



ASA: An Auditory Spatial Attention Dataset with Multiple Speaking Locations

Zijie Lin^{1,2}, Tianyu He¹, Siqi Cai^{*3}, Haizhou Li^{2,1,3,4}

¹ School of Data Science, The Chinese University of Hong Kong, Shenzhen, China

² Shenzhen Research Institute of Big Data, China

³ Department of Electrical and Computer Engineering, National University of Singapore, Singapore

⁴ Machine Listening Lab, University of Bremen, Germany

z.j.lin2001@outlook.com, tianyuhe@link.cuhk.edu.cn, elesiqi@nus.edu.sg,
haizhouli@cuhk.edu.cn

Abstract

Recent studies have demonstrated the feasibility of localizing an attended sound source from electroencephalography (EEG) signals in a cocktail party scenario. This is referred to as EEG-enabled Auditory Spatial Attention Detection (ASAD). Despite the promise, there is a lack of ASAD datasets. Most existing ASAD datasets are recorded from two speaking locations. To bridge this gap, we introduce a new Auditory Spatial Attention (ASA) dataset, featuring multiple speaking locations of sound sources. The new dataset is designed to challenge and refine deep neural network solutions in real-world applications. Furthermore, we build a channel attention convolutional neural network (CA-CNN) as a reference model for ASA, that serves as a competitive benchmark for future studies ¹.

Index Terms: Auditory spatial attention, EEG, channel attention, cocktail party problem

1. Introduction

The “cocktail party effect” refers to human’s ability to selectively attend to a single speaker in a multi-speaker acoustic environment, a capability significantly diminished in those with hearing impairments [1]. Conventional hearing aids often fall short in amplifying the attended sound source in cocktail party scenarios [2]. Recent advancements in neuroscience have revealed the potential to directly detect human auditory attention from EEG signals [3, 4], opening avenues for brain-controlled hearing aids [5].

Auditory attention detection algorithms can be categorized into two main paradigms: stimulus-reconstruction and auditory spatial attention detection (ASAD). The former correlates acoustic stimulus representations with EEG signals to identify the attended speaker, while the latter, i.e. ASAD, relies solely on EEG signals to determine the location of the attended sound source. ASAD is advantageous in practice because it eliminates the need for clean audio signals and operates accurately on low latency settings [2]. These characteristics make ASAD particularly suitable for real-time attention detection systems, especially in complex acoustic scenarios where hearing aid users shift their focus between different locations. Therefore, our study focuses on EEG-enabled ASAD tasks.

Despite recent advancements, there are limited publicly available EEG databases specifically designed for research on selective auditory attention. Fuglsang et al. [6] introduced the DTU database, which consists of 64-channel EEG signals

recorded from 18 subjects with normal hearing. The participants focused on one of two competing speakers in each trial, with stimuli presented at 60° to the left and right of the subjects. Each trial follows multiple-choice questions and the answers indicate whether the subjects were compliant in the auditory attention task [6]. Independently, KUL database [7] and ESAA database [8] were introduced, involving stimuli presented from ±90° using dichotic or head-related transfer function filtering (HRTF) techniques. It is worth noting that the above three widely used datasets are recorded from two speaking locations of sound sources. Recently, the NJU dataset was released [9]. It has a setup of 14 loudspeakers in an array, presenting a more realistic scenario. 32-channel EEG data were recorded from 21 subjects of normal hearing. However, the subjects listened to each trial twice, one to the left and another to the right, which could have an adverse impact on the attention levels [3].

Therefore, we propose a new Auditory Spatial Attention (ASA) dataset that contains 10 competing speaking locations, i.e., ±90°, ±60°, ±45°, ±30°, and ±5°. Moreover, behavioral indices, such as answers to multiple-choice questions and self-rated attention scores, are incorporated into the dataset, that allows for the study of the correlation between the level of attention and the resulting EEG signals.

Given the nonlinear characteristics and low signal-to-noise ratio of EEG signals, deep neural networks (DNNs) have been employed for the ASAD task to uncover hidden features. Among them, the convolutional neural network (CNN) stands out as a well-adopted method for EEG-enabled ASAD, which is referred to as “CNN-KUL” hereafter. In [10], a basic single-layer 2D CNN showed an average accuracy of around 81% with short decision windows (1-2 seconds) in the KUL dataset. Furthermore, a spatial-temporal attention network (STAnet) [11] was proposed and achieved remarkable results within 1-second decision windows (90.1% accuracy in the KUL dataset). More recently, graph convolution networks (GCNs) have surfaced as an alternative to CNNs for ASAD tasks [12]. However, one should note that these models are designed for subject-dependent conditions, while subject-independent solutions are more desirable.

