum of EEG data. The average squared absolute value in the α-band (8–13 Hz) is used as the individual measurement value of each EEG channel. Second, we convert these measurements of different decision windows into a sequence of 2-D images to take full advantage of the spatial features of EEG signals. Specifically, the locations of EEG channels are projected from the 3-D space to a 2-D plane using Azimuth equidistant projection [38]. Moreover, the Clough–Tocher interpolant [39] is exploited to estimate the values in-between the electrodes over a 32×32 mesh. Thus, a topographical activity image of EEG can be generated to depict the α-band within a time window. And the sequence of EEG images derived from consecutive time windows is capable of reflecting the temporal information, which is taken as the input to the subsequent spiking encoder.

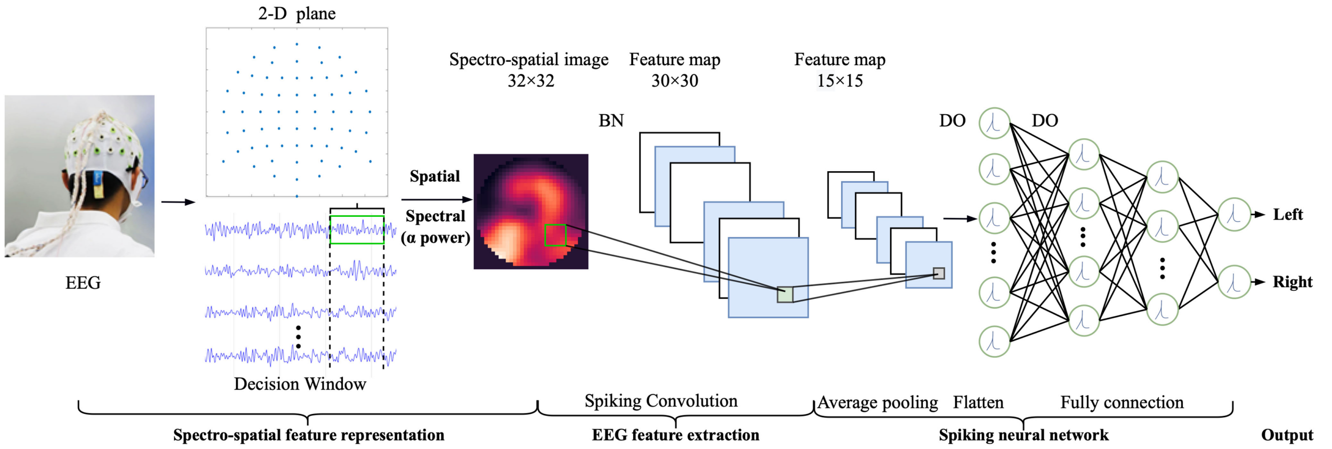

Fig. 1. Proposed neural-inspired architecture for AAD, that is referred to as the NI-AAD model. It includes three modules, a spatio-spectral EEG feature representation module, an EEG feature extraction module, and an SNN classifier. BN = batch normalization, DO = dropout.

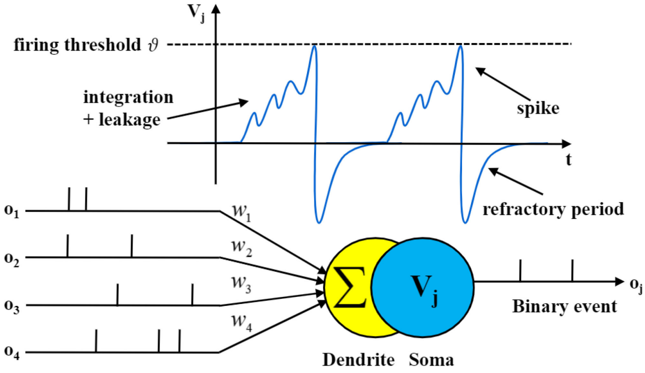

Fig. 2. Model of LIF spiking neuron. Time course of the membrane potential of an LIF neuron is driven by several spike trains $\mathbf{o}_i$. The LIF neuron is composed of Dendrite and Soma. The dendrite plays the role of collecting signals from other neurons and transmitting them to the "central processing unit," called soma, which performs the function of non-linear processing. When the sum of total input exceeds a certain threshold ϑ, an output spike is generated and then delivered to other neurons by the axon. After firing a spike, the neuron stays in the refractory period.

B. EEG Feature Extraction

SNNs are designed to process inputs that are represented as spike trains, which ideally are generated by event-based sensors [40]. Therefore, EEG images generated by the SSF representation method need to be encoded before they are fed into an SNN model. In this section, we first introduce the spiking neuron model that is adopted in our study. We then present the spiking encoder to transform the real-valued samples into discriminative features in the spiking domain.

1) Spiking Neuron Model: Different spiking neuron models have been developed to mimic the natural computing in the brain [41]. The Leaky Integrate-and-Fire (LIF) model is widely used in computational neuroscience for its relatively good biological realism and low computational cost [33]. In this work, a collection of LIF neurons formed the spiking encoder.

