MHANet: Multi-scale Hybrid Attention Network for Auditory Attention Detection

Lu Li, **Cunhang Fan***, **Hongyu Zhang**, **Jingjing Zhang**, **Xiaoke Yang**, **Jian Zhou** and **Zhao Lv**
Anhui Province Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei, 230601, China
{e12314059, e22201103, e22201067, e22201014}@stu.ahu.edu.cn
{cunhang.fan, Jzhou, kjlz}@ahu.edu.cn

Abstract

Auditory attention detection (AAD) aims to detect the target speaker in a multi-talker environment from brain signals, such as electroencephalography (EEG), which has made great progress. However, most AAD methods solely utilize attention mechanisms sequentially and overlook valuable multi-scale contextual information within EEG signals, limiting their ability to capture long-short range spatiotemporal dependencies simultaneously. To address these issues, this paper proposes a multi-scale hybrid attention network (MHANet) for AAD, which consists of the multi-scale hybrid attention (MHA) module and the spatiotemporal convolution (STC) module. Specifically, MHA combines channel attention and multi-scale temporal and global attention mechanisms. This effectively extracts multi-scale temporal patterns within EEG signals and captures long-short range spatiotemporal dependencies simultaneously. To further improve the performance of AAD, STC utilizes temporal and spatial convolutions to aggregate expressive spatiotemporal representations. Experimental results show that the proposed MHANet achieves state-of-the-art performance with fewer trainable parameters across three datasets, 3 times lower than that of the most advanced model. Code is available at: https://github.com/fchest/MHANet.

1 Introduction

In a complex auditory scene where multiple people speak simultaneously, commonly referred to as the cocktail party problem [Haykin and Chen, 2005], humans are capable of directing their auditory attention on one particular speaker. However, individuals with hearing impairments often face significant challenges in identifying and focusing on the attended speaker. Previous neuroscience studies have demonstrated that there exists a connection between brain activity and auditory attention [Mesgarani and Chang, 2012]. The auditory attention detection (AAD) task aims to detect auditory attention from neural activities. Inspired by these findings, researchers have proposed various neurorecording modalities to address this issue, including electroencephalography (EEG) [Choi *et al.*, 2013; O'sullivan *et al.*, 2015], electrocorticography (ECoG) [Mesgarani and Chang, 2012], and magnetoencephalography (MEG) [Ding and Simon, 2012; Akram *et al.*, 2016]. Among these, EEG-based methods stand out as premier solutions due to the non-invasive nature, wearability, and relative affordability of EEG. Therefore, in this paper, we concentrate on utilizing EEG signals for AAD.

According to studies that the cortical responses to an attended speaker are encoded within EEG signals, which correlate with auditory stimulus [Wöstmann *et al.*, 2016; Bednar and Lalor, 2020], the linear methods reconstruct the stimulus from EEG signals to detect the correlation between the reconstructed stimulus and the attended speech envelopes [Biesmans *et al.*, 2017; Katthi *et al.*, 2020]. However, in most real-world scenarios, obtaining clean auditory stimuli is challenging, on which these methods heavily depend. Meantime, the cortical responses have a nonlinear relationship with the acoustic stimuli [Keshishian *et al.*, 2020], and these linear methods have difficulty in capturing this complexity, thus leading to a longer decision window.

Therefore, some studies attempt to employ nonlinear neural networks to directly relate raw EEG signals to the attention detection decision [Ciccarelli *et al.*, 2019; Monesi *et al.*, 2020; Vandecappelle *et al.*, 2021]. For instance, the long short term memory (LSTM) layer is used to capture temporal patterns of EEG signals [Monesi *et al.*, 2020]. Then, a convolutional neural network (CNN) is proposed to extract the locus of auditory attention [Vandecappelle *et al.*, 2021]. Recently, some research has demonstrated that the latent spatial distribution of different EEG channels improves AAD performance. Consequently, some studies project the extracted differential entropy (DE) values on two-dimensional (2D) topological maps [Jiang *et al.*, 2022; Ni *et al.*, 2024; Fan *et al.*, 2025] or transform the original EEG channels into a 2D spatial topological map [Xu *et al.*, 2024]. To capture the dynamic auditory attention activity sensitive to the temporal patterns [Zion Golumbic *et al.*, 2013], the self-attention mechanism, which is proposed in [Vaswani *et al.*, 2017] and widely used in numerous vision tasks [Yan *et al.*, 2023], is also introduced into AAD to learn mutual temporal relationships or spatial distribution features within EEG signals [Su *et al.*, 2022; Pahuja *et al.*, 2023]. However, most methods usu-

*Corresponding author

ally have limited capabilities in comprehensively considering spatiotemporal representations of EEG signals.

To solve the problem, some studies adopt sequential approaches by applying the spatial attention mechanism followed by the temporal attention mechanism [Su *et al.*, 2022]. Nevertheless, EEG signals inherently possess spatiotemporal characteristics, and the separate consideration of spatial and temporal dimensions hinders these models from capturing spatiotemporal dependencies in an integrated manner. Simultaneously, the fusion of representations across multiple temporal and spatiotemporal scales is critical for uncovering latent brain activities related to auditory attention. Unfortunately, most existing models fail to adequately consider multi-scale features of EEG signals, leading to suboptimal feature representation and sensitivity to noise. This oversight significantly restricts their performance of AAD.