Building on the success of the previous studies, we introduce a novel channel attention convolutional neural network (CA-CNN) tailored for EEG-enabled ASAD tasks. CA-CNN serves as the reference model for the proposed ASA dataset, establishing a benchmark for future studies by the ASA user community. The CA-CNN model exhibits exceptional performance and robustness under subject-dependent conditions over the prior approaches.

In the rest of this paper, we first describe the design of the ASA dataset. We then present the performance evaluation of the proposed CA-CNN model on the ASA dataset and discuss

¹The ASA Dataset and reference system are available at <https://zenodo.org/uploads/11541114>

*Corresponding author: Siqi Cai (elesiqi@nus.edu.sg)

the findings.

2. Description of the ASA Dataset

2.1. Participants

Twenty participants (9 male and 11 female) with normal auditory function were recruited from the university in this study. The male participants displayed a mean height of 175.6 ± 6.31 cm and a mean weight of 67.3 ± 12.73 kg, while the female participants exhibited a mean height of 163.5 ± 4.29 cm and a mean weight of 54.0 ± 6.99 kg. Prior to their participation, all volunteers provided written informed consent following ethical guidelines. Approval of all experimental procedures was granted by the ethics committee of the Chinese University of Hong Kong (Shenzhen).

2.2. Auditory stimuli

In this study, all participants, whose native language is Mandarin, were exposed to 40 short Chinese stories as auditory stimuli. Each auditory stimulus ranges from approximately one to one and a half minutes, narrated by a male or a female professional speaker [13]. In line with previous ASAD experimental prototypes [3, 6, 7, 8], each story was accompanied by three multiple-choice questions, with each question offering four possible answers to assess comprehension.

During each trial, two different auditory streams were played in different spatial directions of the listening subject using HRTF filtering. Participants were instructed to concentrate on a single Chinese story presented among the two competing stimuli. The auditory stimuli were delivered at a sample rate of 44.1 kHz.

2.3. Data acquisition

Participants in the study were seated in a soundproof chamber and instructed to wear wired headphones for the auditory stimuli presentation. EEG signals were recorded using the BrainAmp system at a sampling rate of 500 Hz. Specifically, a 64-channel Ag/AgCl electrode cap developed by Easycap was used, adhering to the electrode placement standards of the international 10/20 system. To ensure synchronization between the auditory stimuli and EEG responses, a Python-scripted automated playback and EEG-marking system was implemented.

2.4. ASAD task design

As outlined in Table 1, each participant engaged in a total of 20 trials. The trials were designed to systematically vary the spatial orientation of two auditory stimuli to each participant. Specifically, the orientation angles between the stimuli were adjusted sequentially every four trials, following a predetermined sequence: $\pm 90^\circ$, $\pm 60^\circ$, $\pm 45^\circ$, $\pm 30^\circ$, and $\pm 5^\circ$. Here, a negative angle (“-”) indicates leftward orientation, while a positive angle (“+”) denotes rightward orientation. For instance, an orientation of -90° represents an extreme leftward position.

Participants were instructed to focus on the auditory stimulus emanating from either the left or right side in each trial. The localization of the auditory stimulus was balanced over subjects for the attended side throughout the experiment. Additionally, the narratives delivered by male and female speakers varied across trials, maintaining an equal gender distribution among speakers.

During each trial, participants were instructed to keep their gaze fixed on the centrally positioned crosshair, minimizing eye

blinking frequency. Each trial lasted approximately one to one and a half minutes, resulting in an EEG data collection time of around 24 minutes per participant. Consequently, the cumulative EEG data gathered from all 20 participants amounted to approximately 480 minutes or 8 hours.

2.5. Behavioral indices

Following each trial, participants were required to answer questions related to the presented stimuli. Additionally, they were required to self-assess their attention levels, using a scale ranging from -2 to 2. These assessments not only served to maintain participants’ engagement and motivation toward concentrating on the task but also offered insights into evaluating their attention status. Initially, our dataset incorporated data from 24 participants. However, data from three subjects who didn’t follow the instructions were discarded. Additionally, one subject was excluded from analyses because he/she failed to correctly answer 2/3 of the questions [3, 8]. As a result, our refined dataset comprised data from 20 subjects.

