Brain Topology Modeling With EEG-Graphs for Auditory Spatial Attention Detection

Siqi Cai, *Member, IEEE*, Tanja Schultz, *Fellow, IEEE*, and Haizhou Li, *Fellow, IEEE*

Abstract*—*Objective: **Despite recent advances, the decoding of auditory attention from brain signals remains a challenge. A key solution is the extraction of discriminative features from high-dimensional data, such as multi-channel electroencephalography (EEG). However, to our knowledge, topological relationships between individual channels have not yet been considered in any study. In this work, we introduced a novel architecture that exploits the topology of the human brain to perform auditory spatial attention detection (ASAD) from EEG signals. *Methods:* We propose *EEG-Graph Net*, an EEG-graph convolutional network, which employs a neural attention mechanism. This mechanism models the topology of the human brain in terms of the spatial pattern of EEG signals as a graph. In the EEG-Graph, each EEG channel is represented by a node, while the relationship between two EEG channels is represented by an edge between the respective nodes. The convolutional network takes the multi-channel EEG signals as a time series of EEG-graphs and learns the node and edge weights from the contribution of the EEG signals to the ASAD task. The proposed architecture supports the interpretation of the experimental results by data visualization. *Results:* We conducted experiments on two publicly available databases. The experimental results showed that EEG-Graph Net significantly outperforms the state-of-the-art methods in terms of decoding performance. In addition, the analysis of the learned weight patterns provides insights into the processing of continuous speech in the brain and confirms findings from neuroscientific studies. *Conclusion:* We showed that modeling brain topology with EEG-graphs yields highly competitive results for auditory spatial attention detection. *Significance:* The proposed EEG-Graph Net is more lightweight and accurate than competing baselines and provides explanations for the results. Also, the architecture can be easily transferred to other brain-computer interface (BCI) tasks.**

***Index Terms*—Auditory spatial attention, brain-computer interface, channel-wise attention, electroencephalography, graph convolutional network.**

I. Introduction

HUMANS have the ability to listen to a speaker's voice when surrounded by many other speakers and noises, referred to as the 'cocktail party effect' [1]. Neuroscientific evidence suggests that the listener's auditory attention can be decoded from brain activity. This activity can be captured by methods such as electrocorticography (ECoG) [2], magnetoencephalography (MEG) [3], [4], or electroencephalography (EEG) [5], whereby non-invasive EEG-based auditory attention detection shows particular promise for controlling hearing aids and rehabilitation devices [6]. For this purpose, EEG signals are recorded from EEG electrodes placed on the surface of the skull, where one electrode corresponds to one EEG-channel. In this context, several studies described the decoding [5], [7], [8], [9], [10], the optimization of stimulus/response features [11], [12], [13], and the data acquisition [6], [14], [15]. However, these studies predominantly focused on decoding the speech envelope of the attended speaker from the brain signals of the listener. Also, most studies assumed that the clean speech signals of the attended speaker are available, which unfortunately is not the case in real scenarios.

Recently, a new paradigm called auditory spatial attention detection (ASAD) has been investigated that focuses on the decoding of the spatial locus of the attended speaker [16], [17]. ASAD no longer requires a clean speech stimulus, paving the way towards practical neuro-steered hearing prostheses [18], [19], [20]. Promising works on ASAD, which are popular in the brain-computer interface (BCI) community, include the Common Spatial Pattern (CSP)-based approach by Geirnaert et al. [21] and the geometry-based approach by Riemannian [22]. With the advent of deep learning, convolutional neural networks (CNNs) have been developed to achieve competitive performance especially for short decision windows (around 1 sec) [23], [24], [25], [26] that do not require manual feature crafting. The CNN model is designed to learn local stationary structures determined by the convolutional kernel. Thus, its capability to

Manuscript received 13 July 2022; revised 7 April 2023 and 28 May 2023; accepted 5 July 2023. Date of publication 11 July 2023; date of current version 25 December 2023. This work was supported in part by the A*STAR under its RIE 2020 Advanced Manufacturing and Engineering Programmatic under Grant A1687b0033, in part by the Internal Project of Shenzhen Research Institute of Big Data under Grant T00120220002 in part by the Guangdong Provincial Key Laboratory of Big Data Computing, The Chinese University of Hong Kong, Shenzhen under Grant B10120210117-KP02, in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (University Allowance, EXC 2077, University of Bremen, Germany), and in part by the National Natural Science Foundation of China under Grant 62271432. *(Corresponding author: Haizhou Li.)*

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

Tanja Schultz is with the Cognitive Systems Lab, University of Bremen, Germany.

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

Digital Object Identifier 10.1109/TBME.2023.3294242

characterize complex and irregular local structures, as well as global interaction among input features [27] is rather limited.

An EEG-based auditory (spatial) attention detection pipeline consists of a feature extraction frontend and a classification backend. Previous studies suggest that the locus of auditory attention is neurally encoded. However, the spatially-sensitive neurons are broadly distributed across the scalp [16], [17], [28], [29]. Therefore, in order to recognize auditory attention, it is necessary to model not only the neuronal responses of individual EEG electrodes, but also their interactions and collective activation patterns. State-of-the-art CNN-based ASAD methods have not yet effectively exploited the multivariate information of EEG signals in the spatial domain. The need for modeling the inter-channel relationship led us to investigate graph convolutional networks (GCNs), which offer several advantages.

First, GCNs extend the theory of signal processing to graphs and generalize the convolution operation in the non-Euclidean domain [30]. Therefore, GCNs are better suited than traditional CNNs to process graph-structured data such as EEG signals, which are discrete and discontinuous in the spatial domain [31]. GCNs have proven useful in many tasks where topological relationships between input features matter, such as human pose recognition [32], traffic prediction [33], and disease prediction [34]. Recently, GCNs were studied for brain activity analysis, using e.g. functional magnetic resonance imaging (fMRI) [35], MEG [36], and EEG data [37], [38] but we are not aware of any GCN study for auditory (spatial) attention decoding. For these reasons, we anticipate that investigating how ASAD might benefit from a graph-structured representation of EEG signals will provide new insights.

Second, studies show that some regions in the listener's brain are more closely related to attentional selection than others [16], [17], [39]. As EEG signals are measured from multiple scalp locations, some EEG channels obviously provide more information about auditory attention than others [21], [23]. This has motivated many studies on channel selection in auditory (spatial) attention detection [6], [12], [40]. Regarding auditory attention, it remains an open question about which nodes of the EEG graph are more relevant than others. Moreover, the distribution of effective channels varies from subject to subject [6], [16]. Therefore, it is reasonable to develop a *channel-wise attention* mechanism that dynamically assigns differentiated weights to EEG channels at run-time. Unlike the traditional manual selection of relevant channels, the proposed channel-wise attention mechanism is capable of deriving dynamic weights from the EEG channels across different spatial locations.

In this article, we proposed EEG-Graph Net, which makes auditory attention decisions based on EEG signals and their topological relationship. The main contributions of this work can be summarized as follows: 1) We proposed a way to represent the EEG channels, thus the brain activities, as an EEG Graph where an EEG channel is seen as a node, and a connection between two EEG channels as an edge in a topological graph for the first time. 2) We learned to assign differentiated weights dynamically to the nodes and edges in the graph according to their contributions to the ASAD task. 3) We showed the effectiveness and superiority of the EEG-Graph Net through extensive ablation study, data visualization, and experiments on two publicly available EEG databases. Moreover, the EEG-Graph Net is more interpretable, therefore, potentially revealing the neural mechanisms underlying selective attention processing.

The remainder of this article is organized as follows. In Section II, we elaborate on the proposed EEG-Graph Net pipeline for decoding auditory spatial attention. In Section III, we describe: 1) the used databases and processing; as well as 2) contrastive models and their application to the databases. Details of the experimental results are reported and analyzed in Section IV. In Section V, we discuss our findings and conclude in Section VI.

II. Methods

We would like to represent the input EEG signals as a data structure that reflects the biological topology of the human brain, rather than a collection of independent signals. The data structure is then processed by an EEG-Graph Net, which introduces an attention mechanism to dynamically assign differentiated weights to the EEG channels. As illustrated in Fig. 1, the EEG-Graph Net mainly consists of three modules, namely a graph representation module, a biologically inspired channel-wise attention module, and a graph structure learning mechanism. Unlike the traditional ASAD techniques which involve handcrafted EEG features, the proposed EEG-Graph Net performs in a data-driven end-to-end manner.

