INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

EEG-based Auditory Attention Detection with Spatiotemporal Graph and Graph Convolutional Network

Ruicong Wang[1], *Siqi Cai*[2,*], *and Haizhou Li*[3,2,4]

[1]School of Computing, National University of Singapore, Singapore
[2]Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[3] Shenzhen Research Institute of Big Data, School of Data Science, The Chinese University of Hong Kong, Shenzhen, China
[4] Machine Listening Lab, University of Bremen, Germany

wangruicong@u.nus.edu, elesiqi@nus.edu.sg (*corresponding author), haizhouli@cuhk.edu.cn

Abstract

The ability to detect auditory attention from electroencephalography (EEG) offers many possibilities for brain-computer interface (BCI) applications, such as hearing assistive devices. However, effective feature representation for EEG signals remains a challenge due to the complex spatial and temporal dynamics of EEG signals. To overcome this challenge, we introduce a Spatiotemporal Graph Convolutional Network (ST-GCN), which combines a temporal attention mechanism and a graph convolutional module. The temporal attention mechanism captures the temporal dynamics of EEG segments, while the graph convolutional module learns the spatial pattern of multi-channel EEG signals. We evaluate the performance of our proposed ST-GCN on two publicly available datasets and demonstrate significant improvements over existing state-of-the-art models. These findings suggest that the ST-GCN model has the potential to advance auditory attention detection in real-life BCI applications.
Index Terms: Auditory attention, cocktail party problem, temporal attention, graph convolutional network

1. Introduction

The "cocktail party effect" refers to humans' ability to selectively focus on a particular speaker in the presence of multiple speakers [1]. Recent neuroscience research has shown that auditory attention is a neural activity that can be detected from brain signals, which is referred to as auditory attention detection (AAD) [2]. O'Sullivan et al. [3] first validated the idea of EEG-enabled AAD. Due to its noninvasiveness and ease of use, various methods have been developed to detect auditory attention from EEG, an overview of which can be found in [4]. However, effective feature learning from raw EEG signals remains a challenge in AAD tasks, especially under low-latency conditions.

Previous studies have revealed that the selective auditory task involves spatially separated brain areas [5]. The spatial patterns of neural responses to speech stimuli are crucial for detecting auditory attention. Inspired by this, the common spatial pattern (CSP) method has been used for spatial feature extraction in EEG-enabled AAD [6]. Apart from the traditional method, Vandecappelle et al. [7] applied a convolutional neural network (CNN) to extract spatial features related to auditory attention from EEG signals, which yielded good results for low-latency AAD (around 81% accuracy within 1-2 s). Although CNNs are typically used to extract local features from continuous signals, EEG signals exist in the discrete and discontinuous spatial domain. In this case, graph-based representation methods would provide a more effective approach. One such method is the graph convolutional network (GCN), which extends traditional CNN methods by incorporating spectral theory [8]. GCNs have been successful in various tasks, including human pose recognition, traffic prediction, and disease prediction, where the topological relationship between input features is crucial. Building on this success, we propose a graph-based description method for multi-channel EEG signals that captures the spatial relationships throughout the whole brain.

Moreover, EEG signals are dynamic time-series data, containing rich temporal information [9]. Recent research has revealed that spatiotemporal patterns in the human brain reflect attentional regulation during selective listening [5]. Therefore, we hypothesize that the accuracy of AAD can be improved by capturing the complex temporal dynamics of EEG signals. To explore this hypothesis, we introduced a temporal attention mechanism that assigns varying weights to a sequence of EEG signals, enabling feature extraction in the time domain. In recent years, attention has been incorporated into various neural network architectures, including recurrent neural networks (RNNs) and CNNs, to simulate selective attention in neurophysiological studies [10]. By incorporating the temporal attention module, we aim to capture the complex temporal dynamics of EEG signals and enhance the detection of even subtle changes in attentional states over time.

This study proposes a spatiotemporal graph convolutional network (ST-GCN) to detect auditory attention from EEG signals. By combining spatial and temporal attention mechanisms, ST-GCN can mimic the human ability to selectively focus on specific sounds. The end-to-end framework of ST-GCN extracts more discriminative EEG features, resulting in improved AAD performance.