To explore the relationship between the percentage of questions answered correctly and self-assessed attention levels, we calculated the Pearson correlation coefficient, denoted as r . The results indicate a statistically significant correlation ($r = 0.27$, $p < 0.001$), suggesting a positive association between the question accuracy and the participants’ self-assessed attention levels [3, 8].

Table 1: *Experiment design for a random subject. Trials are numbered according to the order in which they were presented to the subject. The localization of the auditory stimulus was balanced over subjects for the attended side. L=Left, R=Right, F= Female, M=Male*

Trial	Attended Side	Spatial locations	Gender of left speaker	Gender of right speaker
1	L	$-90^\circ, +90^\circ$	F	M
2	R	$-90^\circ, +90^\circ$	M	F
3	R	$-90^\circ, +90^\circ$	F	M
4	L	$-90^\circ, +90^\circ$	M	F
5	L	$-60^\circ, +60^\circ$	M	F
6	R	$-60^\circ, +60^\circ$	M	F
7	L	$-60^\circ, +60^\circ$	F	M
8	R	$-60^\circ, +60^\circ$	F	M
9	L	$-45^\circ, +45^\circ$	M	F
10	R	$-45^\circ, +45^\circ$	M	F
11	R	$-45^\circ, +45^\circ$	F	M
12	L	$-45^\circ, +45^\circ$	F	M
13	L	$-30^\circ, +30^\circ$	M	F
14	R	$-30^\circ, +30^\circ$	F	M
15	L	$-30^\circ, +30^\circ$	F	M
16	R	$-30^\circ, +30^\circ$	M	F
17	L	$-5^\circ, +5^\circ$	M	F
18	R	$-5^\circ, +5^\circ$	M	F
19	R	$-5^\circ, +5^\circ$	F	M
20	L	$-5^\circ, +5^\circ$	F	M

3. Methods

As depicted in Fig. 1, the proposed CA-CNN integrates a 1D-CNN module, a channel attention module, and a classifier, specifically designed to efficiently extract EEG features across various subjects for ASAD tasks.

3.1. 1D-CNN module

EEG signals are dynamic time-series data that contain valuable temporal information. Recently, 1D-CNN has proven effective in capturing these temporal patterns within EEG signals in several applications, such as motor imagery [14] and seizure detection [15]. Inspired by this, we adopt a 3-layer 1D-CNN module to extract temporal representations from EEG signals in our CA-CNN architecture.

The CA-CNN takes $\mathbf{E} \in \mathbb{R}^{L \times C}$, a segment of EEG signals as the input, where L denotes the number of samples within the segment, and C is the number of EEG channels. The first two layers are identical, each consisting of 16 filters of size 3, followed by a batch normalization (BN) layer and an average pooling layer with a pool size of 2. The third layer employs 32 filters of size 3, followed by a BN layer. Throughout these layers, the leaky ReLU is employed as the activation function. The resulting feature, denoted as $\mathbf{E}_f \in \mathbb{R}^{L/4 \times C_f}$, is extracted after the 1D-CNN module, where $C_f = 32$.

3.2. Channel attention module

EEG signals contain information originating from diverse channels, and the patterns within EEG responses to auditory stimuli can vary widely among individuals. The incorporation of attention mechanisms into the analysis of EEG signals offers several advantages, enhancing model performance and adaptability through the dynamic emphasis or de-emphasis of specific channels based on their respective significance [11, 16].

A channel attention module is employed in our study to extract spatial information from EEG signals better. This approach can be expressed as follows:

$$\begin{aligned} \mathbf{W} &= FC_{\sigma}(FC_{\zeta}(GAP(\mathbf{E}_f))) \\ \mathbf{E}_{ca} &= \mathbf{W} \otimes \mathbf{E}_f \end{aligned} \quad (1)$$

Here, $\mathbf{W} \in \mathbb{R}^{C_f}$ represents the channel attention weights. FC_{σ} refers to a fully-connected (fc) layer with a Sigmoid activation function, while FC_{ζ} denotes a fc layer with leaky ReLU function. GAP stands for global average pooling, and \otimes represents element-wise multiplication.

3.3. Classifier

Taking $\mathbf{E}_{ca} \in \mathbb{R}^{L/4 \times C_f}$ as input, the classifier integrates a global average pooling layer to condense the feature maps, and a fc layer with 2 units utilizing Softmax activation for predicting attended sound source. Training is performed iteratively using the binary cross-entropy loss function and the Adam optimizer:

$$\mathcal{L} = -\frac{1}{L} \sum_{l=1}^L y_l \cdot \log p_l + (1 - y_l) \cdot \log(1 - p_l) \quad (3)$$

where y_l represents the true label and p_l the predicted probability for the l -th decision window.