To address these issues, this paper proposes a novel multi-scale hybrid attention network for AAD, named MHANet, which effectively captures spatiotemporal features at multiple scales and uncovers the latent dependencies between spatial distribution features and temporal features from a global perspective. Specifically, our model comprises two modules: (1) *Multi-scale Hybrid Attention Module*. This module integrates a multi-scale temporal attention (MTA) block with the channel attention (CA) mechanism to combine expressive multi-scale temporal information and spatial distribution features, enhancing the representation capacity of EEG signals. Additionally, we introduce a multi-scale global attention (MGA) block, which treats EEG signals as Euclidean feature maps to simultaneously capture spatial distribution features and temporal patterns at multiple scales. This allows our model to focus on key channels while capturing long-short range temporal contextual information globally. (2) *Spatiotemporal Convolution Module*. In this module, multi-scale spatiotemporal features are aggregated through spatial and temporal layers, followed by global average pooling, to consolidate comprehensive representations. We evaluate the AAD performance of our MHANet on three datasets: KUL, DTU, and AVED. The results demonstrate that MHANet achieves state-of-the-art (SOTA) performance across three datasets. The major contributions of this paper are as follows:

- A novel AAD network, named MHANet, is proposed in this paper. This architecture combines multi-scale temporal features and spatial distribution features to capture long-short range spatiotemporal dependencies simultaneously.
- We introduce MTA to extract temporal information of EEG signals at multiple scales, enabling the construction of comprehensive temporal representations. Moreover, we propose MGA to capture multi-scale spatiotemporal dependencies globally, effectively identifying key channels and temporal patterns within EEG signals.
- The MHANet achieves SOTA decoding accuracy within an extremely short 0.1-second decision window on the KUL dataset, with an accuracy of 95.6%. It outperforms the best model by 6.4%. Moreover, our model is highly efficient with only 0.02M parameters, 3 times fewer than the leading model.

2 The Proposed MHANet

2.1 Overall Architecture

The previous AAD models typically apply attention mechanisms sequentially and neglect long-short range dependencies of spatiotemporal features of EEG signals at multiple scales.

To address these issues, we propose MHANet, as shown in Figure 1, which consists of MHA and STC. Specifically, MHA includes CA with a MTA and a MGA. Our MHANet extracts key spatial distribution features and long-short range temporal information at multiple scales, thus establishing a strong relationship between them.

Firstly, the processed data E is fed into MHA to extract comprehensive spatiotemporal features. Then, the output data is aggregated through STC. Finally, a fully connected (FC) layer is applied to generate the final prediction p.

$$p = w(STC(MHA(E)) + b \tag{1}$$

where w and b are the weight and bias parameters of the FC layer. We apply the cross-entropy loss function to supervise the network's training.

In the following, we introduce the proposed MHA, MTA, MGA and STC.

2.2 Multi-scale Hybrid Attention Module

As demonstrated in previous studies [Arvaneh *et al.*, 2011], EEG signals capture the brain's neuronal electrical activity, which varies over time and reflects activity patterns and connectivity across different brain regions. This temporal and spatial variability makes it possible to analyze the brain's response patterns to auditory stimuli by deeply extracting spatiotemporal features from EEG signals.

However, previous studies solely apply sequential attention mechanisms and focus on isolated spatial or temporal information. To address these shortcomings, we combine a MTA with CA [Zamir *et al.*, 2022] to comprehensively extract multi-scale spatiotemporal features from EEG signals. This architecture ensures that both temporal dynamics and spatial relationships are robustly captured, enhancing the decoding of auditory attention.

Firstly, we employ a convolutional layer to increase the channel resolution, enhancing the representation of spatial relationships of EEG signals. Following this, a depth-wise convolution is applied to extract temporal features E' within each individual channel, focusing on the temporal dynamics within EEG signals. This can be formulated as follows:

$$E' = DWConv(Conv(E)) \in \mathbb{R}^{3C\times 1\times T} \tag{2}$$

where $Conv(\cdot)$ represents a 1×1 convolutional layer and $DWConv(\cdot)$ denotes a depth-wise convolutional layer. This two-step convolution ensures that the spatial and temporal features of EEG signals are effectively preserved and processed for the next steps.

Subsequently, we split the tensor E' along the channel dimension to obtain the query Q, key K and value V components used for the self-attention mechanism [Vaswani *et al.*, 2017].