The structure of this paper is outlined as follows. The second section outlines the structure of the proposed ST-GCN. The third section provides detailed information about the datasets, including preprocessing, model training, and evaluation. In the fourth section, we present the results of our experiments and analyze our findings. Finally, we conclude this paper in the fifth section.

2. Methods

The proposed ST-GCN model presents an innovative end-to-end framework that analyzes raw EEG signals and effectively

10.21437/Interspeech.2023-620

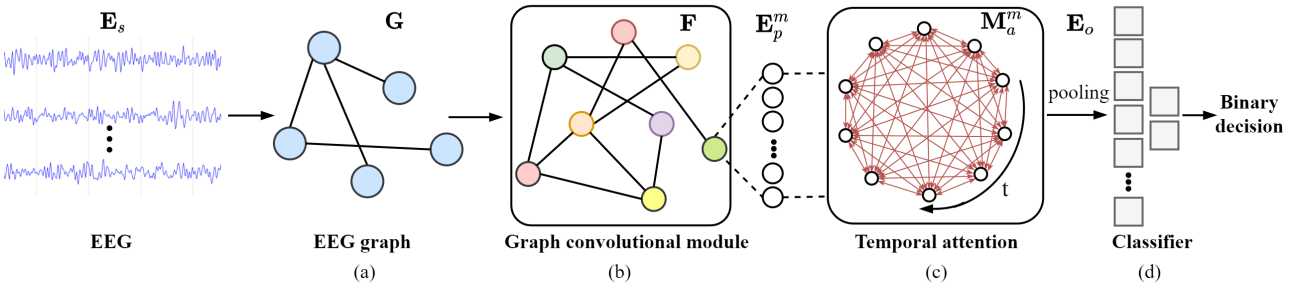

Figure 1: *The architecture of the spatiotemporal graph convolutional network, i.e., ST-GCN, which is an end-to-end AAD solution. It takes EEG signals as input and performs binary decision-making to detect auditory attention.*

detects auditory attention, as depicted in Fig. 1. First, multi-channel EEG is encoded into an EEG graph to preserve the topological information (Fig. 1 (a)). Then, a graph convolutional module is applied to learn the inherent relationship between different EEG channels (Fig. 1 (b)). To further enhance the model's performance, a temporal attention module is introduced to explore the attentive temporal dynamics across diverse EEG graphs, thereby creating a comprehensive spatiotemporal representation of the EEG signals (Fig. 1 (c)). Finally, a backend classifier is designed to detect auditory attention by leveraging the derived EEG features (Fig. 1 (d)).

2.1. EEG Graph

Assuming the raw EEG signal comprises N channels, each channel is considered a node in the graph representation. Therefore, the EEG signal $\mathbf{E}s$ can be transformed into an undirected graph $G = (V, E)$ in a non-Euclidean space. Specifically, V represents the set of nodes, where $|V| = N$, and $(V_i, V_j) \in E$ denotes the set of edges connecting these channels. To capture the intrinsic relationships between the EEG channels, an adjacency matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$ is introduced. The elements of this matrix are pre-determined based on the spatial relationship of the EEG channels, as shown in Fig. 2. The entry of the adjacency matrix $a_{i,j}$ measures the level of connection between the channels i and j.

