Interspeech 2024
1-5 September 2024, Kos, Greece

Auditory Spatial Attention Detection Based on Feature Disentanglement and Brain Connectivity-Informed Graph Neural Networks

Yixiang Niu[1], *Ning Chen*[1], *Hongqing Zhu*[1], *Zhiying Zhu*[1], *Guangqiang Li*[1], *Yibo Chen*[1]

[1]East China University of Science and Technology, China
y10200092@mail.ecust.edu.cn, chenning_750210@163.com, hqzhu@ecust.edu.cn,
zyzhu@ecust.edu.cn, y10220098@mail.ecust.edu.cn, 20002027@mail.ecust.edu.cn

Abstract

Auditory spatial attention detection (ASAD) aims to determine which speaker in a surround sound field a listener is focusing on from a single-trial electroencephalogram (EEG). Latest studies have represented non-Euclidean structured EEGs by graph-based modeling, but how to better incorporate brain connectivity into EEG graphs remains a great challenge. Moreover, due to inter-subject distribution shifts in EEGs, most existing models perform well only on specific subjects. To address these issues, we propose a new ASAD model. EEG graphs are constructed based on brain effective connectivity, and then mapped into embedding spaces by a graph neural network architecture. Meanwhile, feature disentanglement combined with correlation alignment is utilized to learn subject-invariant EEG patterns relevant to ASAD tasks. Experiments on open datasets demonstrate that in cross-subject scenarios, the proposed model outperforms state-of-the-art models, and the algorithmic complexity is relatively low.

Index Terms: auditory spatial attention detection, electroencephalogram, brain effective connectivity, feature disentanglement, graph neural network

1. Introduction

People with normal hearing can easily focus on one sound source of interest and ignore the others in noisy environments, which is known as the "cocktail party effect" [1]. However, it may be challenging for people suffering from hearing impairments. They hope to distinguish the sound source they actually want to hear from background noises with the help of hearing aids or cochlear implants. Neural entrainment in the listening brain makes it possible. Recent studies have achieved auditory attention detection (AAD) from a single-trial electroencephalogram (EEG) [2], leading to a novel brain-computer interface called "neuro-steered hearing device" [3].

The traditional AAD seeks to establish a possible linear or nonlinear relationship between a stimulus and the evoked brain response [4, 5]. It always requires a stimulus representation of each candidate sound source as a reference, e.g., an envelope, however, which is often inaccessible in the real world. Fortunately, hemispheric lateralization indicates that the auditory attention direction is spatio-temporally encoded in brain activities, leading to an alternative solution called auditory spatial attention detection (ASAD) [6, 7, 8]. It aims to detect the locus of auditory attention with no need for stimuli, but requires a difference in the spatial position of each sound source relative to the listener. Fortunately, the real world is exactly a three-dimensional sound field. In addition, ASAD is more suitable for shorter decision windows, because it utilizes instantaneous spatial features in the human brain, rather than temporal features based on stimulus-response correlations [7].

Although existing ASAD studies have made great progress, there are still some unconsidered aspects. First, typical ASAD studies treat EEGs as general matrices (i.e., independent EEG channels) or map them to two-dimensional images to approximate spatial relationships among EEG channels [9, 10]. However, both methods conflict with non-Euclidean properties of EEGs and ignore human functional brain networks. Although some advanced ASAD studies have modeled an EEG as a graph [11, 12, 13], prior metrics of human brain connectivity, especially effective connectivity, have not been fully taken into account. Second, but more importantly, real-world applications often require general ASAD models that can work robustly in cross-subject scenarios. However, the functional organization of the human brain differs between individuals, resulting in inter-subject distribution shifts in EEGs. Existing models may thus overfit source subjects' EEGs during training, which prevents their generalizability to target subjects in testing. To this end, domain adaptation [14] has been introduced into ASAD, but it cannot be applied to new subjects who are unseen in training, nor can it make predictions for short-time EEG segments. In addition, data generation [15] can reduce overfitting in ASAD by improving the quantity and diversity of training data, especially for small datasets, but obviously, it does not positively deal with inter-subject variability.

