EEG-Based Auditory Attention Detection via Frequency and Channel Neural Attention

Siqi Cai, *Student Member, IEEE*, Enze Su, Longhan Xie, *Member, IEEE*, and Haizhou Li, *Fellow, IEEE*

***Abstract*—Humans have the ability to pay attention to one of the sound sources in a multispeaker acoustic environment. Auditory attention detection (AAD) seeks to detect the attended speaker from one's brain signals that will enable many innovative human–machine systems. However, effective representation learning of electroencephalography (EEG) signals remains a challenge. In this article, we propose a neural attention mechanism that dynamically assigns differentiated weights to the subbands and the channels of EEG signals to derive discriminative representations for AAD. In the nutshell, we would like to build a computational attention mechanism, i.e., neural attention, to model the auditory attention in human brain. We incorporate the proposed neural attention into an AAD system, and validate the neural attention mechanism through comprehensive experiments on two publicly available datasets. The experimental results demonstrate that the proposed system significantly outperforms the state-of-the-art reference baselines.**

***Index Terms*—Auditory attention, brain–computer interface (BCI), channel attention, electroencephalography (EEG), frequency attention.**

I. Introduction

HUMANS have the ability to focus their auditory attention on one speaker, and ignore other sound sources in a multispeaker acoustic environment [1], which is described as *cocktail party effect* [2]. Benefiting from the research progress in related areas, such as psychoacoustic, biophysiological, and neuroscience, regarding the brain activity of auditory attention, we are inspired to develop a computational model to detect the attention activities manifested in brain signals.

Recent studies have demonstrated that auditory attention can be decoded from the recordings of brain activity, such as electrocorticography (ECoG) [1], magnetoencephalography [3], [4], and electroencephalography (EEG) [5]–[12] in a multispeaker scenario. Auditory attention detection (AAD) opens up many possibilities for human–machine systems, such as the cognitive control of hearing aids, i.e., neurosteered hearing aids [13], [14], and rehabilitation medicine through neural feedback. As EEG provides a noninvasive means of investigating cortical activity with high temporal resolution for brain–computer interface applications [15], we are particularly interested in decoding the auditory attention from EEG signals in this article.

In general, the algorithms for AAD can be grouped into linear and nonlinear decoders [16]. The design of linear decoders follows the idea of stimulus reconstruction that the cortical responses to an attended speaker, as encoded in EEG signals, correlate with the auditory stimulus. To this end, it first predicts the cortical responses by reconstructing the stimulus from EEG signals, then detects the correlation between the reconstructed stimulus and the attended speech envelopes [7]–[11], [17], [18]. This process can be seen as a regression approach toward the AAD problem. However, the stimulus reconstruction algorithms are not directly optimized for attention detection accuracy. They seek to transform multichannel EEG signals into one single envelope, which potentially results in loss of detailed channel specific information useful for attention detection. The canonical correlation analysis (CCA) method [10] is one of the successful linear models that achieves reliable AAD with a decision window of around 10 s. However, its performance degrades rapidly as the decision window narrows [16], [19]. We note that a 10-s decision window means a much higher latency than what humans require, i.e., around 1 s [20], to switch attention from one speaker to another. Nonetheless, the important finding that there exists a correlation between EEG responses and the envelope of the attended stimulus [7]–[11] inspires many subsequent studies.

Studies in neuroscience have revealed that cortical responses have a nonlinear relationship with the acoustic stimuli [21]–[23]. De Taillez *et al.* [24] first studied a nonlinear neural network to map EEG signals to speech envelopes in a cocktail party scenario that outperforms the linear model baseline. Along the

Manuscript received March 8, 2021; revised June 1, 2021, August 13, 2021, September 22, 2021, and October 20, 2021; accepted October 23, 2021. Date of publication December 2, 2021; date of current version March 14, 2022. This work was supported by in part by the Agency for Science, Technology and Research (A$*$STAR) under its AME Programmatic Funding Scheme under Project A18A2b0046 and in part by A$*$STAR under its RIE 2020 Advanced Manufacturing and Engineering Human (AME) under Programmatic Grant A1687b0033, and in part by the Singapore Government's Research, Innovation and Enterprise 2020 plan (Advanced Manufacturing and Engineering domain) under Programmatic Grant A18A2b0046 and Grant A1687b0033. The work of Longhan Xie was supported by the National Natural Science Foundation of China under Grant 52075177. The work of Haizhou Li was supported by the Deutsche Forschungsgemeinschaft (German Research Foundation) under Germany's Excellence Strategy (University Allowance, EXC 2077, University of Bremen, Germany). This article was recommended by Associate Editor T. H. Falk. *(Corresponding author: Longhan Xie.)*

Siqi Cai is with the Shien-Ming Wu School of Intelligent Engineering, South China University of Technology, Guangzhou 510460, China, and also with the Department of Electrical and Computer Engineering, National University of Singapore 119077, Singapore (e-mail: 2447379329@qq.com).

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

Haizhou Li is with the School of Data Science, The Chinese University of Hong Kong, Shenzhen, Shenzhen 518172, China, and also with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 119077, Singapore, and also with the Machine Listening Lab, University of Bremen, 28359 Bremen, Germany (e-mail: haizhou.li@nus.edu.sg).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/THMS.2021.3125283.

Digital Object Identifier 10.1109/THMS.2021.3125283

same idea, convolutional neural network (CNN) [12], [25], [26] was studied by directly relating both the raw EEG signals and the speech stimulus to the attention detection decision, without reconstructing the auditory stimulus. Let us call this an end-to-end classification approach. In this article, we would like to further the idea of nonlinear CNN decoder [26] and end-to-end solution in several aspects, with a focus on effective EEG representation learning and low latency implementation.

First, numerous different and often spatially distant neuronal populations interact quickly and dynamically when processing the speech stimulus [27]. Previous studies revealed that linear decoders work the best for low-frequency EEG because envelope frequencies between 1 and 8 Hz are linearly relatable to their corresponding EEG signals [28], whereas nonlinear decoders can profit from a wider EEG frequency range (1–32 Hz) [24]. However, different EEG frequency bands are likely to differ both in their physiological genesis and their role in selective auditory attention [5], [28], [29]. They all reflect the attention decision process in human brain. It makes sense to devise an AAD mechanism that learns the differentiated contributions of EEG rhythms.

Second, the positions of electrodes relate the EEG signals to the activities of the related brain areas. In addition, some EEG channels are more informative than others in terms of informing the decision making process in the brain [30]. Furthermore, such differentiated property may vary from subject to subject [8]. We are motivated to study a channel neural attention mechanism that assigns weights dynamically to the EEG channels across different spatial locations over the cortex.

Third, representation learning is one of the successful machine learning techniques that extracts salient information from raw data and greatly improves pattern classification [31]. We note that most of the linear [7], [9], [11] and nonlinear [24]–[26] AAD decoders of multichannel EEG signals have not benefited from representation learning. Among the few feature extraction explorations, Horton *et al.* [6] first hand-crafted features from the neural measurements, others studied common spatial pattern method for EEG enhancement, which improves AAD accuracy [12], [32]. While the aforementioned feature extraction techniques usually see performance gains, their benefits are limited due to the fact that they do not participate in the optimization of the attention classifiers. There is another school of thought that is to select a subset of channels [8], [30] or frequency subbands [5], [28], [33] for attention detection. They generally seek to reduce the dimensionality of raw data, instead of improving the accuracy.

