ListenNet: A Lightweight Spatio-Temporal Enhancement Nested Network for Auditory Attention Detection

Cunhang Fan , **Xiaoke Yang** , **Hongyu Zhang** , **Ying Chen** , **Lu Li** , **Jian Zhou** and **Zhao Lv**[*]
Anhui Province Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University
{cunhang.fan, jzhou, kjlz}@ahu.edu.cn,
{e22201014, e22201103, e23201035, e12314059}@stu.ahu.edu.cn

Abstract

Auditory attention detection (AAD) aims to identify the direction of the attended speaker in multi-speaker environments from brain signals, such as Electroencephalography (EEG) signals. However, existing EEG-based AAD methods overlook the spatio-temporal dependencies of EEG signals, limiting their decoding and generalization abilities. To address these issues, this paper proposes a Lightweight Spatio-Temporal Enhancement Nested Network (ListenNet) for AAD. The ListenNet has three key components: Spatio-temporal Dependency Encoder (STDE), Multi-scale Temporal Enhancement (MSTE), and Cross-Nested Attention (CNA). The STDE reconstructs dependencies between consecutive time windows across channels, improving the robustness of dynamic pattern extraction. The MSTE captures temporal features at multiple scales to represent both fine-grained and long-range temporal patterns. In addition, the CNA integrates hierarchical features more effectively through novel dynamic attention mechanisms to capture deep spatio-temporal correlations. Experimental results on three public datasets demonstrate the superiority of ListenNet over state-of-the-art methods in both subject-dependent and challenging subject-independent settings, while reducing the trainable parameter count by approximately 7 times. Code is available at: https://github.com/fchest/ListenNet.

1 Introduction

In multi-speaker environments, humans with normal hearing have the ability to focus on a specific speaker while ignoring interference from other sound sources, the phenomenon known as the cocktail party effect [Cherry, 1953]. The mechanism behind it is commonly referred to as selective auditory attention. This inherent ability plays a crucial role in human communication and has attracted growing interest in auditory attention detection (AAD), which aims to localize the attended speaker using brain signals [Dai *et al.*, 2018]. AAD could potentially enhance the design of human-centered intelligent interaction systems, such as hearing aids.

Neuroscientific studies have demonstrated a nonlinear relationship between auditory attention and brain activity [Choi *et al.*, 2013; Mesgarani and Chang, 2012], which involves higher cognitive processing in the cerebral cortex. Electroencephalography (EEG) signals are widely used due to their non-invasive nature, ease of acquisition, and high temporal resolution [De Taillez *et al.*, 2020; Fan *et al.*, 2024a]. Spatio-temporal patterns of EEG reveal attentional regulation during selective listening [Tune *et al.*, 2021]. Findings of the inter-subject correlation (ISC) suggest that EEG signals are synchronized across subjects during perception of the same naturalistic visual and narrative speech stimuli [Dmochowski *et al.*, 2014; Shen *et al.*, 2022]. Taking this perspective, EEG signals exhibit temporal correlations, spatial correlations across channels, and spatio-temporal dependencies, which could provide valuable information for discriminating different attention states and advancing robust AAD methods.

Despite significant progress made by existing EEG-based AAD methods, three major challenges still limit their performance and practical application. Firstly, many existing methods have made substantial strides in spatio-temporal modeling, effectively capturing dynamic spatial patterns, leading to improved detection performance. These methods typically treat space and time separately, as shown in Figure 1 (a) and (b). Spatial dependencies are captured independently, and temporal dependencies are subsequently extracted. However, these methods overlook the temporal context under dynamic time conditions, as well as the spatio-temporal dependencies across different channels during auditory stimulus processing. Secondly, the individual differences and the non-stationary characteristics of EEG signals lead to significant performance degradation when applying AAD methods across subjects. [Cai *et al.*, 2024; Fan *et al.*, 2024b] effectively leverage individual-specific features to demonstrate strong performance in the subject-dependent setting, but they lack good generalization ability, which makes it difficult to develop subject-independent robust methods. Lastly, the pursuit of accuracy in current methods [Jiang *et al.*, 2022; Ni *et al.*, 2024] leads to large model sizes and high computational complexity, which are often attributed to complex feature extraction methods and transformer attention mechanisms, making them impractical for low-power devices.

[*] Corresponding Author

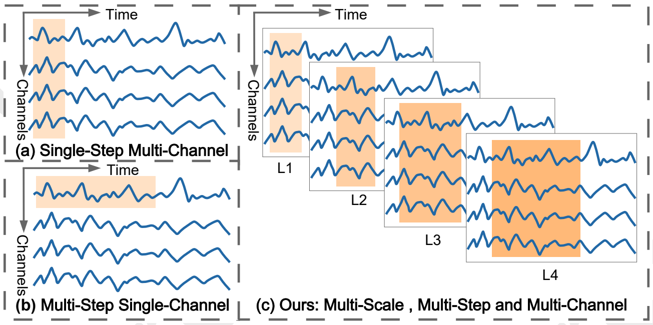
Time
Channels
(a) Single-Step Multi-Channel
Time
Channels
(b) Multi-Step Single-Channel
Time
Channels
L1
L2
L3
L4
(c) Ours: Multi-Scale , Multi-Step and Multi-Channel

Figure 1: Spatio-temporal modeling methods for AAD. Existing methods typically treat space and time separately, processing them from (a) to (b). The proposed ListenNet introduces a multi-scale of temporal patterns, as shown in (c), by considering cross-channel dependencies, temporal dynamics, and spatio-temporal dependencies for a more comprehensive modeling approach.

To address these issues, this paper proposes a **Li**ghtweight **S**patio-**T**emporal **E**nhancement **N**ested **Net**work (ListenNet) with low parameter count and computational complexity. As shown in Figure 1 (c), it captures multi-channel spatio-temporal dependencies and multi-scale dynamic temporal patterns, ensuring high accuracy and strong generalization. Specifically, the proposed ListenNet consists of three components: (1) *Spatio-temporal Dependency Encoder (STDE)* captures consecutive time steps and multi-channel features, differing from previous studies that first focus on channel features. It expands the input EEG signals within each channel to capture temporal dependencies and extracts spatial features both within and across channels, enhancing spatio-temporal representation capacity. (2) *Multi-scale Temporal Enhancement (MSTE)* captures temporal dependencies at multiple time scales, adding dynamic temporal context to build robust temporal embeddings. (3) *Cross-Nested Attention (CNA)* groups spatio-temporal features in parallel, extracts sub-feature context, and recalibrates weights by encoding global information, enhancing deep spatio-temporal correlations. Finally, the effective features are passed to a classifier to predict the subject's attended speaker. The major contributions of this work are summarized as follows:

- The proposed ListenNet overcomes the performance and efficiency limitations of existing methods for AAD by efficiently capturing spatio-temporal dependencies in both subject-dependent and subject-independent settings.
- A novel MSTE module is designed to efficiently extract multi-channel dependencies across multiple scales and time steps to integrate multi-level features, enhancing and complementing robust temporal representations.
- Experimental results show that ListenNet achieves outstanding accuracy while reducing the trainable parameter count by approximately 7 times. Specifically, it surpasses the best baseline by 6.1% on the DTU dataset under the subject-dependent setting and by 8.2% on the KUL dataset under the subject-independent setting, all within a 1-second decision window.