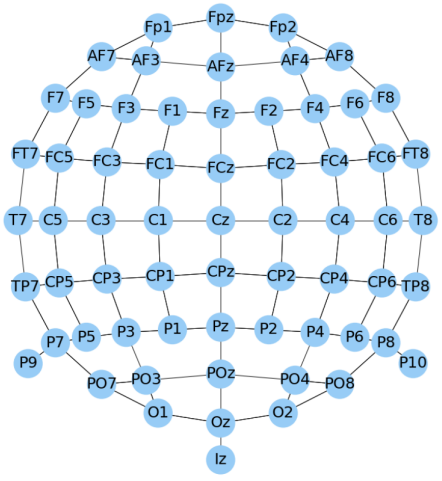

Figure 2: *The spatial relationship of 64-channel EEG.*

2.2. Graph Convolution

The GCN theory extends the traditional convolutional operation, commonly used on grid data in Euclidean space, to graph data in non-Euclidean space [8]. Specifically, the graph convolution aggregates features of a vertex and its neighboring vertices to generate a comprehensive representation of the vertex. As defined in [11], the convolution operation is realized by computing the eigendecomposition of the graph Laplacian in the Fourier domain. The Laplacian matrix of a graph can be expressed as $\mathbf{L} = \mathbf{D} - \mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$, where $\mathbf{D}$ represents the degree matrix, $\mathbf{U}$ is the matrix of eigenvectors, and $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues. The graph convolutional operation can be formulated as the multiplication of a signal $\mathbf{x}$ with a filter $g_\theta = \mathrm{diag}(\theta)$, which is parameterized by $\theta \in \mathbb{R}^N$:

$$\mathbf{x}^{'} = \mathbf{U} g_\theta \mathbf{U}^T \mathbf{x} \tag{1}$$

Here we apply a graph convolutional module for representation learning of the EEG graph G, as shown in Fig. 1 (b), and obtain the feature representation $\mathbf{F} = [\mathbf{E}_p^1, \ldots, \mathbf{E}_p^m, \ldots, \mathbf{E}_p^C] \in \mathbb{R}^{C\times N\times T}$, where C denotes the number of graph convolutional kernels.

2.3. Temporal Attention

Attention mechanisms have been extensively studied in deep neural networks [10], and are commonly used in various machine learning tasks to selectively focus on relevant information. At run-time inference, attention modulation assigns differentiated weights to samples at discrete instants of time. To enhance the representation ability of the model, a temporal attention module is employed, as shown in Fig. 1(c), to concentrate on important components and suppress unnecessary ones over time. In this study, attention modulation is executed in the following manner.

First, $\mathbf{E}_p^m$ is transformed into query $\mathbf{E}_q^m$, key $\mathbf{E}_k^m$, and value $\mathbf{E}_v^m$ using linear projection.

$$\begin{aligned}
\mathbf{E}_q^m &= \beta((\mathbf{E}_p^m)^T \mathbf{W}_q) \in \mathbb{R}^{T\times d_k} \\
\mathbf{E}_k^m &= \beta((\mathbf{E}_p^m)^T \mathbf{W}_k) \in \mathbb{R}^{T\times d_k} \\
\mathbf{E}_v^m &= \beta((\mathbf{E}_p^m)^T \mathbf{W}_v) \mathbf{W}_v^{'} \in \mathbb{R}^{T\times N}
\end{aligned} \tag{2}$$

where $\mathbf{W}_q \in \mathbb{R}^{N\times d_k}$, $\mathbf{W}_k \in \mathbb{R}^{N\times d_k}$, $\mathbf{W}_v \in \mathbb{R}^{N\times d_k}$, $\mathbf{W}_v^{'} \in \mathbb{R}^{d_k\times N}$. $\beta(\cdot)$ denotes the $tanh$ activation function.

Second, the relationship between the query $\mathbf{E}_q^m$ and key $\mathbf{E}_k^m$ can be computed using a dot product,

$$\mathbf{M}_a^m = \mathbf{E}_q^m (\mathbf{E}_k^m)^T \in \mathbb{R}^{T \times T} \quad (3)$$

where $\mathbf{M}_a^m$ is the temporal attention weights.

Finally, attention weights are dynamically assigned to EEG over time, and the attention-weighted summation $\mathbf{E}_a^m$ can be obtained:

$$\mathbf{E}_a^m = (\mathbf{E}_v^m)^T \mathbf{M}_a^m \in \mathbb{R}^{N \times T} \quad (4)$$

2.4. AAD Classifier

The back-end classifier first incorporates the optimized feature $\mathbf{E}_a^m$ into $\mathbf{E}_p^m$ by

$$\mathbf{E}_o^m = \mathbf{E}_p^m \bigotimes \mathbf{E}_a^m \in \mathbb{R}^{N \times T} \quad (5)$$

where $\bigotimes$ denotes a point-wise multiplication.