The implementation of representation learning in this article is also motivated by the idea of feature extraction and selection, where we discover the benefits of weighting the contributions of different EEG channels [8], [30] and frequency subbands [5], [28], [29], [32], [33]. The end-to-end approach in the previous nonlinear AAD models can be seen as a way of learning the weights in a data-driven manner. Typically, such weights are pretrained as part of the model parameters and fixed at run-time, independently of the input EEG signals [12], [24]–[26]. However, neuroscience studies have provided convincing evidence that human auditory attention is a dynamic temporal process [5], [27]. For example, the EEG signals for auditory attention vary with the temporal content of speech stimulus [34]–[36]. In other words, the contributions of EEG channels and frequency subbands to AAD performance may vary over time. This prompts us to study a nonlinear, dynamic weighting mechanism that is known as the neural attention mechanism in deep neural networks. The dynamic weighting mechanism is a departure from the pretrained weighting mechanism in the sense that the former learns to assign weights dynamically according to input signals. As the weights are of continuous values, the dynamic weighting mechanism is also different from the traditional feature selection techniques that simply select a subset of features.

Not to confuse between neural attention and auditory attention in this article, neural attention refers to a network implementation, whereas AAD is the task we would like to perform. We make three main contributions to EEG-based AAD, which are as follows.

1) We motivate and propose a frequency and channel neural attention mechanism for EEG representation learning.
2) We validate the hypothesis of differentiated frequency and channel contributions of EEG signals through data visualization and comprehensive experiments on two publicly available EEG datasets.
3) We formulate an end-to-end nonlinear decoder with the frequency-channel neural attention mechanism for low-latency attention detection.

To the best of our knowledge, this is the first study of a frequency-channel neural attention mechanism that learns to dynamically assign differentiated weights to incoming EEG signals.

The rest of this article is organized as follows. In Section II, we formulate a novel EEG-based AAD model with neural attention. In Section III, we present the experimental setup. In Section IV, we report the experiment results. In Section V, we perform an analytical study on the proposed neural attention. Finally, Section VI concludes this article.

II. Neural Attention for AAD

A window of EEG signal can be considered as a three-dimensional (3-D) feature, which has the frequency bands, the EEG channels, and the time indices of data samples as its three dimensions. A 3-D EEG feature is illustrated in Fig. 1(C_1) as the input to an AAD system. Let us formulate EEG-based AAD as a binary classification problem for a two-speaker scenario [24]–[26].

We propose a neural attention mechanism that performs frequency and channel attention on the 3-D EEG features, and can be easily incorporated into existing CNN architecture [26]. We also propose to arrange the frequency and channel attention in sequence, which involves less computation and parameters overhead than a full 3-D attention solution. The sequential arrangement with two separable modules also allows us to produce a modularized output feature of the same size as the input feature, which facilitates the subsequent ablation study. There are two possible sequential permutations between the frequency

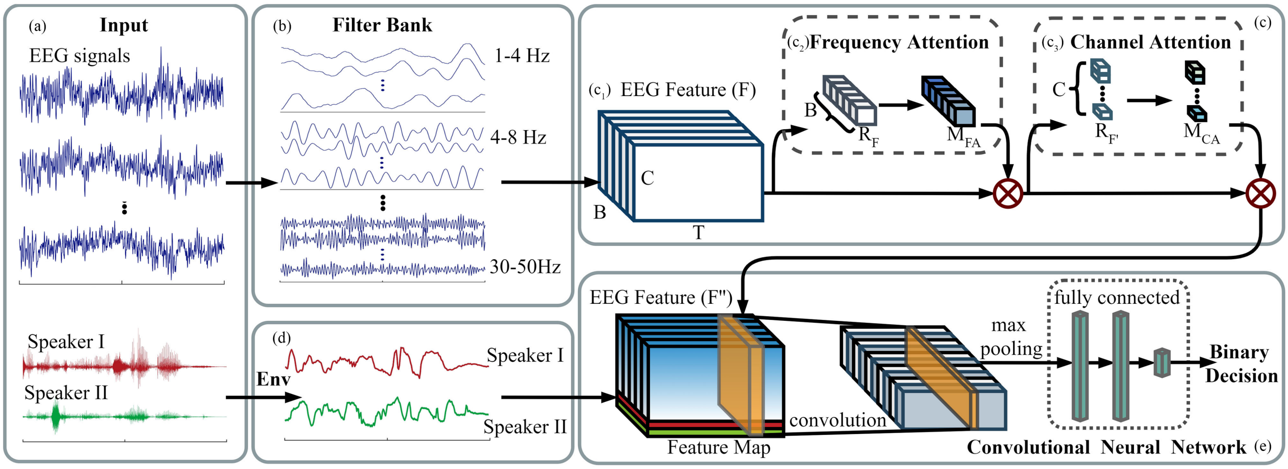

Fig. 1. Schematic diagram of the proposed CNN classifier of five components with frequency and channel neural attention (CNN-FC): (A) input multichannel EEG signals and speech envelopes as the auditory stimulus references; (B) filter bank for EEG signals; (C_1) 3-D feature extraction of multichannel EEG signals, (C_2) frequency attention module, and (C_3) channel attention module; (D) envelope extraction for speech streams; and (E) a CCN. The CNN-FC model is trained to detect the attended speaker, either speaker I or II, from the EEG signals. Note: speech streams of speakers I and II are denoted in red and green, respectively, whereas the EEG signals of the listener are denoted in blue.

and channel attention modules. For brevity, we only discuss the frequency-channel attention sequence in detail, as illustrated in Fig. 1. The channel-frequency architecture can be formulated in a similar way.

The proposed AAD system consists of a signal processing front-end and a back-end classifier. The AAD system with a frequency-channel neural attention is referred to as CNN-FC, whereas its channel-frequency counterpart is referred to as CNN-CF hereafter. We illustrate the frequency-channel attention module for EEG representation learning in Fig. 1(C), which aims to automatically discover the representations needed for attention detection from raw EEG data.

The neural attention is implemented through a masking mechanism, where a feedforward network is employed to predict a mask of differentiated weights [37], [38] for EEG frequency bands and channels, respectively. In frequency attention, the mask represents the selective auditory attention on EEG frequency bands, whereas in channel attention, the mask represents the differentiated contributions of individual EEG channels. The neural attention module is expected to improve the separation between EEG signals of opposite attention, therefore, reduce the required decision window size. Finally, a CCN serves as a binary back-end classifier for decision making. Meanwhile, we apply the power-law subbands method [9] to improve the speech envelope extraction process.

A. *Frequency Attention*

Humans have the ability to pay selective attention in many everyday situations. Auditory attention in the cocktail party is a typical example [1], [2], which can be described as the modulation between a bottom-up sensory-driven stimulus and a top-down attention task. The modulation is achieved rapidly in a cognitive process [39], [40]. It results in a receptive field in response to the input stimulus, which is the attended voice in this case. Simply speaking, a receptive field works like a mask that only lets the attended voice pass. Similar to the attentional modulation in the physiological studies, the attentional modulation in deep neural networks has been implemented in different ways [37], [38], [41]–[43]. The idea is to model the top-down and bottom-up modulation by dynamically assigning differentiated weights to the composition of the input stimulus at run-time. The differentiated weights form a receptive field, which is also called an attention mask. As the attention mask is dynamically generated by a neural attention mechanism, it is not a set of pretrained weights. The neural attention mechanism is capable of biasing the allocation of available resources toward the most informative components of a signal.