2 Related Works

For spatial dependency modeling, existing methods are divided into physical and dynamic dependencies. [Cai *et al.*, 2021; Jiang *et al.*, 2022] project differential entropy (DE) features in the frequency domain onto 2D topological maps using the known electrode positions to calculate spatial dependency based on physical distance and achieve good performance. Although physical dependency conforms to prior physiological paradigms, the electrode positions relations between channels cannot be directly equated to their functional connections [Liu *et al.*, 2024]. Currently, some researchers autonomously learn spatial dependency relationships during training. [Fan *et al.*, 2024b] extracts DE features as nodes to construct graph neural networks (GNN) and utilize an updated parameter matrix to represent spatial dependency. [Su *et al.*, 2021; Cai *et al.*, 2023; Cai *et al.*, 2024] design channel-wise attention mechanisms that learn to assign distinct weights to capture spatial patterns. [Ni *et al.*, 2024] utilizes a dual-branch approach to extract features from the temporal and frequency domains in parallel. For the frequency branch, it projects DE onto 2D maps and uses their topological patterns. For the temporal branch, the transformer encoder embeds a single cross-channel time step as an input token to autonomously learn features. The current state-of-the-art (SOTA) study [Yan *et al.*, 2024] employs spatial convolution operations across all channels to effectively capture spatial dependencies, resulting in competitive AAD performance.

For temporal dependency modeling, existing methods typically capture temporal dependencies using convolutional neural networks (CNN) and attention mechanisms. [Monesi *et al.*, 2020] independently uses long short-term memory (LSTM) networks to capture dependencies within EEG signals and achieve decent decoding performance. [Vandecappelle *et al.*, 2021] applies a simple one layer CNN model to directly process EEG data, where the time series are reduced to a single value. [Su *et al.*, 2022] sequentially processes temporal information after spatial attention, multiplying attention maps with EEG signals for adaptive feature refinement. [Wang *et al.*, 2023] utilizes a temporal attention mechanism after GNN that assigns varying weights to a sequence of EEG signals, enabling the capture of the complex temporal dynamics and enhancing the detection of even subtle changes in attentional states over time. Recently, [EskandariNasab *et al.*, 2024] employs gated recurrent units (GRU) and CNN to consider both historical and new temporal information when calculating the current state value, thereby inferring the temporal dependencies between time steps.

The methods mentioned above often focus separately on spatial and temporal features, or adopt a two-step processing strategy in which spatial dependencies are captured, followed by the modeling of temporal dependencies. However, these approaches tend to overlook the rich temporal contextual information under dynamic time conditions, as well as the spatio-temporal distribution characteristics of different brain regions during the reception, processing, and response to auditory stimuli. As a result, the failure to capture critical spatio-temporal dependencies significantly limits model performance.

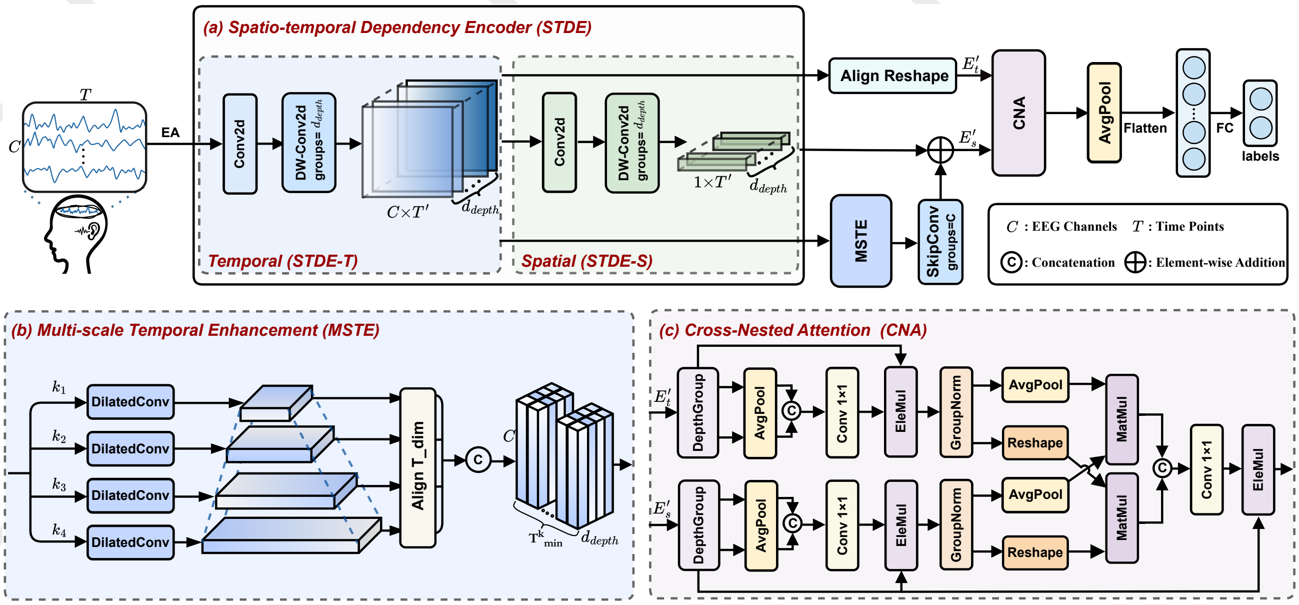

Figure 2: The overall structure of our ListenNet for AAD consists of three modules: (a) STDE module, (b) MSTE module, where k_i ($i \in \{1, 2, 3, 4\}$) represents the kernel size used in the dilated convolution, and (c) CNA module, where E_t' and E_s' are depth-aligned input feature maps. The model inputs are normalized and Euclidean-aligned EEG signals, and the outputs are two predicted labels related to auditory attention obtained through a classifier applied to the CNA output features.

3 The Proposed ListenNet Method

The proposed ListenNet is designed to comprehensively integrate spatio-temporal dependencies in EEG signals, addressing the limitations of existing methods by modeling dependencies across both multiple channels and time scales. Figure 2 illustrates the overall structure of ListenNet. The method will be specified in the following subsections.

