STAnet: A Spatiotemporal Attention Network for Decoding Auditory Spatial Attention From EEG

Enze Su, Siqi Cai, *Member, IEEE*, Longhan Xie, *Member, IEEE*, Haizhou Li, *Fellow, IEEE*, and Tanja Schultz, *Fellow, IEEE*

Abstract*—*Objective: **Humans are able to localize the source of a sound. This enables them to direct attention to a particular speaker in a cocktail party. Psycho-acoustic studies show that the sensory cortices of the human brain respond to the location of sound sources differently, and the auditory attention itself is a dynamic and temporally based brain activity. In this work, we seek to build a computational model which uses both spatial and temporal information manifested in EEG signals for auditory spatial attention detection (ASAD).** ***Methods:*** **We propose an end-to-end spatiotemporal attention network, denoted as STAnet, to detect auditory spatial attention from EEG. The STAnet is designed to assign differentiated weights dynamically to EEG channels through a spatial attention mechanism, and to temporal patterns in EEG signals through a temporal attention mechanism.** ***Results:*** **We report the ASAD experiments on two publicly available datasets. The STAnet outperforms other competitive models by a large margin under various experimental conditions. Its attention decision for 1-second decision window outperforms that of the state-of-the-art techniques for 10-second decision window. Experimental results also demonstrate that the STAnet achieves competitive performance on EEG signals ranging from 64 to as few as 16 channels.** ***Conclusion:*** **This study provides evidence suggesting that efficient low-density EEG online decoding is within reach.** ***Significance:*** **This study also marks an important step towards the practical implementation of ASAD in real life applications.**

***Index Terms*—Auditory attention, brain-computer interface, electroencephalography, spatial attention, temporal attention.**

Manuscript received September 1, 2021; revised November 30, 2021; accepted December 28, 2021. Date of publication January 4, 2022; date of current version June 20, 2022. The work of Haizhou Li was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (University Allowance, EXC 2077, University of Bremen, Germany), and the work of Longhan Xie was supported by the National Natural Science Foundation of China under Grant 52075177. This work was supported in part by the Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Funding Scheme under Project A18A2b0046, and in part by A*STAR under its RIE 2020 Advanced Manufacturing and Engineering Human (AME) Programmatic Grant under Grant A1687b0033. *(Corresponding author: Longhan Xie.)*

Enze Su is with the Shien-Ming Wu School of Intelligent Engineering, South China University of Technology, China.

Siqi Cai is with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore.

Longhan Xie is with the Shien-Ming Wu School of Intelligent Engineering, South China University of Technology, Guangzhou, Guangdong 510460, China (e-mail: melhxie@scut.edu.cn).

Haizhou Li is with the School of Data Science, The Chinese University of Hong Kong (Shenzhen), China, and with the Machine Listening Lab, University of Bremen, Germany, and also with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore.

Tanja Schultz is with the Cognitive Systems Lab, University of Bremen, Germany.

Digital Object Identifier 10.1109/TBME.2022.3140246

I. Introduction

HUMANS have the ability to focus the auditory attention on one speaker in a multi-speaker environment, or "cocktail party scenario" [1]. However, people with hearing loss will find such situations are particularly difficult. Modern hearing aids are developed for a better experience by applying noise suppression, however, these devices often fail in practice for unable to single out and enhance the attended speech stream. The studies in neuroscience show that auditory attention can be directly detected from neural activities [2]–[4], which is known as auditory attention detection (AAD). Such progress motivates us to develop engineering solutions to AAD, that in turn opens up many possibilities for the cognitive control of hearing aids, also called neuro-steered hearing aids [5], [6].

To develop a neurophysiologically plausible brain-computer interface (BCI), many studies have been devoted to discovering the relationship between neural responses and speech stimuli for AAD. Mesgarani and Chang [2] have demonstrated that speech spectrograms reconstructed from cortical responses to a mixture of speakers are dominated by the salient spectro-temporal features of the attended speaker. Along this line of thought, a stimulus-reconstruction method is studied, where neural responses are used to approximate the envelope of the speech stream heard by the subject. The reconstructed envelope is then compared with the original speech stimulus to detect the speaker's attention [4], [7]–[11]. Unfortunately, such correlation between the reconstructed and the attended speech envelopes is generally weak. This could be due to the over-simplified linear computational model. Considering the inherent non-linear processing of acoustic signals along the auditory pathway [12], [13], Taillez *et al.* [14] firstly studied a non-linear neural network to map the EEG signals to speech envelopes in a cocktail party scenario, that outperforms the linear model baseline. Recently convolutional neural network (CNN) models [15], [16] were studied to detect the attended speakers directly when both the EEG signals and speech stimuli are available.