In this paper, we propose a new ASAD model to solve the above problems. First, we model a multi-channel EEG as a graph based on brain effective connectivity that is quantified by symbolic phase transfer entropy (SPTE) [16]. Then, we map EEG graphs to embedding spaces using a graph neural network (GNN) architecture based on convolutional auto-regressive moving average (ARMA) filters [17]. It is a kind of spectral graph convolutions that provides a more flexible frequency response and better captures the global graph structure. It can avoid the risk of over-smoothing node features by using skip connections. Second, we assume that distribution shifts in EEGs mostly result from covariate shifts, i.e., only marginal distributions change. Thus, we adopt the domain generalization (DG) [18, 19] technique to extract subject-invariant EEG patterns relevant to ASAD tasks and thus achieve subject-independent ASAD. Specifically, we introduce feature disentanglement (FD) combined with correlation alignment (CORAL) [20] to align the marginal distributions of different subjects' EEGs in the embedding space. We specifically adopt class-level alignment that aligns EEG segments belonging to each class of auditory spatial attention separately, to prevent different classes from being close to each other. A series of experiments on publicly available datasets demonstrate that the proposed model outperforms state-of-the-art ASAD models, and both FD and CORAL play positive roles. In addition, the algorithmic complexity of the proposed model is relatively low.

10.21437/Interspeech.2024-1940

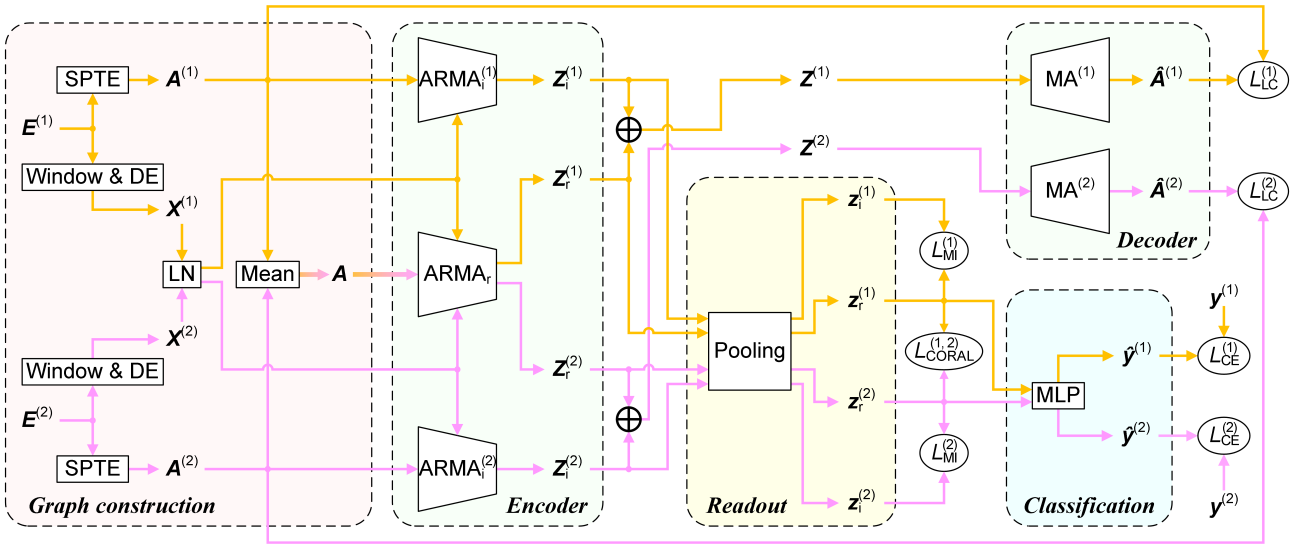

Figure 1: *Block diagram of the proposed model. The orange and pink lines represent source subjects 1 and 2, respectively.*

2. Methods

As illustrated in Figure 1, during training, the proposed model consists of five modules: *graph construction*, *encoder*, *decoder*, *readout*, and *classification*, which are all detailed below. Although there are only $N = 2$ source subjects in Figure 1, the proposed model can be easily extended to more source subjects. The changes of the proposed model during testing versus training are explained at the end of each module specification.