Moreover, previous studies suggest that different EEG frequency bands have different functional roles in speech processing [5], [28], [29], [44]. We consider that the frequency bands of EEG signals manifest the brain activities as far as auditory attention is concerned. It makes sense that we develop a neural attention mechanism over the different frequency bands, also referred to as frequency attention, as shown in Fig. 1(C_2). We would like to use the predefined frequency bands in neuroscience study, i.e., δ, θ, α, β, and low-γ, due to two considerations. First, the data-driven frequency analysis is very much data dependent. The resulting subbands are not as interpretable as those in neuroscience study [5], [27]–[29], [32], [33] in terms of spectral coverage. Second, an independent frequency analysis front-end, as shown in Fig. 1(B), allows us to study a dynamic weighting mechanism explicitly.

The frequency attention is implemented in following three steps.

1) We filter the original EEG into subband EEG signals, as shown in Fig. 1(B).
2) We predict an attention mask with a frequency attention mechanism, as shown in Fig. 1(C),
3) We modulate the EEG signals with the attention mask.

In the first step, we decompose the EEG signal from each channel into five classic frequency bands, namely δ (1–4 Hz),

θ (4–8 Hz), α (8–12 Hz), β (12–30 Hz), and low-γ (30–50 Hz) bands [45]. They have been extensively studied for their differential effects in auditory attention in cocktail party scenarios [1], [3], [5], [28], [33], [44], [46].

By applying a sliding window to an EEG signal, we obtain a sequence of decision windows. Each window frame is characterized by a 3-D feature. The 3-D EEG feature is denoted as $F \in \mathbb{R}^{B \times C \times T}$, where B denotes the number of the frequency bands, C is the number of the EEG channels, and T is the number of EEG samples in a decision window.

In the second step, a frequency attention mechanism learns to predict an attention mask for the EEG frequency bands. We first aggregate spectral information of F by using a convolution layer as follow:

$$R_F = \text{Max}(\text{ELU}(\text{Conv}(F))) \tag{1}$$

where Conv$(\cdot)$ denotes a convolution operation with the size of the convolution kernel being $1 \times C \times T$. Exponential linear unit (ELU$(\cdot)$) is the activation function in the convolution operations [47]. Max$(\cdot)$ represents a max pooling layer, which is employed to reduce the number of parameters and to further explore the temporal information of the EEG signals. After this, we obtain a spectral context descriptor to represent each EEG frequency band, i.e., $R_F \in \mathbb{R}^B$.

Then, a gating mechanism, which focuses on enhancing the representational power of the network by modeling the relationships among different EEG frequency bands in a computationally efficient manner, was adopted [37]. To reduce model complexity and improve generalization, two fully connected (*fc*) layers around the nonlinearity were employed to parameterize the gating mechanism of frequency attention as follows:

$$M_{\text{FA}} = \tan h(\mathbf{w}_2 \cdot \tan h(R_F \cdot \mathbf{w}_1 + \mathbf{b}_1) + \mathbf{b}_2) \tag{2}$$

where $\mathbf{w}_1$ and $\mathbf{w}_2$ denote the parameters of the first and second *fc* layers, respectively. $\mathbf{b}_1$ and $\mathbf{b}_2$ denote the first and second biases, respectively. tan$h$ activation function is applied after each *fc* layer. And M_{FA} represents the attention mask generated by the frequency attention module.

Finally, the frequency attention mechanism modulates the input EEG signals by applying the frequency attention mask $F' = M'_{\text{FA}} \bigotimes F$, where $\bigotimes$ denotes a pointwise multiplication. M'_{FA} is obtained by broadcasting the frequency attention M_{FA} along the channel and temporal dimension, i.e., $M'_{\text{FA}} \in \mathbb{R}^{B \times C \times T}$.

B. Channel Attention

In ECoG signal analysis, it is shown that the strength of attentive effect, manifested in various regions of auditory cortices in response to a speech stimulus, varies significantly [28], [46]. For high-density EEG, which is a global measure of cortical activity, the responses of individual EEG channels to auditory stimuli are different. Some channels are more informative than others in terms of decoding the auditory attention in the brain [7]. The channel selection strategy [8], [30], [48] is motivated by this finding. It reduces the number of channels by turning some of them completely OFF. Unlike the traditional channel selection, we propose a soft channel attention mechanism, which seeks to capture the interchannel relationship of EEG signals and adaptively assign differentiated weights to individual channels according to the EEG signals and the speech envelopes. As shown in Fig. 1(C_3), the channel attention mechanism consists of three steps, which is similar to those in frequency attention.

First, we adopt a convolution layer to aggregate channel information of F' as follows:

$$R_{F'} = \text{Ave}(\text{ELU}(\text{Conv}(F'))) \tag{3}$$

where Conv$(\cdot)$ denotes the convolution operation with the size of the convolution kernel being $B \times 1 \times T$. Ave$(\cdot)$ denotes an average pooling layer.

Second, the gating mechanism of channel attention can be expressed as follows:

$$M_{\text{CA}} = \tan h(\mathbf{w}_4 \cdot \tan h(R_{F'} \cdot \mathbf{w}_3 + \mathbf{b}_3) + \mathbf{b}_4) \tag{4}$$

where $\mathbf{w}_3$ and $\mathbf{w}_4$ denote the parameters of the first and second *fc* layers, respectively. $\mathbf{b}_3$ and $\mathbf{b}_4$ denote the first and second biases, respectively. And M_{CA} represents the attention mask generated by the channel attention mechanism.

Finally, the channel attention mechanism modulates the input EEG feature F' by applying the channel attention mask $F'' = M'_{\text{CA}} \bigotimes F'$, where $\bigotimes$ denotes a pointwise multiplication. Similarly, the attention value ($M'_{\text{CA}} \in \mathbb{R}^{B \times C \times T}$) is obtained by broadcasting M_{CA} along the frequency and temporal dimension.

C. Aligning Speech Envelopes With EEG Features

The frequency-channel neural attention mechanism produces F'' as part of the input to a back-end classifier. As will be described in detail in Section III, in the KUL and DTU datasets, two concurrent speech streams are presented simultaneously to the listeners, each of which is associated with one of the two binary outputs. We align the envelopes of two speech streams with the EEG features F'' to form a 3-D feature map, which allows the classifier to examine the correlation between EEG features and speech envelopes.

The studies show that CNN clearly outperforms linear models in detecting such correlation, especially excel in low latency settings [12], [16], [26]. In practice, CNN models detect the correlation between EEG features and speech envelopes without explicitly reconstructing speech envelopes from EEG features, but rather learn to discover the correlation through a network architecture, followed by a binary classifier.

D. Back-End Classifier

The CNN back-end classifier takes the feature map as input and makes a binary decision, as illustrated in Fig. 1(E). Inspired by Vandecappelle *et al.* [26], the CNN architecture consists of a convolution layer with a $5 \times 66 \times 9$ kernel, i.e., 5 frequency bands by 64 EEG plus 2 speech channels, and by 9 samples width for the 3-D feature map, a max pooling. Rectifying linear unit activation function is adopted in convolution layers. Finally, two *fc* layers with the sigmoid activation function are added for binary decision. We employ the weighted cross-entropy loss function as the cost function. During training, the stochastic gradient descent technique is implemented for network updating, with a learning rate of 0.1 and a self-adaptive learning rate reduction strategy.