The early studies of engineering solutions to AAD confronts two challenges. 1) There is a trade-off between the accuracy and

the AAD decision window size. The decoding accuracy drops as the decision window narrows, i.e., temporal resolution increases, because the low-frequency envelopes of small window contain little information of speech [17], [18]. 2) The AAD methods require clean speech stimuli as the reference, which are not always available in real-world applications, such as hearing prostheses or robotic voice acquisition, where a system is expected to perform in a complex acoustic environment of multiple speakers. While speaker extraction and speech enhancement techniques can be explored to derive such speech stimuli [5], [19], [20], they add overhead to the AAD system, and their quality may impact AAD accuracy, that becomes another issue [17].

Inspired by the findings that the locus of the auditory attention is neurally encoded [21]–[23], auditory spatial attention detection (ASAD) from EEG signals has been studied recently [17], [24]–[27], where the spatial location of the attended speaker is decoded from brain activities alone, without the need of clean speech stimuli as the reference. This is highly desirable for neuro-steered hearing aids as clean speech stimuli are not always available. Moreover, the ASAD approach is based on brain lateralization [28], which is an instantaneous feature, as opposed to the low-frequency speech envelope, which requires a temporal observation window of reasonable size. We hypothesize that the ASAD approach will perform more accurately than the AAD approach in low-latency settings. Vandecappelle *et al.* [24] developed a CNN-based ASAD model, which achieves a competitive accuracy of around 80% for a decision window of 1 s. Unfortunately, as the EEG signals are reduced from a time series to a single value in this approach, the temporal information is not exploited. We consider that the dynamics of the EEG signals contain valuable information for decoding auditory spatial attention [29], that will be a focus of this paper.

Effective feature representation is a crucial step for pattern classification due to the low signal-to-noise ratio of raw EEG signals [30]–[33]. The EEG signals contain multivariate information in space and time. In space, the EEG channels reflect different functional roles of human brain in speech processing [34], [35], therefore the EEG signals are essentially non-linear time series data [36]. We consider that the EEG channels provide differentiated contributions to the encoding of spatial attention in human brain; in time, we hope to leverage the information encoded in the temporal progression of EEG signals. An early study by Bednar *et al.* [29] shows that the spatiotemporal pattern of the EEG features is critical for successfully decoding different spatial locations. In this paper, we propose a neural attention mechanism that is inspired by the findings in the spatiotemporal analysis of human AAD, and implement an engineering solution for the first time.

First, previous studies show that the human responses to speech stimuli differ in different brain regions in a cocktail party task [21]–[23], [37]–[39]. The EEG signals are recorded from multiple sites of the scalp, therefore, some EEG channels are more informative than others in terms of informing the decision-making process in the brain [4], [17], [34], [35]. At the same time, the distribution of effective channels may vary from subject to subject [7], [28]. To extract discriminative features from the spatial information, some employ channel selection techniques to choose more relevant channels [40], [41]. Unlike the channel selection techniques, in this study, we propose a spatial attention mechanism, that derives weights dynamically from the input EEG channels across different spatial locations over the cortex, just like how human brains selectively attend to input acoustic stimuli.

Second, any attentional neural mechanism that respects the temporal structure of auditory sensory inputs must therefore be dynamic over time, including spatial attention [22]. In view of the fact that brain activity is a temporal process, and the EEG signals are essentially non-linear time series data [36], we propose a temporal attention mechanism to capture the temporal characteristics of EEG. The temporal attention mechanism is designed to assign differentiated weights temporally to EEG signals over a decision window to form a final representation [42].