Given the EEG data split by a moving window, a series of decision windows is obtained, each containing a short time segment of EEG signals. Consider the original EEG data of a decision window represented by $X = [x_1, ..., x_i, ..., x_T] \in \mathbb{R}^{C\times T}$, where C is the number of EEG channels and T is the length of the decision window. Here, $x_i \in \mathbb{R}^{C\times 1}$ is the EEG data at the i-th window of X. We aim to learn a representation $F(\cdot)$, which maps x to the corresponding label $y = F(x)$. Here, y denotes the locus (i.e., left or right) of auditory attention. Before inputting the EEG data into ListenNet, a Euclidean alignment (EA) method [Miao *et al.*, 2022] is employed, which standardizes the EEG data by calculating the average covariance matrix to extract shared features from the data across different brain states. $\tilde{X} \in \mathbb{R}^{C\times T}$ is obtained by normalizing and aligning X.

3.1 Spatio-temporal Dependency Encoder (STDE)

EEG signals are derived from different brain regions and exhibit dynamic changes in connectivity patterns between brain regions over time. Previous studies neglect the spatio-temporal characteristics of EEG signals. Meanwhile, as networks become increasingly complex [Zhang *et al.*, 2023; Chen *et al.*, 2023; Niu *et al.*, 2024], the limited size of EEG data makes these networks prone to overfitting. CNN-based networks have demonstrated sufficient feature extraction capabilities in brain-computer interface (BCI) tasks [Lawhern *et al.*, 2018; Miao *et al.*, 2023]. Considering these characteristics, we design a spatio-temporal dependency encoder to extract robust dynamic patterns using depthwise separable convolutions, which consists of the temporal feature component (STDE-T) and the spatial feature component (STDE-S), as shown in Figure 2 (a).

Firstly, STDE-T extracts dynamic features from EEG signals through temporal convolution layers, capturing temporal dependencies and constructing the temporal patterns E_t. This can be expressed as:

$$E_t = GELU(DepthwiseConv(Conv(\tilde{X}))) \tag{1}$$

where $E_t \in \mathbb{R}^{d_{\text{depth}}\times C\times T'}$, $Conv(\cdot)$ represents convolutional filters with a 1×1 kernel size to perform spatio-temporal reshaping on the input signals. $DepthwiseConv(\cdot)$ performs convolution independently on each input channel along the time dimension with a kernel size $1 \times k_0$ and a group size d_{depth}, followed by the $GELU(\cdot)$ activation function.

Subsequently, STDE-S encodes the spatial distribution information across all channels through spatial convolution layers, capturing the spatial distribution features E_s from EEG signals, which facilitate a comprehensive understanding of the brain's activity patterns in response to various auditory stimuli. This can be expressed as:

$$E_s = GELU(DepthwiseConv(Conv(E_t))) \tag{2}$$

where $E_s \in \mathbb{R}^{d_{\text{depth}}\times 1\times T'}$, $Conv(\cdot)$ represents convolutional filters with a 1×1 kernel size for initial channel mapping and achieving channel-wise feature fusion. $DepthwiseConv(\cdot)$ performs convolution to capture inter-channel dependencies with a $C \times 1$ and a group size d_{depth}, with the $GELU(\cdot)$ activation function. We integrate the spatial distribution features with the temporal patterns to form a comprehensive spatio-temporal embedding E_s.

3.2 Multi-scale Temporal Enhancement (MSTE)

The auditory system is sensitive to the temporal patterns [Puffay *et al.*, 2022]. Inspired by the concept of multi-scale modeling [Wu *et al.*, 2020; Fan *et al.*, 2024c], we propose a novel MSTE module. As shown in Figure 2 (b), the module captures dynamic brain activity across multiple time scales, offering a comprehensive representation of temporal patterns.

MSTE integrates dilated convolutions with the inception strategy to capture temporal features across multiple scales, thereby enabling a more comprehensive representation of multi-level temporal dependencies and enhancing the modeling of complex temporal patterns. The dilated convolution filters use four different kernel sizes to capture patterns at different time scales, with same dilation factor progressively expanding the effective receptive field. This enables the module to more efficiently capture both fine-grained and long-term temporal dependencies without increasing the number of parameters. Formally, the Inception strategy is combined with dilated convolutions to capture multi-scale temporal features. Given the input from the temporal convolution layers, the module applies four convolutional filters, each with a fixed dilation factor, to extract multi-scale temporal features. The outputs are truncated to match the size of the largest kernel, concatenated along the channel dimension, and normalized using batch normalization, ultimately generating the multi-scale feature map. The above process can be formulated as:

$$U = [\text{DilatedConv}_{1\times k}(E_t) \mid k \in \{k_1, k_2, k_3, k_4\}] \tag{3}$$

where $U \in \mathbb{R}^{d_{\text{depth}}\times C\times T^{k}_{\min}}$, and $T^{k}_{\min}$ represents the minimum time dimension among the outputs. $[\cdot]$ represents concatenation operation, and $DilatedConv_{1\times k}(\cdot)$ is implemented as a set of dilated convolutions with $k \in \{k_1, k_2, k_3, k_4\}$. For each kernel size k, a convolution is applied along the temporal dimension with a fixed dilation factor d.

The skip connection is implemented using a depthwise convolution with a kernel size of $C \times 1$ and a group size d_{depth}. These transform spatial information while preserving channel structure and standardize the sequence length for consistent transmission to the output module. The features are resized via bilinear interpolation to match the dimensions required by the subsequent layer, resulting in $S \in \mathbb{R}^{d_{\text{depth}}\times 1\times T'}$, which is then added to E_s, producing a robust representation of spatio-temporal dynamics $E'_s \in \mathbb{R}^{d_{\text{depth}}\times 1\times T'}$.

3.3 Cross-Nested Attention (CNA)

The multi-head attention mechanism in transformer models achieves significant results but incurs high computational cost. Inspired by the parallel strategy for cross-dimensional spatial information aggregation [Wang *et al.*, 2020; Ouyang *et al.*, 2023], we propose a novel cross-nested attention module that efficiently integrates hierarchical spatio-temporal features and reduces computational cost.