$$Q, K, V = Split(E') \in \mathbb{R}^{C\times 1\times T} \tag{3}$$

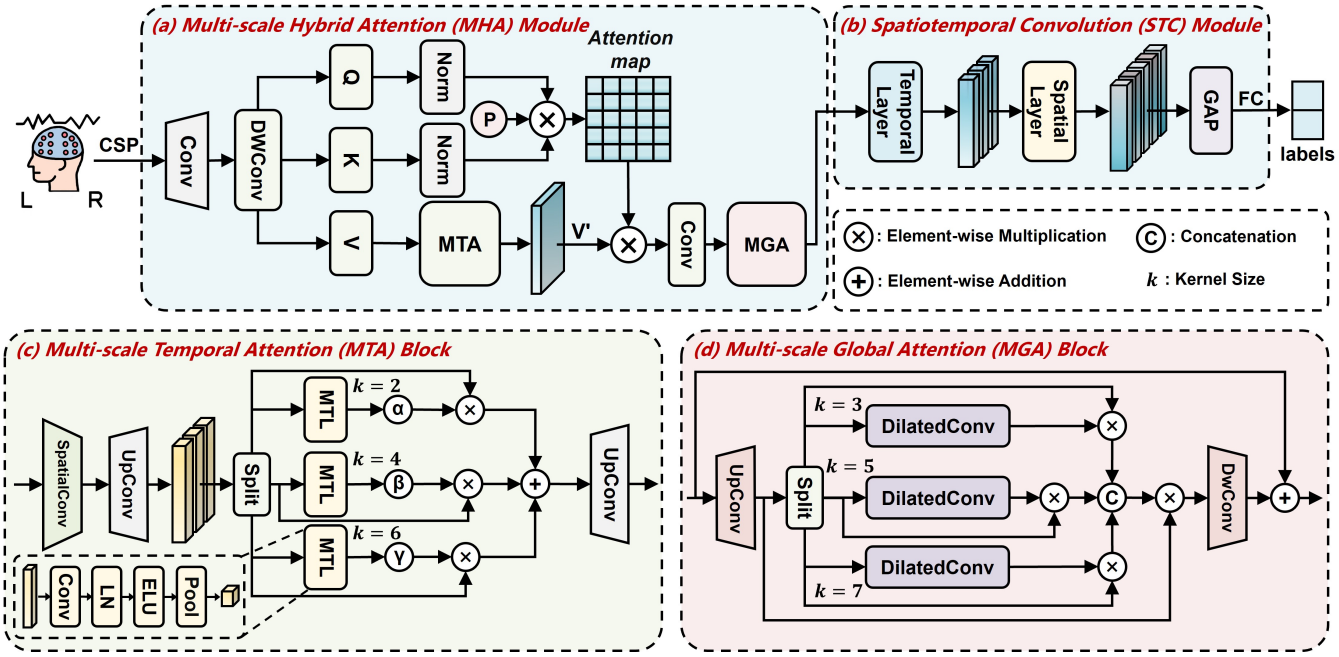

Figure 1: The overview architecture of our MHANet model for AAD, which mainly consists of two modules: (a) multi-scale hybrid attention (MHA) module and (b) spatiotemporal convolution (STC) module.

where $Split(\cdot)$ denotes the operation of splitting E' along the channel dimension.

Then, we introduce a MTA to comprehensively extract latent long-short range information and construct expressive temporal features V' with richer contextual details. The purpose of MTA is to further refine the value V representation by capturing diverse temporal dependencies at multiple scales.

$$V' = MTA(V) \in \mathbb{R}^{C\times 1\times T} \tag{4}$$

The detailed design and functionality of MTA will be elaborated upon in the following subsection.

Then, by employing the effective self-attention operation across all channels, our method extracts the key spatial distribution features and multi-scale temporal features. Meantime, we apply a 1×1 convolutional layer to further improve the nonlinear expression ability and obtain more robust spatiotemporal features H for the next block. The process can be formalized as:

$$H = Conv(Attention(Q, K, V')) \tag{5}$$

$$Attention(Q, K, V') = Softmax(\frac{QK^T}{t})V' \tag{6}$$

where $Conv(\cdot)$ represents a 1×1 convolutional layer without bias. $Softmax(\cdot)$ denotes the softmax function and t is a learnable scaling parameter.

Finally, the EEG data H is passed into MGA for further processing, constructing the refined spatiotemporal features F.

$$F = MGA(H) \in \mathbb{R}^{1\times C\times T} \tag{7}$$

The MGA will be discussed in detail in the following subsection.

2.3 Multi-scale Temporal Attention Block

Previous studies have demonstrated that the use of temporal attention mechanisms can effectively capture the time-varying nature of EEG signals [Su *et al.*, 2022; Ni *et al.*, 2024]. However, they usually overlook the importance of multi-scale information and fail to fully consider long-short range dependencies inherent in temporal features. To address this limitation, we propose MTA to fully leverage time-series information by incorporating multiple temporal scales.

Firstly, a spatial filter is used to reduce the channel dimension to prevent an excessive number of trainable parameters. Next, we use a 1×1 convolutional layer to increase the channel again and split EEG signals into three parts X, Y, Z for operations of different scales. It can be formally described as follows:

$$X, Y, Z = Split(UpConv(SpatialConv(V))) \tag{8}$$

where $SpatialConv(\cdot)$ denotes a $C\times 1$ convolutional layer. $UpConv(\cdot)$ denotes a 1×1 convolutional layer with 3 output channels. $Split(\cdot)$ performs the split operation along the channel dimension.

Next, X, Y and Z are passed through different multi-scale temporal layers to compute multiple attention values α, β, γ.

$$\{\alpha, \beta, \gamma\} = AAP(ELU(LN(Conv_i(\{X, Y, Z\})))) \tag{9}$$

where $Conv_i(\cdot)$ refers to a convolutional layer with a kernel size of $i \in \{2, 4, 6\}$. $LN(\cdot)$ denotes the layer normalization technique [Ba *et al.*, 2016]. $ELU(\cdot)$ is the exponential linear unit (ELU) activation function [Clevert *et al.*, 2016] and $AAP(\cdot)$ denotes an adaptive average pooling layer.

Then, we multiply each attention value with X, Y, Z separately and add them up to obtain the final fused temporal

features V' with multi-scale temporal patterns. Finally, it is put into a 1×1 convolutional layer to recover the number of channels and reshaped for further processing in the next stage.

$$V' = Conv(\alpha \odot X + \beta \odot Y + \gamma \odot Z) \tag{10}$$

where $\odot$ denotes the element-wise multiplication and $Conv(\cdot)$ represents a 1×1 convolutional layer with C output channels. α, β, γ are the learnable attention weight values.

2.4 Multi-scale Global Attention Block

Most previous studies usually overlook the dynamic relationship between spatial distribution features and time-series data of EEG signals. Enlightened by [Wang *et al.*, 2024], we propose MGA to further extract latent spatiotemporal dependencies within EEG signals from a global perspective.