The LIF neuron model is introduced by the concept of membrane potential V_j, as shown in Fig. 2. The membrane potentials of the presynaptic neurons contribute to the postsynaptic neurons by the positive correlation with the firing time of presynaptic spikes. More concretely, at time-step t, the membrane potential of neuron j in layer l is formulated by

$$V_j^l[t] = \lambda V_j^l[t-1] + I_j^l[t] - \vartheta o_j^l[t-1] \tag{1}$$

with

$$I_j^l[t] = \sum_i w_{ji} o_i^{l-1}[t] + b_j^l \tag{2}$$

where λ is a leak factor and ϑ denotes the firing threshold. $I_j^l[t]$ indicates current contributions from presynaptic neurons to the neuron j. w_{ji} is the connection weight between presynaptic neuron i and postsynaptic neuron j. b_j^l denotes the constant injecting current to the neuron j. The spikes generated by neurons are defined as follows:

$$o_j^l[t] = \begin{cases} 1, & \text{if} V_j^l[t] > \vartheta \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

The event, viz., spike, is triggered if the membrane potential exceeds the firing threshold ϑ at t, in general, $\vartheta = 1$. The membrane potential $V_j[t]$ is reset to the rest potential V_{rest} after firing, and stays at the refractory period for a period of time.

2) Spiking Encoder: Generally, effective feature extraction is a crucial step for pattern classification. It is the data transformation from a high-dimensional space into a low-dimensional one, and therefore, the low-dimensional representation retains some meaningful, underlying properties from the raw data, ideally close to its intrinsic dimension. Motivated by a generalizable solution to dimensionality reduction, we proposed a spiking encoder with a spiking convolutional layer, as shown in Fig. 1. Compared to the conventional CNNs, the proposed spiking convolution extracts EEG features with spiking events. In this study, we take the real-valued inputs as the time-dependent input currents and directly apply them in (1) at the first time-step.

The spike count from neuron i at layer l can be computed as follows:

$$c_i^l = \sum_{t=1}^{N_t} o_i^l[t]. \tag{4}$$

where N_t is the length of spike trains $o_i^l[t]$ and referred to as the encoding time window hereafter. In practice, the discrete spike counts accumulated over an encoding time window N_t will be used for decoding.

C. Spiking Neural Network Classifier

CNNs make use of "convolution" and "pooling" techniques to reduce a large amount of input data into their essential features for effective classification. Nevertheless, with the increase of wider and deeper neural networks, the power consumption of such networks becomes demanding. Moreover, it is a long-standing problem in computational neuroscience to understand how the plasticity dynamics are organized in multilayer (deep) biological neural networks to achieve efficient data-driven learning [42]. In general, the success of previous CNN models for EEG-based AAD tasks [30], [31] gives rise to questioning whether the factors to their success are compatible with their biological ingredients, viz., SNNs.

As shown in Fig. 1, an SNN decoder is applied to decode the auditory spatial attention. It contains three hidden layers, two of which consist of LIF spiking neurons. The first and second neural layers are composed of 512 and 32 spiking neurons, respectively. Finally, an SNN back end is applied to decode the output spike counts of the second layer into pattern classes (i.e., the spatial focus of auditory attention).

III. Experiments

A. Data Specifications

In this study, experiments were carried out on two publicly available AAD databases, namely, KUL [43] and DTU databases [44].

1) *KUL database:* 64-channel EEG is recorded from 16 normal-hearing subjects while they are instructed to pay selective attention to one of two competing speakers. EEG is recorded at a sample rate of 8192 Hz using a BioSemi ActiveTwo system. Four Dutch short stories, narrated by different male speakers, are used as stimuli. The auditory stimuli are presented from 90° to the left and 90° to the right of the subject. Throughout the experiments, 48 min of EEG for each subject is collected and 12.8 h of EEG is recorded in total. More details of the experiment can be found in [43].
2) *DTU database:* This database consists of 18 normal-hearing subjects who selectively attend to one of two simultaneous speakers. 64-channel EEG is recorded using a BioSemi ActiveTwo system at a sampling rate of 512 Hz. Speech stimuli are excerpts taken from Danish audiobooks that are narrated by male and female speakers. The speech mixtures are presented binaurally from 60° to the left and 60° to the right of the subject. The positions of the target speech and the gender of speakers are randomized across trials. Each subject listens to 60 trials in which they are presented by 50 s of the speech mixtures. Therefore, the DTU database includes 50 min of EEG for each subject and 15 h of EEG for all 18 subjects. Further details can be found in [44].

B. Data Preprocessing

EEG signals are firstly processed to filter out 50 Hz line noise and harmonics [45]. Then, each channel data is rereferenced to the average response of the mastoid electrodes. As the proposed NI-AAD is expected to function in an end-to-end manner, no artifacts removal operation is involved in the data processing. All EEG data are first down-sampled to 128 Hz, and subsequently bandpass-filtered between 1 and 50 Hz. Then, all EEG channels are normalized for each trial.

For each subject, the EEG data is randomly split into a training set (60%), a validation set (20%), and a test set (20%). For each set, the EEG data is split into segments of smaller duration by a moving window (viz., *decision window*), with an overlap of 50%. The tail of a window in training may end up in an overlapping segment of the test set. This kind of segment is referred to as "repeated segment." All the repeated segments are discarded to keep the training, validation, and test sets mutually exclusive. Finally, we utilize SSF representation to convert each decision window into an EEG image.