In this paper, we propose an end-to-end spatiotemporal attention network, denoted as STAnet, to detect auditory spatial attention solely based on EEG signals. To the best of our knowledge, this is the first study of a spatiotemporal attention mechanism for EEG-based ASAD tasks. The main contributions of this work can be summarized as follows. (1) We propose a spatial attention mechanism to automatically assign differentiated modulation weights to EEG channels, and a temporal attention mechanism to learn temporal feature representation that is relevant to ASAD. (2) We propose an end-to-end pipeline architecture that optimizes spatial and temporal representation explicitly and in a logical order. (3) We validate the effectiveness of the spatiotemporal attention mechanism through extensive ablation study, data visualization, and comprehensive experiments on two publicly available EEG datasets.

The remainder of this paper is organized as follows. Section II elaborates on the proposed STAnet pipeline for decoding auditory spatial attention. In Section III, we describe: 1) the used datasets and processing; as well as 2) contrastive models and their application to the datasets. In Section IV, we report on experimental results and compare our proposed approach to competing ASAD models. In Section V, we discuss our findings and conclude in Section VI.

II. Methods

A traditional EEG-based ASAD pipeline consists of a feature extraction frontend and a pattern classification backend. The STAnet is a novel end-to-end architecture, as illustrated in Fig. 1, that features a spatial attention and temporal attention mechanism. It is different from a traditional pipeline in many ways. The end-to-end architecture learns to automatically discover spatial and temporal representations needed for attention detection from raw EEG data, therefore, without the need of hand-crafted feature extraction. The attention mechanism learns to dynamically pay attention to specific channels and temporal patterns during run-time inference.

By applying a moving window to raw EEG data, we obtain a sequence of *decision windows*, each of which has a small duration, and is used for feature representation. Let $\mathbf{E} = [c_1, \ldots, c_i, \ldots, c_N] \in \mathbb{R}^{T \times N}$ be the EEG signals of a decision window, where $c_i \in \mathbb{R}^{T \times 1}$ is a time series of T samples from the

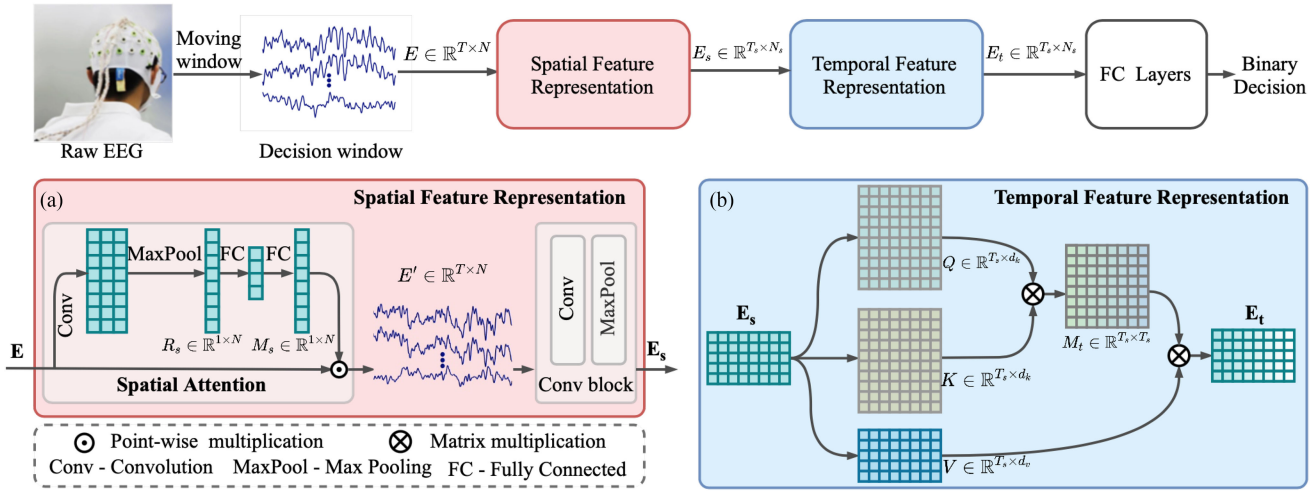

Fig. 1. A schematic diagram of the proposed spatiotemporal attention network, i.e., STAnet, which mainly consists of three components: spatial feature representation, temporal feature representation, and classification module. Taking EEG data as input, the network is trained to detect auditory spatial attention by making a binary decision.

i-th EEG channel, and N is the number of the EEG channels. The STAnet takes $\mathbf{E}$ as the input and makes auditory spatial attention decision.