III. Experiment Setup

A. AAD Dataset

In this article, experiments are carried out on two publicly available AAD datasets. The first dataset was recorded at KU Leuven, which is referred to as KUL dataset [49]. The second dataset is referred to as DTU dataset [50].

1) KUL Dataset: In this dataset, EEG data were collected from eight male and eight female normal-hearing subjects, who were instructed to attend to one particular speaker and ignore the other in the presence of two simultaneous speakers. The EEG data were split into eight trials of 6-min duration each. In total, EEG data from 16 normal-hearing subjects were collected and there was 8×6 min $= 48$ min of data for each subject. The EEG data were recorded at a sampling rate of 8192 Hz using a BioSemi ActiveTwo system, in which 64 electrodes were placed on the head according to international 10–20 standards. Four Dutch short stories, narrated by different male speakers, were used as the auditory stimuli. All silences longer than 500 ms in the audio files were truncated to 500 ms. The auditory stimuli were low-pass filtered with a cutoff frequency of 4 kHz and presented at 60 dB through a pair of insert earphones (Etymotic ER3).

2) DTU Dataset: This dataset consists of 18 normal-hearing subjects who selectively attended to one of the two simultaneous speakers. Stimuli were excerpts taken from Danish audiobooks that were narrated by a male and a female speaker. Each subject listened to 60 trials in which they were presented 50 s of the speech mixtures. 64-channel EEG was recorded using a BioSemi ActiveTwo system at a sampling rate of 512 Hz. More details could be found in [50].

B. Data Preparation

The recorded EEG signals are first high-pass filtered with cutoff frequency at 0.5 Hz to remove the dc component and electrode drift. Then, EEG signals are rereferenced to a common average reference and resampled to 128 Hz [51]. We prepare the EEG data for two sets of experiments. The EEG data were first bandpass-filtered between 1 and 50 Hz, which is referred to as broadband EEG hereafter. They are then decomposed into five frequency bands, as stated in Section II-A, which is referred to as multiband EEG hereafter.

To extract the envelopes from the speech stimuli, we adopt an auditory filterbank with power law compression, denoted as Env in box (D) in Fig. 1. This method resembles the nonlinear transformation process of the speech streams in human auditory system. It has been proven effective in previous AAD research work [9]. In practice, the speech stream is first fed to a gammatone filterbank ranging from 150 to 8000 Hz. Then, the absolute value of each subband is processed with a power-law compression with an exponent of 0.6. The subbands are bandpass-filtered between 1 to 50 Hz, and finally they are combined with equal weights, and downsampled to 128 Hz to match the EEG data [11], [26].

C. Training and Evaluation

The data of each subject are randomly split into a training set (60%) and a validation set (20%), and a test set (20%) for cross-validation.[1] For each partition, we apply a sliding window on the data by shifting half a window size along the time axis. Each segment is used as a decision window. All the repeated segments are discarded to keep the training, validation, and test sets mutually exclusive. To avoid data bias, we perform the experiments with ten random splits of data for each subject and report the average results. In addition, dropout with a probability of 0.5 is used in all *fc* layers. We also use early stopping to avoid overfitting. The training stops as soon as no loss reduction is found for tenconsecutive training epochs. Following the protocol in the previous studies [12], [24]–[26], we train subject-dependent systems and report both subject average accuracy and overall average accuracy in this study.

As near real-time responses are required in real-world applications, we are interested in the attention detection accuracy with short decision windows. Specifically, experiments are carried out with 1- and 2-s decision windows that are approximately the human time lags when switching attention [20].

IV. Experiment Results

We first design two reference baselines for broadband EEG data on KUL dataset, i.e., CNN with and without channel attention, denoted as CNN(s)-C and CNN(s). In this study, we would like to search for an effective configuration for channel attention. We adopt the same CNN architecture [26] for the two baseline models, where the CNN includes a convolution layer with a 66×9 kernel, i.e., 64 EEG plus 2 speech channels by 9 samples width, a max pooling, and two *fc* layers (Input:10, hidden:10, output:2), with sigmoid activation function and weighted cross-entropy as the loss function. Both models take the $[C, T]$ matrix, i.e., C channels by T samples of the broadband EEG data, and the envelopes of two speech streams as inputs.

We then conduct extensive ablation experiments on KUL dataset with multiband EEG, where we would like to observe the contributions of the frequency and channel neural attention. The experiments involve four models, namely CNN [26], CNN with frequency attention (CNN-F), CNN-FC, and CNN-CF. They share the same CNN architecture as depicted in Fig. 1(E), except that CNN model takes F as the EEG features, CNN-F only involves frequency attention, CNN-FC and CNN-CF involve both frequency and channel attention modules in a different order. We summarize the model configuration in Table I and describe them in detail next.

Finally, we perform experiments on DTU dataset to explore the generalization ability of our model.

[1] The code for our model and experiments can be found in https://github.com/SCUT-IEL/CNN-FC.

TABLE I
SUMMARY OF MODEL CONFIGURATION WITH CNN AS THE BACK-END CLASSIFIER IN THE EEG-BASED ATTENTION DETECTION EXPERIMENTS

Model	Front-end	Frequency Attention	Channel Attention
CNN(s)	broadband EEG	No	No
CNN(s)-C	broadband EEG	No	Yes
CNN	multi-band EEG	No	No
CNN-F	multi-band EEG	Yes	No
CNN-CF	multi-band EEG	Yes	Yes
CNN-FC	multi-band EEG	Yes	Yes

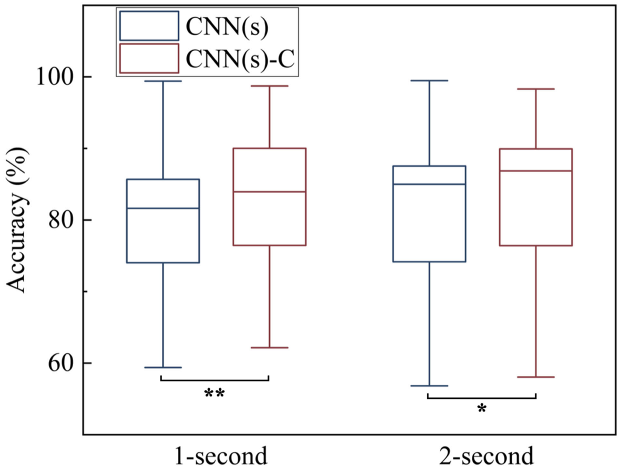

Fig. 2. AAD accuracy (%) for 1- and 2-s decision windows reported in broadband EEG evaluation on KUL dataset. Statistically significant differences: $^{*}p < 0.05$, $^{**}p < 0.01$.