Firstly, a 1×1 convolutional layer is applied to increase the channel dimension of the input data, producing H'. This is followed by cloning the data to facilitate the final multi-scale fusion. Then, we split the data into three distinct parts P, S, R for further processing.

$$H' = UpConv(H) \in \mathbb{R}^{3 \times C \times T} \tag{11}$$

$$P, S, R = Split(H') \in \mathbb{R}^{1 \times C \times T} \tag{12}$$

where $UpConv(\cdot)$ denotes a 1×1 convolutional layer that outputs 3 channels. $Split(\cdot)$ refers to a split operation along the channel dimension.

Subsequently, we separately compute the spatiotemporal attention maps. To capture multi-level spatiotemporal dependencies, we utilize dilated convolutional layers with varying kernel sizes and dilation rates. Each attention map is multiplied with its corresponding original features to obtain local refined attention maps δ, φ, μ.

$$\{\delta, \varphi, \mu\} = DConv_i(\{P, S, R\}) \odot \{P, S, R\} \tag{13}$$

where $DConv_i(\cdot)$ represents i-th dilated convolutional layer with varying kernel size (3×3, 5×5 and 7×7). The dilation rate of each layer is determined by the length of the decision window.

Next, we concatenate the resulting feature maps to obtain a multi-scale global attention map. It is then multiplied with H' to comprehensively capture spatiotemporal dependencies. Finally, a 1×1 convolutional layer is employed to restore its original shape. Simultaneously, a residual connection [He *et al.*, 2016] is introduced to enhance the performance and robustness of the block. It can be formulated as follows:

$$F = DwConv(H' \odot [\delta, \varphi, \mu]) + H \tag{14}$$

where $[\cdot]$ denotes the concatenate operation, which combines multiple attention maps along the channel dimension. $DwConv(\cdot)$ represents a 1×1 convolutional layer with 1 output channel.

2.5 Spatiotemporal Convolution Module

Through the comprehensive extraction of spatiotemporal features performed by the previous module, we obtain expressive information from EEG signals. To further improve the performance of AAD, STC is introduced at the end of our model to aggregate the extracted features. In STC, we apply a temporal layer and a spatial layer, followed by an adaptive average pooling layer to further process and downsample the data.

$$F' = ELU(BN(TemporalConv(F))) \tag{15}$$

$$F'' = ELU(BN(SpatialConv(F'))) \tag{16}$$

$$O = AdaptiveAvgPool(F'')) \tag{17}$$

where $TemporalConv(\cdot)$ denotes a 2D convolutional layer with a 1×2 kernel size, and $SpatialConv(\cdot)$ represents a 2D convolutional layer with a $C \times 1$ kernel size across all the channels. $BN(\cdot)$ is the batch normalization layer [Ioffe and Szegedy, 2015] and $AdaptiveAvgPool(\cdot)$ denotes an adaptive average pooling layer.

3 Experiments

3.1 Datasets

In this paper, we conduct extensive experiments on three publicly available datasets, namely KUL [Das *et al.*, 2016; Das *et al.*, 2019], DTU [Fuglsang *et al.*, 2017; Fuglsang *et al.*, 2018] and AVED [Fan *et al.*, 2024b], as shown in Table 1. The KUL and DTU datasets are among the most commonly used for AAD with the audio-only scene. The AVED dataset, provided for the ISCSLP Chinese AAD Challenge 2024, includes both audio-only and audio-visual scenes.

1) **KUL Dataset:** In this dataset, 64-channel EEG data were collected from 16 normal-hearing subjects (8 males and 8 females) at a sampling rate of 8192 Hz. The stimuli comprised four Dutch short stories narrated by three male speakers from 90° to the left or right. Each subject completed 8 trials, with each trial lasting 6 minutes.
2) **DTU Dataset:** In this dataset, 64-channel EEG data were collected from 20 normal-hearing subjects at a sampling rate of 512 Hz. The auditory stimuli consisted of Danish audiobooks narrated by a male and a female speaker. Each subject completed 60 trials, with each trial lasting 50 seconds.
3) **AVED Dataset:** In this dataset, 32-channel EEG data were collected from 20 normal-hearing subjects (14 males and 6 females) at a sampling rate of 1 kHz. The subjects were divided into two groups of 10. One group underwent audio-only experiments, and the other group underwent audio-video ones. All auditory stimuli were derived from 16 stories selected from a collection of Chinese short stories narrated by a male and a female speaker. Each subject completed 16 trials, with each trial lasting 152 seconds.

3.2 Data Processing

To ensure a fair comparison of the performance of our MHANet, specific preprocessing steps are applied to each dataset. For the KUL dataset, EEG data are initially re-referenced to the average response of mastoid electrodes, followed by bandpass filtering between 0.1 Hz and 50 Hz. The data are then down-sampled to 128 Hz. For the DTU dataset, EEG data are processed to remove 50 Hz linear noise and its

Dataset	Subjects	Scene	Duration (minutes)	Language
KUL	16	audio-only	48	Dutch
DTU	18	audio-only	10	Danish
AVED	10	audio-only	40	Mandarin
	10	audio-visual	40	Mandarin

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

harmonics. Eye artefacts are removed through joint decorrelation, and then the data are re-referenced to the average response of mastoid electrodes. Finally, the data are downsampled to 64 Hz. For the AVED dataset, a notch filter is applied first to eliminate powerline interference at 50 Hz. Next, a finite impulse response (FIR) filter is used for high-pass and low-pass filtering to remove the noise. Subsequently, the EEG data are downsampled to 128 Hz, followed by independent component analysis (ICA) for further noise removal. Finally, a re-referencing process is performed across all EEG channels to ensure consistency and comparability.