C. Model Implementation

We conducted subject-dependent experiments on both DTU and KUL databases. Here, the cross-entropy loss function is adopted as the objective function. All features are encoded within a short encoding time window, namely, $N_t = 10$ time-steps [46], for SNN simulations.

Due to the discrete and nondifferentiable nature of SNNs in spike generation, the powerful error backpropagation method cannot directly be applied in the training process [47]. To that end, we adopted the conversion technique proposed in [48], which is called *Tandem Learning* (TL), to train the NI-AAD model. In brief, the TL training approach is capable of linking an SNN to a coupled ANN for parameter optimization. The coupled ANN is an auxiliary structure that facilitates the error backpropagation for the training of the SNN at the spike-train level.

For the coupled ANN, the convolutional layer is conditioned with a kernel size of 3×3, followed by a Rectified Linear Unit (ReLU) activation function, and an average pooling layer. The pooling operation is performed with a pooling size of 2×2. The training is performed by RMSprop optimizer with a learning rate of 3×10^{-4}. In addition, a dropout layer with a probability of 0.3 is applied after the pooling layer and the first fully connected (*fc*) layer, respectively. Meanwhile, a batch normalization layer is exploited after the convolution layer to reduce the effect of internal neuron distribution [49]. As well, an early stopping scheme is utilized to avoid overfitting, where the training stops when no loss reduction is found for 10 consecutive training epochs. To avoid data bias, we perform the experiments with 10 random splits of data for each subject. All hyperparameters given above were determined by running a grid search over a set of reasonable values. Performance during this grid search was measured on the validation set. Consistent with previous

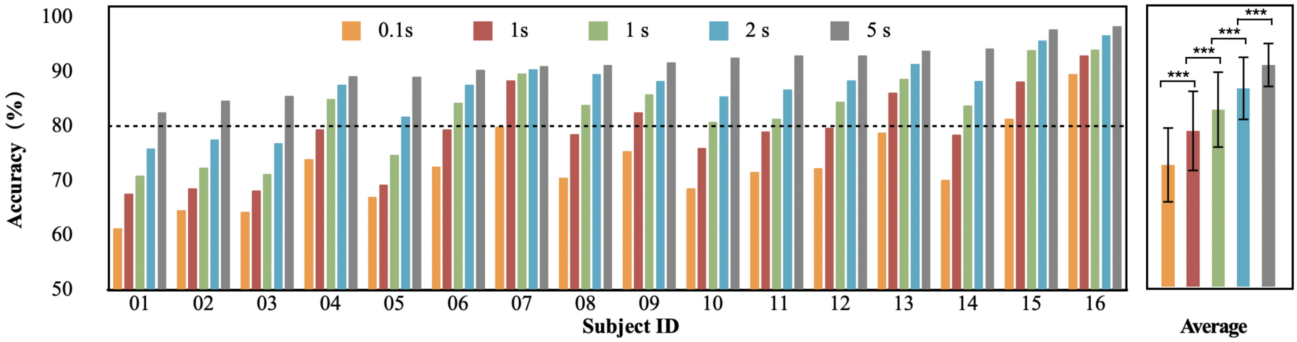

Fig. 3. AAD accuracy (%) of the neural-inspired model for five different decision windows across all 16 subjects in the KUL database. These subjects are ranked according to the accuracy for the 5 s decision window. The dashed line is a reference at 80% of detection accuracy. Statistically significant differences: $^{***}p < 0.001$.

studies [30], [31], the AAD accuracy is defined as the percentage of correctly classified decision windows on the test set.

IV. Results

A. Attention Decoding Accuracy

To evaluate how the proposed NI-AAD model performs on AAD tasks, we trained our model using decision windows of length 0.1, 0.5, 1, 2, and 5 s, respectively. For the KUL database, the overall average AAD accuracy and the average AAD accuracy per subject for five different decision window sizes are presented in Fig. 3. The proposed NI-AAD model shows a relatively high decoding performance with an accuracy of 82.8% (SD: 7.49) for 1s, 87.1% (SD: 6.17) for 2 s, and 91.2% (SD: 5.13) for 5s decision window. Consistent with previous studies [6], [30], [31], the AAD accuracy generally decreases for shorter decision window lengths. However, we are encouraged by the relatively high AAD result on the 0.1 s decision window 73.1% (SD: 7.26), which could potentially be suitable for real-time neuro-steered hearing aids. It is noted that the AAD accuracy for the 5 s decision window is slightly lower than for the 2 s decision window in the DTU database, which is consistent with a previous study [32]. One possible explanation would be that these nonlinear AAD models used direct classification instead of stimulus reconstruction approaches. Specifically, auditory attention is directly predicted without explicitly reconstructing the speech envelope. Therefore, the tradeoff between AAD accuracy and decision window length is improved, which could be beneficial for low-latency AADs [50].