TABLE II
AAD ACCURACY (%) IN A COMPARATIVE STUDY OF DIFFERENT MODELS ON KUL DATASET

Model	Front-end	Decision window	
		1-second	2-second
Stimulus-reconstruction [7]	broadband EEG	58.1^{a_3}	61.3^{a_3}
Deep CCA [17]	broadband EEG	62.8^{a_3}	66.4^{a_3}
Match-mismatch [18]	broadband EEG	61.6^{a_3}	65.9^{a_3}
DNN [25]	broadband EEG	71.3^{a_3}	75.2^{a_3}
CNN(s) [26]	broadband EEG	78.9^{a_2}	$80.4\ ^{a_1}$
CNN(s)-C	broadband EEG	81.1	82.1
CNN [26]	multi-band EEG	78.4^{b_3}	79.6^{b_3}
CNN-F	multi-band EEG	81.7^{b_2}	83.7^{b_3}
CNN-CF	multi-band EEG	**83.8**	**87.1**
CNN-FC	multi-band EEG	**83.6**	**86.9**

Note: ^{a_1}, ^{a_2}, and ^{a_3} denote the significant drops of detection accuracy over the CNN(s)-C model with $P < 0.05$, $P < 0.01$, and $P < 0.001$, respectively. ^{b_2} and ^{b_3} denote the significant drop of detection accuracy over the CNN-FC model with $P < 0.01$ and $P < 0.001$, respectively.

A. *Channel Attention With Broadband EEG*

We report the detection accuracy of CNN(s) and CNN(s)-C models for 1- and 2-s decision windows with broadband EEG data, as shown in Fig. 2 and Table II. With 1-s decision window, CNN(s)-C model outperforms CNN(s) model with an average improvement of 2.2% (mean: 81.1% versus 78.9%, SD: 11.87 versus 11.56). With 2-s decision window, CNN(s) model obtains an average accuracy of 80.4% (SD: 11.67). CNN(s)-C model achieves better performance, with an average accuracy of 82.1% (SD: 11.96). Statistical analyses are performed using IBM SPSS statistics software (ver. 24.0, IBM Corporation, Armonk, NY, USA) and a level of significance of 0.05 is selected. The descriptive statistics are used for means and standard deviations. The Kolmogorov–Smirnov test is used to confirm the normality of the distribution of the data, prior to the selection of appropriate statistical tests. Paired *t*-tests are employed to compare AAD performance of different models to identify which model gains a significant improvement. CNN(s)-C model significantly outperforms CNN model for both 1-s decision window ($p = 0.007$) and 2-s decision window ($p = 0.019$).

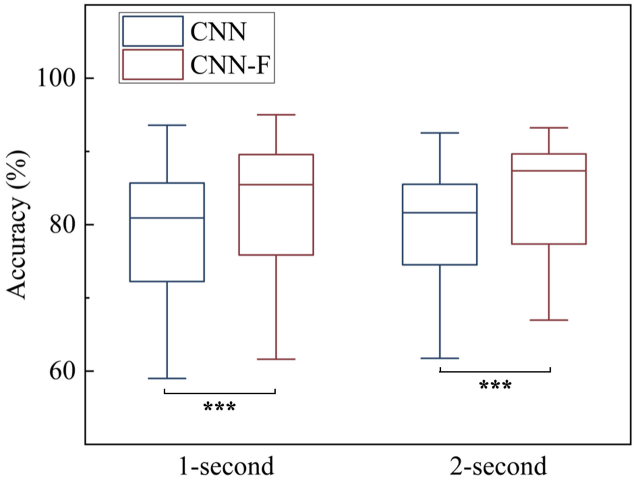

Fig. 3. AAD accuracy (%) 1- and 2-s decision windows reported in multiband EEG evaluation on KUL dataset. *** represents statistically significant difference at $p < 0.001$ level.

These results clearly validate the effectiveness of the proposed channel attention mechanism and the resulting representations.

B. *Frequency Attention With Multiband EEG*

To evaluate the effect of frequency attention, we move on to the multiband EEG data. We compare CNN and CNN-F models for 1- and 2-s decision windows, as shown in Fig. 3 and Table II. With 1-s decision window, CNN-F model outperforms CNN model with an average improvement of 3.3% (mean: 81.7% versus 78.4%, SD: 10.28 versus 10.14). With 2-s decision window, CNN model obtains an average accuracy of 79.6% (SD: 11.67) and CNN-F model gains an improvement of 4.1% (mean: 83.7%, SD: 8.29). The CNN-F model significantly outperforms the CNN model for both 1-s decision window ($p < 0.001$) and 2-s decision window ($p < 0.001$).

In addition, we perform *t*-test to examine the detection accuracy between two CNN experiments with broadband and multiband EEG data, i.e., CNN(s) and CNN models in Table I. We find no statistically significant differences between these two models for either 1-s decision window ($p = 0.41$) or 2-s decision window ($p = 0.28$). The CNN-F model with multiband EEG significantly outperforms the CNN model in both broadband and multiband EEG evaluations. These results clearly validate the effectiveness of the proposed frequency attention mechanism and the resulting representations.

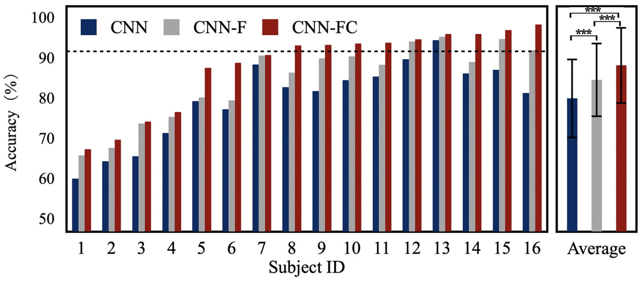

Fig. 4. AAD accuracy (%) of different models for 2-s decision window reported in multiband EEG evaluation on KUL dataset. The subjects are ordered according to the accuracy of CNN-FC model. The horizontal dotted line is a reference at 90% of detection accuracy. *** represents statistically significant difference at p <0.001 level.

C. Frequency-Channel Attention With Multiband EEG

In Fig. 4, we observe that the CNN-FC model, which employs both frequency and channel neural attention, attains the best performance with an average accuracy of 86.9% (SD: 8.83) in multiband EEG evaluation for 2-s decision window. 6.25% (1 of 16), 25.0% (4 of 16), and 56.25% (9 of 16) of subjects are reported with over 90% detection accuracy for CNN, CNN-F, and CNN-FC models, respectively. Furthermore, the detection accuracy of CNN-FC is significantly higher than CNN model (p <0.001) and CNN-F model (p <0.001).

We further compare the three multiband EEG models, as summarized in Table II. Either with 1- or 2-s decision window, the CNN-FC model significantly outperforms the CNN-F and CNN models, at 83.6% accuracy for 1-s decision window, which is on par with the CNN-F model at 83.7% accuracy for 2-s decision window. It is noted that the size of the decision window represents the minimum response latency of an attention detection system. A 1-s decision window is not far from the expectation by real-world applications in terms of response latency, such as neurosteered hearing prostheses [13], [14].

We also compare the results of CNN-FC and CNN-CF models in Table II. The experiments show that the CNN-CF model slightly outperforms CNN-FC. However, we find no statistically significant differences for either 1-s decision window (p = 0.49) or 2-s decision window (p = 0.43) between CNN-FC and CNN-CF. Both CNN-FC and CNN-CF significantly outperform the linear models [7], [17], [18] and the nonlinear models [25], [26], as well as the CNN models with only the channel or frequency attention module.

To summarize, the proposed AAD system benefits from both the frequency and channel neural attention, and exhibits a high level of performance. We have shown a successful first attempt to exploit frequency-channel EEG representations for AAD.

D. Speech Envelopes as References

In linear AAD models, such as stimulus reconstruction method [7]–[11], [17], [18], the speech envelopes play a role as they are used to compare with the reconstructed stimulus for attention detection. However, in the proposed CNN-FC network architecture, the speech envelopes are taken as the input together with other EEG features for decision making. The question arises as to whether or how much the speech envelopes actually contribute to the decision making.