3.3 Implementation Details

In previous research on AAD, classification accuracy has been utilized as the standard metric for evaluating model performance. Following this convention, we assess our proposed MHANet using the KUL, DTU, and AVED datasets. To illustrate implementation details, including training settings and network configuration, we use the KUL dataset as an example, with a 1-second decision window.

The dataset is initially divided into training, validation, and test sets in a ratio of 8:1:1. For each subject in the KUL dataset, we allocate 4,600 decision windows for training, 576 for validation, and 576 for testing. The training process uses a batch size of 32, with a maximum of 100 epochs. An early stopping strategy is employed, halting training if there is no decrease in the validation set's loss function value for 15 consecutive epochs. The model is trained using the AdamW optimizer with a learning rate of 5e-3 and weight decay of 3e-4.

Initially, we employ the common spatial patterns (CSP) algorithm [Ramoser *et al.*, 2000; Blankertz *et al.*, 2007] to extract raw features from the EEG signals and rearrange them into E. Then, through MHA, we get the refined spatiotemporal features F. It is then sent to STC. After convolutional layers and global average pooling, we obtain O. Finally, the final binary AAD classification result p is achieved through a fully connected layer (input: 5, output: 2).

4 Results on AAD

4.1 Performance of MHANet

To comprehensively evaluate our proposed MHANet, we assess its AAD performance under different decision windows and compare our model with other outstanding methods, as shown in Table 2. For open-access models, we replicate their architectures, while results from other studies are cited accordingly.

Our MHANet demonstrates significant improvement and superior performance on KUL, DTU, and AVED datasets. On the DTU dataset, the accuracies are 75.5% (SD: 5.68%), 82.2% (SD: 8.13%), 83.0% (SD: 7.14%) under the 0.1-second, 1-second and 2-second decision windows, respectively. Similar accuracy trends are observed on the KUL and AVED datasets at different lengths of the decision window.

Overall, our model's performance decreases as the decision window length shortens. However, we observe that on the KUL dataset, the performance under 0.1-second is nearly as good as under 1-second. This could be because, under a 0.1-second decision window, the number of windows significantly increases, providing the model with more training samples, which aids in its learning. At the same time, the performance on the DTU dataset is generally lower compared to the KUL dataset. This could be attributed to differences in stimulus source locations, the number of speakers, their genders, and variations in data processing.

4.2 Ablation Experiment

To thoroughly analyze our model, we conduct extensive ablation experiments by removing the CA mechanism, MTA, MGA, both MTA and CA, and STC. All experiments are conducted under the same conditions as the previous settings. The results of these ablation experiments are presented in Table 3.

Experimental results show that on the DTU dataset, removing CA from MHANet leads to a decrease in average accuracy by 8.6% under the 1-second decision window. Removing the MTA also causes a similar accuracy drop: 3.7% for the 1-second decision window. When the MGA is removed, the accuracy decreases by 0.5% for the 1-second decision window. After removing both MTA and CA, the average accuracy decreases by 11.2% under the 1-second decision window. The removal of STC results in accuracy drops of 2.5% for the 1-second decision window. Similar trends of decreased accuracy are observed on the KUL and AVED datasets and at different lengths of the decision window after removing the aforementioned modules or blocks.

Overall, the complete MHANet demonstrates the best performance compared to versions with individual modules or blocks removed. On the DTU dataset with a 1-second decision window, MHANet outperforms the version without CA by 8.6%, highlighting the importance of focusing on spatial distribution features within EEG signals. It also achieves a 3.7% improvement over the version without MTA, indicating the effectiveness of capturing multi-scale temporal patterns. Additionally, MHANet shows a 0.5% improvement compared to the version without MGA, demonstrating the value of considering spatiotemporal dependencies from a global perspective. Finally, it outperforms the version without STC by 2.5%, emphasizing the necessity of integrating and refining spatiotemporal features of EEG signals from different dimensions.

5 Analysis and Discussion

5.1 Comparison with the SOTA Models

We compare the performance of our MHANet with other advanced AAD models, as presented in Table 2. The results