As shown in Fig. 4, the proposed NI-AAD model obtains an average accuracy of 59.7% (SD: 3.25) for 0.1 s, 60.2% (SD: 3.30) for 0.5 s, 61.6% (SD: 3.12) for 1 s, 63.2% (SD: 2.96) for 2 s, and 61.5% (SD: 3.06) for 5 s decision window in the DTU database, respectively. The AAD accuracy obtained for subjects belonging to the DTU database is significantly lower than that in KUL database, which is in line with the observation made in [32] and [50]. One possible explanation could be that the two speech streams arrive 60° to the left and 60° to the right of the subject in the DTU database [44], while the speech streams come from ± 90° in the KUL database [43]. Therefore, it is more challenging to differentiate the spatial locations of the target speaker in the DTU database. Another major difference of the DTU database compared to the KUL database is that the auditory stimuli are presented with varying amounts of reverberation, which might reduce the cortical speech tracking in brain [51] and decrease the differential responses between attended and unattended speakers [52].

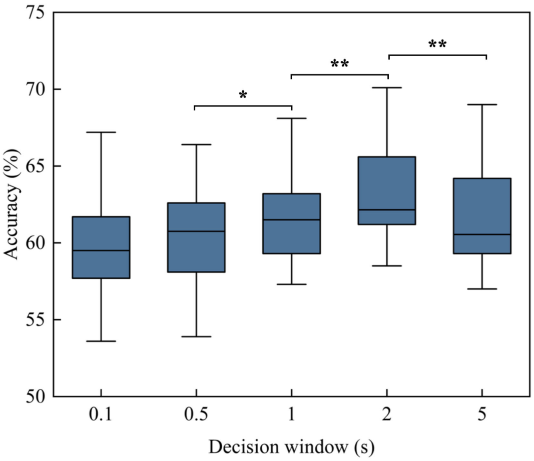

Fig. 4. AAD accuracy (%) of the neural-inspired model for five different decision windows across all 18 subjects in the DTU database. Statistically significant differences: $^{*}p < 0.05$, $^{**}p < 0.01$.

B. Comparative Study

To validate the effectiveness of the proposed NI-AAD model, we start by comparing our model with the classical linear AAD model [6]. The linear model is reimplemented in the DTU and KUL databases, in which the EEG signals are utilized to restore the attended speech envelope. Note that clean individual speech envelopes are required for the stimulus reconstruction approach. As summarized in Table I, the NI-AAD model is markedly superior to the linear model with an average improvement of 24.7% from 1 s to 5 s decision windows in the KUL database. For the DTU database, the NI-AAD model is also significantly better than the linear model with an average increase of 7.2%. Statistical analyses are performed using IBM SPSS statistics software and a level of significance of 0.05 is selected. Descriptive statistics are employed for means and standard deviations. The Kolmogorov–Smirnov test is used to confirm the normality

TABLE I
AAD ACCURACY (%) COMPARISON OF DIFFERENT MODELS ON KUL DATABASE AND DTU DATABASE FOR FIVE DIFFERENT DECISION WINDOW LENGTHS

Database	Model	Decision window (second)				
		0.1	0.5	1	2	5
KUL	linear [6]	-	-	58.1	61.3	67.5
	CNN [30]	65.9	73.4	80.8	82.1	83.6
	SSF-CNN [31]*	77.1	84.7	88.7	92.4	96.7
	NI-AAD	**73.1**	**79.4**	**82.8**	**87.1**	**91.2**
DTU	linear [6]	-	-	52.1	54.6	57.9
	CNN [30]	-	-	55.9	57.8	58.5
	SSF-CNN [31]*	65.1	67.9	70.2	72.6	68.6
	NI-AAD	**59.7**	**60.2**	**61.6**	**63.2**	**61.5**

Linear model denotes the setting in [6], while CNN model denotes the setting in [30].
*Here, we reimplement the SSF-CNN model in [31] with our experimental setup for comparison.

TABLE II
SPIKING RATES OF THE PROPOSED NI-AAD MODEL ON KUL AND DTU DATABASES FOR FIVE DIFFERENT DECISION WINDOW LENGTHS

Database	Decision window (second)				
	0.1	0.5	1	2	5
KUL	18.04%	11.90%	11.64%	10.24%	9.67%
DTU	7.06%	5.80%	5.77%	4.38%	4.52%

of the data distribution, prior to selection of appropriate statistical tests. AAD accuracy is significantly different between the NI-AAD and linear models in both the KUL database (paired *t*-test: $p < 0.001$) and DTU database ($p < 0.001$).

As stated by previous studies [14], [30], [50], [53], [54], nonlinear methods based on deep neural networks work much better than linear models, especially in low latency settings. We then compare the proposed NI-AAD model with the nonlinear CNN-based AAD model in [30]. In brief, the CNN architecture includes a convolution layer with a kernel size of 64×17, an average pooling, and two *fc* layers (Input: 5, hidden: 5, output: 2). The activation function is ReLU and the loss function is the cross-entropy. The implementation code of this CNN model is available online. For a fair comparison, we tuned the hyperparameters of the CNN model for both databases in the same way we did for our NI-AAD model.

The results in Table I show that the NI-AAD model outperforms the model in [30] with an average increase of 5.6% across five different decision windows in the KUL database. We also observe a statistically significant difference between these two models in terms of AAD accuracy (paired *t*-test: $p = 0.0025$). Similarly, the NI-AAD model obtains a consistent improvement of 4.7% in AAD accuracy in comparison with the CNN model in the DTU database.