Attention, as a human ability, plays a fundamental role in our everyday behavior in spatially and temporally fast-varying environments [22]. Auditory attention, which enhances the target speech and attenuates the interfering speech in a cocktail party, is a typical example [1], [2]. Inspired by the human cognitive process, computational attention mechanisms are employed widely in deep learning architecture [42]–[44]. Briefly, the attention mechanism can be interpreted as a means of dynamically assigning differentiated weights to the components of a signal at run-time. These differentiated weights form a receptive field, which can be in the form of an attention mask, or a regression function. The differentiated weights are dynamically generated by a neural attention mechanism, as opposed to a set of pre-trained weights. With the neural attention mechanism, we hope to extract salient information from EEG signals with respect to the ASAD task.

We employ feature representation not only for dimension reduction, but also to explicitly capture salient spatial and temporal information. The spatial and temporal attention mechanisms, as shown in Fig. 1(a) and (b), learn "where to attend" and "when to attend" to the EEG signals in a decision window. As the temporal attention mechanism is expected to explore the interaction among the EEG signals across channels. It is logical to place spatial attention module in front of temporal attention module in a pipeline architecture.

A. Spatial Feature Representation

Studies show that several cortical regions of human brain are involved in spatial auditory processing [28], [29], [38]. The multi-channel EEG recorded from different scalp regions reflect the brain responses to auditory stimulus [7], [40], [41]. We are motivated by this finding to design an attention mechanism, that learns to assign differentiated weights to EEG channels dynamically according to their individual contributions. Such attention mechanism is referred to as spatial attention. It is implemented in three steps as illustrated in Fig. 1(a),

First, we aggregate channel-wise EEG signals c_i through a convolutional layer followed by a max pooling layer. This step is equivalent to a frequency analysis front-end that performs feature extraction from the time-domain signals c_i as described next,

$$r_i = Max(elu(Conv(c_i))) \tag{1}$$

where $Conv(\cdot)$ denotes the convolution operation with an exponential linear unit $elu(\cdot)$ as the activation function [45]. $Max(\cdot)$ denotes a max pooling layer. $r_i \in \mathbb{R}^{1\times 1}$ is the representation of the i-th channel c_i and therefore $R_s = [r_1, \ldots, r_i, \ldots, r_N] \in \mathbb{R}^{1\times N}$ is the representation of the multi-channel EEG signals.

Second, a gating mechanism, which models the interaction among the EEG channels, is adopted [43]. The gating mechanism learns to assign differentiated weights to EEG channels on the fly. As a trade-off between model complexity and generalizability, two fully-connected (fc) layers are employed to parameterize the gating mechanism, and to achieve a non-linear mapping, as follows:

$$M_s = elu(\mathbf{w}_2(elu(\mathbf{w}_1 R_s + \mathbf{b}_1)) + \mathbf{b}_2) \tag{2}$$

where $\mathbf{w}_1$ and $\mathbf{w}_2$ is the parameter of the first and the second fc layers, respectively. $\mathbf{b}_1, \mathbf{b}_2$ are the bias terms of two fc layers. M_s is the attention mask generated by the spatial attention mechanism. The EEG signals $\mathbf{E}$ is then modulated by the attention mask M_s as follows,

$$\mathbf{E}' = M_s \bigotimes \mathbf{E} \tag{3}$$

where $\bigotimes$ denotes a point-wise multiplication. During the multiplication, the attention value is broadcast along the temporal dimension, i.e., $M_s \in \mathbb{R}^{T\times N}$.