CNA employs dual-branch decomposition and interactive enhancement, extracting deep spatio-temporal features through attention weighting. Prior to processing, the input temporal feature E_t is depth-aligned with E'_s to produce $E'_t \in \mathbb{R}^{d_{\text{depth}}\times d_{\text{depth}}\times T'}$. As shown in Figure 2(c), both E'_t and E'_s are divided into G groups along the depth dimension, where $G = \lfloor d_{\text{depth}}/2 \rfloor$, and $\lfloor\cdot\rfloor$ denotes the floor operation. The dimension-adjusted features are denoted as F_t and F_s, respectively. Then, a dual-branch spatio-temporal module is applied to decompose and capture global information in both directions, producing two enhanced features, F_1 and F_2, as formulated below:

$$\begin{aligned} F_1 &= GN\left(F_t \odot \sigma\left(AAP_S(F_t)\right) \odot \sigma\left(AAP_T(F_t)\right)\right) \\ F_2 &= GN\left(F_s \odot \sigma\left(AAP_S(F_s)\right) \odot \sigma\left(AAP_T(F_s)\right)\right) \end{aligned} \tag{4}$$

where, $GN(\cdot)$ denotes the group normalization operation, $AAP_S(\cdot)$ denotes the spatial adaptive average pooling operation, $AAP_T(\cdot)$ denotes the temporal adaptive average pooling operation, $\sigma(\cdot)$ denotes the sigmoid activation function, and $\odot$ denotes the element-wise multiplication operation.

To capture long-range dependencies and global context, global average pooling and softmax are applied to each input branch to produce attention vectors. These are reshaped and used to compute cross-attention maps with features from the opposite branch via matrix multiplication. The resulting maps are concatenated and passed through a shared 1×1 convolution for feature fusion and dimensionality reduction, yielding the final attention weights $W \in \mathbb{R}^{(B\times G)\times 1\times 1\times T'}$, with B denoting the batch size. Finally, the output deep spatio-temporal features $E \in \mathbb{R}^{d_{\text{depth}}\times 1\times T'}$ are obtained by applying element-wise multiplication between F_s and the sigmoid-activated W.

3.4 Classifier

The classifier is designed to provide the final auditory attention results. Global average pooling is applied to reduce the dimensions of the features output by the CNA module. Then, the normalized feature maps are flattened into a 1D vector and fed into a fully connected layer to produce the final result. In the training stage, we apply the binary cross-entropy function to update the parameters.

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \cdot \log Q_i + (1 - y_i) \cdot \log(1 - Q_i)\right] \tag{5}$$

where y_i means the ground-truth label of i-th decision window, N means the number of samples, and Q_i is the corresponding possibility of predicted direction label with softmax function processing.

4 Experiment

4.1 Datasets

We evaluate ListenNet on three publicly available datasets: KUL [Das *et al.*, 2019; Das *et al.*, 2016], DTU [Fuglsang *et al.*, 2018; Fuglsang *et al.*, 2017], and AVED [ZHANG *et al.*, 2024]. We summarize the details of the above datasets in Table 1.

1) **KUL:** This dataset consists of 16 normal-hearing subjects, with 64-channel EEG data recorded. Each subject was instructed to attend to one of two competing voices from either the 90° left or right. Each subject completed 8 trials, each lasting 6 minutes.

Dataset	Scene	Subjects	Channels	Stimulus Direction	Duration (minutes)
KUL	audio-only	16	64	±90°	48
DTU	audio-only	18	64	±60°	50
AVED	audio-only	10	32	±90°	40
	audio-visual	10	32	±90°	40

Table 1: Details of three EEG datasets used in experiments.

2) **DTU:** This dataset consists of 18 normal-hearing subjects, with 64-channel EEG data recorded. Each subject was instructed to perform a target speaker tracking task in an environment with reverberation and dynamic background noise interference, attending to one of two competing voices from speakers positioned at a 60° relative to the subject. Each subject completed 60 trials, each lasting 50 seconds.

3) **AVED:** This dataset consists of 20 normal-hearing subjects, with 32-channel EEG data recorded. Subjects were evenly divided into two experimental conditions: audio-only and audio-visual, with 10 subjects in each condition. Each subject was instructed to attend to one of two competing voices from either the 90° left or right. In the audio-visual condition, subjects not only listened to the stories but also watched the video of the narrator they were instructed to focus on. Each subject completed 16 trials, each lasting 152 seconds.

4.2 Data Processing

To eliminate artifact noise and obtain cleaner EEG signals, specific preprocessing steps are applied to the three datasets to ensure consistency and comparability across experiments. For the KUL dataset, EEG signals are band-pass filtered (0.1–50 Hz) to remove irrelevant frequencies and downsampled to 128 Hz. For the DTU dataset, 50 Hz line noise and power line interference are filtered out, followed by downsampling to 128 Hz and high-pass filtering at 0.1 Hz. Eye artifacts are removed using joint decorrelation, and data are re-referenced to the average EEG channel response. For the AVED dataset, 50 Hz power line interference is removed, and the signals are band-pass filtered (0.1–50 Hz) and downsampled to 128 Hz. Subsequently, ocular and muscle artifacts are eliminated using independent component analysis (ICA). Finally, all EEG channels were re-referenced.

To evaluate ListenNet, we compare it with other SOTA AAD methods under both subject-dependent and the more challenging subject-independent settings. Specifically, four open-source models are selected as baselines: SSF-CNN [Cai *et al.*, 2021], MBSSFCC [Jiang *et al.*, 2022], DBPNet [Ni *et al.*, 2024], and DARNet [Yan *et al.*, 2024].

4.3 Implementation Details

We evaluate the performance of ListenNet on KUL, DTU, and AVED datasets under both subject-dependent and subject-independent settings. For the subject-dependent condition, each subject's data is split into training, validation, and test sets in an 8:1:1 ratio. The batch size is set to 32, the maximum number of epochs to 100, and an early stopping strategy is employed. Moreover, the model is trained using an Adam optimizer with a learning rate of 5e-4 and weight decay of 3e-4. For the subject-independent condition, the leave-one-subject-out (LOSO) cross-validation strategy is used. Namely, one subject's EEG data constituted the testing data, and the remaining subjects' EEG data constituted the training data. Here, the batch size is set to 128, with a maximum of 100 epochs. An Adam optimizer is also used with a learning rate of 1e-3 and a weight decay of 3e-4.

The following describes the implementation details, including the training settings and network configuration. The hyperparameters of ListenNet are consistently fixed across the three datasets to ensure a fair comparison of its generalizability. For STDE, the kernel size k_0 is set to 8, and the group size d_{depth} is set to 16. For MSTE, the kernel sizes used in the 2D dilated convolutional filter are $k \in \{1, 2, 3, 5\}$, and the dilation factor d is set to 1. Consequently, the number of groups G in CNA is configured as 8. All experiments are conducted using PyTorch on an RTX 4090 GPU.

5 Results

5.1 Comparison with Prior Art

In this work, we maintain the same subject-dependent setup as most existing models and evaluate our model in a more challenging subject-independent setup to better align with real-world applications, as detailed in Table 2.