Recently, we proposed a spatio-spectral feature representation method, i.e., SSF representation, to extract more discriminative features for EEG-based AAD. With the SSF representation of EEG signals, the CNN classifier [30] can achieve better AAD performance. The combination of these two components is referred to as the SSF-CNN model [31]. Table I provides the overall AAD accuracies of the NI-AAD and SSF-CNN models across all different window lengths in both KUL and DTU databases. The proposed NI-AAD performs better than the linear and CNN models among all decision windows, and yet is inferior to SSF-CNN in terms of accuracy. As tabulated in Table II, we evaluate the sparsity of the proposed NI-AAD model on KUL and DTU databases, where the sparsity (viz., spiking rate) is defined as the number of spikes over the number of neurons. Herein, we observe that the proposed NI-AAD achieved average spiking rates of 12.3% and 5.5% for KUL and DTU databases, respectively.

Compared with the dense SSF-CNN, although our model performs slightly worse at the same architecture, it takes full advantage of event-driven manner and distributed connection to show its latent superiority on computational efficiency.

V. DISCUSSION

A. Comparison of Computational Cost

In this section, we further compare the proposed NI-AAD and the SSF-CNN model in terms of computational cost. The total computational cost is proportional to the total amount of floating-point (FP) operations per second (Flops), and the total inference cost is computed based on the standard 45 nm CMOS process [55]. In the proposed NI-AAD model, a neuron is activated only when it receives enough input spikes to pass over the threshold, hence inactive neurons can be put into low-power mode to save power. Moreover, the computation of the NI-AAD implementation is event-driven by binary spike $\{1, 0\}$ processing manner, and thus, the MAC operation reduces to just an FP addition. While for SSF-CNN implementation, it still requires one FP addition and one FP multiplication to conduct the same MAC operation, which suffers from low computational efficiency.

As summarized in Table III, the computational cost of our NI-AAD implementation is significantly lower than the SSF-CNN implementation (paired *t*-test: $p < 0.001$) in two databases. Compared to SSF-CNN, the NI-AAD model achieves an average computational cost reduction of 99.27% and 99.68% in KUL and DTU databases, respectively.

In brief, the existing EEG-based AAD architectures are computationally too expensive, which is not suitable for devices with limited resources. Our proposed NI-AAD architecture offers tremendous energy benefits for efficient intelligent information processing applications, such as neuro-steered hearing aids.

B. Bio-Plausible Visualization for NI-AAD

In order to obtain the visual interpretation of SNN and enhance the understanding of the network, we employ spike activation map (SAM) technique to visualize different time-steps after convolution [56]. Considering that short inter-spike-interval (ISI) spikes have more information in a neurological system [57], [58], SAM computes a neuronal contribution score across the channel axis of input data to get a 2-D spatial heatmap, i.e., attention map. The attention map highlights neurons that carry more information for classification over different time-steps.

TABLE III
TOTAL COMPUTATIONAL COST COMPARISON OF THE NEURAL-INSPIRED MODEL AND THE SSF-CNN MODEL [31] ON TWO DATABASES FOR DIFFERENT DECISION WINDOW SIZES

Database	Computational Cost (pJ)	Decision window (second)					
		0.1	0.5	1	2	5	Average
KUL	$E_{SSF-CNN}$	2.00×10^{11}	3.99×10^{10}	1.99×10^{10}	9.91×10^{9}	3.95×10^{9}	5.46×10^{10}
	E_{NI-AAD}	9.78×10^{8}	4.17×10^{8}	3.03×10^{8}	1.93×10^{8}	1.15×10^{8}	4.01×10^{8}
	$E_{SSF-CNN}/E_{NI-AAD}$	0.0049	0.0105	0.0152	0.0194	0.0292	0.0073
DTU	$E_{SSF-CNN}$	2.08×10^{11}	4.14×10^{10}	2.06×10^{10}	1.02×10^{10}	3.95×10^{9}	5.67×10^{10}
	E_{NI-AAD}	4.46×10^{8}	1.44×10^{8}	1.11×10^{8}	1.28×10^{8}	1.07×10^{8}	1.87×10^{8}
	$E_{SSF-CNN}/E_{NI-AAD}$	0.0021	0.0035	0.0054	0.0125	0.0271	0.0032

$E_{\text{SSF-CNN}}$ denotes the total computational cost of the SSF-CNN model, while $E_{\text{NI-AAD}}$ denotes the total computational cost of the NI-AAD model.