Decision Window	Model	Dataset			
		KUL	DTU	AVED (AO)	AVED (AV)
0.1-second	SSF-CNN* [Cai *et al.*, 2021]	76.3 ± 8.47	62.5 ± 3.40	53.3 ± 1.91	54.2 ± 2.00
	MBSSFCC* [Jiang *et al.*, 2022]	79.0 ± 7.34	66.9 ± 5.00	57.6 ± 2.87	58.9 ± 2.60
	EEG-Graph Net [Cai *et al.*, 2024]	-	72.5 ± 7.41	-	-
	DBPNet* [Ni *et al.*, 2024]	85.3 ± 6.22	74.0 ± 5.20	53.6 ± 2.93	55.7 ± 2.45
	DARNet* [Yan *et al.*, 2024]	89.2 ± 5.50	74.6 ± 6.09	51.3 ± 3.50	50.3 ± 0.60
	MHANet(ours)	**95.6 ± 4.83**	**75.5 ± 5.68**	**67.9 ± 2.10**	**67.4 ± 3.24**
1-second	SSF-CNN* [Cai *et al.*, 2021]	84.4 ± 8.67	69.8 ± 5.12	57.1 ± 3.54	59.2 ± 5.13
	MBSSFCC* [Jiang *et al.*, 2022]	86.5 ± 7.16	75.6 ± 6.55	70.5 ± 3.92	69.5 ± 5.77
	DGSD [Fan *et al.*, 2024a]	90.3 ± 7.29	79.6 ± 6.76	-	-
	EEG-Graph Net [Cai *et al.*, 2024]	-	78.7 ± 6.47	-	-
	DenseNet-3D [Xu *et al.*, 2024]	94.3 ± 4.3	-	-	-
	DBPNet* [Ni *et al.*, 2024]	94.4 ± 4.62	79.8 ± 6.91	58.7 ± 3.60	62.0 ± 4.92
	DARNet* [Yan *et al.*, 2024]	94.8 ± 4.53	80.1 ± 6.85	80.6 ± 15.69	83.1 ± 11.64
	MHANet(ours)	**95.8 ± 4.29**	**82.2 ± 8.13**	**87.1 ± 4.48**	**86.0 ± 5.32**
2-second	SSF-CNN* [Cai *et al.*, 2021]	87.8 ± 7.87	73.3 ± 6.21	59.8 ± 4.72	63.4 ± 5.13
	MBSSFCC* [Jiang *et al.*, 2022]	89.5 ± 6.74	78.7 ± 6.75	76.2 ± 3.64	74.3 ± 7.04
	DGSD [Fan *et al.*, 2024a]	93.3 ± 6.53	82.4 ± 6.86	-	-
	EEG-Graph Net [Cai *et al.*, 2024]	-	79.4 ± 7.16	-	-
	DenseNet-3D [Xu *et al.*, 2024]	95.9 ± 4.3	-	-	-
	DBPNet* [Ni *et al.*, 2024]	95.3 ± 3.50	80.2 ± 6.79	62.2 ± 6.27	63.3 ± 4.56
	DARNet* [Yan *et al.*, 2024]	95.5 ± 4.89	81.2 ± 6.34	91.3 ± 2.73	87.6 ± 13.19
	MHANet(ours)	**96.6 ± 3.67**	**83.0 ± 7.14**	**92.9 ± 3.93**	**92.0 ± 3.84**

Table 2: Auditory attention detection accuracy (%) comparison on the KUL, DTU, and AVED datasets. The KUL and DTU datasets consist of the audio-only scene. The AVED dataset includes both audio-only (AO) and audio-visual (AV) scenes. – indicates that no corresponding experiments are conducted or no results are provided in the respective paper. The results of the baseline models marked with * have been reproduced.

demonstrate that MHANet achieves significant improvements over the current SOTA methods.

On the DTU dataset, our MHANet achieves relative improvements of 13.0%, 8.6%, 3.0%, 1.5%, and 0.9% under the 0.1-second decision window compared to the SSF-CNN, MBSSFCC, EEG-Graph Net, DBPNet and DARNet models, respectively. For the 1-second decision window, MHANet achieves relative improvements of 12.4%, 6.6%, 2.6%, 3.5%, 2.4%, and 2.1% compared to the SSF-CNN, MBSSFCC, DGSD, EEG-Graph Net, DBPNet, and DARNet models. Similarly, the relative improvements under the 2-second decision window are 9.7%, 4.3%, 0.6%, 3.6%, 2.8%, and 1.8%, respectively.

On both the KUL and AVED datasets, MHANet achieves comparable improvements over other models. Notably, on the KUL dataset, MHANet achieves a 6.4% relative improvement under the 0.1-second decision window compared to the current SOTA method. At the same time, on the AVED dataset, the accuracies of other models are generally lower under the 0.1-second decision window, while our model shows a significant improvement. These all suggest that our model has an advantage in tasks that require adaptation to extremely short time windows.

Overall, the outstanding performance of MHANet across different datasets and decision windows highlights its strong real-world practicality and applicability, making it a promising solution for realistic hearing aids.

5.2 Ablation Analysis

As shown in Table 3, the results demonstrate that removing the CA, MTA, MGA, MTA and CA or STC leads to performance degradation. This highlights the effectiveness of each block and module in contributing to the advanced performance of our MHANet.

Effectiveness of CA

The distribution of EEG channels reflects the spatial relationships between them. Therefore, it is crucial to exploit the spatial coherence across different EEG channels [Geirnaert *et al.*, 2021]. The incorporation of the channel attention mechanism enables our model to focus on key spatial information within EEG signals, thus capturing the activities of different brain regions, which is essential for effectively analyzing EEG signals.

Effectiveness of MTA

Our MTA effectively captures multi-scale temporal information and generates more expressive and robust temporal features through convolutional attention at different scales. Therefore, it further improves the model's ability to understand latent temporal context at different levels and identify long-short range temporal patterns within EEG signals.