2.1. Modules

Graph construction. For the n-th source subject, we divide the C-channel EEG signal $\boldsymbol{E}^{(n)}$ into several segments with a sliding decision window of a pre-set length. Obviously, EEG segments belonging to the same trial share the same label of the attention locus $\boldsymbol{y}^{(n)}$. Each EEG segment is then modeled as a graph, where EEG channels are regarded as nodes, and differential entropy (DE) [21] serves as node features. On each channel, the EEG signal is decomposed into B sub-bands with a step of 2 Hz via the maximal overlap discrete wavelet packet transform, and DE is subsequently calculated within each sub-band. Thus, the node feature matrix of an EEG graph can be constructed as

$$\boldsymbol{X}^{(n)} = \begin{bmatrix} \mathrm{DE}_{1,1}^{(n)} & \cdots & \mathrm{DE}_{1,B}^{(n)} \\ \vdots & \ddots & \vdots \\ \mathrm{DE}_{C,1}^{(n)} & \cdots & \mathrm{DE}_{C,B}^{(n)} \end{bmatrix} \in \mathbb{R}^{C\times B}. \tag{1}$$

Note that subject-shared layer normalization (LN) will be adopted later to normalize node feature matrices.

The inter-channel connections in the n-th source subject's adjacency matrix $\boldsymbol{A}^{(n)} \in \mathbb{R}^{C\times C}$ are derived from brain effective connectivity that is quantified by SPTE [16]. Thus, each element in $\boldsymbol{A}^{(n)}$ except for those on the main diagonal is initialized as the SPTE from the i-th to the j-th channel:

$$A_{ij}^{(n)} = \begin{cases} 0, & i = j \\ \mathrm{SPTE}_{ij}^{(n)}, & i \neq j \end{cases} \tag{2}$$

where $i, j = 1, \ldots, C$. It is worth mentioning that SPTE is calculated within each whole trial rather than each EEG segment, and then averaged across trials to obtain $\boldsymbol{A}^{(n)}$ shared by all EEG segments belonging to the n-th source subject. In addition, a public adjacency matrix $\boldsymbol{A}$ is generated by averaging each source subject's $\boldsymbol{A}^{(n)}$. To maximize the efficiency of the network topology, we retain the top 20% values in $\boldsymbol{A}$ and each $\boldsymbol{A}^{(n)}$, and set the rest to zeros [22]. During testing, only the node feature matrices need to be constructed for the target subject.

Encoder. The EEG graphs of the n-th source subject are mapped to embedding spaces to achieve FD. It relies on GNN implementation with convolutional ARMA filters [17]:

$$\begin{cases} \boldsymbol{Z}_{\mathrm{r}}^{(n)} = \mathrm{ARMA}_{\mathrm{r}}\left(\mathrm{LN}\left(\boldsymbol{X}^{(n)}\right), \tilde{\boldsymbol{A}}\right) \\ \boldsymbol{Z}_{\mathrm{i}}^{(n)} = \mathrm{ARMA}_{\mathrm{i}}^{(n)}\left(\mathrm{LN}\left(\boldsymbol{X}^{(n)}\right), \tilde{\boldsymbol{A}}^{(n)}\right) \end{cases} \tag{3}$$

where $\tilde{\boldsymbol{A}}$ and $\tilde{\boldsymbol{A}}^{(n)}$ are normalized $\boldsymbol{A}$ and $\boldsymbol{A}^{(n)}$, respectively. $\mathrm{ARMA}_{\mathrm{r}}$ is the subject-shared ARMA graph convolutional layer that extracts the task-relevant feature $\boldsymbol{Z}_{\mathrm{r}}^{(n)}$; $\mathrm{ARMA}_{\mathrm{i}}^{(n)}$ is the n-th source subject's private one that extracts the task-irrelevant feature $\boldsymbol{Z}_{\mathrm{i}}^{(n)}$. During testing, we only need to utilize the target subject's node feature matrices and $\tilde{\boldsymbol{A}}$ that has been computed in training to extract only the task-relevant feature.

Decoder. We add the task-relevant and task-irrelevant features of the n-th source subject as $\boldsymbol{Z}^{(n)} = \boldsymbol{Z}_{\mathrm{r}}^{(n)} + \boldsymbol{Z}_{\mathrm{i}}^{(n)}$. Since there are only 20% non-zero elements in $\boldsymbol{A}^{(n)}$, we utilize a subject-private masked attention (MA) mechanism [23] to reconstruct the n-th source subject's adjacency matrix $\hat{\boldsymbol{A}}^{(n)}$ with elements $\hat{A}_{ij}^{(n)}$:

$$\begin{cases} \hat{A}_{ij}^{(n)} = \mathrm{Mask}^{(n)}\left(\mathrm{ReLU}\left(\tanh\left(\xi e_{ij}^{(n)}\right)\right)\right) \\ e_{ij}^{(n)} = \boldsymbol{a}^{(n)}\left(\boldsymbol{W}^{(n)}\boldsymbol{Z}_{j,:}^{(n)} \,\|\, \boldsymbol{W}^{(n)}\boldsymbol{Z}_{i,:}^{(n)}\right) \end{cases} \tag{4}$$