By applying a moving window to the raw EEG data, we obtained a sequence of small *decision windows*, each of which was processed independently for feature representation and detection decision. Let $\mathbf{C} = [c_1, \ldots, c_i, \ldots, c_N] \in \mathbb{R}^{T \times N}$ be a $T \times N$ matrix of EEG signals for a decision window T, where $c_i \in \mathbb{R}^{T \times 1}$ is a time series of T samples from the i-th of N EEG channels.

A. EEG Graph

We first encode the multi-channel EEG signals $\mathbf{C}$ into a graph data structure, as shown in Fig. 1(a). where a node is associated with an EEG channel whereas an edge represents the connection between two nodes in the graph. The graph representation reflects the spatial distribution of the EEG electrode placement, which captures the topological structure of the human brain. We expect that the relationship between EEG channels, i.e. the nodes, that an EEG Graph Net learns, also describes the complex relationship among the brain regions.

Let $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ denote a connected graph, where $\mathcal{V} = \{\mathcal{V}_1, \ldots, \mathcal{V}_i, \ldots, \mathcal{V}_N\}$ is instantiated by a set of N nodes $\mathbf{V} = \{v_1, \ldots, v_i, \ldots, v_N\}$. In practice, $\mathbf{V}$ takes the values of the input EEG signals $\mathbf{C}$. For a series of T graph instances within a decision window of T samples, we have $\mathbf{V} \in \mathbb{R}^{T \times N}$. We also have $\mathcal{E}_{i,j} = (\mathcal{V}_i, \mathcal{V}_j)$ and $\mathcal{E}_{i,j} \in \mathcal{E}$ to represent the edges of the graph, that is instantiated by an adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$. In practice, $\mathbf{A}$ is initialized by the connectivity of a set of nodes based on the international 10–20 standards [41] as depicted in Fig. 1(a). The adjacency matrix $\mathbf{A}$ reflects the spatial relationship of the EEG channels. Specifically, each element a_{ij} of the adjacency matrix $\mathbf{A}$ indicates the strength of the connection between the i-th and j-th nodes.

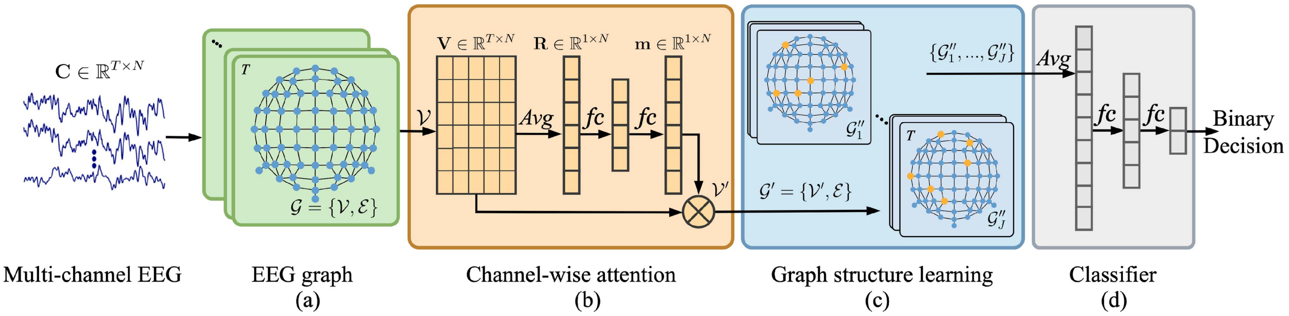
$\mathbf{C} \in \mathbb{R}^{T\times N}$
T
$\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$
$\mathbf{V} \in \mathbb{R}^{T\times N}$
$\mathbf{R} \in \mathbb{R}^{1\times N}$
$\mathbf{m} \in \mathbb{R}^{1\times N}$
$\mathcal{V}$
Avg
fc
fc
$\mathcal{V}'$
$\{\mathcal{G}''_1, \ldots, \mathcal{G}''_J\}$
$\mathcal{G}''_1$
T
$\mathcal{G}''_J$
$\mathcal{G}' = \{\mathcal{V}', \mathcal{E}\}$
Avg
fc
fc
Binary Decision
Multi-channel EEG
EEG graph
(a)
Channel-wise attention
(b)
Graph structure learning
(c)
Classifier
(d)

Fig. 1. A schematic diagram of the proposed EEG-Graph Net, which consists of four modules: an EEG graph representation module, a channel-wise attention module, a graph structure learning module, and a classifier. Taking multi-channel EEG data as input, the network is trained to detect auditory spatial attention by making a binary decision. *Avg* denotes an average pooling layer, *fc* denotes a fully-connected layer.

A graph $\mathcal{G}$ is then instantiated by $\mathbf{G} = \{\mathbf{V}, \mathbf{A}\}$. The series of $\mathbf{G}$ instances are modulated subsequently by a channel-wise attention mechanism that learns to assign differentiated weights to the nodes during run-time inference, as shown in Fig. 1(b). It is further modulated by a graph convolutional layer, as shown in Fig. 1(c), according to the topological relationship of nodes. Finally, a classifier is employed to decode the auditory spatial attention, as shown in Fig. 1(d).

B. Channel-Wise Attention

Neural attention occurs in human brain, which selects one of the acoustic stimuli and enhances/prioritizes its processing over that of others [42], [43]. Auditory attention allows us to precisely listens to a sound of interest and attenuates others in a cocktail party [1], [29], [44]. The study of computational neural attention is motivated by this human ability [45], [46], [47]. Briefly, the idea is to model the neural attentional modulation by assigning differentiated weights to relevant and irrelevant inputs. Note that the differentiated weights are dynamically generated by a computational attention mechanism at run-time, as opposed to a set of pre-trained weights.

Multi-channel EEG signals recorded from different scalp regions manifest different functional roles of human brain in spatial auditory processing [4], [16], [48]. They contribute differently to the decoding of auditory spatial attention in a listening brain [6], [12], [24], [40]. We design a *channel-wise attention* mechanism [46], [47], that learns to assign differentiated weights to EEG channels, i.e., nodes, dynamically according to their individual contributions to the ASAD task. While the channel-wise attention takes a graph as input, it only operates on the nodes without involving the edges. The detailed implementation is illustrated in Fig. 1(b) and further explained next.

We first aggregate a series of T samples in a decision window for $\mathcal{V}_i$ by average-pooling, generating a spatial descriptor $\mathbf{R} = [r_1, \ldots, r_i, \ldots, r_N] \in \mathbb{R}^{1\times N}$ for the input EEG signals $\mathbf{C}$, where r_i is for the i-th node $\mathcal{V}_i$, which can be obtained as follows:

$$r_i = Avg(v_i) \tag{1}$$

where $Avg(\cdot)$ denotes an average-pooling layer.

Second, a simple gating mechanism is adopted to make use of the aggregated information, which is parameterized by two fully-connected (fc) layers [46]. The gating mechanism is expected to dynamically generate differentiated weights for individual nodes based on importance.

$$\mathbf{m} = \mathbf{w}_2(tanh(\mathbf{w}_1\mathbf{R} + \mathbf{b}_1)) + \mathbf{b}_2 \tag{2}$$

where $\mathbf{w}_1$ and $\mathbf{w}_2$ is the parameter of the first and the second fc layers, respectively. $\mathbf{b}_1, \mathbf{b}_2$ are the bias terms of two fc layers. $\mathbf{m} \in \mathbb{R}^{1\times N}$ is the attention mask generated by the spatial attention mechanism, which is broadcast repeatedly along the temporal axis to form a matrix $\mathbf{M}$, we have $\mathbf{M} \in \mathbb{R}^{T\times N}$. $\mathbf{V}$ is then modulated by the attention mask $\mathbf{M}$ as follows,

$$\mathbf{V}' = \mathbf{M} \bigotimes \mathbf{V} \tag{3}$$

where $\bigotimes$ denotes a point-wise multiplication. With the channel-wise attention mechanism, we modulate the series of T graphs for a decision window. As we only modulate the values of the nodes, the modulated graph $\mathcal{G}' = \{\mathcal{V}', \mathcal{E}\}$ is instantiated as $\mathbf{G}' = \{\mathbf{V}', \mathbf{A}\}$.

C. Graph Structure Learning

With EEG-Graph, we seek to learn a graph structure, that is optimized for auditory spatial attention decisions. We would like to modulate the value of a graphical node $\mathbf{V}'$ according to its topological connections with other nodes and their values. It is apparent that a stronger edge between two nodes denotes a higher level of inter-node dependency. In this way, the values of the nodes and edges not only vary with the input signals but are also modulated by the underlying graph structure. Here we would like to learn a set of parameters that describe the underlying graph structure.