The EEG feature $\mathbf{E}_o$, which is represented as $[\mathbf{E}_o^1, \ldots, \mathbf{E}_o^m, \ldots, \mathbf{E}_o^C] \in \mathbb{R}^{C \times N \times T}$, is then passed through a temporal pooling layer and fc layers with $sigmoid$ activation function to produce a probability vector $\mathbf{E}_o^{'} \in \mathbb{R}^2$.

Finally, the binary cross-entropy loss is used for the AAD task:

$$\mathcal{L} = -\frac{1}{L} \sum_{l=1}^{L} y_l \cdot \log p_l + (1 - y_l) \cdot \log(1 - p_l) \quad (6)$$

where y_l denotes the label of l-th decision window.

3. Experiments

3.1. AAD Datasets

To facilitate a comparative study, we perform the experiments on two independent and public datasets, which are referred to as DTU [12, 13] and KUL [7, 14].

1) DTU dataset: The study involved 18 individuals with normal hearing who were presented with two different auditory streams, each consisting of one male and one female speaker talking simultaneously. The two streams were positioned 60 degrees to the left and right, and each participant listened to a total of 60 trials, with each trial lasting approximately 50 seconds. EEG data were recorded using 64 channels, resulting in a total of 15.0 hours of data.

2) KUL dataset: A total of 16 participants with self-reported normal hearing were included in the study. EEG data were recorded using 64 channels to capture brain activity. The speech stimuli consisted of a collection of Dutch stories narrated by various male speakers. During each trial, participants were presented with two auditory streams positioned at 90 degrees to the left and right of their listening position. They were instructed to selectively focus on one of the two competing streams. Each participant contributed approximately 48 minutes of EEG data, resulting in a cumulative EEG dataset of 12.8 hours across all subjects.

3.2. Data Preprocessing

The EEG signals were preprocessed following the methods used in previous AAD studies [7, 15], which involved band-pass filtering between 1 and 32 Hz and downsampling to 128 Hz. To segment the data into shorter durations, referred to as *decision windows*, we applied an overlapping time window. Given that humans can switch their auditory attention between speakers in approximately 1 second [6], low-latency AAD solutions are needed for real-world applications. Thus, this study focused on decision windows of 0.1 s, 0.2 s, 0.5 s, and 1 s to achieve low-latency AAD performance.

3.3. Model Implementation and Evaluation

The AAD models were trained and evaluated using a 5-fold cross-validation approach with nested cross-validation loops [16]. The accuracy of AAD is measured as the proportion of correctly identified windows compared to the total number of decision windows in the test set [7, 15]. The average accuracy over all the testing folds is reported as the final result. All hyperparameters are grid-searched on the validation set to choose reasonable values. During training, the network is updated using Adam optimizer with a learning rate of 10^{-3}. To prevent overfitting and enhance generalization, dropout and batch normalization techniques are implemented. The PyTorch framework was utilized to implement all models, and the training process was performed on 2 NVIDIA TITAN Xp Pascal GPUs.

Here we present the proposed model with a case study of a 1-second decision window. The EEG data $\mathbf{E}s \in \mathbb{R}^{128 \times 64}$, which represents 128 samples by 64 channels, is first encoded into an EEG graph G. Then, the graph convolutional module uses a trainable weight matrix $g_\theta \in \mathbb{R}^{5 \times 64 \times 64}$ to learn the spatial information from the EEG graph G. In the temporal attention module, the scale factor d_k is set to 8, which is used to derive an output $\mathbf{E}_a^m \in \mathbb{R}^{128 \times 64}$ from the EEG graph. The original features are then combined with the output from the temporal attention module to obtain the final feature representation $\mathbf{E}_o \in \mathbb{R}^{5 \times 128 \times 64}$. Finally, a global average pooling layer is applied along the temporal dimension, and the resulting data is flattened into a one-dimensional vector for use as inputs to *fc* layers (input: 8, output: 2) to detect auditory attention.