where $i, j = 1, \ldots, C$. $\xi = 0.02$ is a small constant, and $\|$ denotes concatenation. $\boldsymbol{Z}_{i,:}^{(n)}$ and $\boldsymbol{Z}_{j,:}^{(n)}$ are the transposes of the i-th and j-th rows of $\boldsymbol{Z}^{(n)}$, respectively. $\boldsymbol{a}^{(n)}$ is a row vector that parametrizes the attention mechanism, and $\boldsymbol{W}^{(n)}$ is a weight matrix. The attention coefficient $e_{ij}^{(n)}$ indicates the importance of the i-th node's embedding to the j-th node. According to the

positions of zeros in $\boldsymbol{A}^{(n)}$, Mask$^{(n)}$ sets the matching elements in $\hat{\boldsymbol{A}}^{(n)}$ to zeros. During testing, the decoder module is directly ignored due to the lack of the task-irrelevant feature.

Readout. For the n-th source subject, we aggregate the updated node feature matrices $\boldsymbol{Z}_{\mathrm{r}}^{(n)}$ and $\boldsymbol{Z}_{\mathrm{i}}^{(n)}$ into graph-level representations $\boldsymbol{z}_{\mathrm{r}}^{(n)}$ and $\boldsymbol{z}_{\mathrm{i}}^{(n)}$ through a global average pooling layer that pools a graph by computing the arithmetic mean along the node axis. During testing, the same operation is performed to pool the task-relevant feature of the target subject.

Classification. The classifier is a multi-layer perceptron (MLP), which is composed of two fully connected layers and one batch normalization layer between them. It outputs a one-hot encoded label $\hat{\boldsymbol{y}}^{(n)}$ from $\boldsymbol{z}_{\mathrm{r}}^{(n)}$ to indicate the spatial locus of the n-th source subject's auditory attention within a decision window. During testing, the well-trained classifier is directly applied to the target subject, to make predictions within a decision window of the same length as adopted during training.

2.2. Loss function

As exhibited in Figure 1, the proposed model contains four loss functions during training. First, the cross-entropy (CE) loss can quantify the difference between the prediction and the ground truth; second, the mutual information (MI) loss helps to separate the task-relevant feature from the task-irrelevant one; third, to avoid generating a task-irrelevant feature that tends to be random and meaningless, we specially employ the log-cosh (LC) loss to constrain the similarity between the reconstructed adjacency matrix and the original one. The above three loss functions are all averaged across source subjects:

$$\begin{cases} \mathcal{L}_{\mathrm{CE}} = \overline{\mathcal{L}}_{\mathrm{CE}}^{(n)} = \dfrac{1}{N}\displaystyle\sum_{n=1}^{N} \mathrm{CE}\left(\boldsymbol{y}^{(n)}, \hat{\boldsymbol{y}}^{(n)}\right) \\ \mathcal{L}_{\mathrm{MI}} = \overline{\mathcal{L}}_{\mathrm{MI}}^{(n)} = \dfrac{1}{N}\displaystyle\sum_{n=1}^{N} \mathrm{MI}\left(\boldsymbol{z}_{\mathrm{r}}^{(n)}, \boldsymbol{z}_{\mathrm{i}}^{(n)}\right) \\ \mathcal{L}_{\mathrm{LC}} = \overline{\mathcal{L}}_{\mathrm{LC}}^{(n)} = \dfrac{1}{N}\displaystyle\sum_{n=1}^{N} \mathrm{LC}\left(\boldsymbol{A}^{(n)}, \hat{\boldsymbol{A}}^{(n)}\right) \end{cases} . \tag{5}$$