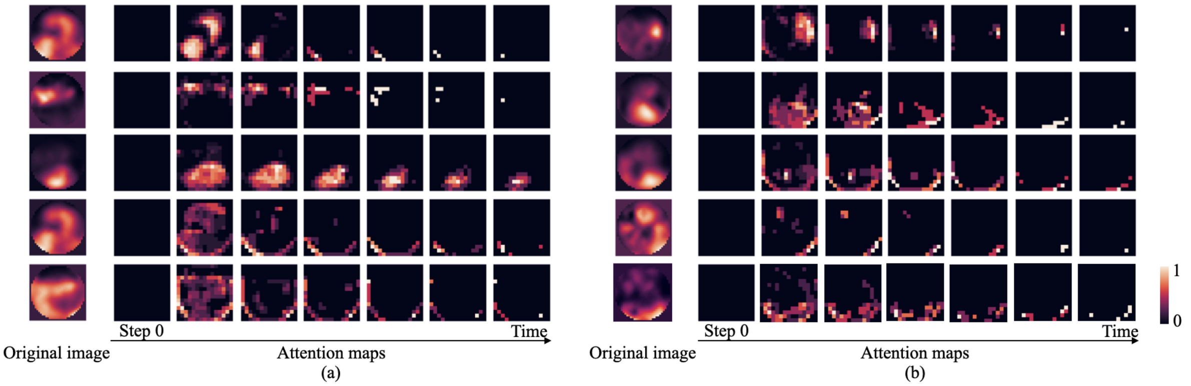

Fig. 5. Visualization of the internal spike representation of the neural-inspired model. (a) Attention maps of EEG signals for leftward auditory attention. Original images are obtained by the SSF representation of EEG signals. Here, we show the attention maps of five randomly selected subjects. The attention maps are calculated by monitoring neurons that carry more information (i.e., spikes) over different time-steps. The visualization highlights the discriminative region of the image. The color of the cells denotes the weights with lighter color corresponding to larger weight. (b) Attention maps of EEG images for rightward auditory attention from five randomly selected subjects.

Fig. 5 illustrates attention maps of EEG images for leftward and rightward auditory attention, respectively. Findings of previous research have demonstrated that the parietal alpha power increases over the hemisphere ipsilateral to attentional focus compared to the contralateral hemisphere [26], [28]. The difference in alpha power at parietal sites across hemispheres indicates the auditory spatial attention to speech [24]. Our results are in agreement with previous research that the attention map of our NI-AAD model highlights the important regions, i.e., the parieto-occipital region, in the EEG image for decoding the auditory attention. Specifically, when visualizing for "leftward auditory attention," the NI-AAD model identifies the discriminative image regions, i.e., the left parieto-occipital region, as shown in Fig. 5(a). When visualizing for "rightward auditory attention," the right parieto-occipital region is highlighted, as shown in Fig. 5(b). Further, we note that the visualization varies across each time-step underlying the fact that the NI-AAD model looks at different regions of the same input over time to make a binary decision.

VI. CONCLUSION

Real-world BCI systems, such as neuro-steered hearing aids, call for fast, accurate, and energy-efficient AAD architecture. In this study, we developed a neural-inspired model to parallel the energy efficiency and computing functionality of the brain to detect auditory attention, which is termed as NI-AAD. Comprehensive experiments show that the NI-AAD achieves relatively high average accuracies in both the KUL and DTU databases, especially in low latency settings. In addition, the visualization results explain the internal spike behavior of the NI-AAD and unleash its bio-plausible characteristics. Sparsity evaluation and energy computation on our model demonstrate that the NI-AAD can tolerate around 10% sparsity without considerable deterioration in performance and yield around two-order of magnitude energy efficiency of ANNs. Moreover, the proposed NI-AAD method does not require clean speech signals. This study could pave way for the practical implementation of AAD in real-life.

Although the performance-complexity tradeoff has not been well optimized, the merits of the proposed NI-AAD are obvious from the viewpoints of biological plausibility and low-power consumption, which is also attractive to investigate tricks to leverage their gap for future investigation.

ACKNOWLEDGMENT

Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the institutions and companies supporting the joint lab.

REFERENCES

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

[2] L. L. Cunningham and D. L. Tucci, "Hearing loss in adults," *New England J. Med.*, vol. 377, no. 25, pp. 2465–2473, 2017.

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

[4] N. Ding and J. Z. Simon, "Neural coding of continuous speech in auditory cortex during monaural and dichotic listening," *J. Neurophysiol.*, vol. 107, no. 1, pp. 78–89, 2012.

[5] S. Akram, J. Z. Simon, and B. Babadi, "Dynamic estimation of the auditory temporal response function from MEG in competing-speaker environments," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 8, pp. 1896–1905, Aug. 2017.

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

[7] B. Mirkovic, S. Debener, M. Jaeger, and M. De Vos, "Decoding the attended speech stream with multi-channel EEG: Implications for online, daily-life applications," *J. Neural Eng.*, vol. 12, no. 4, 2015, Art. no. 046007.

[8] 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.

[9] N. Das, A. Bertrand, and T. Francart, "EEG-based auditory attention detection: Boundary conditions for background noise and speaker positions," *J. Neural Eng.*, vol. 15, no. 6, 2018, Art. no. 066017.

[10] S. Zhao *et al.*, "Decoding auditory saliency from brain activity patterns during free listening to naturalistic audio excerpts," *Neuroinformatics*, vol. 16, no. 3, pp. 309–324, 2018.

[11] S. Miran, S. Akram, A. Sheikhattar, J. Z. Simon, T. Zhang, and B. Babadi, "Real-time tracking of selective auditory attention from M/EEG: A Bayesian filtering approach," *Front. Neurosci.*, vol. 12, 2018, Art. no. 262.

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

[13] A. Aroudi and S. Doclo, "Cognitive-driven binaural beamforming using EEG-based auditory attention decoding," *IEEE/ACM Trans. Audio Speech, Lang. Process.*, vol. 28, pp. 862–875, Jan. 2020.

[14] T. de 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, 2020.