4. Results and Discussion

4.1. Analysis of Graph Learning

The effectiveness of graph learning for EEG signals is explored by comparing the performance of GCN and CNN models. To ensure a fair comparison, the CNN-based AAD model from [7] is re-implemented under the same experimental settings. It is worth noting that the GCN model has fewer parameters than the CNN model, with only 2,930 parameters compared to 5,500 parameters.

As shown in Fig. 3, the CNN model achieves an average accuracy of 63.3% (SD: 5.9) on the DTU dataset with a 1-second decision window. In comparison, the GCN model significantly outperforms the CNN model with an average accuracy of 73.1% (SD: 7.61), resulting in a large margin of 9.8%. Similarly, the GCN model (mean: 89.4%, SD: 6.35) outperforms the CNN model (mean: 84.1%, SD: 10.16) by an average improvement of 5.3% on the KUL dataset. These results indicate that GCNs can learn the spatial features of EEG signals more effectively than CNN models, resulting in better AAD performance. One possible explanation is that CNNs are restricted in their ability to capture the complex neighborhood information of EEG due to their focus on local regions with fixed connections, while GCNs can preserve the rich topology information of the brain, leading to a more effective representation of EEG signals.

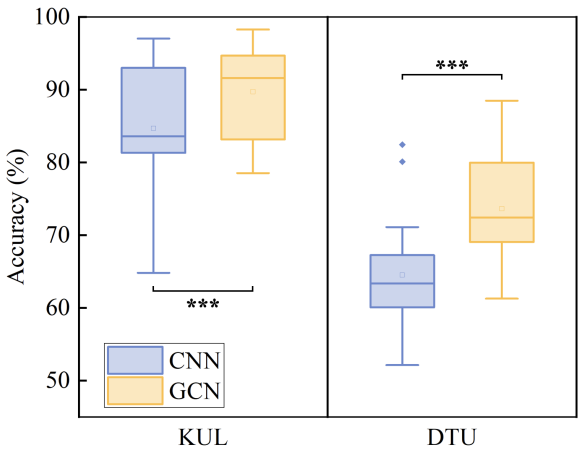

Figure 3: *AAD accuracy (%) for CNN and GCN models on the KUL and DTU datasets using a decision window of 1 second.*

4.2. Analysis of Temporal Attention

To assess the effectiveness of the temporal attention module, an ablation analysis was conducted on both the KUL and DTU datasets, using a 1-second decision window. Specifically, we compared the AAD performance of two models: the GCN model and the ST-GCN model that incorporates the temporal attention module.

Our results demonstrate that the ST-GCN model consistently outperforms the GCN model on both datasets. On the KUL dataset with a 1-second decision window, the ST-GCN model achieves a mean accuracy of 92.5% (SD: 5.56), which is 3.1% higher than the GCN model. Similarly, on the DTU dataset, the ST-GCN model achieves a mean accuracy of 77.3% (SD: 6.80), which is 4.2% higher than the GCN model. Additionally, the statistical analysis demonstrates that the temporal attention module contributes significantly to the improvement in AAD performance (paired t-test: $p < 0.001$).

The effectiveness of the temporal attention module can be attributed to the fact that EEG signals are inherently dynamic and complex, with different temporal patterns carrying different amounts of information. By allowing the ST-GCN model to focus on the most informative segments, the temporal attention module is able to extract more discriminative features from the EEG signals, leading to improved AAD performance.

4.3. Comparative study

Table 1 summarizes the comparison between the proposed ST-GCN and the state-of-the-art STAnet [15]. Both models exhibit similar sizes, each consisting of around 5000 parameters.

The results demonstrate the superior performance of the ST-GCN model compared to STAnet on both the KUL and DTU datasets across all four different decision windows. For the DTU dataset, the ST-GCN model achieves an average AAD accuracy of 77.3% for the 1-second decision window, which is a notable improvement of 5.4% compared to STAnet. Similarly, for the KUL dataset, the ST-GCN model achieves an average AAD accuracy of 92.5% for the 1-second decision window, surpassing STAnet with a gain of 2.4%. Furthermore, ST-GCN outperforms STAnet for all other decision windows as well, with an average accuracy improvement of 4.1% and 4.2% on the KUL and DTU datasets, respectively.