Finally, the CORAL [20] loss is applied to align marginal distributions of all source subjects' EEGs in the embedding space. It is defined as the distance between the second-order statistics, covariances, of the EEG representations belonging to two source subjects. Thus, it needs to be calculated between each pair of source subjects and then averaged:

$$\mathcal{L}_{\mathrm{CORAL}} = \overline{\mathcal{L}}_{\mathrm{CORAL}}^{(n_1,n_2)} = \frac{1}{{}_N C_2} \sum_{1 \le n_1 < n_2 \le N} \left(\frac{1}{Q} \sum_{q=1}^{Q} \mathrm{CORAL}\left(\boldsymbol{z}_{\mathrm{r}(q)}^{(n_1)}, \boldsymbol{z}_{\mathrm{r}(q)}^{(n_2)}\right) \right) \tag{6}$$

where ${}_N C_2$ denotes the number of combinations of two subjects chosen from all N source subjects. We assume that there are Q classes of attention loci, and $\boldsymbol{z}_{\mathrm{r}(q)}^{(n_1)}$ (or $\boldsymbol{z}_{\mathrm{r}(q)}^{(n_2)}$) is a sub-set of $\boldsymbol{z}_{\mathrm{r}}^{(n_1)}$ (or $\boldsymbol{z}_{\mathrm{r}}^{(n_2)}$) that corresponds to the EEG segments of the n_1-th (or n_2-th) subject belonging to the q-th class.

The total loss function is computed as

$$\mathcal{L} = w_1\mathcal{L}_{\mathrm{CE}} + w_2\mathcal{L}_{\mathrm{MI}} + w_3\mathcal{L}_{\mathrm{LC}} + w_4\mathcal{L}_{\mathrm{CORAL}} \tag{7}$$

where loss weights are initialized as $w_1 = 1$ and $w_2 = w_3 = w_4 = 0$. They remain unchanged in the first two epochs. Starting from the third epoch, they are adjusted automatically at the beginning of each epoch according to the dynamic weight average scheme [24], so as to find the correct balance between them.

3. Experimental setup

3.1. Datasets and preprocessing

KUL [25]. In a dual-speaker scenario, 16 normal-hearing subjects were instructed to focus on one speaker and ignore the other. Two speech sources were played after being filtered with a head-related transfer function (HRTF) to simulate their positions at ±90° to the subjects. A total of approximately 36 minutes of EEG signals with 64 channels, consisting of ten trials, were recorded for each subject. The EEG signals had been downsampled to 128 Hz. See [26] for more details.

SNHL [27]. A total of 22 subjects with normal hearing were selectively listening to one of two simultaneous speech streams, which were spatially separated at ±90° to the subjects by an HRTF. The 64-channel EEG signals were recorded at a sampling rate of 512 Hz. Each subject participated in 32 trials, each of which was about 50 s long. See [28] for more details.

For both datasets, artifacts in EEG signals were marked through visual inspection, and then removed trial-by-trial using multi-channel Wiener filtering [29]. Then, the EEG signals were re-referenced to a common average, and 50 Hz power-line noises were attenuated by a notch filter. Additionally, the EEG signals in the SNHL dataset were low-pass filtered with a cutoff frequency of 64 Hz, and then downsampled to 128 Hz.

3.2. Implementation

Based on the premise of subject-independent ASAD, we adopt *leave-one-subject-out* cross-validation to simulate cross-subject scenarios. In this way, each subject in a dataset participates in testing in turn as the target subject, while all remaining subjects in that dataset participate in training as source subjects.

During training, the proposed model consistently runs for 80 epochs (~ 2.5 minutes per epoch for both datasets) with an early stopping strategy. An AdamW optimizer with a weight decay of 0.005 is adopted. The batch size is set to 64. The learning rate is initialized to 0.005, and a cosine decay schedule with warm restarts is applied. The proposed model has 220761 and 302169 parameters on the KUL and SNHL datasets, respectively. Both hyperparameters and weight initialization in the proposed model are specified in the source code available at `https://github.com/Yixiang-Niu/ASAD-DG-ARMA`. The proposed model is implemented with the TensorFlow 2.10 framework and the Spektral 1.3 library on an NVIDIA GeForce GTX 1650 GPU.

The ASAD accuracy is adopted to evaluate the model performance, which is computed as the percentage of correctly classified EEG segments of the target subject. We also report the minimal expected switch duration (MESD) for the proposed model, which is defined as the estimated time required to switch predictions after attention switches of the target subject [30].

4. Results and discussion

4.1. Performance of the proposed model

Generally, real-world applications require an ASAD model to make predictions at very short time intervals, in order to detect attention switches in time. We thus choose four sliding decision windows with lengths of 1, 2, 3, and 4 s, but with a consistent hop length of 1 s. In this way, two adjacent EEG segments may overlap each other, making the total numbers of EEG segments under different decision window lengths approximately equal.