TABLE III
AAD Accuracy (%) of Different Models on DTU Dataset

Model	Front-end	Decision window	
		1-second	2-second
CNN(s) [26]	broadband EEG	69.2^{b_3}	71.2^{b_3}
CNN [26]	multi-band EEG	70.7^{b_3}	72.4^{b_3}
CNN-FC	multi-band EEG	**79.3**	**82.9**

Note: ^{b_3} denotes the significant drop of detection accuracy over the CNN-FC model with P <0.001.

We further perform an ablation experiment on KUL dataset where we remove the speech envelopes, as illustrated in Fig. 1(D), from the input. The model is referred to as CNN-FC (no-se) hereafter. Since the speakers and directions are labeled separately, the KUL dataset allows us to perform two tasks, i.e., the detection of attended speakers regardless of directions, and the detection of attended directions regardless of speakers. The former task is performed with the CNN-FC model having both speech envelopes and EEG signals as the input, the latter is performed only with EEG signals with the CNN-FC (no-se) model. In the absence of speech envelopes input, the CNN-FC (no-se) model practically decodes the attended direction based on the differences of the EEG activity between both directions as in [34] and [52].

Our results show that the attended speaker detection with CNN-FC model only marginally outperforms CNN-FC (no-se) model by 2.3% (mean: 83.6% versus 81.3%, SD: 10.33 versus 11.45), and 2.1% (mean: 86.9% versus 84.8%, SD: 8.83 versus 9.18) for 1- and 2-s decision window, respectively. In other words, the detection results for attended direction and attended speaker are comparable. Our finding is consistent with the previous studies [26], [32], which suggests that it is possible to decode the spatial focus of auditory attention from the EEG signals alone, without the need of speech envelopes as references.

E. Experiments on DTU Dataset

Finally, we report the experiments on DTU dataset in Table III. As there are no significant differences between CNN-FC and CNN-CF, we only compare CNN-FC with the baselines. For 1-s decision window, it is observed that the CNN-FC model achieves an average AAD accuracy of 79.3% (SD: 8.17) that significantly outperforms 70.7% (SD: 8.64) of the CNN model (p <0.001); for 2-s decision window, the CNN-FC model outperforms the CNN model by 10.5%. We are encouraged by the results on the DTU dataset that corroborate the findings on the KUL dataset.

V. Empirical Analysis of Neural Attention

We hypothesize that the frequency and channel neural attention provides a more effective representation than raw EEG data. To validate our hypotheses, next we compare the proposed CNN-FC model with other competing models in the literature, and understand the neural attention masks from the perspective of neuroscience.

A. Comparative Study

Due to different experimental setup, it is not straightforward to compare our AAD results directly with those in the previous studies. We reimplemented the reported systems on KUL dataset as the reference baselines.

For broadband EEG evaluation, we reimplemented the stimulus reconstruction [7], deep CCA [17], and match-mismatch [18] models as the linear decoder baselines. We also reimplemented the DNN model [25] and state-of-the-art CNN model [26] as the nonlinear decoder baselines. In Table II, we report the overall average detection accuracy across all subjects. The CNN(s)-C model outperforms the stimulus reconstruction model by a large margin of 23.0% and 20.8% for 1- and 2-s decision windows, respectively. It also shows a consistent and significant performance gain over deep CCA [17] and match-mismatch [18] models. The difference between CNN(s)-C and linear models is significant for both 1-s (p <0.001) and 2-s (p <0.001) decision windows. As expected, the nonlinear models show a clear advantage over linear models especially in low latency settings [12], [16], [24]–[26]. In comparison with DNN model [25], the CNN(s)-C model gains an increase of 9.8% and 6.9% for 1- and 2-s decision windows in terms of AAD accuracy. It also outperforms the CNN(s) model [26] with consistent improvements of 2.2% and 1.7% for 1- and 2-s decision windows. The difference between CNN(s)-C and CNN(s) is that CNN(s)-C employs a channel neural attention with dynamic weighting to the 3-D feature map, whereas CNN(s) employs pretrained fixed weights. The improvement of CNN(s)-C over CNN(s) model clearly validates the effectiveness of the dynamic weighting scheme.

For multiband EEG evaluation, we reimplemented the CNN model [26] as the state-of-the-art baseline. We observe the following from Table II: First, the CNN model for multiband EEG does not show any improvement over the CNN(s) for broadband EEG, whereas CNN-F for multiband EEG clearly outperforms CNN(s). The results suggest that subband decomposition does not contribute by itself, it takes effect only with the frequency neural attention. Second, the fact that CNN-FC model significantly outperforms the CNN(s)-C and CNN-F models suggests that channel neural attention is more effective in combination with frequency neural attention.

We are not aware of other models that have the capability to perform the same level of detection accuracy as CNN-FC with such low latency settings. These results support that the hypothesis that CNN-FC model can learn *what* and *where* to attend through the frequency and channel neural attention, and the resulting representation is highly effective.

B. Analysis of Channel Attention Mask

To gain a better insight into the underlying reasoning processes that CNN-FC learns to perform, we would like to study the attention mask that is predicted by the neural attention. The attention mask is a set of differentiated weights assigned dynamically to the channels that is expected to reflect the actual neural activities in brain signals.

To analyze the distributions of the differentiated masking weights M_{CA} for individual channels, we plot the attention weights, which are greater than 0.5, channel by channel for all subjects in Fig. 5. It is observed that the average weights assigned by the channel attention mechanism vary with the channels, which supports the theory that EEG signals across the channels are not equally informative as far as AAD is concerned [8], [30], [48]. It is expected that the locations indicative of neural activity contributing to speech processing have higher weights.

While the attention weights generally reflect the functional organization in human brain, we note from the channelwise boxplot (see Fig. 5) that the attention weights of individual EEG channels vary across subjects within a small range, which can be interpreted as the subject variability of the brain signals [33], [53]. These results support our hypothesis that EEG-based AAD tasks will benefit from the differentiated weights over individual channels that vary from subject to subject. The proposed neural attention mechanism is designed to find such weights dynamically during run-time inference.

Generally, the channel attention mechanism is shown to derive differentiated weights dynamically across different spatial locations over the cortex, which effectively improves detection accuracy. The visualization of attention mask corroborates the findings in neuroscience.

C. Visualization of Frequency Attention Mask

To relate our neural attention model with different EEG frequency bands, we visualize the attention distributions produced by the frequency attention for multiband EEG. Fig. 6 depicts the subject average attention mask M_{FA} for the EEG frequency bands, where we aggregate the subject-dependent masking weights for 2-s decision windows. We also aggregate the masking weights M_{FA} for 2-s decision windows across all 16 subjects.

As expected, the assigned weights vary across these five EEG frequency bands. In general, the average weights of lower EEG frequencies, i.e., 1–8 Hz at δ and θ bands, are obviously greater than those of other frequency bands that conform to the previous studies [3], [7], [28], [35], [44], [54], [55]. It was found that selective attention would act via a gain increase on the low-frequency EEG signals (<8 Hz) for the attended speech. The strength of the δ and θ bands may be a reflection of the neural computations in human brain that takes advantage of the high power and signal-to-noise ratio in speech at slow envelope frequencies [3], [7], [29]. Another possible explanation could be that improving the representation of the temporal modulations is important for speech intelligibility, in which the modulations below 8 Hz are perhaps the most essential [44], [55].