Performance of Subject-Dependent

The comparison of subject dependence AAD performance between the ListenNet model and other baselines on the KUL, DTU and AVED datasets is presented in Tables 2. Our method significantly outperforms the current SOTA on both the KUL and DTU datasets. Specifically, on the KUL dataset, ListenNet demonstrates higher accuracies by 3.3%, 2.1%, and 1.8% for the 0.1-second, 1-second, and 2-second decision windows, respectively. Similarly, on the DTU dataset, it achieves improvements of 4.8%, 6.1%, and 5.4% in the same decision windows. On the AVED dataset, ListenNet performs slightly worse than DARNet in the 1-second and 2-second decision windows, but still achieves optimal performance in the very short 0.1-second window. One possible explanation is that DARNet's transformer attention outperforming by capturing long-range cross-modal dependencies in the AVED dataset.

We observe that ListenNet's decoding accuracy increases with the enlargement of decision windows, due to longer decision windows providing more information. The proposed ListenNet exhibits satisfactory performance at a temporal resolution of 1-second, which is approximately close to the time lag necessary for humans to switch attention. Moreover, our advantages are further enhanced under the highly challenging short 0.1-second decision window length, thereby contributing to the subsequent realization of real-time decoding of auditory attention.

Performance of Subject-Independent

Apart from excellent results in the subject-dependent setup, the proposed ListenNet also demonstrates comprehensive leading classification performance in the more challenging subject-independent setup across three datasets for the commonly used two detection window sizes. ListenNet benefits from better results by more comprehensively and effectively

Dataset	Scene	Model	Subject-Dependent			Subject-Independent	
			0.1-second	1-second	2-second	1-second	2-second
KUL	audio-only	CNN [Vandecappelle *et al.*, 2021]	74.3	84.1	85.7	56.8 ± 5.58	59.5 ± 8.21
		SSF-CNN [Cai *et al.*, 2021]	76.3 ± 8.47	84.4 ± 8.67	87.8 ± 7.87	59.3 ± 6.69	60.8 ± 8.40
		MBSSFCC [Jiang *et al.*, 2022]	79.0 ± 7.34	86.5 ± 7.16	89.5 ± 6.74	62.7 ± 8.08	64.7 ± 8.62
		EEGraph [Cai *et al.*, 2023]	88.7 ± 6.59	96.1 ± 3.22	96.5 ± 3.34	-	-
		DGSD [Fan *et al.*, 2024b]	-	90.3 ± 7.29	93.3 ± 6.53	63.6 ± 8.00	-
		DBPNet [Ni *et al.*, 2024]	85.3 ± 6.22	94.4 ± 4.62	95.3 ± 4.63	61.1 ± 8.26	62.3 ± 7.37
		DARNet [Yan *et al.*, 2024]	89.2 ± 5.50	94.8 ± 4.53	95.5 ± 4.89	69.9 ± 11.82	71.9 ± 13.01
		ListenNet (ours)	**92.5 ± 5.24**	**96.9 ± 3.01**	**97.3 ± 2.62**	**78.1 ± 13.50**	**79.6 ± 14.60**
DTU	audio-only	CNN [Vandecappelle *et al.*, 2021]	56.7	63.3	65.2	51.8 ± 3.03	52.9 ± 3.42
		SSF-CNN [Cai *et al.*, 2021]	62.5 ± 3.40	69.8 ± 5.12	73.3 ± 6.21	52.3 ± 3.50	53.4 ± 4.16
		MBSSFCC [Jiang *et al.*, 2022]	66.9 ± 5.00	75.6 ± 6.55	78.7 ± 6.75	52.5 ± 4.35	53.9 ± 5.80
		EEGraph [Cai *et al.*, 2023]	72.5 ± 7.41	78.7 ± 6.47	79.4 ± 7.16	-	-
		DGSD [Fan *et al.*, 2024b]	-	79.6 ± 6.76	82.4 ± 6.86	55.2 ± 4.07	-
		DBPNet [Ni *et al.*, 2024]	74.0 ± 5.20	79.8 ± 6.91	80.2 ± 6.79	55.5 ± 6.33	55.8 ± 6.11
		DARNet [Yan *et al.*, 2024]	74.6 ± 6.09	80.1 ± 6.85	81.2 ± 6.34	55.6 ± 4.13	55.6 ± 4.04
		ListenNet (ours)	**79.4 ± 7.00**	**86.2 ± 5.55**	**86.6 ± 4.82**	**56.8 ± 7.32**	**57.2 ± 5.83**
AVED	audio-only	SSF-CNN [Cai *et al.*, 2021]	53.4 ± 1.47	58.4 ± 3.79	58.9 ± 5.35	51.7 ± 0.85	52.5 ± 1.55
		MBSSFCC [Jiang *et al.*, 2022]	55.9 ± 1.80	70.2 ± 4.10	74.2 ± 7.24	52.2 ± 1.52	52.7 ± 1.87
		DBPNet [Ni *et al.*, 2024]	53.6 ± 2.65	58.9 ± 3.65	62.8 ± 5.93	52.1 ± 1.19	53.3 ± 1.88
		DARNet [Yan *et al.*, 2024]	49.7 ± 1.05	**80.2 ± 14.67**	**83.6 ± 12.10**	51.3 ± 0.21	52.1 ± 1.54
		ListenNet (ours)	**57.7 ± 1.71**	74.6 ± 3.36	77.1 ± 5.31	**52.8 ± 1.30**	**53.8 ± 1.98**
	audio-visual	SSF-CNN [Cai *et al.*, 2021]	54.5 ± 1.79	59.2 ± 5.44	63.1 ± 6.55	52.4 ± 2.29	53.8 ± 2.27
		MBSSFCC [Jiang *et al.*, 2022]	57.5 ± 2.75	69.6 ± 5.57	75.5 ± 4.34	52.8 ± 1.57	54.1 ± 1.86
		DBPNet [Ni *et al.*, 2024]	56.1 ± 2.68	61.5 ± 4.33	64.1 ± 6.09	53.3 ± 2.39	54.0 ± 1.61
		DARNet [Yan *et al.*, 2024]	50.3 ± 0.77	**83.6 ± 12.10**	**88.7 ± 13.15**	51.4 ± 0.32	52.6 ± 0.29
		ListenNet (ours)	**57.9 ± 2.16**	74.9 ± 4.63	76.5 ± 5.07	**53.7 ± 1.60**	**54.1 ± 1.83**

Table 2: Comparison of accuracy (%) on KUL, DTU and AVED datasets. The subject-dependent setup is conducted with three decision windows (0.1-second, 1-second, 2-second), with the results for AVED being reproduced, and the remaining results replicated from the corresponding papers. The subject-independent setup is conducted with two decision windows (1-second, 2-second), with DGSD from the original paper and others reproduced. Best results are highlighted in bold.

integrating dynamic temporal patterns and spatio-temporal dependencies, enabling the model to flexibly utilize subject-invariant representations. The results further confirm this capability. Especially on the KUL dataset, ListenNet achieves notable performance, demonstrating accuracy increases of 8.2% and 7.7% over the current SOTA model for the 1-second and 2-second decision windows, respectively. Furthermore, ListenNet outperforms baselines for DTU and AVED as well.