Decision Window	Model	Dataset			
		KUL	DTU	AVED (AO)	AVED (AV)
0.1-second	w/o CA	80.2 ± 12.03	66.1 ± 6.69	49.5 ± 1.50	50.0 ± 0.15
	w/o MTA	95.3 ± 3.81	72.8 ± 5.89	55.9 ± 6.35	55.9 ± 5.91
	w/o MGA	95.2 ± 4.87	74.6 ± 5.85	67.4 ± 2.28	67.1 ± 3.22
	w/o MTA and CA	89.1 ± 5.89	71.4 ± 8.12	50.6 ± 1.82	51.3 ± 2.69
	w/o STC	94.4 ± 5.72	74.7 ± 5.71	67.7 ± 2.38	67.3 ± 3.44
	MHANet(ours)	**95.6 ± 4.83**	**75.5 ± 5.68**	**67.9 ± 2.10**	**67.4 ± 3.24**
1-second	w/o CA	82.8 ± 12.10	73.6 ± 9.91	48.4 ± 3.65	50.6 ± 3.87
	w/o MTA	95.5 ± 4.69	78.5 ± 8.62	48.1 ± 4.72	68.4 ± 13.18
	w/o MGA	95.5 ± 4.44	81.7 ± 8.74	86.9 ± 3.96	85.3 ± 4.94
	w/o MTA and CA	93.1 ± 4.86	71.0 ± 9.60	49.6 ± 1.75	50.1 ± 1.69
	w/o STC	95.6 ± 4.01	79.7 ± 8.32	87.0 ± 3.43	84.4 ± 4.81
	MHANet(ours)	**95.8 ± 4.29**	**82.2 ± 8.13**	**87.1 ± 4.48**	**86.0 ± 5.32**
2-second	w/o CA	81.5 ± 11.99	69.9 ± 11.65	50.7 ± 2.22	50.2 ± 0.67
	w/o MTA	96.1 ± 4.45	78.9 ± 7.64	60.7 ± 17.86	60.1 ± 15.92
	w/o MGA	95.9 ± 4.51	82.4 ± 7.25	91.4 ± 2.60	91.3 ± 5.35
	w/o MTA and CA	94.6 ± 4.23	72.8 ± 8.33	49.7 ± 0.94	49.6 ± 1.07
	w/o STC	95.7 ± 4.60	79.7 ± 8.46	89.8 ± 5.68	88.6 ± 4.26
	MHANet(ours)	**96.6 ± 3.67**	**83.0 ± 7.14**	**92.9 ± 3.93**	**92.0 ± 3.84**

Table 3: Ablation Study on KUL, DTU, and AVED dataset. The KUL and DTU datasets consist of the audio-only scene. The AVED dataset includes both audio-only (AO) and audio-visual (AV) scenes. CA represents the channel attention. MTA denotes the multi-scale temporal attention block. MGA represents the multi-scale global attention block. STC denotes the spatiotemporal convolution module.

Effectiveness of MGA

Our MGA fully leverages the spatiotemporal feature maps to extract more valuable spatial distribution features and temporal patterns at multiple scales. By treating the EEG data as Euclidean data, it can capture spatiotemporal dependencies of EEG signals from a global perspective. The results show that our MGA further enhances performance to some extent, demonstrating the effectiveness of simultaneous spatiotemporal consideration.

Effectiveness of STC

The STC effectively aggregates spatiotemporal features along the temporal and channel dimensions through convolutional operations, respectively, resulting in more robust features with reduced noise and outliers. Meanwhile, the integration enhances our model's ability to understand brain activity, further improving its overall robustness and generalization ability.

5.3 Comparison of Computational Cost

We compare the training parameter counts of our MHANet, SSF-CNN [Cai *et al.*, 2021], MBSSFCC [Jiang *et al.*, 2022], DBPNet [Ni *et al.*, 2024] and DARNet [Yan *et al.*, 2024]. The results are shown in Table 4. The parameter count of MHANet is 209.5 times lower than that of SSF-CNN, 4194.5 times lower than MBSSFCC, 44.5 times lower than DBPNet, and 3 times lower than that of DARNet. MHANet showcases excellent parameter efficiency by achieving competitive performance with a significantly reduced number of parameters. This suggests our model's strong real-world applicability and practicality, making it well-suited for use in real-world low-resource hearing aids.

Model	Trainable Parameters
SSF-CNN [Cai *et al.*, 2021]	4.21 M
MBSSFCC [Jiang *et al.*, 2022]	83.91 M
DBPNet [Ni *et al.*, 2024]	0.91 M
DARNet [Yan *et al.*, 2024]	0.08 M
MHANet (ours)	**0.02 M**

Table 4: The training parameter counts of our MHANet and four open-source models. "M" denotes a million.

6 Conclusion

This paper proposes MHANet, a novel multi-scale hybrid attention network, to address the oversight of multi-scale spatiotemporal dependencies in AAD. We introduce MHA to capture long-short range spatiotemporal dependencies of EEG signals. The MHA combines CA with MTA to construct comprehensive and robust EEG features. Then, the MGA extracts the dynamic relationship between key spatial distribution features and temporal patterns. Subsequently, STC aggregates expressive EEG signals, improving our model's overall robustness and generalization ability. We evaluate the performance of the proposed MHANet on three datasets: KUL, DTU, and AVED, which demonstrate that MHANet achieves SOTA performance with fewer trainable parameters across all three datasets. This highlights its strong real-world practicality and applicability for realistic hearing aids. For future research, we plan to incorporate time-frequency analysis to further explore the spatiotemporal dependencies within EEG signals and enhance the performance of AAD.

Acknowledgments

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.

Contribution Statement

Lu Li and Cunhang Fan contributed equally to this work.

References

[Akram *et al.*, 2016] Sahar Akram, Jonathan Z Simon, and Behtash Babadi. Dynamic estimation of the auditory temporal response function from meg in competing-speaker environments. *IEEE Transactions on Biomedical Engineering*, 64(8):1896–1905, 2016.

[Arvaneh *et al.*, 2011] Mahnaz Arvaneh, Cuntai Guan, Kai Keng Ang, and Chai Quek. Optimizing the channel selection and classification accuracy in eeg-based bci. *IEEE Transactions on Biomedical Engineering*, 58(6):1865–1873, 2011.

[Ba *et al.*, 2016] Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *ArXiv*, abs/1607.06450, 2016.

[Bednar and Lalor, 2020] Adam Bednar and Edmund C. Lalor. Where is the cocktail party? decoding locations of attended and unattended moving sound sources using eeg. *NeuroImage*, 205:116283, 2020.