In practice, we apply a graph convolution on the input graph $\mathbf{G}'$ to modulate the value of a graphical node $\mathbf{V}'$. The graph convolution, also called spectral graph filtering, extends the convolutional operation to the graph domain using the spectral filters computed from the normalized graph Laplacian [49]. The Laplacian matrix of graph $\mathbf{G}'$ can be represented as:

$$\mathbf{L} = \mathbf{D} - \mathbf{A} \tag{4}$$

where $\mathbf{D}$ is the degree matrix of graph, $d_{ij} = \sum_j a_{ij}$. Here, the ranges of i and j for the summation of d_{ij} are from 1 to N, where N is the total number of nodes in the graph. As the adjacency matrix $\mathbf{A}$ is symmetric positive semi-definite, $\mathbf{L} \in \mathbb{R}^{N\times N}$ can be orthogonalized and diagonalized via eigen-decomposition as follows:

$$\mathbf{L} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T \tag{5}$$

where $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues (spectrum), and $\mathbf{U}$ is the corresponding eigenvector.

For a given spatial signal $\mathbf{x} \in \mathbb{R}^{1\times N}$, its graph Fourier transformation can be expressed as follows:

$$\hat{\mathbf{x}} = \mathbf{U}^T\mathbf{x} \tag{6}$$

where $\hat{\mathbf{x}}$ represents the transformed signal in the frequency domain. The inverse graph Fourier transform is defined as:

$$\mathbf{x} = \mathbf{U}\mathbf{U}^T\mathbf{x} = \mathbf{U}\hat{\mathbf{x}} \tag{7}$$

Following the definition by Shuman et al. [50], the graph convolution operator $\star_g$ can be applied with the spectral graph convolution as follows:

$$g_\theta *_g \mathbf{x} = g_\theta(\mathbf{L})\mathbf{x} = g_\theta(\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T)\mathbf{x} = \mathbf{U}g_\theta(\mathbf{\Lambda})\mathbf{U}^T\mathbf{x} \tag{8}$$

where $g_\theta(\cdot)$ denotes the filter function. $g_\theta(\mathbf{\Lambda})$ is a diagonal matrix filled with a set of learnable parameters Θ to describe the graph structure.

At run-time, the convolutional process, as parameterized by Θ, takes the instantiation of $\mathcal{V}'$, i.e. $\mathbf{x} = \mathbf{V}'$, and the adjacency matrix $\mathbf{A}$ as input, and generates $\mathcal{G}'' = \{\mathcal{V}'', \mathcal{E}'\}$ as the output. We note that $\mathcal{G}'$ is the weighted graph of the original EEG graph structure $\mathcal{G}$. Here $\mathcal{G}'$ is further modulated by the graph convolutional layer to generate $\mathcal{G}''$. The graph convolution is applied to the instantiation of $\mathcal{G}''$, i.e. $\mathbf{G}'' = \{\mathbf{V}'', \mathbf{A}'\}$. Assuming that we employ J filters in the graph convolution, we obtain an output of the graph structure as $\{\mathcal{G}''_1, \ldots, \mathcal{G}''_J\}$, as illustrated in Fig. 1(c), that will be taken by the back-end classifier for a binary decision.

D. Classifier

As shown in Fig. 1, the neural architecture features a data-driven end-to-end solution. First, we present a decision window of EEG signals, T samples across N channels, as T graphs. Second, an adaptive channel-wise attention module applies differentiated weights to derive $\mathbf{G}'$. During training, we update the parameters for the gating mechanism, i.e., $\Omega = \{\mathbf{w}_1, \mathbf{w}_2, \mathbf{b}_1, \mathbf{b}_2\}$. Third, the graph structure learning module applies the convolution operation to $\mathbf{G}'$ in the non-Euclidean space, and derives $\mathbf{G}''$. During training, we update the parameters for the filters, i.e., Θ. We adopt a binary cross-entropy loss as the learning objective:

$$Loss = -\frac{1}{K}\sum_{k=1}^{K} y_k \cdot log p_k + (1 - y_k) \cdot log(1 - p_k) \tag{9}$$

where y_k is the ground-truth label of the k-th decision window, while p_k is the predicted probability of the k-th decision window.

The detailed training algorithm is summarized in Algorithm 1. For an ablation study, we also implement a reduced version of the EEG-Graph Net, that is referred to as GCN, by skipping the channel-wise attention module in Fig. 1. In other words, the graph structure learning module, Fig. 1(c), takes the EEG graph directly as input.

Algorithm 1: A Training Algorithm of the EEG-Graph Net.

```
Input: Multi-channel EEG data $\mathbf{C}$, the class labels $y_j$
  corresponding to the EEG, the adjacency matrix $\mathbf{A}$, the
  learning rate $\rho$, the number of filters $J$, and other model
  hyper-parameters
Output: The model parameters $\Omega$, $\Theta$, and parameters of the
  classifier
    1: Randomly initialize the model parameters
    2: Assign the EEG signals, $\mathbf{C}$, to EEG graph $\mathcal{G}$
    3: Calculate the Laplacian matrix $\mathbf{L}$
    4: Calculate the eigenvector matrix $\mathbf{U}$
  Repeat
      Forward Pass:
      a. Calculate the spatial descriptor $\mathbf{R}$ based on (1)
      b. Generate the attention mask $\mathbf{M}$ based on (2)
      c. Generate the graph $\mathcal{G}'$ by applying the attention
      mask based on (3)
      d. Update the adjacency matrix by the graph
      convolution based on (4)-(8)
      e. Calculate the modulated graphs $\{\mathcal{G}''_1, \ldots, \mathcal{G}''_J\}$ from
      $J$ convolution filters
      f. Calculate the results of the fully-connected layers in
      Fig. 1(d)
      g. Calculate the loss function according to (9)
      Backward Pass:
      a. Update $\Theta$, $\Omega$, and other parameters through
      back-propagation
  until The iteration satisfies the convergence condition
```

III. Experiments

A. EEG Databases

In this study, experiments are conducted on two publicly available databases, which are denoted as KUL [51] and DTU databases [52].

1) KUL Database [23], [51]: This database consists of 16 normal-hearing subjects, who were instructed to selectively attend to one of the two simultaneous speakers. The speech stimuli consist of four Dutch stories, narrated by three male Flemish speakers. The stimuli were either presented dichotically (one speaker per ear) or after head-related transfer function (HRTF) filtering to simulate speech from 90° to the left and 90° to the right of the subject. Throughout the experiments, the order of presentation of both conditions was randomized over different subjects. There are equal amounts of left-attended and right-attended trials. 64-channel EEG signals were recorded using a BioSemi ActiveTwo device at a sampling rate of 8,192 Hz. In this study, we used the downsampled EEG data at 128 Hz. In total, 8×6 minutes of EEG data were collected for each subject, accumulating to 12.8 hours of EEG data for all 16 subjects. The original experiment also included 12 additional trials of 2

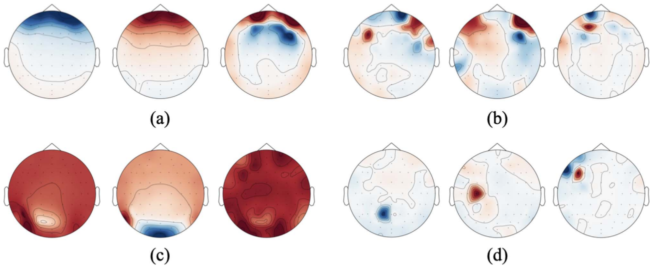

Fig. 2. Four types of artifacts, (a) eye blinks, (b) muscle activities, (c) heartbeat, and (d) generic discontinuities (sudden amplitude fluctuations in a channel).

minutes each, but these trials were repetitions of earlier stimuli and were not used in this study.

2) DTU Database [52], [53]: The EEG data were collected from 18 normal-hearing subjects, who attend to one target speaker and ignore the other in the presence of two competing speakers. The speech stimuli consist of speech by a male and a female native speaker who simultaneously speak in anechoic or reverberant rooms. The speech mixtures were presented to the subjects, with the two speech streams lateralized at respectively -60° and +60° along the azimuth direction. The position and the gender of the target speaker were randomized across the trials, resulting in equal amounts of left-attended and right-attended trials. 64-channel EEG data were recorded at a sample rate of 512 Hz using a BioSemi Active system. The data was then downsampled to 128 Hz to be the same as that of the KUL database. Each subject listened to 60 trials in total, and each trial contained auditory stimuli with a duration of 50 seconds. In total, 50 minutes of EEG data were collected for each subject, accumulating to 15 hours of EEG data for all 18 subjects.