Compared to the widely-used KUL dataset, the DTU and AVED datasets pose a more challenging AAD task. Specifically, DTU presents speech at a narrower angle, and its recording environment includes reverberation and background noise, whereas AVED introduces complex multi-modal stimulus materials. The results show that ListenNet outperforms the baseline methods across diverse datasets, with lower variability in its results, further highlighting the stability and reliability of our approach across different decision windows. It learns the common pattern of feature distribution from subjects, thereby more effectively simulating real-world scenarios. These results highlight the robustness and generalization capabilities of the proposed model, emphasizing its potential superiority in EEG-based applications.

5.2 Ablation Analysis

Ablation studies are conducted on three datasets using a 1-second window setting, which most closely aligns with human attention switching [Jiang *et al.*, 2022; Fan *et al.*, 2025]. ListenNet constructs robust spatio-temporal representations. This enables the model to capture the full spatiotemporal information in EEG signals, thereby improving the interpretation of brain activity. Table 3 presents a comparison between the full ListenNet model and these four variants across the three datasets.

STDE-T and STDE-S are each removed to disrupt the integrity of STDE, thereby assessing the critical role of these components in the model's performance. Removing the STDE-T module for spatio-temporal dependency encoding has the most significant impact on the model's performance. The effectiveness of STDE-T can be attributed to the fact that EEG signals, as high temporal-resolution time series, exhibit strong temporal dependencies. Prioritizing the modeling of temporal continuity allows for the extraction of more effective and accurate spatio-temporal feature embeddings. Removing the STDE-S module results in accuracy decline, as full-channel spatial convolution captures inter-channel dependencies and establishes a robust spatio-temporal feature framework.

The removal of the MSTE module results in the loss of multi-scale temporal information and disrupts potential dependencies between temporal segments, thereby increasing the risk of overlooking critical temporal features essential for accurate recognition. Similarly, eliminating the CNA module diminishes the model's ability to dynamically assign feature weights and enhance spatio-temporal representations, weakening the extraction and integration of multi-level spatio-temporal features and ultimately reducing overall accuracy.

Dataset	Model	Subject-Dependent	Subject-Independent
KUL	w/o STDE-T	91.1 ± 6.05	62.6 ± 11.10
	w/o STDE-S	94.6 ± 6.18	76.0 ± 15.05
	w/o MSTE	96.7 ± 3.46	77.8 ± 13.39
	w/o CNA	96.3 ± 2.76	77.7 ± 14.74
	ListenNet	**96.9 ± 3.01**	**78.1 ± 13.50**
DTU	w/o STDE-T	72.5 ± 5.53	52.3 ± 2.01
	w/o STDE-S	84.3 ± 5.89	54.3 ± 8.36
	w/o MSTE	84.9 ± 6.59	56.7 ± 7.91
	w/o CNA	85.8 ± 5.75	56.5 ± 5.83
	ListenNet	**86.2 ± 5.55**	**56.8 ± 7.32**
AVED (audio-only)	w/o STDE-T	64.2 ± 6.62	51.1 ± 1.43
	w/o STDE-S	66.2 ± 4.50	52.6 ± 1.71
	w/o MSTE	71.8 ± 3.00	52.5 ± 1.48
	w/o CNA	74.3 ± 3.36	52.5 ± 1.32
	ListenNet	**74.6 ± 3.36**	**52.8 ± 1.30**
AVED (audio-visual)	w/o STDE-T	64.9 ± 5.30	53.3 ± 3.03
	w/o STDE-S	66.2 ± 5.27	53.2 ± 2.14
	w/o MSTE	72.8 ± 3.40	53.6 ± 1.80
	w/o CNA	74.6 ± 3.08	53.2 ± 2.53
	ListenNet	**74.9 ± 4.63**	**53.7 ± 1.60**

Table 3: Ablation study on all three datasets. The subject-dependent and subject-independent setups are conducted with 1-second decision windows, and "w/o" means without.

Model	Params (M)	MACs (M)
MBSSFCC [Jiang *et al.*, 2022]	83.91	89.15
DBPNet [Ni *et al.*, 2024]	0.91	96.55
DARNet [Yan *et al.*, 2024]	0.08	16.36
ListenNet (ours)	**0.01**	**12.16**

Table 4: The training parameter counts (Params) and multiply-accumulates (MACs) comparison on the KUL dataset.

5.3 Computational Cost

Table 4 compares the parameter counts and MACs of ListenNet with those of MBSSFCC, DBPNet, and DARNet on the KUL dataset. With only 0.01 M trainable parameters, ListenNet achieves remarkable parameter efficiency, requiring approximately 8390 times fewer parameters than MBSSFCC, 90 times fewer parameters than DBPNet, 7 times fewer parameters than DARNet. Additionally, ListenNet's computational demand is also markedly reduced, with its MACs only 12.16 M, approximately 86% lower than MBSSFCC, 87% lower than DBPNet and 26% lower than DARNet. These substantial reductions in both parameter count and computational complexity highlight ListenNet's enhanced efficiency, making it especially suitable for deployment on devices with limited computational resources.

5.4 Visualization Analysis

To assess the effect of extracting subject-invariant features, we randomly select 30 samples from each subject in the KUL dataset and visualize them using t-SNE [Van der Maaten and Hinton, 2008]. The resulting plots are shown in Figure 3. Different colors represent subjects, with circles and squares indicating attention to the left or right speaker, respectively. In Figure 3 (a), the raw features are scattered with significant overlap between subjects and labels, lacking clear structure and separability. In Figure 3 (b), preprocessing improves feature quality to some extent, but notable overlap and insufficient separability still remain. In Figure 3 (c), features extracted using STDE form clearer attention-related subgroups. By capturing spatio-temporal cross dependencies, the STDE module learns dynamic patterns and enhances feature separability, though some class boundaries remain indistinct. In Figure 3 (d), features extracted by ListenNet exhibit more distinct clustering for attention labels across subjects, and the distributions become more organized. This demonstrates that ListenNet learns subject-invariant features while maintaining clear boundaries between attention categories. These results further confirm the effectiveness of our method in enhancing the model's ability to decode attention states accurately while improving generalization across different subjects.