[15] P. Stegman, C. S. Crawford, M. Andujar, A. Nijholt, and J. E. Gilbert, "Brain–computer interface software: A review and discussion," *IEEE Trans. Hum.-Mach. Syst.*, vol. 50, no. 2, pp. 101–115, Apr. 2020.

[16] R. Abiri, S. Borhani, J. Kilmarx, C. Esterwood, Y. Jiang, and X. Zhao, "A usability study of low-cost wireless brain-computer interface for cursor control using online linear model," *IEEE Trans. Hum.-Mach. Syst.*, vol. 50, no. 4, pp. 287–297, Aug. 2020.

[17] N. Robinson, T. W. J. Chester, and K. G. Smitha, "Use of mobile EEG in decoding hand movement speed and position," *IEEE Trans. Hum.-Mach. Syst.*, vol. 51, no. 2, pp. 120–129, Apr. 2021.

[18] 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.

[19] R. Zink, S. Proesmans, A. Bertrand, S. Van Huffel, and M. De Vos, "Online detection of auditory attention with mobile EEG: Closing the loop with neurofeedback," to be published, doi: 10.1101/218727.

[20] 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.

[21] 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, Feb. 2016.

[22] N. Das, J. Zegers, H. Van Hamme, T. Francart, and A. Bertrand, "EEG-informed speaker extraction from noisy recordings in neuro-steered hearing aids: Linear versus deep learning methods," to be published, doi: 10.1101/2020.01.22.915181.

[23] E. Ceolini *et al.*, "Brain-informed speech separation (BISS) for enhancement of target speaker in multitalker speech perception," *NeuroImage*, vol. 223, 2020, Art. no. 117282.

[24] J. R. Kerlin, A. J. Shahin, and L. M. Miller, "Attentional gain control of ongoing cortical speech representations in a "cocktail party"," *J. Neurosci.*, vol. 30, no. 2, pp. 620–628, 2010.

[25] J. N. Frey, N. Mainy, J.-P. Lachaux, N. Müller, O. Bertrand, and N. Weisz, "Selective modulation of auditory cortical alpha activity in an audiovisual spatial attention task," *J. Neurosci.*, vol. 34, no. 19, pp. 6634–6639, 2014.

[26] M. Wöstmann, B. Herrmann, B. Maess, and J. Obleser, "Spatiotemporal dynamics of auditory attention synchronize with speech," *Proc. Nat. Acad. Sci.*, vol. 113, no. 14, pp. 3873–3878, 2016.

[27] A. Bednar and E. C. Lalor, "Where is the cocktail party? Decoding locations of attended and unattended moving sound sources using EEG," *NeuroImage*, vol. 205, 2020, Art. no. 116283.

[28] Y. Deng, I. Choi, and B. Shinn-Cunningham, "Topographic specificity of alpha power during auditory spatial attention," *Neuroimage*, vol. 207, 2020, Art. no. 116360.

[29] 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.*, 2021, pp. 1115–1119.

[30] 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, 2021, Art. no. e56481.

[31] 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.*, 2021, pp. 5812–5815.

[32] I. Kuruvila, J. Muncke, E. Fischer, and U. Hoppe, "Extracting the locus of attention at a cocktail party from single-trial EEG using a joint CNN-LSTM model," *Front. Physiol.*, vol. 12, 2021, Art. no. 1178.

[33] E. M. Izhikevich, "Simple model of spiking neurons," *IEEE Trans. Neural Netw.*, vol. 14, no. 6, pp. 1569–1572, Nov. 2003.

[34] M. Pfeiffer and T. Pfeil, "Deep learning with spiking neurons: Opportunities and challenges," *Front. Neurosci.*, vol. 12, 2018, Art. no. 774.

[35] G. Guo, Z. Liu, S. Zhao, L. Guo, and T. Liu, "Eliminating indefiniteness of clinical spectrum for better screening COVID-19," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 5, pp. 1347–1357, May 2021.

[36] A. Sengupta, Y. Ye, R. Wang, C. Liu, and K. Roy, "Going deeper in spiking neural networks: VGG and residual architectures," *Front. Neurosci.*, vol. 13, 2019, Art. no. 95.

[37] L. Deckers, N. Das, A. H. Ansari, A. Bertrand, and T. Francart, "EEG-based detection of the attended speaker and the locus of auditory attention with convolutional neural networks," to be published, doi: 10.1101/475673.

[38] J. P. Snyder, "Map projections–A working manual," U.S. Government Printing Office, 1987, vol. 1395, pp. 191–197.

[39] I. Amidror, "Scattered data interpolation methods for electronic imaging systems: A survey," *J. Electron. Imag.*, vol. 11, no. 2, pp. 157–176, 2002.

[40] S.-C. Liu, A. van Schaik, B. A. Minch, and T. Delbruck, "Asynchronous binaural spatial audition sensor with 2 × 64 × 4 channel output," *IEEE Trans. Biomed. Circuits Syst.*, vol. 8, no. 4, pp. 453–464, Aug. 2014.

[41] W. Gerstner and W. M. Kistler, *Spiking Neuron Models: Single Neurons, Populations, Plasticity*. Cambridge, U.K.: Cambridge Univ. Press, 2002.