B. Data Preprocessing

The EEG data were firstly re-referenced to the average response of all channels. The previous ASAD studies suggest that β-band (12-30 Hz) is the most informative EEG frequency band as far as auditory spatial attention is concerned [21], [22], [23]. EEG data were then all bandpass filtered in the β-band by a 6th-order Chebyshev Type II bandpass filter. Unless otherwise stated, the ASAD performance is evaluated with the β-band EEG data in our study. To make the topological graph of EEG as interpretable as possible, artifacts were removed by performing the independent component analysis (ICA) with EEGLAB toolbox [54], [55]. Several kinds of artifacts were detected and removed. A set of examples is illustrated in Fig. 2. Finally, the EEG data of each decision window were converted into a graph.

Considering that humans are able to switch attention from one speaker to another within 2-second [56], the real-world applications call for low-latency ASAD solutions. Therefore, we are interested in the study of short decision windows, e.g. 0.1-second, 0.2-second, 0.5-second, 1-second, and 2-second decision windows. After pre-processing, we obtained a total of 2,880 decision windows per subject, resulting in 46,080 decision windows for the 1-second case in the KUL database. In the DTU database, we obtained 2,940 decision windows per subject, totaling 52,920 decision windows for the 1-second case. The attention label, which represents the ground truth, is provided in both the KUL and DTU databases.

TABLE I
SETTINGS OF EEG-GRAPH NET HYPER-PARAMETERS AND THE HYPER-PARAMETER SEARCH GRID

Hyper-parameter	Value	Grid
Attention parameter	4	[2, 4, 8, 16]
filter number J	5	[3, 5, 10]
Hidden size	8	[4, 8, 16, 32]
Learning rate ρ	10^{-3}	[10^{-1}, 10^{-2}, 10^{-3}]
Batch size	20	[10, 20, 50]
Epochs	100	[50, 100, 150, 200]

C. Model Implementation and Evaluation

We evaluated the performance of the proposed and baseline methods subject by subject. For each subject, we split the data into five grand folds. We used one grand fold for testing, and the rest four grand folds for hyper-parameter tuning using an inner 5-fold cross-validation [57]. We repeated the above process five times over the five grand folds and calculated the average results. In line with previous studies [23], [24], [25], the ASAD accuracy is defined as the percentage of correctly classified decision windows on the test set. The average performance over all the testing folds is reported as the final result.

The hyper-parameters were chosen via a grid search over a set of reasonable values on a validation set, as summarized in Table I. The Adaptive Moment Estimation (Adam) optimizer [58] was employed to minimize the cross-entropy loss function with a learning rate of 10^{-3}. In addition, dropout [59] and batch normalization [60] were applied to prevent over-fitting and improve generalization. The batch size was set as 20. In our experiments, we trained the model for a total of 100 epochs with an early stopping scheme, that is, to terminate the iterations as soon as no significant improvement in the loss function was detected for 10 consecutive epochs. All models in this study were implemented with the TensorFlow framework and trained on an NVIDIA TITAN Xp Pascal GPU.

Taking a 1-second decision window as an example, we describe the network configuration of EEG-Graph Net in detail as follows. The 1-second EEG signals $\mathbf{C} \in \mathbb{R}^{128\times64}$, i.e. 128 samples by 64 channels, are firstly taken as the input values $\mathbf{V}$ of the set of nodes $\mathcal{V}$. The channel-wise attention module, which consists of two fc layers (input: 64, hidden: 4, output: 64), derives an output $\mathbf{V}' \in \mathbb{R}^{128\times64}$. In the graph structure learning module, the size of the graph convolution kernel is $N = 64$ and the number of filters is $J = 5$. The output of the graph convolution layer is therefore $\mathbf{V}'' \in \mathbb{R}^{5\times128\times64}$. Then, an average pooling layer is applied along the temporal dimension with the size of 5×64. The data are flattened into a one-dimensional vector as inputs for the classification decision. Specifically, two fc layers (input: 320, hidden: 8, output: 2) are employed.

D. Statistical Analysis

For statistical analyses, descriptive statistics were used for means and standard deviations (SDs). The Kolmogorov-Smirnov test was used to confirm the normality of the distribution of the data, prior to the selection of appropriate statistical tests. Paired *t*-tests with a 0.05 significance level were employed to compare differences between ASAD performance of two different models. All analyses were carried out with IBM SPSS statistics software in this study.

IV. Results

We conducted extensive experiments on the two publicly available datasets, namely KUL and DTU. Firstly, we compared the ASAD performance of the GCN model and CNN model, where we would like to observe the contributions of the graph representation. Secondly, we conducted extensive ablation experiments, where we would like to observe the contributions of the channel-wise attention mechanism. Then, we tested the EEG-Graph Net with five different detection window sizes and report the ASAD performance on such low-latency settings. Finally, we also evaluated the EEG-Graph Net with the low-density setting of EEG signals that is more suitable for practical applications.

A. GCN versus CNN Decoder

We compared the GCN model with the CNN model by Vandecappelle et al. [23] on the same ASAD task. 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 activate function is ReLU and the loss function is the cross-entropy. For 1-second decision window, the CNN model takes $\mathbf{C} \in \mathbb{R}^{128\times 64}$ as the input and makes a binary decision. For a fair comparison, we re-implemented the CNN model based on the published codes [23] with our experimental setup for both KUL and DTU databases and tuned the hyperparameters of the CNN model in the same way as we did for the GCN model.

As shown in Fig. 3, the CNN model attains a mean ASAD accuracy of 63.3% (SD: 5.96%) on the DTU database and 84.1% (SD: 10.16%) on the KUL database with 1-second decision window, respectively. The proposed GCN model consistently outperforms CNN model by a large margin on both databases. Specifically, the GCN model achieves an average improvement of 9.4% (mean: 72.7%, SD: 7.39%) or an error reduction of 25.6%, i.e. from 36.7% to 27.3%, on the DTU database, 7.6% (mean: 91.7%, SD: 5.54%) or an error reduction of 47.8%, i.e. from 15.9% to 8.3%, on the KUL database, respectively. In addition, it is worth noting that the number of parameters of the GCN model is clearly lower than that of the CNN model. As stated in [23], the CNN model consists of approximately 5,500 parameters, whereas our GCN model consists of around 3,500 parameters. Considering EEG signals' typically limited dataset size, the simplicity makes the GCN an excellent choice.

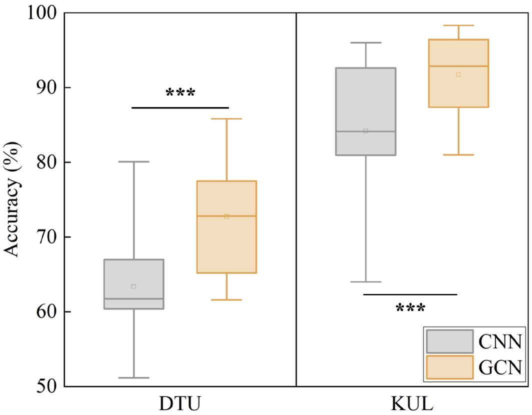

Fig. 3. Accuracy of CNN and GCN model for decoding auditory spatial attention among all subjects on DTU and KUL databases with 1-second decision window. Statistically significant difference: $^{***}p$ <0.001.

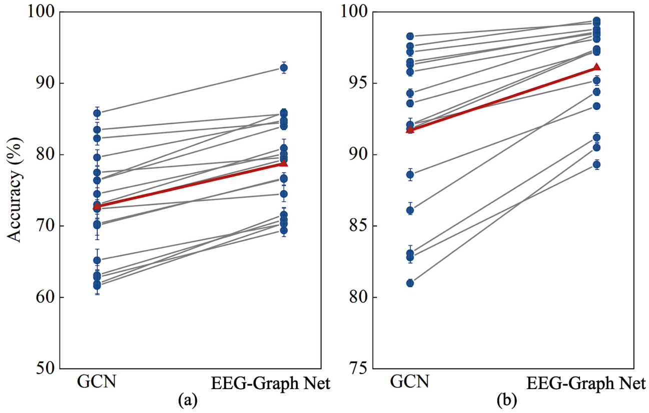

Fig. 4. Auditory spatial attention detection accuracy of the GCN and EEG-Graph Net with 1-second decision window on two databases. Blue dots: individual results (mean ± SD). Gray lines: same subjects. Red triangles: mean accuracies of all subjects. (a) DTU database (b) KUL database.