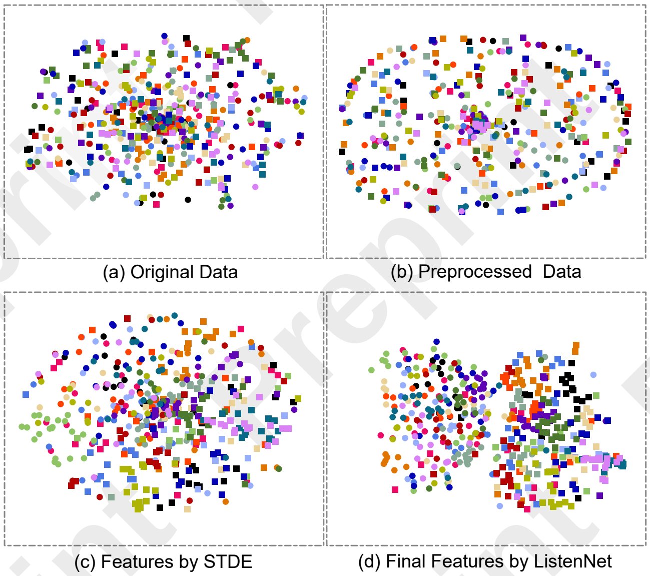
(a) Original Data
(b) Preprocessed Data
(c) Features by STDE
(d) Final Features by ListenNet

Figure 3: The t-SNE visualization of different types of features on the KUL dataset under the subject-independent condition. Different colors represent different subjects. Circles and squares denote attention to the left or right speaker, respectively.

6 Conclusion

This paper introduces ListenNet, a lightweight, highly accurate, and generalizable network for AAD. By combining spatio-temporal convolution operations across time steps and all channels, it effectively utilizes spatial information embedded in temporal EEG signals. Additionally, it captures temporal patterns at multiple scales, previously overlooked, by using multi-scale dilated convolutions. It integrates hierarchical spatio-temporal features through cross-nested attention mechanisms. Subject-dependent and subject-independent experiments are conducted on three AAD datasets. Experimental results show that our ListenNet exhibits competitive accuracy, especially in the very short 0.1-second decision window and across subjects. Furthermore, the compact size of our model and the reduced computational costs open new possibilities for deployment on low-power devices. For future work, we intend to extend ListenNet to streaming architectures, integrating incremental learning for real-time adaptation to AAD scenarios.

Acknowledgements

This work is supported by the STI 2030—Major Projects (No.2021ZD0201500), the National Natural Science Foundation of China (NSFC) (No.62201002, 6247077204), Excellent Youth Foundation of Anhui Scientific Committee (No. 2408085Y034), Distinguished Youth Foundation of Anhui Scientific Committee (No.2208085J05), Special Fund for Key Program of Science and Technology of Anhui Province (No.202203a07020008), Cloud Ginger XR-1.

References

[Cai *et al.*, 2021] Siqi Cai, Pengcheng Sun, Tanja Schultz, and Haizhou Li. Low-latency auditory spatial attention detection based on spectro-spatial features from eeg. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 5812–5815. IEEE, 2021.

[Cai *et al.*, 2023] Siqi Cai, Tanja Schultz, and Haizhou Li. Brain topology modeling with eeg-graphs for auditory spatial attention detection. *IEEE Transactions on Biomedical Engineering*, 2023.

[Cai *et al.*, 2024] Siqi Cai, Ran Zhang, and Haizhou Li. Robust decoding of the auditory attention from eeg recordings through graph convolutional networks. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2320–2324. IEEE, 2024.

[Chen *et al.*, 2023] Xiaoyu Chen, Changde Du, Qiongyi Zhou, and Huiguang He. Auditory attention decoding with task-related multi-view contrastive learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6025–6033, 2023.

[Cherry, 1953] E Colin Cherry. Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America*, 25(5):975–979, 1953.

[Choi *et al.*, 2013] Inyong Choi, Siddharth Rajaram, Lenny A Varghese, and Barbara G Shinn-Cunningham. Quantifying attentional modulation of auditory-evoked cortical responses from single-trial electroencephalography. *Frontiers in human neuroscience*, 7:115, 2013.

[Dai *et al.*, 2018] Bohan Dai, Chuansheng Chen, Yuhang Long, Lifen Zheng, Hui Zhao, Xialu Bai, Wenda Liu, Yuxuan Zhang, Li Liu, Taomei Guo, et al. Neural mechanisms for selectively tuning in to the target speaker in a naturalistic noisy situation. *Nature communications*, 9(1):2405, 2018.

[Das *et al.*, 2016] Neetha Das, Wouter Biesmans, Alexander Bertrand, and Tom Francart. The effect of head-related filtering and ear-specific decoding bias on auditory attention detection. *Journal of neural engineering*, 13(5):056014, 2016.

[Das *et al.*, 2019] Neetha Das, Tom Francart, and Alexander Bertrand. Auditory attention detection dataset kuleuven. *Zenodo*, 2019.

[De Taillez *et al.*, 2020] Tobias De Taillez, Birger Kollmeier, and Bernd T Meyer. Machine learning for decoding listeners' attention from electroencephalography evoked by continuous speech. *European Journal of Neuroscience*, 51(5):1234–1241, 2020.

[Dmochowski *et al.*, 2014] Jacek P Dmochowski, Matthew A Bezdek, Brian P Abelson, John S Johnson, Eric H Schumacher, and Lucas C Parra. Audience preferences are predicted by temporal reliability of neural processing. *Nature communications*, 5(1):4567, 2014.

[EskandariNasab *et al.*, 2024] MohammadReza Eskandari-Nasab, Zahra Raeisi, Reza Ahmadi Lashaki, and Hamidreza Najafi. A gru–cnn model for auditory attention detection using microstate and recurrence quantification analysis. *Scientific Reports*, 14(1):8861, 2024.

[Fan *et al.*, 2024a] Cunhang Fan, Jinqin Wang, Wei Huang, Xiaoke Yang, Guangxiong Pei, Taihao Li, and Zhao Lv. Light-weight residual convolution-based capsule network for eeg emotion recognition. *Advanced Engineering Informatics*, 61:102522, 2024.

[Fan *et al.*, 2024b] Cunhang Fan, Hongyu Zhang, Wei Huang, Jun Xue, Jianhua Tao, Jiangyan Yi, Zhao Lv, and Xiaopei Wu. Dgsd: Dynamical graph self-distillation for eeg-based auditory spatial attention detection. *Neural Networks*, 179:106580, 2024.

[Fan *et al.*, 2024c] Cunhang Fan, Jingjing Zhang, Hongyu Zhang, Wang Xiang, Jianhua Tao, Xinhui Li, Jiangyan Yi, Dianbo Sui, and Zhao Lv. Msfnet: Multi-scale fusion network for brain-controlled speaker extraction. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1652–1661, 2024.