Table 1: *The AAD accuracy (%) of different models on KUL [14] and DTU [12] with different decision window sizes was compared. Here ST-GCN model shows significantly higher AAD accuracy than STAnet on both datasets ($p < 0.001$).*

Dataset	Model	Decision window (second)			
		0.1	0.2	0.5	1
DTU [12]	STAnet [15]	65.7	68.1	70.8	71.9
	ST-GCN (Ours)	**68.5**	**72.1**	**75.4**	**77.3**
KUL [14]	STAnet [15]	80.8	84.3	87.2	90.1
	ST-GCN (Ours)	**86.2**	**89.1**	**91.0**	**92.5**

In addition, we have also observed that the proposed ST-GCN model achieves impressive AAD accuracy on the KUL dataset, reaching 86.2% and 89.1% for decision windows of 0.1 seconds and 0.2 seconds, respectively. This suggests that the ST-GCN model can enable near real-time detection of auditory attention, which is crucial for practical applications, such as neuro-steered hearing aids. These results demonstrate the potential of the ST-GCN model to be adapted for real-world scenarios where fast responses are required.

In sum, our proposed ST-GCN has shown efficacy in capturing discriminative EEG features, which leads to enhanced performance in AAD. Our findings support that spatiotemporal information is valuable and plays a significant role in AAD.

5. Conclusions

In this study, we present a new AAD approach named ST-GCN that can effectively detect auditory attention from EEG signals. Our proposed ST-GCN can preserve the spatiotemporal dynamics of EEG signals and extract discriminative features in an end-to-end manner. We evaluated the performance of ST-GCN on two publicly available datasets with comparing it with state-of-the-art methods. Our findings confirm that spatiotemporal information is informative and contributes significantly to the AAD task. In future work, we will expand the proposed AAD system to handle more complex scenarios where there are more than two target speakers, such as cocktail party environments.

6. Acknowledgements

The research is supported by National Natural Science Foundation of China (Grant No. 62271432); Guangdong Provincial Key Laboratory of Big Data Computing, The Chinese University of Hong Kong, Shenzhen, China (Grant No. B10120210117-KP02); A*STAR Singapore under its RIE 2020 Advanced Manufacturing and Engineering Human (AME) Programmatic Grant (Grant No. A1687b0033); the DFG Excellence Strategy (University Allowance, EXC 2077), University of Bremen, Germany.

7. References

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the Acoustical Society of America*, 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, p. 233, 2012.

[3] J. A. O'Sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, and E. C. Lalor, "Attentional selection in a cocktail party environment can be decoded from single-trial EEG," *Cerebral Cortex*, vol. 25, no. 7, pp. 1697–1706, 2015.

[4] S.Geirnaert, S. Vandecappelle, E. Alickovic, A. de Cheveigne, E. Lalor, B. T. Meyer, S. Miran, T. Francart, and A. Bertrand, "Electroencephalography-based auditory attention decoding: Toward neurosteered hearing devices," *IEEE Signal Processing Magazine*, vol. 38, no. 4, pp. 89–102, 2021.

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

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

[7] 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, p. e56481, 2021.

[8] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip, "A comprehensive survey on graph neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 4–24, 2020.

[9] M. Wöstmann, B. Herrmann, B. Maess, and J. Obleser, "Spatiotemporal dynamics of auditory attention synchronize with speech," *Proceedings of the National Academy of Sciences*, vol. 113, no. 14, pp. 3873–3878, 2016.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[11] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[12] 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

[13] 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.

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

[15] E. Su, S. Cai, L. Xie, H. Li, and T. Schultz, "STAnet: A spatiotemporal attention network for decoding auditory spatial attention from EEG," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 7, pp. 2233–2242, 2022.

[16] D. D. Wong, S. A. Fuglsang, J. Hjortkjær, E. Ceolini, M. Slaney, and A. De Cheveigne, "A comparison of regularization methods in forward and backward models for auditory attention decoding," *Frontiers in neuroscience*, vol. 12, p. 531, 2018.