Finally, we employ a convolutional layer followed by a max pooling layer, that is referred to as the *Conv* block illustrated

in Fig. 1, to extract the spatial feature representation from the masked EEG signals $\mathbf{E}'$. Without spatial attention, the Conv block was first studied for ASAD in [24]. We believe that the same Conv block would work well for the masked EEG signals as the masked EEG signals carry the similar properties of original EEG signals except that the signals are weighted channel by channel. The height of the convolutional filter is set to N, the same as the number of EEG channels, and the width of the filter is extended to better explore the temporal dynamics. In this way, the masked EEG signals $\mathbf{E}'$ are encoded as $\mathbf{E}'' = Conv(\mathbf{E}')$. We adopt $tanh$ function as the activation function in the convolutional layer, and apply max pooling to reduce the number of parameters. The spatial feature representation can be expressed as $\mathbf{E_s} = Max(\mathbf{E}'') \in \mathbb{R}^{T_s \times N_s}$.

B. Temporal Feature Representation

Psycho-acoustic studies have provided convincing evidence that human attention itself is a dynamic and temporally based activity [46], [47], and the auditory system is sensitive to the temporal patterning [38], [48]. We believe that temporal patterns in EEG signals carry spatial attention information. As shown in Fig. 1(b), a self-attention mechanism is adopted to explore the attentive temporal dynamics of EEG signals. Self-attention is an intra-attention mechanism that relates different positions of a single sequence to generate a representation of the sequence [42]. It is implemented in three steps as follows.

First, the attention mechanism transforms EEG features $\mathbf{E_s}$ into query Q, key K, and value V using linear projections.

$$\begin{aligned} Q &= E_s W_q \\ K &= E_s W_k \\ V &= E_s W_v \end{aligned} \tag{4}$$

Here, the projections are the weight matrices $W_q \in \mathbb{R}^{N_s \times d_k}$, $W_k \in \mathbb{R}^{N_s \times d_k}$, and $W_v \in \mathbb{R}^{N_s \times d_v}$.

Then, dot product is used to calculate the relationship between query and key. And temporal attention mask M_t is calculated by using the $softmax$ function.

$$M_t = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \in \mathbb{R}^{T_s \times T_s} \tag{5}$$

where $\sqrt{d_k}$ is the scaling factor. The inner product values are proportional to the dimension of hidden feature space, thus need to be normalized by the square root of hidden dimension [42].

Finally, the temporal attention mask assigns different weights over the time axis to an EEG sequence, that leads to an attention-weighted summation $\mathbf{E_t}$,

$$\mathbf{E_t} = M_t V \in \mathbb{R}^{T_s \times N_s} \tag{6}$$

C. End-to-End Spatiotemporal Attention Network

An end-to-end neural architecture takes a window of EEG signals as input and makes spatial attention detection decision. It allows the spatial and temporal attention mechanisms to learn the respective feature representations in a way to optimize ASAD performance.

First, we adopt a spatial attention mechanism to dynamically assign differentiated weights to individual EEG channels. The masked EEG signals are processed by a convolutional layer and a max pooling layer to derive spatial feature representation.

Second, we adopt a self-attention mechanism to assign differentiated weights to the EEG signals temporally. In this way, we expect to generate a discriminative spatiotemporal EEG representation $\mathbf{E_t}$ that is optimized for the ASAD task.

Similar to state-of-the-art ASAD approaches [17], [24], we treat the ASAD task as a classification problem. The extracted EEG feature $\mathbf{E_t}$ is then transformed into a probability vector $\mathbf{E'_t}$ by a fc layer with $sigmoid$ activation function.

$$p = \sigma(\mathbf{w}\mathbf{E'_t} + \mathbf{b}) \tag{7}$$

where $\mathbf{w}$ and $\mathbf{b}$ is the weight and the bias of the fc layer, respectively. p represents the predicted probability for a decision window. $\sigma(\cdot)$ denotes the $sigmoid$ activation function. Finally, we apply the binary cross-entropy loss to supervise the network training.

III. Experiments

A. Data Specifications

We conduct the auditory attention detection experiments on two publicly available datasets, which are denoted as KUL [49] and DTU datasets [50]. Details of the datasets are summarized in Table I.