In particular, we found that the average weight of θ-band EEG is the greatest, which aligns well with the findings of previous studies [5], [27]. One explanation could be that θ-band EEG is strongly represented in the speech envelope and important for speech comprehension [29], [56].

After θ and δ bands, β-band EEG is also assigned with greater weights than other EEG frequency bands. Given the established role of β-band EEG in top-down predictive mechanisms [56], [57] and selective attention in a top-down state, higher weights

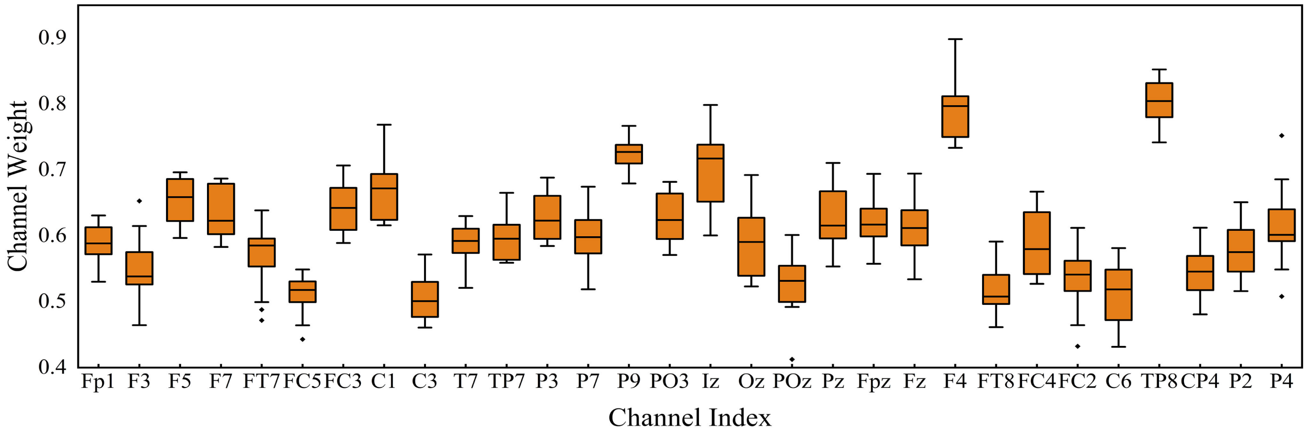

Fig. 5. EEG channelwise attention weights for all 16 subjects of KUL dataset in 2-s decision window experiments. The electrodes are placed according to the international 10–20 system. From left to right, $Fp1$ to $FO3$ are electrodes on the left hemisphere, whereas $F4$ to $P4$ are those on the right.

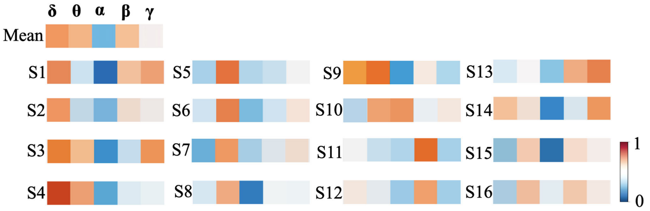

Fig. 6. Average masking weights M_{FA} for different EEG frequency bands for each subject (S1–S16) of KUL dataset and the weights averaged over all 16 subjects (mean). The subjects are ordered according to the pattern of their frequency attention masks. S1–S4 has the largest weights for δ-band EEG; S5–S9 has the largest weights for θ-band EEG. S10 has the largest weight for α-band EEG. S11 and S12 have the largest weights for β-band EEG. S13 and S14 have the largest weights for low-γ band. S15 and S16 have the largest weights for θ and β bands. The color of the cells denotes the weights, with red as 1 and blue as 0.

of β band can be interpreted by its top-down predictive function. It is also worth noting that the average weight of low-γ band is even higher than α band. Likewise, Kerlin *et al.* [5] also observed that low-γ band is helpful in distinguishing the attended utterance in a cocktail party scenario. Additionally, previous studies suggest that low-γ band plays an important role in the underlying physiologic computations and stimulus competition [28], [58], [59]. These findings indicate that low-γ band also carries useful information regarding attention, which is reflected in the frequency attention mask.

Unexpectedly, α-band EEG is assigned with relatively lower weights, which seems to conflict with the findings by Wostmann *et al.* [34], [52]. We note that studies in [34] and [52] were focused on the lateralization (i.e., left-right hemispherical asymmetry) of the α power instead of the overall α-band EEG. Furthermore, they studied the case where both the attended and unattended speech streams have the same temporal structure. Unlike previous auditory spatial attention work [36], [51], we are not interested in the spatial selective attention, but rather the attention to the difference of the temporal structure of the attended and unattended speech streams. In practice, we seek to detect whether the EEG signals are associated with the speech envelopes of attended speaker. Viswanathan *et al.* [33] have also reported that the α-band EEG is less informative about attentional focus than δ, θ, β, and low-γ bands.

In summary, EEG frequency bands show different functional roles in auditory attention. Studies in neuroscience describe the qualitative contributions of attention signals carried by the frequency bands, which motivates the study in this article. The proposed frequency neural attention mechanism provides a way to assign differentiated weights dynamically to the frequency bands in a quantitative manner. In this way, we do not make a hard selection of frequency bands, but rather embrace all information available across the frequency bands. The successful implementation of frequency neural attention is attributed not only to the weights that reflect the contributions of the frequency bands, but also to the dynamic combination of the weights that are adapted for individual subjects.

VI. Conclusion

This study is motivated by the findings in neuroscience that auditory attention is manifested in EEG signals in a differentiated manner across the spatial locations over the cortex and the frequency bands. We propose a novel frequency-channel attention mechanism as a neural approach to AAD, and show that the proposed framework consistently outperforms all state-of-the-art competing models on both KUL and DTU datasets. The proposed frequency-channel representation can be generalized to other EEG-based decoding tasks. In the empirical analysis, we confirm that the channel and frequency masks corroborate the findings in neuroscience. This study marks an important step toward real-time attention detection for neurosteered hearing aids. As future work, we would like to study the AAD in real-world acoustic environments, where the auditory stimuli are corrupted by noises [13], [14], and from more than two competing speakers.

Acknowledgment

The authors would like to thank Y. Song, P. Li, and X. Jia for their valuable comments and technical assistance.

REFERENCES

[1] 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, 2012.

[2] 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, 1953.

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

[4] 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.

[5] 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.

[6] C. Horton, R. Srinivasan, and M. D'Zmura, "Envelope responses in single-trial EEG indicate attended speaker in a 'cocktail party'," *J. Neural Eng.*, vol. 11, no. 4, 2014, Art. no. 046015.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] N. Das *et al.*, "EEG-informed speaker extraction from noisy recordings in neuro-steered hearing aids: Linear versus deep learning methods," 2020, *BioRxiv*.

[15] H.-J. Hwang, S. Kim, S. Choi, and C.-H. Im, "EEG-based brain-computer interfaces: A thorough literature survey," *Int. J. Human- Comput. Interact.*, vol. 29, no. 12, pp. 814–826, 2013.

[16] 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.

[17] J. R. Katthi, S. Ganapathy, S. Kothinti, and M. Slaney, "Deep canonical correlation analysis for decoding the auditory brain," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2020, pp. 3505–3508.