To summarize, the simple GCN model attains a significantly higher average accuracy than the CNN model (p <0.001) on both databases. These results indicate that the topology-aware representations generated by GCN tremendously enhance the discriminative ability of CNN features, therefore, contribute to the enhancement of ASAD performance.

B. Ablation Analysis

To appreciate the contributions of the channel-wise attention mechanism, we conducted an ablation analysis using 1-second decision window as a case study. As shown in Fig. 4, the ASAD accuracy of the GCN and EEG-Graph Net are reported across all subjects on KUL and DTU databases, respectively.

On the KUL database, the EEG-Graph Net achieves a relatively high ASAD accuracy (mean: 96.1%, SD: 3.22%), which significantly outperforms that of the GCN model with an average improvement of 4.4% or an error reduction of 52.4%, i.e. from 8.3% to 3.9%, (p <0.001). The results on the DTU database corroborate the findings on the KUL database. Specifically, the EEG-Graph Net significantly outperforms GCN model with an average improvement of 6.0% (mean: 78.7%, SD: 6.47%) or an error reduction of 22.0%, i.e. from 27.3% to 21.3%, on the

TABLE II
THREE EEG-GRAPH NET CONFIGURATIONS IN A COMPARATIVE STUDY

Model	Description
EEG-Graph Net(NE)	channel attention followed by graph convolution
EEG-Graph Net(EN)	graph convolution followed by channel attention
EEG-Graph Net(P)	channel attention in parallel with graph convolution

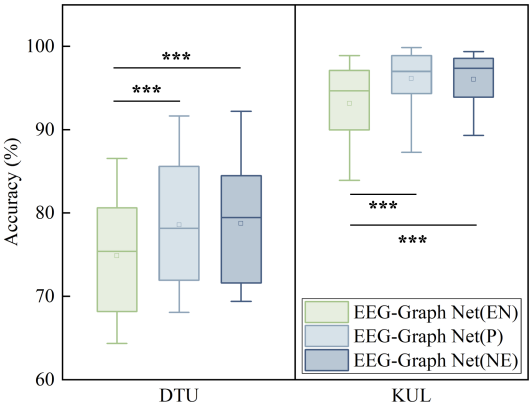

Fig. 5. Auditory spatial attention detection accuracy of three contrastive network configurations with 1-second decision window across all subjects in KUL and DTU databases, respectively. Statistically significant difference: $^{***}p < 0.001$.

DTU database ($p < 0.001$). In sum, the proposed channel-wise attention mechanism contributes to the performance gains of EEG-Graph Net over the GCN model on two publicly available datasets.

Summarising the results of Sections IV-A and IV-B, the EEG-Graph Net obtains an average accuracy improvement of 12.0% and 15.4% over CNN on KUL and DTU databases with 1-second decision window, in which the channel-wise attention mechanism and GCN decoder make equally significant contributions.

C. Effect of Network Configuration

Given that channel-wise attention and graph structure learning modules in our ASAD model focuses on 'node' and 'edge' respectively, we are interested in knowing what is the best way of arranging these two modules. As summarized in Table II, the channel-wise attention and graph structure learning modules can be placed in a parallel or sequential manner. In a sequential manner, we can have either node-edge order EEG-Graph Net(NE) as shown in Fig. 1, or edge-node order EEG-Graph Net(EN). In a parallel manner, the model is referred to as EEG-Graph Net(P). We evaluated these three network configurations with 1-second decision window on both KUL and DTU databases in a comparative study.

As depicted in Fig. 5, the EEG-Graph Net(EN) obtains a mean decoding accuracy across all subjects of 93.1% (SD: 4.94%) on KUL and 74.9% (SD: 6.54%) on DTU with 1-second decision window. We observed that the EEG-Graph Net(NE) (node-edge sequence) outperforms significantly the EEG-Graph Net(EN) (edge-node sequence) ($p < 0.001$).

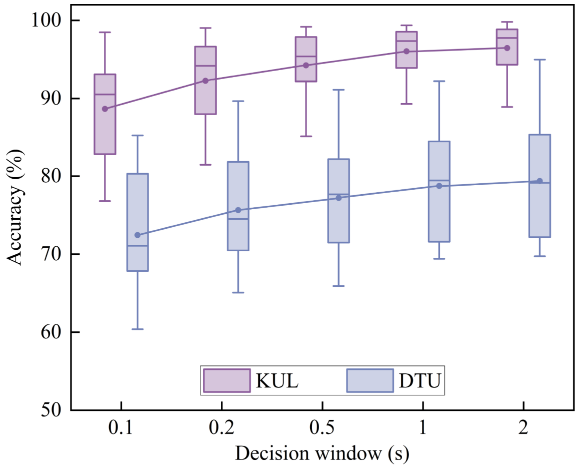

Fig. 6. Auditory spatial attention detection performance of the EEG-Graph Net for five decision window sizes across all subjects in KUL and DTU databases, respectively.

The channel-wise attention only involves the collection of individual channels, i.e. nodes, whereas the graph convolutional layer involves the graph structure, i.e. both nodes and edges. We consider that the node-edge sequence is a more logical order than edge-node sequence because the former makes biological sense. In the human neural attention process, the cortical neurons first encode stimulus properties locally, i.e., bottom-up stimulus responsiveness, then top-down attention modulates the magnitude of these responses across widespread cortical regions, i.e., global connections, according to task demands [42], [43], [44], [48]. It is worth noting that the EEG-Graph Net(EN) still outperforms GCN model by an average accuracy of 1.4% and 2.2% on the KUL and DTU databases. This confirms the contribution of the channel-wise attention module.

In addition, we observed that the EEG-Graph Net(P) performs similarly to EEG-Graph Net(NE) on both KUL database (mean: 96.2%, SD: 3.48%) and DTU database (mean: 78.5%, SD: 6.93%) with 1-second decision window, respectively. We found no statistically significant differences for either KUL ($p = 0.63$), or DTU ($p = 0.59$) between EEG-Graph Net(P) and EEG-Graph Net(NE). As the sequential models involve much less computational and parameters overhead than the parallel model, it is logical to place channel-wise attention and graph structure learning modules in a node-edge sequential manner. Therefore, the EEG-Graph Net is used to denote the EEG-Graph Net(NE) configuration throughout this article.

D. Low-Latency ASAD

We would like to further evaluate the feasibility of EEG-based ASAD in practical BCI applications. We report the ASAD accuracy of the EEG-Graph Net with relatively short decision windows ranging from 0.1-second to 2-second on both databases, as shown in Fig. 6.

On the KUL database, the EEG-Graph Net demonstrates superior ASAD performance with 1-second decision window (mean: 96.1%, SD: 3.22%) and 2-second decision window (mean: 96.5%, SD: 3.04%). Though the accuracy degrades for decision window sizes below 1-second, the EEG-Graph Net is competitive, with a mean accuracy of 94.2% (SD: 4.41%) for 0.5-second decision window and 92.3% (SD: 5.40%) for 0.2-second decision window. In general, a larger decision window provides a better decoding performance, which is in line with the findings in previous ASAD studies [7], [21], [22], [23], [24], [25]. It is worth noting that the EEG-Graph Net achieves a relatively high accuracy (mean: 88.7%, SD: 6.59%) when operating at a high temporal resolution of 0.1-second.

On the DTU database, the EEG-Graph Net achieves a mean accuracy of 72.5% (SD: 7.41%) for 0.1-second, 75.7% (SD: 6.89%) for 0.2-second, 77.2% (SD: 6.71%) for 0.5-second, 78.7% (SD: 6.47%) for 1-second, and 79.4% (SD: 7.16%) for 2-second decision windows, respectively. The ASAD performance on the DTU database is lower than that on the KUL database. This result is consistent with those in other studies [7], [24], [25], [61]. The exact reason for the difference between the ASAD performance of these two databases remains unclear. The major difference between the DTU and the KUL database, that we know, is that the two auditory stimuli arrive 60° to the left and 60° to the right of the listening participants on the DTU database [52], but arrive from ± 90° instead on the KUL database [51]. Moreover, in the DTU dataset, the auditory stimuli are presented with a room reverberation at different levels, which might adversely affect the cortical tracking of attended speech streams [62]. In contrast, in the KUL database, the auditory stimuli are presented to the listeners in an anechoic chamber. Therefore, it is more challenging for the listeners to attend to the target speech stimulus, thus, the attention detection, in the DTU database than in the KUL database.