As shown in Figure 2, the accuracy of the proposed model gradually improves with the extension of the decision window. And the proposed model achieves median MESDs across subjects of 4.6 s and 22.6 s on the KUL and SNHL datasets, respectively. However, the proposed model fails to generalize ideally to several subjects, which seems to encounter the problem of negative transfer. We suppose that a public feature space shared by all subjects can be learned through FD with the constraint of the CORAL loss. However, it is not absolutely solid, especially considering that neither the KUL nor SNHL datasets contain a huge number of subjects per training.

In addition, the proposed model performs better on the KUL dataset than on the SNHL dataset. This may be due to the fact that in the SNHL dataset, the subjects are instructed to fixate their eye gaze at a cross hair presented on a computer screen. But in the KUL dataset, the subjects do not receive any instruction. Thus, their eye gaze may naturally move toward the direction of the attended sound source. Although we have removed artifacts from EEG signals as much as possible, it is possible for the proposed model to unintentionally leverage the residual electrooculogram, to enhance ASAD performance [31].

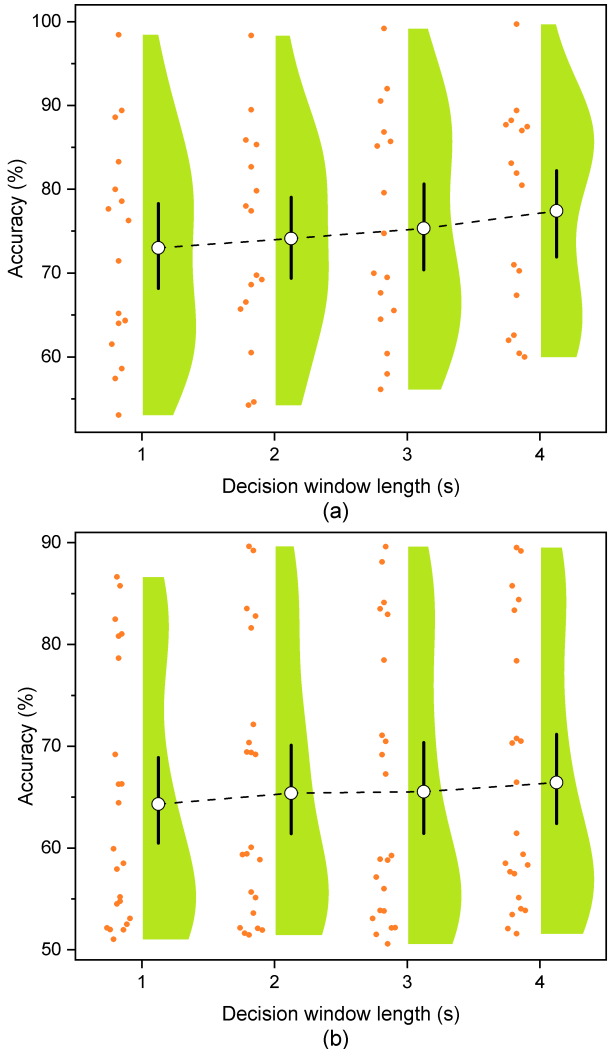

Figure 2: *ASAD accuracies of the proposed model on the (a) KUL and (b) SNHL datasets. Each orange dot represents a subject, while the means, 90% confidence intervals, and kernel density plots are given on the right.*

4.2. Comparative analyses

A series of ablation experiments are conducted on 2 s decision windows to verify the roles of two DG approaches, namely FD and CORAL, in the proposed model. First, we re-evaluate the proposed model after simplifying FD to domain-invariant representation learning. That is, only a subject-shared ARMA graph convolutional layer is utilized to learn the task-relevant feature, while the subject-private feature extraction, the MI and LC loss terms in Equation (7), and the decoder module are all removed from the proposed model. As shown in Table 1, the performance of the proposed model is undoubtedly degraded without FD. When the distribution shifts are quite serious, it is a great challenge for the domain-invariant representation learning to constrain the feature space to be subject-invariant. By contrast, the disentangled representation learning can better facilitate DG by weakening the influence of the subject-private feature. Second, we only remove the CORAL loss term in Equation (7) from the proposed model, and then retrain and test it. Table 1 shows that the proposed model suffers severer performance degradation, because it is likely to overfit source subjects' EEG signals during training, rather than learn a public feature space shared by all subjects.