[18] A. de Cheveigné, M. Slaney, S. A. Fuglsang, and J. Hjortkjaer, "Auditory stimulus-response modeling with a match-mismatch task," *J. Neural Eng.*, vol. 18, no. 4, 2021, Art. no. 046040.

[19] 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.

[20] 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.

[21] P. Faure and H. Korn, "Is there chaos in the brain? I. Concepts of nonlinear dynamics and methods of investigation," *Comptes Rendus de l'Académie des Sciences-Series III-Sciences de la Vie*, vol. 324, no. 9, pp. 773–793, 2001.

[22] H. Korn and P. Faure, "Is there chaos in the brain? II. Experimental evidence and related models," *Comptes Rendus Biologies*, vol. 326, no. 9, pp. 787–840, 2003.

[23] M. Keshishian, H. Akbari, B. Khalighinejad, J. L. Herrero, A. D. Mehta, and N. Mesgarani, "Estimating and interpreting nonlinear receptive field of sensory neural responses with deep neural network models," *Elife*, vol. 9, 2020, Art. no. e53445.

[24] 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.

[25] 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.

[26] 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.

[27] A.-L. Giraud and D. Poeppel, "Cortical oscillations and speech processing: Emerging computational principles and operations," *Nat. Neurosci.*, vol. 15, no. 4, 2012, Art. no. 511.

[28] 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.

[29] 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, 2018.

[30] A. M. Narayanan and A. Bertrand, "Analysis of miniaturization effects and channel selection strategies for EEG sensor networks with application to auditory attention detection," *IEEE Trans. Biomed. Eng.*, vol. 67, no. 1, pp. 234–244, Jan. 2020.

[31] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[32] 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.

[33] V. Viswanathan, H. M. Bharadwaj, and B. G. Shinn-Cunningham, "Electroencephalographic signatures of the neural representation of speech during selective attention," *eNeuro*, vol. 6, no. 5, 2019, pp. 1–14.

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

[35] O. Ghitza, A.-L. Giraud, and D. Poeppel, "Neuronal oscillations and speech perception: Critical-band temporal envelopes are the essence," *Front. Hum. Neurosci.*, vol. 6, 2013, Art. no. 340.

[36] A. J. Power, J. J. Foxe, E.-J. Forde, R. B. Reilly, and E. C. Lalor, "At what time is the cocktail party? A late locus of selective attention to natural speech," *Eur. J. Neurosci.*, vol. 35, no. 9, pp. 1497–1503, 2012.

[37] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[38] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[39] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Netw.*, vol. 19, no. 9, pp. 1395–1407, 2006.

[40] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5706–5722, Dec. 2015.

[41] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.

[42] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[43] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[44] G. M. Di Liberto, J. A. O'Sullivan, and E. C. Lalor, "Low-frequency cortical entrainment to speech reflects phoneme-level processing," *Curr. Biol.*, vol. 25, no. 19, pp. 2457–2465, 2015.

[45] G. Buzsáki and A. Draguhn, "Neuronal oscillations in cortical networks," *Science*, vol. 304, no. 5679, pp. 1926–1929, 2004.

[46] B. N. Pasley *et al.*, "Reconstructing speech from human auditory cortex," *PLoS Biol.*, vol. 10, no. 1, 2012, Art. no. e1001251.

[47] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," 2015, *arXiv:1511.07289*.

[48] M. Arvaneh, C. Guan, K. K. Ang, and C. Quek, "Optimizing the channel selection and classification accuracy in EEG-based BCI," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 6, pp. 1865–1873, Jun. 2011.

[49] N. Das, T. Francart, and A. Bertrand, "Auditory Attention Detection Dataset KULeuven," Version 1.1.0, Aug. 2020. *[Online]*. Available: https://doi.org/10.5281/zenodo.3997352

[50] 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

[51] N. Das, W. Biesmans, A. Bertrand, and T. Francart, "The effect of head-related filtering and ear-specific decoding bias on auditory attention detection," *J. Neural Eng.*, vol. 13, no. 5, 2016, Art. no. 056014.

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

[53] H. Ramoser, J. Muller-Gerking, and G. Pfurtscheller, "Optimal spatial filtering of single trial EEG during imagined hand movement," *IEEE Trans. Rehabil. Eng.*, vol. 8, no. 4, pp. 441–446, Dec. 2000.

[54] J. Obleser and C. Kayser, "Neural entrainment and attentional selection in the listening brain," *Trends Cogn. Sci.*, vol. 23, no. 11, pp. 913–926, 2019.

[55] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *J. Acoust. Soc. Amer.*, vol. 95, no. 2, pp. 1053–1064, 1994.

[56] 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, 2017.

[57] L. Fontolan, B. Morillon, C. Liegeois-Chauvel, and A.-L. Giraud, "The contribution of frequency-specific activity to hierarchical information processing in the human auditory cortex," *Nat. Commun.*, vol. 5, no. 1, pp. 1–10, 2014.

[58] P. Lakatos, G. Karmos, A. D. Mehta, I. Ulbert, and C. E. Schroeder, "Entrainment of neuronal oscillations as a mechanism of attentional selection," *Science*, vol. 320, no. 5872, pp. 110–113, 2008.

[59] C. Börgers, S. Epstein, and N. J. Kopell, "Gamma oscillations mediate stimulus competition and attentional selection in a cortical network model," *Proc. Nat. Acad. Sci. USA*, vol. 105, no. 46, pp. 18023–18028, 2008.

Siqi Cai (Student 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 2021, and the Workshop Chair of the 47th IEEE International Conference on Acoustics, Speech, and Signal Processing, 2022.

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.

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. His research interests include biomedical engineering and robotics.

Dr. Xie is a member of ASME.

Haizhou Li (Fellow, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in electrical and electronic engineering from the South China University of Technology, Guangzhou, China, in 1984, 1987, and 1990, respectively.

He is currently a Professor with the School of Data Science, The Chinese University of Hong Kong, Shenzhen, Shenzhen, China, and the Department of Electrical and Computer Engineering, National University of Singapore (NUS), Singapore. Prior to joining NUS, he taught at The University of Hong Kong from 1988 to 1990 and South China University of Technology from 1990 to 1994. He was a Visiting Professor with CRIN in France from 1994 to 1995, a Research Manager with the Apple-ISS Research Centre from 1996 to 1998, a Research Director with Lernout & Hauspie Asia Pacific from 1999 to 2001, a Vice President with InfoTalk Corporation Ltd., from 2001 to 2003, and the Principal Scientist and Department Head of Human Language Technology with the Institute for Infocomm Research, Singapore, from 2003 to 2016. His research interests include automatic speech recognition, speaker and language recognition, and natural language processing.

Dr. Li served as the Editor-in-Chief for the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING from 2015 to 2018, a member of the Editorial Board for the *Computer Speech and Language* from 2012 to 2018. He was an elected Member of the IEEE Speech and Language Processing Technical Committee from 2013 to 2015, 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, and the President of Asian Federation of Natural Language Processing from 2017 to 2018. He was the General Chair of ACL 2012, INTERSPEECH 2014, ASRU 2019, and ICASSP 2022. He is a Fellow of the ISCA. He was a recipient of the National Infocomm Award 2002 and the President's Technology Award 2013 in Singapore. He was named one of the two Nokia Visiting Professors in 2009 by the Nokia Foundation, and Bremen Excellence Chair Professor in 2019.