4. Experiments

4.1. Data preprocessing

The EEG data were preprocessed using MNE-Python[17]. Initially, a high-pass FIR filter with a 1 Hz cutoff frequency was utilized to eliminate electrode drift. Subsequently, the EEG data were re-referenced to the average of all scalp channels. Following previous AAD studies [16, 2], we used an FIR low-pass filter to bandpass the EEG data to a frequency range of 1-50

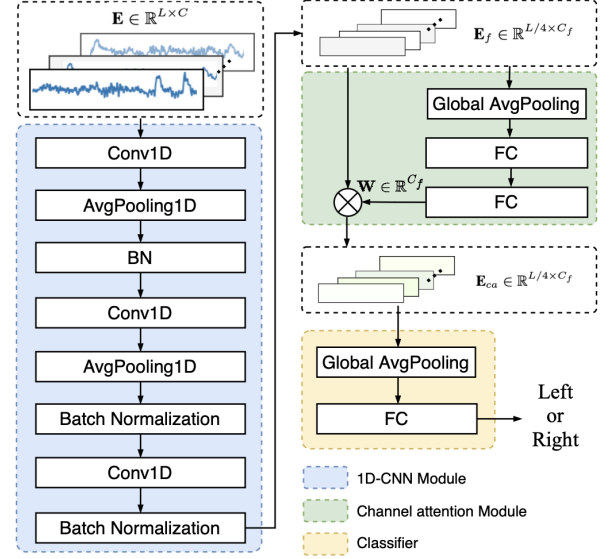


Figure 1: The architecture of the proposed channel attention convolutional neural network, i.e. CA-CNN. Conv1D= 1D convolution, BN=batch normalization, and FC= fully-connected

Hz. The data were then resampled to a frequency of 128 Hz. For analysis, we utilized a sliding decision window with a 50% overlap rate to segment the EEG data, denoted as a “decision window.” Considering the approximate 1-second lag in human attention shifting, the decision window length was set to 1 second in this study. Normalization was applied to each decision window to minimize trial-related biases.

4.2. Training and evaluation

We evaluated the proposed models under subject-independent conditions using a 5-fold cross-validation approach [9]. The average results across all subjects were reported as the model’s performance. Moreover, to prevent potential biased outcomes resulting from data leakage, we ensure that EEG responses from different subjects to the same speech stimulus were not present in both the training and testing datasets during data splitting. As illustrated in Fig. 2, the EEG data from each trial were segmented into five consecutive folds. Each fold comprised EEG data organized in a continuous time-series across all subjects. Notably, we also discarded potential overlaps, referred to as “repeated segments,” where the tail end of a training window might overlap with the test set.

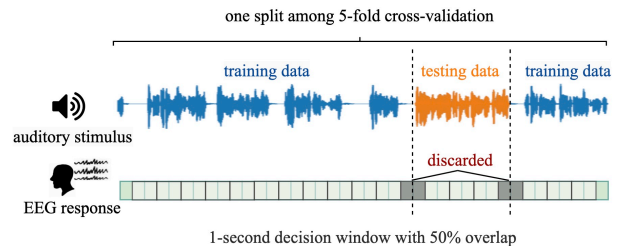


Figure 2: The procedure of splitting EEG data

The model implementation utilized TensorFlow version 2.13.0, employing the Adam optimizer with a learning rate of