[42] J. Kaiser, H. Mostafa, and E. Neftci, "Synaptic plasticity dynamics for deep continuous local learning (DECOLLE)," *Front. Neurosci.*, vol. 14, 2020, Art. no. 424.

[43] N. Das, T. Francart, and A. Bertrand, "Auditory attention detection dataset KULeuven," Aug. 2020, *Version 1.1.0*. [Online]. Available: https://doi.org/10.5281/zenodo.3997352

[44] 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

[45] A. de Cheveigné and D. Arzounian, "Robust detrending, rereferencing, outlier detection, and inpainting for multichannel data," *NeuroImage*, vol. 172, pp. 903–912, 2018.

[46] B. Rueckauer, I.-A. Lungu, Y. Hu, M. Pfeiffer, and S.-C. Liu, "Conversion of continuous-valued deep networks to efficient event-driven networks for image classification," *Front. Neurosci.*, vol. 11, 2017, Art. no. 682.

[47] Z. Zhang and Q. Liu, "Spike-event-driven deep spiking neural network with temporal encoding," *IEEE Signal Process. Lett.*, vol. 28, pp. 484–488, Feb. 2021.

[48] J. Wu, E. Yılmaz, M. Zhang, H. Li, and K. C. Tan, "Deep spiking neural networks for large vocabulary automatic speech recognition," *Front. Neurosci.*, vol. 14, 2020, Art. no. 199.

[49] Y. Kim and P. Panda, “Revisiting batch normalization for training low-latency deep spiking neural networks from scratch,” *Front. Neurosci.*, vol. 15, 2020, Art. no. 1638.

[50] 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.

[51] 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, 2013.

[52] J. M. Rimmele, E. Z. Golumbic, E. Schröger, and D. Poeppel, “The effects of selective attention and speech acoustics on neural speech-tracking in a multi-talker scene,” *Cortex*, vol. 68, pp. 144–154, 2015.

[53] 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, 2019.

[54] 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*, 2020, pp. 2772–2776.

[55] M. Horowitz, “1.1 computing’s energy problem (and what we can do about it),” in *Proc. IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, 2014, pp. 10–14.

[56] Y. Kim and P. Panda, “Visual explanations from spiking neural networks using inter-spike intervals,” *Sci. Rep.*, vol. 11, no. 1, pp. 1–14, 2021.

[57] D. S. Reich, F. Mechler, K. P. Purpura, and J. D. Victor, “Interspike intervals, receptive fields, and information encoding in primary visual cortex,” *J. Neurosci.*, vol. 20, no. 5, pp. 1964–1974, 2000.

[58] J. Y. Shih, C. A. Atencio, and C. E. Schreiner, “Improved stimulus representation by short interspike intervals in primary auditory cortex,” *J. Neurophysiol.*, vol. 105, no. 4, pp. 1908–1917, 2011.

Siqi Cai (Member, IEEE) received the Ph.D. degree in mechanical engineering 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 with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. Her research interests include brain-computer interface, and biosignal processing.

Dr. Cai has served as the local arrangement chair of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDial) 2021, and the workshop chair of the 47th IEEE International Conference on Acoustics, Speech, & Signal Processing 2022.

Peiwen Li received the B.E. degree in mechanical engineering, in 2019, from the South China University of Technology, Guangzhou, China, where he is currently working toward the M.S. degree in mechanical engineering with the Department of Shien-Ming Wu School of Intelligent Engineering.

Enze Su received the B.E. degree in mechanical engineering, in 2019, from the South China University of Technology, Guangzhou, China, where he is currently working toward the M.S. degree in mechanical engineering with the Department of Shien-Ming Wu School of Intelligent Engineering.

Qi Liu (Member, IEEE) received the bachelor’s and master’s degrees in electrical engineering from Harbin Engineering University, Harbin, China, in 2013 and 2016, respectively, and the Ph.D degree in electrical engineering from City University of Hong Kong, Hong Kong, in 2019.

He is currently a Professor with the School of Future Technology, South China University of Technology, Guangzhou, China. During 2018–2019, he was a Visiting Scholar with University of California Davis, CA, USA. From 2019 to 2022, he worked as a Research Fellow with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. His research interests include machine learning, optimization methods, and neuromorphic computing with applications to image/video/speech signal processing.

Dr. Liu has been an Associate Editor for the IEEE SYSTEMS JOURNAL (2022-), and Digital Signal Processing (2022-). He was also the Guest Editor for the *International Journal of Antennas and Propagation, and Wireless Communications and Mobile Computing*. He was the recipient of the Best Paper Award of IEEE International Conference on Signal, Information and Data Processing (ICSIDP) in 2019.

Longhan Xie (Member, IEEE) received the B.S. and M.S. degrees in mechanical engineering from Zhejiang University, Hangzhou, China, in 2002 and 2005, respectively, and the Ph.D. degree in mechanical and automation engineering from the Chinese University of Hong Kong, Hong Kong, in 2010.

From 2010 to 2016, he was an Assistant Professor and Associate Professor with the School of Mechanical and Automotive Engineering, South China University of Technology, Guangzhou, China. Since 2017, he has been a Professor with Shien-Ming Wu School of Intelligent Engineering, South China University of Technology, Guangzhou, China. His research interests include biomedical engineering and robotics.

Prof. Xie is a member of ASME.