Overall, the proposed EEG-Graph Net performs reasonably well at high temporal resolutions, which is comparable to the time required for auditory attention-switch by humans. We are not aware of other auditory (spatial) attention detection models that perform similarly in such low latency settings, i.e., around 100 ms. These results suggest that the real-time decoding of auditory (spatial) attention is within reach, which paves the way for daily-life neuro-steered hearing prostheses.

E. Low-Density EEG-Based ASAD

As low-density EEG systems show great potential in portable EEG-based BCI devices, such as neuro-steered hearing devices [6], [12], we were interested in how the number of EEG channels, i.e. the density of EEG signals, has an impact on the ASAD performance.

Both KUL and DTU databases were collected with a 64-channel BioSemi ActiveTwo system. In this study, 32-channel and 16-channel EEG were selected following the electrode locations of the international 10/20 system [41]. Fig. 7 summarizes the performance of different ASAD approaches with 1-second decision windows based on 16-channel and 32-channel EEG signals over all subjects on the KUL and DTU databases.

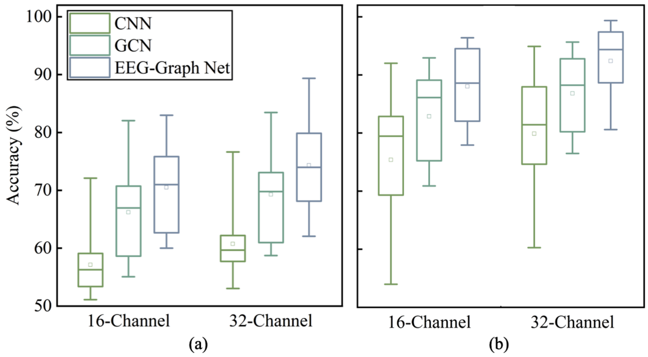

Fig. 7. Auditory spatial attention detection accuracy of the EEG-Graph Net, GCN, and CNN models with 32-channel EEG and 16-channel EEG, respectively. (a) DTU database and (b) KUL database.

On the KUL database, the ASAD accuracy of CNN model degrades from 64-channel (mean: 84.1%, SD: 10.16%) to 32-channel (mean: 79.9%, SD: 10.46%), and further to 16-channel (75.4%, SD: 11.01%). The ASAD performance for the GCN remains competitive (16-channel, mean: 82.9%, SD: 7.46%; 32-channel, 86.8%, SD: 6.31%), which significantly outperforms the CNN model (paired *t*-test: p <0.001). These results also support the claim that the topology-aware representations learned by the GCN are more effective than the CNN features. It is encouraging to see that the EEG-Graph Net still decodes auditory spatial attention accurately with 32-channel (92.4%, SD: 5.91%) and 16-channel (88.0%, SD: 6.23%).

On the DTU database, the ASAD accuracy of CNN model also decreases significantly from 64-channel EEG (mean: 63.3%, SD: 5.96%) to 32-channel (mean: 60.2%, SD: 5.84%), and further to 16-channel (mean: 56.7%, SD: 5.18%). For GCN, we observe a modest accuracy drop of 3.4% and 6.5% for 32-channel (mean: 69.3%, SD: 7.43%) and 16-channel (mean: 66.2%, SD: 7.75%) over 64-channel EEG, respectively. The mean accuracy for the EEG-Graph Net remains competitive with 16-channel EEG (mean: 70.5%, SD: 6.82%) and 32-channel EEG (mean: 74.3%, SD: 6.79%), which significantly outperforms the CNN model (p <0.001) and GCN model (p <0.001). These results again demonstrate that the graph-based representation and channel-wise attention modules yield a significant improvement in ASAD performance.

To conclude, the EEG-Graph Net works well with relatively low-density EEG systems, which makes itself a perfect candidate for neuro-steered hearing-assistive devices.

V. Discussions

We believe that the bio-inspired EEG-Graph Net exploits the topological structure of multi-channel EEG through three contributing modules, namely EEG graph data representation, channel-wise attention mechanism, and graph structure learning. We further validated our proposal through comparative studies.

We first compared the performance of our proposed EEG-Graph Net with other competing models in the literature. Besides its superior performance, the EEG-Graph Net is biologically motivated. By visualizing the attention weights assigned by the EEG-Graph Net at run-time, we are able to explain the contributions of EEG channels and their inter-channel relations from the perspective of the human brain topology.

TABLE III
AUDITORY SPATIAL ATTENTION DETECTION ACCURACY OF A NUMBER OF MODELS ON KUL DATABASE [51] FOR FIVE DIFFERENT DECISION WINDOW SIZES

Database	Model	Decision window (second)				
		0.1	0.2	0.5	1	2
KUL [51]	CSP (Geirnaert et al. [21])	-	74.8%	77.6%	79.1%	80.3%
	RGC (Geirnaert et al. [22])	-	72.2%	78.1%	79.4%	81.6%
	CNN[1] (Vandecappelle et al. [23])	74.3 ± 10.92%	78.2 ± 10.10%	80.6 ± 10.37%	84.1 ± 10.16%	85.7 ± 9.72%
	CNN[2] (Su et al. [24])	77.2 ± 8.24%	80.6 ± 8.33%	84.3 ± 8.56%	86.5 ± 7.99%	88.3 ± 7.89%
	CNN[3] (Su et al. [25])	80.8 ± 9.87%	84.3 ± 9.73%	87.2 ± 9.77%	90.1 ± 8.95%	91.4 ± 8.22%
	GCN (This work)	80.6 ± 7.36%	85.1 ± 6.51%	89.3 ± 6.33%	91.7 ± 5.54%	92.5 ± 5.28%
	EEG-Graph Net(This work)	**88.7** ± **6.59%**	**92.3** ± **5.40%**	**94.2** ± **4.41%**	**96.1** ± **3.22%**	**96.5** ± **3.04%**
DTU [52]	CNN[1] (Vandecappelle et al. [23])	56.7 ± 4.87 %	58.4 ± 5.94 %	61.7 ± 6.68 %	63.3 ± 5.96%	65.2 ± 7.51 %
	CNN[2] (Su et al. [24])	60.8 ± 5.02 %	63.7 ± 6.83 %	67.2 ± 7.74%	67.9 ± 7.41 %	69.5 ± 8.94 %
	CNN[3] (Su et al. [25])	65.7 ± 5.50%	68.1 ± 7.08%	70.8 ± 8.04%	71.9 ± 8.94%	73.7 ± 9.59%
	GCN (This work)	66.1 ± 7.55%	68.7 ± 7.61%	71.2 ± 7.57%	72.7 ± 7.39%	74.4 ± 7.41%
	EEG-Graph Net (This work)	**72.5** ± **7.41%**	**75.7** ± **6.89%**	**77.2**± **6.71%**	**78.7** ± **6.47%**	**79.4** ± **7.16%**

Note that the EEG-Graph net significantly outperforms other models in terms of decoding accuracy (P <0.001). CSP= Common Spatial Pattern, RGC= Riemannian Geometry Classifier, CNN=Convolutional Neural Network.

TABLE IV
THE PROPOSED MODELS WITH EEG GRAPH AND THREE OTHER CONTRASTIVE CNN MODELS

Model	Graph	Attention		
		Channel	Time	Frequency
CNN[1] (Vandecappelle et al. [23])	×	×	×	×
CNN[2] (Su et al. [24])	×	✓	×	✓
CNN[3] (Su et al. [25])	×	✓	✓	×
GCN (This work)	✓	×	×	×
EEG-Graph Net (This work)	✓	✓	×	×

Channel, Time, and frequency respectively denote channel-wise, Temporal, and frequency-wise attention mechanisms.

A. Comparative Study

As summarized in Table III, we performed a comparative study on a number of models on DTU [52] and KUL [51] databases.

On the KUL database, we started by comparing with two competitive traditional models, namely CSP-based [21] and RGC-based [22], which are known to give good performance. The GCN model obtains an average accuracy gain of 11.7% and 8.9% across all decision window sizes over CSP-based and RGC-based models, respectively. Similarly, the EEG-Graph Net outperforms these two classic models by a large margin (>15%) across all decision window sizes. These results, along with the previous studies [7], [9], [10], demonstrate that non-linear machine learning methods could be beneficial for rapid and reliable decoding of auditory attention (spatial) attention.