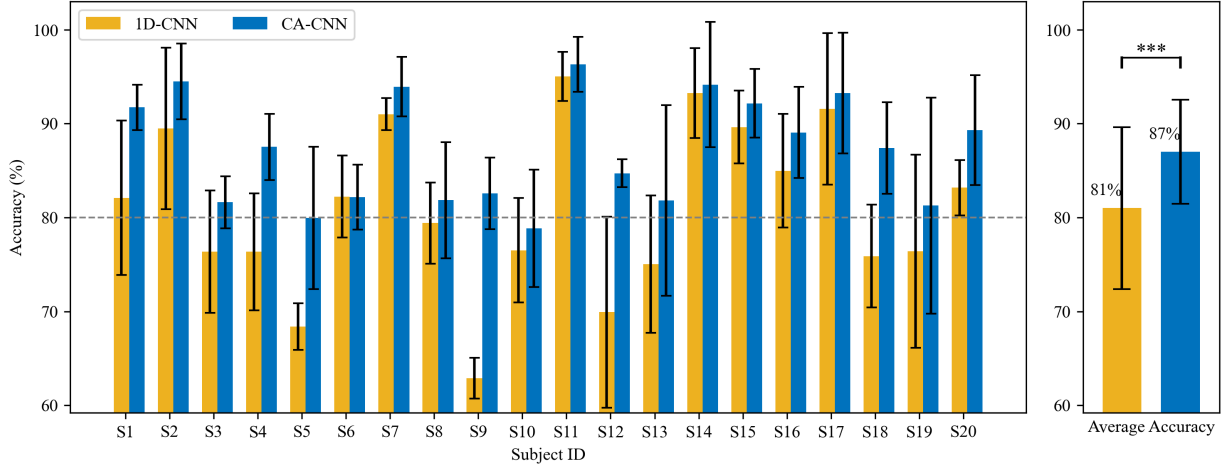


Figure 3: ASAD Accuracy of our proposed 1D-CNN and CA-CNN models across all subjects in the ASA data set, respectively. *** $p < 0.001$ (paired t -test)

5×10^{-4} . To prevent overfitting, early stopping was applied, monitoring accuracy with a minimum required improvement of 0.01 to sustain training. The patience parameter was set to 8 epochs, and an automatic mode selection was employed. The training process was limited to a maximum of 100 epochs.

5. Results and Discussion

5.1. Effect of channel attention

To assess the effectiveness of channel attention, we conduct an ablation analysis, comparing the performance of 1D-CNN with CA-CNN. Specifically, 1D-CNN refers to CA-CNN without the channel attention module. As shown in Fig. 3, the CA-CNN not only achieves a superior average accuracy of 87.20% but also exhibits enhanced robustness across subjects, as indicated by a lower standard deviation (SD) of 5.56% and a competitive minimum accuracy of 79.29%. In contrast, the 1D-CNN shows an average accuracy of 80.97%, accompanied by a higher SD of 8.62% and a less competitive minimum accuracy of 69.55%. The findings underscore the crucial role of channel attention for a more robust and discriminative representation of EEG for ASAD tasks, showcasing its potential to enhance the model’s sensitivity to individual auditory attention patterns.

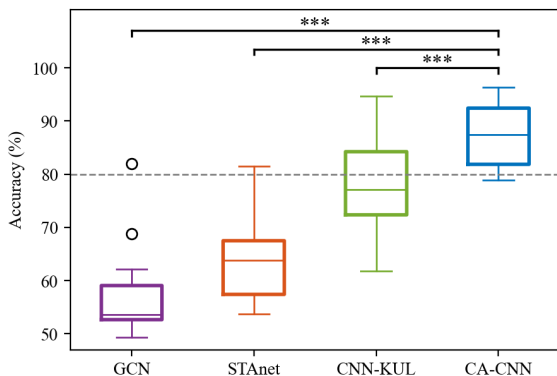


Figure 4: ASAD accuracy of the proposed CA-CNN and other comparative models. *** $p < 0.001$ (paired t -test)

5.2. Comparative analysis

We conducted a comprehensive evaluation of various models, including GCN [12], STAnet [11], the CNN-KUL model [10], and our proposed CA-CNN, assessing the ASAD accuracy under the subject-independent scenario in ASA dataset. Each model is comparable in size, with parameter counts as follows: GCN with 3510, STAnet with 3929 parameters, CNN-KUL with 5487, and CA-CNN with 6834. This ensures a fair comparison of their ASAD performance in our study.

As depicted in Fig. 4, GCN achieves an average accuracy of 56.59% (SD: 7.30%), STAnet exhibits an average accuracy of 63.63% (SD: 7.17%), CNN-KUL attains an average accuracy of 77.76% (SD: 9.13%), and CA-CNN significantly outperforms all models (paired t -test, $p < 0.001$), achieving a remarkable accuracy of 87.20% (SD: 5.56%). Notably, GCN, STAnet, and CNN-KUL were initially evaluated on scenarios with only two competing speaker directions. The performance decline observed in all models under this more challenging experimental setup emphasizes the robustness of our CA-CNN model. It also underscores the necessity for additional ASAD datasets that encompass more realistic scenarios, as the existing models experience limitations in such conditions. The incorporation of expanded datasets is crucial to fully exploit the practical application potential of DNNs.