[Fan *et al.*, 2025] Cunhang Fan, Hongyu Zhang, Qinke Ni, Jingjing Zhang, Jianhua Tao, Jian Zhou, Jiangyan Yi, Zhao Lv, and Xiaopei Wu. Seeing helps hearing: A multi-modal dataset and a mamba-based dual branch parallel network for auditory attention decoding. *Information Fusion*, page 102946, 2025.

[Fuglsang *et al.*, 2017] Søren Asp Fuglsang, Torsten Dau, and Jens Hjortkjær. Noise-robust cortical tracking of attended speech in real-world acoustic scenes. *Neuroimage*, 156:435–444, 2017.

[Fuglsang *et al.*, 2018] Søren A Fuglsang, Daniel DE Wong, and Jens Hjortkjær. Eeg and audio dataset for auditory attention decoding. *Zenodo*, 2018.

[Jiang *et al.*, 2022] Yifan Jiang, Ning Chen, and Jing Jin. Detecting the locus of auditory attention based on the spectro-spatial-temporal analysis of eeg. *Journal of Neural Engineering*, 19(5):056035, 2022.

[Lawhern *et al.*, 2018] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.

[Liu *et al.*, 2024] Chenyu Liu, Xinliang Zhou, Jiaping Xiao, Zhengri Zhu, Liming Zhai, Ziyu Jia, and Yang Liu. Vsgt:

variational spatial and gaussian temporal graph models for eeg-based emotion recognition. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 3078–3086, 2024.

[Mesgarani and Chang, 2012] Nima Mesgarani and Edward F Chang. Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, 485(7397):233–236, 2012.

[Miao *et al.*, 2022] Zhengqing Miao, Xin Zhang, Carlo Menon, Yelong Zheng, Meirong Zhao, and Dong Ming. Priming cross-session motor imagery classification with a universal deep domain adaptation framework. *arXiv preprint arXiv:2202.09559*, 2022.

[Miao *et al.*, 2023] Zhengqing Miao, Meirong Zhao, Xin Zhang, and Dong Ming. Lmda-net: A lightweight multi-dimensional attention network for general eeg-based brain-computer interfaces and interpretability. *NeuroImage*, 276:120209, 2023.

[Monesi *et al.*, 2020] Mohammad Jalilpour Monesi, Bernd Accou, Jair Montoya-Martinez, Tom Francart, and Hugo Van Hamme. An lstm based architecture to relate speech stimulus to eeg. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 941–945. IEEE, 2020.

[Ni *et al.*, 2024] Qinke Ni, Hongyu Zhang, Cunhang Fan, Shengbing Pei, Chang Zhou, and Zhao Lv. Dbpnet: Dual-branch parallel network with temporal-frequency fusion for auditory attention detection. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 3115–3123, 2024.

[Niu *et al.*, 2024] Yixiang Niu, Ning Chen, Hongqing Zhu, Zhiying Zhu, Guangqiang Li, and Yibo Chen. Auditory spatial attention detection based on feature disentanglement and brain connectivity-informed graph neural networks. In *Proc. Interspeech 2024*, pages 887–891, 2024.

[Ouyang *et al.*, 2023] Daliang Ouyang, Su He, Guozhong Zhang, Mingzhu Luo, Huaiyong Guo, Jian Zhan, and Zhijie Huang. Efficient multi-scale attention module with cross-spatial learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[Puffay *et al.*, 2022] Corentin Puffay, Jana Van Canneyt, Jonas Vanthornhout, Hugo Van Hamme, and Tom Francart. Relating the fundamental frequency of speech with eeg using a dilated convolutional network. *arXiv preprint arXiv:2207.01963*, 2022.

[Shen *et al.*, 2022] Xinke Shen, Xianggen Liu, Xin Hu, Dan Zhang, and Sen Song. Contrastive learning of subject-invariant eeg representations for cross-subject emotion recognition. *IEEE Transactions on Affective Computing*, 14(3):2496–2511, 2022.

[Su *et al.*, 2021] Enze Su, Siqi Cai, Peiwen Li, Longhan Xie, and Haizhou Li. Auditory attention detection with eeg channel attention. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 5804–5807. IEEE, 2021.

[Su *et al.*, 2022] Enze Su, Siqi Cai, Longhan Xie, Haizhou Li, and Tanja Schultz. Stanet: A spatiotemporal attention network for decoding auditory spatial attention from eeg. *IEEE Transactions on Biomedical Engineering*, 69(7):2233–2242, 2022.

[Tune *et al.*, 2021] Sarah Tune, Mohsen Alavash, Lorenz Fiedler, and Jonas Obleser. Neural attentional-filter mechanisms of listening success in middle-aged and older individuals. *Nature Communications*, 12(1):4533, 2021.

[Van der Maaten and Hinton, 2008] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[Vandecappelle *et al.*, 2021] Servaas Vandecappelle, Lucas Deckers, Neetha Das, Amir Hossein Ansari, Alexander Bertrand, and Tom Francart. Eeg-based detection of the locus of auditory attention with convolutional neural networks. *Elife*, 10:e56481, 2021.

[Wang *et al.*, 2020] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11534–11542, 2020.

[Wang *et al.*, 2023] Ruicong Wang, Siqi Cai, and Haizhou Li. Eeg-based auditory attention detection with spatiotemporal graph and graph convolutional network. In *Proceedings of INTERSPEECH*, pages 1144–1148, 2023.

[Wu *et al.*, 2020] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, Xiaojun Chang, and Chengqi Zhang. Connecting the dots: Multivariate time series forecasting with graph neural networks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 753–763, 2020.

[Yan *et al.*, 2024] Sheng Yan, Cunhang Fan, Hongyu Zhang, Xiaoke Yang, Jianhua Tao, and Zhao Lv. Darnet: Dual attention refinement network with spatiotemporal construction for auditory attention detection. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, pages 31688–31707, 2024.

[Zhang *et al.*, 2023] Yuanming Zhang, Haoxin Ruan, Ziyan Yuan, Haoliang Du, Xia Gao, and Jing Lu. A learnable spatial mapping for decoding the directional focus of auditory attention using eeg. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[ZHANG *et al.*, 2024] Hongyu ZHANG, Jingjing ZHANG, DONG Xingguang, LÜ Zhao, TAO Jianhua, ZHOU Jian, WU Xiaopei, and FAN Cunhang. Based on audio-video evoked auditory attention detection electroencephalogram dataset. *Journal of Tsinghua University (Science and Technology)*, 64(11):1919–1926, 2024.