We also compared the proposed models with EEG graph, i.e. GCN and EEG-Graph Net, with three CNN models in the prior work. Their network configurations are summarised in Table IV for ease of reference. The GCN model outperforms the CNN models [23], [24], [25] by 7.3%, 4.5%, and 1.2% across all decision window sizes in terms of ASAD accuracy. The improvement of GCN over CNNs clearly validates that the topology-aware representations of EEG generated by the GCN significantly enhance the discriminative ability of 2D or 3D matrices used by regular CNNs. In addition, the EEG-Graph Net further enhances the ASAD accuracy and significantly outperforms the CNN models with consistent improvements of 10.2%, 6.8%, and 5.7%, respectively. The promising results, especially on such low-latency settings, confirm the effectiveness of our methods.

Like on the KUL database, GCN-based models consistently outperform CNN-based models on the DTU database too (p <0.001 for [23] and [24], $p = 0.002$ for [25]). The EEG-Graph Net improves over CNN-based decoders across all decision windows by a large margin of 15.6%, 10.9%, and 6.7%, respectively.

In summary, the graph-based representation and channel-wise attention mechanisms significantly improve the performance on both databases over the state-of-the-art ASAD algorithms. These results support our claim that EEG-Graph Net reliably decodes the auditory spatial attention for short decision windows, and confirm that the results are reproducible on independent databases.

B. Analysis of Channel-Wise Attention

Multi-channel EEG signals collected from various positions of the scalp are not equally informative as far as auditory attention is concerned [6], [12], [21]. The proposed channel-wise attention module provides a tool for interpreting how the spatially differentiated weights on EEG channels contribute to the performance gain. We visualized the attention mask $\mathbf{M}$ in the channel-wise attention module by aggregating over all 1-second decision windows for each individual on the KUL database in Fig. 8.

As expected, EEG channels indicative of neural activities related to speech processing function have higher weights. Consistent with the findings of previous ASAD studies [21], [23], [63], the spatial activation patterns of the β-band activity are mainly above the frontotemporal cortex. Specifically, higher weights are assigned to electrodes placed over the frontal and temporal regions than elsewhere by the channel-wise attention mechanism. In addition, we found that the spatial activation shows higher weights at electrodes placed over the left hemisphere. These results are in line with functional specialization in the human brain that the processing of continuous speech is reliant on the cortical regions of the left hemisphere [2], [64].

It is also clearly observed that the attention mask reflects an individual's attentional focus. They vary from subject to subject. These findings support that EEG signals exhibit subject-specific

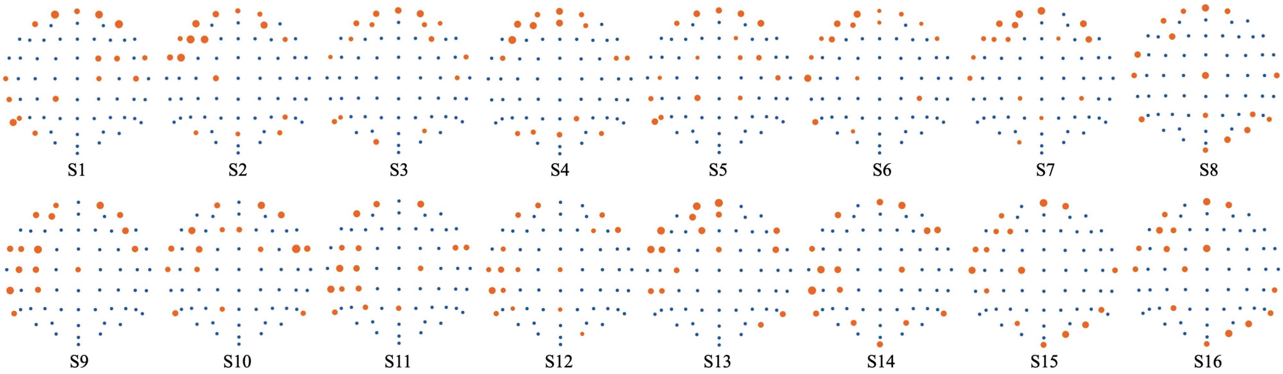

Fig. 8. Visualization of the channel-wise attention weights associated with the EEG electrodes in the KUL database for 16 subjects. The blue dots mark the 64 EEG electrodes, whereas the orange dots correspond to the top 16 electrodes. The sizes of orange dots are scaled by the attention weights. A larger orange dot denotes a higher weightage.

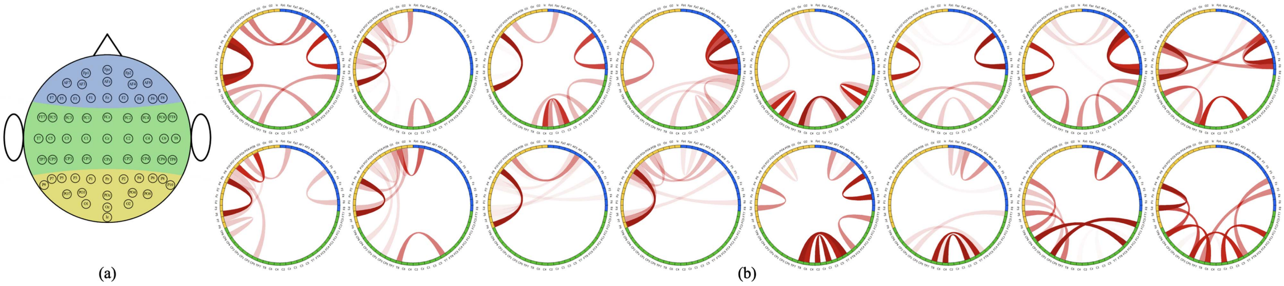

Fig. 9. Visualization of the inter-channel relationship in the KUL database. (a) The blue, green, and yellow area represents the electrodes in the frontal, central, and parieto-occipital groups [16]. (b) Inter-channel relationship for all individual subjects. The EEG channels are arranged by color and according to their topological positions. The edge with darker color corresponds to a higher weightage.

patterns due to the physiological and psychological individuality [16], [65]. Considering the individuality, the handcrafted features from EEG signals in previous studies would not be the best because we cannot expect one setting to work for all subjects. In contrast, the channel-wise attention module is able to assign differentiated weights dynamically to channels during run-time inference, which effectively addresses the subject individuality challenge.

C. Analysis of Inter-Channel Relationship

With the graph structure, the proposed EEG-Graph Net relates the feature representations with topological relationship of EEG channels, and makes the decision process more explainable. One unique property of the EEG-Graph Net is its ability to discover and model the intrinsic relationship between EEG channels. To look into such inter-channel relationship in the modulation of auditory attention, we visualized the top 10 connections between channels that have the largest edge weights in the adjacency matrix for each subject in the KUL database, as illustrated in Fig. 9.

It is clear from Fig. 9(b) that the connections between the right-hemisphere electrodes and left-hemisphere electrodes are of higher weights, for instance, channel pairs (AF7, AF8), (F3, F4), (F5, F6), (C1, C2),(C5, C6),(T7, T8),(TP7, TP8), (P5, P6), (P7, P8), and (P9, P10), indicating that neural dynamics between the left and the right hemisphere are essential for decoding auditory attention. This result is consistent with previous literature [21], [66], [67], which observes that the direction of auditory spatial attention is related to attentional lateralization, i.e., asymmetric changes of neural oscillations in left versus right hemisphere.

To explore whether the global inter-channel relationship contributes to the performance gain, EEG electrodes are divided into frontal, central, and parieto-occipital groups [16], as illustrated in Fig. 9(a). The learned edge importance localizes meaningful functional connections as far as ASAD concerned, including channel pairs (AF3, PO3), (AF4, PO4), (F1, T7), (F2, T8), (F5, P5), (F6, P6), (TP7, P5), and (CP6, P4). It is worth noting that global inter-channel relations are among the stronger connections for most subjects, which is in agreement with recent studies on the auditory attention modulation related functional brain networks [67], [68].

The inter-channel relationship is also subject-dependent to some extent, which can be explained by the fact that brain signals from different subjects are highly variable, discriminative, and semantic [16], [65]. Adding dynamic weights to the inter-channel relationship leads to significant improvements in decoding performance, which again suggests that auditory attention detection may benefit from feature representations and end-to-end learning as opposed to handcrafted feature extraction.

In sum, the EEG-Graph Net takes advantage of the inter-channel relations for the ASAD task. It also allows for improved explainability of detection results, which provides a means of study from the perspective of human brain topology.