The performance of the proposed model is compared with those of three baseline models, STAnet [32], MBSSFCC [10], and Densenet [33]. For a fair comparison, they are all evaluated on 2 s decision windows with the same data splits as used for the proposed model. As shown in Table 1, the proposed model consistently performs the best, which is primarily attributed to its DG-oriented optimization for cross-subject scenarios. Although both MBSSFCC and Densenet outperform STAnet by approximating spatial relationships between EEG electrodes, neither of them takes brain connectivity into account, which should also lead to their lower performance than that of the proposed model. Moreover, both MBSSFCC and Densenet have more than twice the training runtime and at least three times the number of parameters than the proposed model.

Table 1: *Mean ASAD accuracies (%) across subjects along with 90% confidence intervals*

Model	Dataset	
	KUL	SNHL
STAnet [32]	68.9 (64.1, 73.9)	57.8 (55.9, 60.2)
Densenet [33]	72.4 (67.8, 76.6)	61.8 (59.0, 65.3)
MBSSFCC [10]	72.2 (67.1, 77.2)	62.0 (58.6, 66.0)
Ours w/o CORAL	72.4 (67.1, 78.3)	63.1 (59.1, 67.7)
Ours w/o FD	72.9 (67.6, 78.6)	63.7 (59.6, 68.4)
Ours	**74.1 (69.2, 79.2)**	**65.4 (61.3, 70.3)**

5. Conclusions and future work

In this paper, we propose a new ASAD model for "neuro-steered hearing devices". Through a graph neural network architecture informed by brain effective connectivity and subject-invariant EEG pattern extraction based on the combination of FD and CORAL, the proposed model achieves higher accuracy than state-of-the-art ASAD models in cross-subject scenarios, with relatively low algorithmic complexity.

In the future, we will collect ultra-high-density EEG signals and perform EEG source connectivity analyses to further explore interactions across brain regions. We also consider improving the expressive power of the proposed model by fine-tuning pre-trained large graph models to replace shallow GNNs.

6. Acknowledgements

This work was supported in part by the National Natural Science Foundation of China [grant numbers 61771196, 61872143].

7. References

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

[2] 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, Jul. 2015.

[3] S. Geirnaert, S. Vandecappelle, E. Alickovic, A. de Cheveigné, 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, Jul. 2021.

[4] G. Ciccarelli, M. Nolan, J. Perricone, P. T. Calamia, S. Haro, J. O'Sullivan, N. Mesgarani, T. F. Quatieri, and C. J. Smalt, "Comparison of two-talker attention decoding from EEG with nonlinear neural networks and linear methods," *Scientific Reports,* vol. 9, 11538, Aug. 2019.

[5] Y. Niu, N. Chen, H. Zhu, J. Jin, and G. Li, "Music-oriented auditory attention detection from electroencephalogram," *Neuroscience Letters,* vol. 818, 137534, Jan. 2024.

[6] 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, no. 15, 116283, Jan. 2020.

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

[8] S. Cai, P. Sun, T. Schultz, and H. Li, "Low-latency auditory spatial attention detection based on spectro-spatial features from EEG," in *Proc. EMBC 2021 – 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society*, Guadalajara, Mexico, Nov. 2021, pp. 5812–5815.

[9] 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, e56481, Apr. 2021.

[10] Y. Jiang, N. Chen, and J. Jin, "Detecting the locus of auditory attention based on the spectro-spatial-temporal analysis of EEG," *Journal of Neural Engineering,* vol. 19, no. 5, 056035, Oct. 2022.

[11] S. Cai, T. Schultz, and H. Li, "Brain topology modeling with EEG-graphs for auditory spatial attention detection," *IEEE Transactions on Biomedical Engineering,* vol. 71, no. 1, pp. 171–182, Jan. 2024.

[12] R. Wang, S. Cai, and H. Li, "EEG-based auditory attention detection with spatiotemporal graph and graph convolutional network," in *Proc. INTERSPEECH 2023 – 24th Annual Conference of the International Speech Communication Association*, Dublin, Ireland, Aug. 2023, pp. 1144–1148.

[13] C. Fan, H. Zhang, W. Huang, J. Xue, J. Tao, J. Yi, Z. Lv, and X. Wu, "DGSD: Dynamical graph self-distillation for EEG-based auditory spatial attention detection," *arXiv preprint arXiv: 2309.07147*, 2023.

[14] J. Wilroth, B. Bernhardsson, F. Heskebeck, M. A. Skoglund, C. Bergeling, and E. Alickovic, "Improving EEG-based decoding of the locus of auditory attention through domain adaptation," *Journal of Neural Engineering,* vol. 20, no. 6, 066022, Dec. 2023.