6. Conclusion

We introduced a new ASA dataset, encompassing auditory stimuli featuring spatial variations across multiple angles ranging from -90° to 90° . This dataset serves as a valuable complement to existing data repositories in the field of EEG-enabled ASAD tasks. Additionally, we have developed a reference model, i.e. CA-CNN, for detecting auditory spatial attention. Through comparative analysis and experiments, we show that CA-CNN consistently outperforms the state-of-the-art models, particularly under subject-dependent conditions. The ASA database and the source code of the reference model are to be publicly released for research purposes.

7. Acknowledgement

This work is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (University Allowance, EXC 2077, University of Bremen, Germany). The research is also supported by National Natural Science Foundation of China (Grant No. 62271432); Internal Project of Shenzhen Research Institute of Big Data (Grant No. T00120220002).

We thank all the individuals who took part in these experiments, and extend our thanks to Xinyi Chen and Duo Ma for their valuable assistance in collecting the data.

8. References

- [1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *Journal of the Acoustical Society of America*, vol. 25, pp. 975–979, 1953.
- [2] S. Cai, H. Zhu, T. Schultz, and H. Li, "EEG-based Auditory Attention Detection in Cocktail Party Environment," *APSIPA Transactions on Signal and Information Processing*, vol. 12, no. 3, p. e22, 2023.
- [3] J. A. O'Sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, and E. C. Lalor, "Attentional selection in a cocktail party environment can be decoded from single-trial EEG," *Cerebral Cortex*, vol. 25, no. 7, pp. 1697–1706, 2015.
- [4] S. Cai, P. Li, E. Su, Q. Liu, and L. Xie, "A neural-inspired architecture for EEG-based auditory attention detection," *IEEE Transactions on Human-Machine Systems*, vol. 52, no. 4, pp. 668–676, 2022.
- [5] S. Geirnaert, S. Vandecappelle, E. Alickovic, A. de Cheveigné, E. C. Lalor, B. T. Meyer, S. Miran, T. Francart, and A. Bertrand, "Electroencephalography-based auditory attention decoding: Toward neurosteered hearing devices," *IEEE Signal Processing Magazine*, vol. 38, pp. 89–102, 2020.
- [6] S. A. Fuglsang, T. Dau, and J. Hjortkjær, "Noise-robust cortical tracking of attended speech in real-world acoustic scenes," *NeuroImage*, vol. 156, pp. 435–444, 2017.
- [7] N. Das, W. Biesmans, A. Bertrand, and T. Francart, "The effect of head-related filtering and ear-specific decoding bias on auditory attention detection," *Journal of Neural Engineering*, vol. 13, no. 5, p. 056014, 2016.
- [8] P. Li *et al.*, "ESAA: An EEG-Speech Auditory Attention Detection Database," in *2022 25th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*. IEEE, 2022.
- [9] Y. Zhang, H. Ruan, Z. Yuan, H. Du, X. Gao, and J. 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)*, 2023, Conference Proceedings, pp. 1–5.
- [10] S. Vandecappelle, L. Deckers, N. Das, A. H. Ansari, A. Bertrand, and T. Francart, "EEG-based detection of the locus of auditory attention with convolutional neural networks," *Elife*, vol. 10, p. e56481, 2021.
- [11] E. Su, S. Cai, L. Xie, H. Li, and T. Schultz, "STAnet: a spatiotemporal attention network for decoding auditory spatial attention from EEG," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 7, pp. 2233–2242, 2022.
- [12] R. Wang, S. Cai, and H. Li, "EEG-based auditory attention detection with spatiotemporal graph and graph convolutional network," *INTERSPEECH 2023*, 2023.
- [13] Chinese Proficiency Test, "HSK," <http://www.chinesetest.cn/godownload.do>, 2020, accessed: 25 June 2022.
- [14] F. Mattioli, C. Porcaro, and G. Baldassarre, "A 1D CNN for high accuracy classification and transfer learning in motor imagery EEG-based brain-computer interface," *Journal of Neural Engineering*, vol. 18, no. 6, p. 066053, 2022.
- [15] F. Hassan, S. F. Hussain, and S. M. Qaisar, "Epileptic seizure detection using a hybrid 1D CNN-machine learning approach from EEG data," *Journal of Healthcare Engineering*, 2022.
- [16] S. Cai, E. Su, L. Xie, and H. Li, "EEG-Based Auditory Attention Detection via Frequency and Channel Neural Attention," *IEEE Transactions on Human-Machine Systems*, vol. PP, pp. 1–11, 2021.
- [17] A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen, and M. S. Hämäläinen, "MEG and EEG data analysis with MNE-Python," *Frontiers in Neuroscience*, vol. 7, no. 267, pp. 1–13, 2013.