VI. Conclusion

In this article, we proposed a novel data-driven model for decoding auditory spatial attention, which preserves the topology information of the brain and effectively learned discriminative EEG representations in the spatial domain. We confirmed that the proposed model benefits from the idea of the topologically aware EEG-graph. Extensive experiments showed that the proposed EEG-Graph Net achieved an average accuracy of 96.1% and 78.7% within 1 s on the KUL and DTU databases, respectively. The proposed model significantly outperformed the state-of-the-art ASAD approaches on both databases. In addition, it offers a higher level of explainability in EEG signal decoding, thus, marking a significant step toward explaining the attentional selection mechanism in the human brain. As a future work, it would be interesting to study model generalization to unseen subjects in a subject-independent setup.

Acknowledgment

The authors would like to thank Enze Su for the invaluable discussions and suggestions that improved this article.

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] 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.

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

[4] M. Wöstmann et al., "Spatiotemporal dynamics of auditory attention synchronize with speech," in *Proc. Nat. Acad. Sci. USA*, vol. 113, no. 14, pp. 3873–3878, 2016.

[5] 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.

[6] B. Mirkovic et al., "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.

[7] 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.

[8] A. de Cheveigné et al., "Decoding the auditory brain with canonical component analysis," *NeuroImage*, vol. 172, pp. 206–216, 2018.

[9] 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.

[10] 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.

[11] W. Biesmans et al., "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.

[12] 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.

[13] S. Cai et al., "Low latency auditory attention detection with common spatial pattern analysis of EEG signals," in *Proc. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 2772–2776.

[14] B. Mirkovic et al., "Target speaker detection with concealed EEG around the ear," *Front. Neurosci.*, vol. 10, 2016, Art. no. 349.

[15] L. Fiedler et al., "Single-channel in-ear-EEG detects the focus of auditory attention to concurrent tone streams and mixed speech," *J. Neural Eng.*, vol. 14, no. 3, 2017, Art. no. 036020.

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

[17] 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.

[18] 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.

[19] 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.

[20] 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, 2020.

[21] 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.

[22] S. Geirnaert and 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.

[23] S. Vandecappelle et al., "EEG-based detection of the locus of auditory attention with convolutional neural networks," *Elife*, vol. 10, 2021, Art. no. e56481.

[24] E. Su et al., "Auditory attention detection with EEG channel attention," in *Proc. IEEE 43rd Annu. Int. Conf. Eng. Med. Biol. Soc.*, 2021, pp. 5804–5807.

[25] E. Su et al., "STAnet: A spatiotemporal attention network for decoding auditory spatial attention from EEG," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 7, pp. 2233–2242, Jul. 2022.

[26] S. Cai et al., "Low-latency auditory spatial attention detection based on spectro-spatial features from EEG," in *Proc. IEEE 43rd Annu. Int. Conf. Eng. Med. Biol. Soc.*, 2021, pp. 5812–5815.

[27] Z. Wu et al., "A comprehensive survey on graph neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 4–24, Jan. 2021.

[28] A. Bednar, F. M. Boland, and E. C. Lalor, "Different spatio-temporal electroencephalography features drive the successful decoding of binaural and monaural cues for sound localization," *Eur. J. Neurosci.*, vol. 45, no. 5, pp. 679–689, 2017.

[29] 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.

[30] M. M. Bronstein et al., "Geometric deep learning: Going beyond Euclidean data," *IEEE Signal Process. Mag.*, vol. 34, no. 4, pp. 18–42, Jul. 2017.

[31] S. Zhang et al., "Graph convolutional networks: A comprehensive review," *Comput. Social Netw.*, vol. 6, no. 1, pp. 1–23, 2019.

[32] Z. Zou and W. Tang, "Modulated graph convolutional network for 3D human pose estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 11477–11487.

[33] M. Lv et al., "Temporal multi-graph convolutional network for traffic flow prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 6, pp. 3337–3348, Jun. 2021.

[34] P. Xuan et al., "Graph convolutional network and convolutional neural network based method for predicting lncRNA-disease associations," *Cells*, vol. 8, no. 9, 2019, Art. no. 1012.

[35] S. I. Ktena et al., "Metric learning with spectral graph convolutions on brain connectivity networks," *NeuroImage*, vol. 169, pp. 431–442, 2018.

[36] Y. Guo, H. Nejati, and N.-M. Cheung, "Deep neural networks on graph signals for brain imaging analysis," in *Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 3295–3299.

[37] T. Zhang et al., "GCB-Net: Graph convolutional broad network and its application in emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 13, no. 1, pp. 379–388, First quarter 2022.

[38] Q. Li et al., "Residual GCB-net: Residual graph convolutional broad network on emotion recognition," *IEEE Trans. Cogn. Devel. Syst.*, early access, Jan. 31, 2022, doi: 10.1109/TCDS.2022.3147839.

[39] 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.

[40] M. Arvaneh et al., "Optimizing the channel selection and classification accuracy in EEG-based BCI," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 6, pp. 1865–1873, Jun. 2011.

[41] R. W. Homan, J. Herman, and P. Purdy, "Cerebral location of international 10–20 system electrode placement," *Electroencephalogr. Clin. Neuriophysiol.*, vol. 66, no. 4, pp. 376–382, 1987.

[42] T. J. Buschman and E. K. Miller, "Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices," *Science*, vol. 315, no. 5820, pp. 1860–1862, 2007.

[43] S. Tune et al., "Neural attentional-filter mechanisms of listening success in middle-aged and older individuals," *Nature Commun.*, vol. 12, no. 1, pp. 1–14, 2021.

[44] 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.

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

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

[47] S. Woo et al., "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[48] A.-L. Giraud and D. Poeppel, "Cortical oscillations and speech processing: Emerging computational principles and operations," *Nature Neurosci.*, vol. 15, no. 4, pp. 511–517, 2012.

[49] F. Scarselli et al., "The graph neural network model," *IEEE Trans. Neural Netw.*, vol. 20, no. 1, pp. 61–80, Jan. 2009.

[50] D. I. Shuman et al., "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 83–98, May 2013.

[51] 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

[52] S. A. Fuglsang, D. D. Wong, and J. Hjortkjær, "EEG and audio dataset for auditory attention decoding," Mar. 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1199011

[53] S. A. Fuglsang, T. Dau, and J. Hjortkjær, "Noise-robust cortical tracking of attended speech in real-world acoustic scenes," *NeuroImage*, vol. 156, pp. 435–444, 2017.

[54] A. Delorme and S. Makeig, "EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *J. Neurosci. Methods*, vol. 134, no. 1, pp. 9–21, 2004.

[55] L. Wang et al., "Cold pressor pain assessment based on EEG power spectrum," *SN Appl. Sci.*, vol. 2, 2020, Art. no. 1976.

[56] S. Miran et al., "Real-time tracking of selective auditory attention from M/EEG: A bayesian filtering approach," *Front. Neurosci.*, vol. 12, 2018, Art. no. 262.

[57] D. D. Wong et al., "A comparison of regularization methods in forward and backward models for auditory attention decoding," *Front. Neurosci.*, vol. 12, 2018, Art. no. 531.

[58] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.

[59] N. Srivastava et al., "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[60] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[61] I. Kuruvila et al., "Extracting the auditory attention in a dual-speaker scenario from EEG using a joint CNN-LSTM model," *Front. Physiol.*, vol. 12, 2021, Art. no. 700655.

[62] J. M. Rimmele et al., "The effects of selective attention and speech acoustics on neural speech-tracking in a multi-talker scene," *Cortex*, vol. 68, pp. 144–154, 2015.

[63] Y. Gao et al., "Selective attention enhances beta-band cortical oscillation to speech under "cocktail-party" listening conditions," *Front. Hum. Neurosci.*, vol. 11, 2017, Art. no. 34.

[64] S. K. Scott and I. S. Johnsrude, "The neuroanatomical and functional organization of speech perception," *Trends Neurosciences*, vol. 26, no. 2, pp. 100–107, 2003.

[65] I. Choi et al., "Individual differences in attentional modulation of cortical responses correlate with selective attention performance," *Hear. Res.*, vol. 314, pp. 10–19, 2014.

[66] M. Bauer et al., "Attentional modulation of alpha/beta and gamma oscillations reflect functionally distinct processes," *J. Neurosci.*, vol. 34, no. 48, pp. 16117–16125, 2014.

[67] N. Ding et al., "Cortical tracking of hierarchical linguistic structures in connected speech," *Nature Neurosci.*, vol. 19, no. 1, pp. 158–164, 2016.

[68] B. Tóth et al., "Attention and speech-processing related functional brain networks activated in a multi-speaker environment," *PLoS One*, vol. 14, no. 2, 2019, Art. no. e0212754.