[Biesmans *et al.*, 2017] Wouter Biesmans, Neetha Das, Tom Francart, and Alexander 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*, 25(5):402–412, 2017.

[Blankertz *et al.*, 2007] Benjamin Blankertz, Ryota Tomioka, Steven Lemm, Motoaki Kawanabe, and Klaus-Robert Muller. Optimizing spatial filters for robust eeg single-trial analysis. *IEEE Signal processing magazine*, 25(1):41–56, 2007.

[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.*, 2024] Siqi Cai, Tanja Schultz, and Haizhou Li. Brain topology modeling with eeg-graphs for auditory spatial attention detection. *IEEE transactions on bio-medical engineering*, 71(1):171–182, 2024.

[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.

[Ciccarelli *et al.*, 2019] Gregory Ciccarelli, Michael Nolan, Joseph Perricone, Paul T. Calamia, Stephanie Haro, James O'Sullivan, Nima Mesgarani, Thomas F. Quatieri, and Christopher J. Smalt. Comparison of two-talker attention decoding from eeg with nonlinear neural networks and linear methods. *Scientific Reports*, 9:11538, 2019.

[Clevert *et al.*, 2016] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). In *Proceedings of the 4th International Conference on Learning Representations (ICLR 2016)*, 2016.

[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.

[Ding and Simon, 2012] Nai Ding and Jonathan Z Simon. Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of neurophysiology*, 107(1):78–89, 2012.

[Fan *et al.*, 2024a] 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.*, 2024b] Cunhang Fan, Jingjing Zhang, Hongyu Zhang, Xiang Wang, 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 (MM '24)*, 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, DD Wong, and Jens Hjortkjær. Eeg and audio dataset for auditory attention decoding. *Zenodo*, 2018.

[Geirnaert *et al.*, 2021] Simon Geirnaert, Servaas Vandecappelle, Emina Alickovic, Alain de Cheveigne, Edmund Lalor, Bernd T. Meyer, Sina Miran, Tom Francart, and Alexander Bertrand. Electroencephalography-based auditory attention decoding: Toward neurosteered hearing devices. *IEEE Signal Processing Magazine*, 38(4):89–102, 2021.

[Haykin and Chen, 2005] Simon Haykin and Zhe Chen. The cocktail party problem. *Neural Computation*, 17(9):1875–1902, 2005.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[Ioffe and Szegedy, 2015] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning (ICML'15)*, pages 448–456, 2015.

[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.

[Katthi *et al.*, 2020] Jaswanth Reddy Katthi, Sriram Ganapathy, Sandeep Kothinti, and Malcolm Slaney. Deep canonical correlation analysis for decoding the auditory brain. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 3505–3508, 2020.

[Keshishian *et al.*, 2020] Menoua Keshishian, Hassan Akbari, Bahar Khalighinejad, Jose L Herrero, Ashesh D Mehta, and Nima Mesgarani. Estimating and interpreting nonlinear receptive field of sensory neural responses with deep neural network models. *eLife*, 9:e53445, 2020.

[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.

[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*, pages 941–945, 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 IJCAI 2024*, pages 3115–3123, 2024.

[O'sullivan *et al.*, 2015] James A O'sullivan, Alan J Power, Nima Mesgarani, Siddharth Rajaram, John J Foxe, Barbara G Shinn-Cunningham, Malcolm Slaney, Shihab A Shamma, and Edmund C Lalor. Attentional selection in a cocktail party environment can be decoded from single-trial eeg. *Cerebral cortex*, 25(7):1697–1706, 2015.

[Pahuja *et al.*, 2023] Saurav Pahuja, Siqi Cai, Tanja Schultz, and Haizhou Li. XAnet: Cross-attention between eeg of left and right brain for auditory attention decoding. In *2023 11th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 1–4, 2023.

[Ramoser *et al.*, 2000] Herbert Ramoser, Johannes Muller-Gerking, and Gert Pfurtscheller. Optimal spatial filtering of single trial eeg during imagined hand movement. *IEEE transactions on rehabilitation engineering*, 8(4):441–446, 2000.

[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.

[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.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017)*, pages 6000–6010, 2017.

[Wang *et al.*, 2024] Yan Wang, Yusen Li, Gang Wang, and Xiaoguang Liu. Multi-scale attention network for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5950–5960, 2024.

[Wöstmann *et al.*, 2016] Malte Wöstmann, Björn Herrmann, Burkhard Maess, and Jonas Obleser. Spatiotemporal dynamics of auditory attention synchronize with speech. *Proceedings of the National Academy of Sciences*, 113(14):3873–3878, 2016.

[Xu *et al.*, 2024] Xiran Xu, Bo Wang, Yujie Yan, Xihong Wu, and Jing Chen. A densenet-based method for decoding auditory spatial attention with eeg. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1946–1950. IEEE, 2024.

[Yan *et al.*, 2023] Peilei Yan, Xuehu Liu, Pingping Zhang, and Huchuan Lu. Learning convolutional multi-level transformers for image-based person re-identification. *Visual Intelligence*, 1(1):24, 2023.

[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*, 2024.

[Zamir *et al.*, 2022] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming–Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5718–5729, 2022.

[Zion Golumbic *et al.*, 2013] Elana M. Zion Golumbic, Nai Ding, Stephan Bickel, Peter Lakatos, Catherine A. Schevon, Guy M. McKhann, Robert R. Goodman, Ronald Emerson, Ashesh D. Mehta, Jonathan Z. Simon, David Poeppel, and Charles E. Schroeder. Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". *Neuron*, 77(5):980–991, 2013.