[15] S. Pahuja, G. Ivucic, F. Putze, S. Cai, H. Li, and T. Schultz, "Enhancing subject-independent EEG-based auditory attention decoding with WGAN and Pearson correlation coefficient," in *Proc. SMC 2023 – IEEE International Conference on Systems, Man, and Cybernetics*, Honolulu, USA, Oct. 2023, pp. 3715–3720.

[16] N. Zhang, A. Lin, and P. Shang, "Symbolic phase transfer entropy method and its application," *Communications in Nonlinear Science and Numerical Simulation,* vol. 51, pp. 78–88, Oct. 2017.

[17] F. M. Bianchi, D. Grattarola, L. Livi, and C. Alippi, "Graph neural networks with convolutional ARMA filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 44, no. 7, pp. 3496–3507, Jul. 2022.

[18] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, and P. S. Yu, "Generalizing to unseen domains: A survey on domain generalization," *IEEE Transactions on Knowledge and Data Engineering,* vol. 35, no. 8, pp. 8052–8072, Aug. 2023.

[19] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 45, no. 4, pp. 4396–4415, Apr. 2023.

[20] B. Sun and K. Saenko, "Deep CORAL: Correlation alignment for deep domain adaptation," in *Proc. ECCV 2016 Workshops – 14th European Conference on Computer Vision Workshops*, Amsterdam, The Netherlands, Nov. 2016, pp. 443–450.

[21] L. Shi, Y. Jiao, and B. Lu, "Differential entropy feature for EEG-based vigilance estimation," in *Proc. EMBC 2013 – 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Osaka, Japan, Jul. 2013, pp. 6627–6630.

[22] P. Zhong, D. Wang, and C. Miao, "EEG-based emotion recognition using regularized graph neural networks," *IEEE Transactions on Affective Computing,* vol. 13, no. 3, pp. 1290–1301, May 2020.

[23] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proc. ICLR 2018 – 6th International Conference on Learning Representations*, Vancouver, Canada, Feb. 2018, pp. 1–12.

[24] S. Liu, E. Johns, and A. J. Davison, "End-to-end multi-task learning with attention," in *Proc. CVPR 2019 – IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, USA, Jun. 2019, pp. 1871–1880.

[25] N. Das, T. Francart, and A. Bertrand. "Auditory attention detection dataset KULeuven", Aug. 2019, Version 1.1.0 [Online]. Available: https://doi.org/10.5281/zenodo.4004271.

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

[27] S. A. Fuglsang, J. Märcher-Rørsted, T. Dau, and J. Hjortkjær. "Selective auditory attention in normal-hearing and hearing-impaired listeners", Jan. 2020, Version 1 [Online]. Available: https://doi.org/10.5281/zenodo.3618205.

[28] S. A. Fuglsang, J. Märcher-Rorsted, T. Dau, and J. Hjortkjær, "Effects of sensorineural hearing loss on cortical synchronization to competing speech during selective attention," *Journal of Neuroscience,* vol. 40, no. 12, pp. 2562–2572, Mar. 2020.

[29] B. Somers, T. Francart, and A. Bertrand, "A generic EEG artifact removal algorithm based on the multi-channel Wiener filter," *Journal of Neural Engineering,* vol. 15, no. 3, 036007, Jun. 2018.

[30] S. Geirnaert, T. Francart, and A. Bertrand, "An interpretable performance metric for auditory attention decoding algorithms in a context of neuro-steered gain control," *IEEE Transactions on Neural Systems and Rehabilitation Engineering,* vol. 28, no. 1, pp. 307–317, Jan. 2020.

[31] I. Rotaru, S. Geirnaert, N. Heintz, I. Van de Ryck, A. Bertrand, and T. Francart, "What are we really decoding? Unveiling biases in EEG-based decoding of the spatial focus of auditory attention," *Journal of Neural Engineering,* vol. 21, no. 1, 016017, Feb. 2024.

[32] 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, Jul. 2022.

[33] X. Xu, B. Wang, Y. Yan, X. Wu, and J. Chen, "A DenseNet-based method for decoding auditory spatial attention with EEG," in *Proc. ICASSP 2024 – 49th IEEE International Conference on Acoustics, Speech and Signal Processing*, Seoul, Korea, Apr. 2024, pp. 1946–1950.