1) KUL dataset: The EEG data were collected from 16 normal-hearing subjects, who were instructed to attend to one particular speaker and ignore the other in the presence of two simultaneous speakers. The speech stimuli consist of four Dutch stories, narrated by three male Flemish speakers. All stimuli were normalized to have the same root mean squared (RMS) intensities and were perceived as equally loud. The stimuli were either presented dichotically (one speaker per ear) or after head-related transfer function (HRTF) filtering to simulate speech from 90° to the left and 90° to the right of the subject. Each subject listened to 8 trials of 6 minutes each. Throughout the experiments, the order of presentation of both conditions was randomized over the different subjects. The 64-channel EEG data were collected using a BioSemi ActiveTwo device at a sampling rate of 8,192 Hz in an electromagnetically shielded and soundproof room. More details of the KUL dataset can be found in [24], [49].

2) DTU dataset: The EEG data were collected from 18 normal-hearing subjects, who selectively attended to one of the two simultaneous speakers. The speech stimuli consist of speech by a male and a female native speaker who simultaneously speak in anechoic or reverberant rooms. The two concurrent speech streams were normalized to have the same RMS value. The speech mixtures were presented to the subjects, with the two speech streams lateralized at respectively −60° and +60° along the azimuth direction. Each subject listened to 60 trials in total, and each trial contained auditory stimuli with a duration of 50 seconds. The position and the gender of the target speaker were randomized across the trials. 64-channel EEG data were recorded at a sample rate of 512 Hz using a BioSemi Active

TABLE I
DETAILS OF TWO EEG DATASETS, KUL AND DTU, USED IN THE EXPERIMENTS

Dataset	# subjects	Language	Spatial locus of the stimuli	Duration per subject (minutes)	Total duration (hours)
KUL [49]	16	Flemish	90° to the left and 90° to the right	48	12.8
DTU [50]	18	Danish	60° to the left and 60° to the right	50	15.0

TABLE II
THE PROPOSED STANET AND THREE CONSTRASTIVE MODELS IN THE EXPERIMENTS

Model	Spatial Attention	Temporal Attention
CNN [24]	×	×
SAnet	✓	×
TAnet	×	✓
STAnet	✓	✓

system. Further details of the DTU dataset can be found in [50], [51].

B. Data Processing

The EEG data of each channel are firstly re-referenced to the average response of all electrodes. As the EEG signals analyzed are collected at different sampling frequencies, they are all bandpass filtered between 1 and 32 Hz by a 6th-order Chebyshev Type II bandpass filter, and downsampled to 128 Hz sampling rate. The frequency range is chosen based on previous non-linear AAD studies [14]–[16], [24]. Finally, EEG data channels are normalized to ensure zero mean and unit variance for each trial. As the proposed STAnet is a data-driven solution, that is expected to function in an end-to-end manner, no artifacts removal operation is involved in the data processing. The simplified end-to-end process greatly facilitates the implementation of real-time BCI systems, such as neuro-steered hearing aids.

We analyze seven decision window sizes in this study, i.e., 0.1, 0.2, 0.5, 1, 2, 5, and 10 seconds. After preprocessing, we obtain a total of 2,864 decision windows per subject, totaling 45,824 decision windows for 1-second case in the KUL dataset. In the DTU dataset, the test set results in 2,880 decision windows per subject, totaling 51,840 decision windows for 1-second case.

C. Contrastive Models

To validate the effectiveness of the spatial attention and temporal attention mechanisms, we conduct experiments on four models, that include 1) a CNN model as in [24], that is a reduced version of STAnet (see Fig. 1) by only keeping the Conv block, and FC layers in the pipeline in Fig. 1, and removing the spatial attention mechanism and the temporal feature representation module; 2) a SAnet by removing the temporal feature representation module from the STAnet; 3) a TAnet by removing the spatial attention mechanism in spatial feature representation module from the STAnet; and 4) the proposed STAnet. The composition of the models are summarized in Table II.

D. Network Configuration

We evaluate the performance of the proposed and baseline methods using 5-fold cross-validation (CV) over the collection of decision windows. We make a decision for each decision window and report the overall ASAD accuracy and by subject. For each subject, the feature data and attention labels are divided into five groups equally. Four of them are used to train the classifiers. The remaining group is used for evaluation. This process is repeated five times until all data are tested once. The models are implemented with the TensorFlow framework and trained on an NVIDIA TITAN Xp Pascal GPU.

We take 1-second decision window as an example to describe the network configuration. As the input to the systems, 1-second EEG signals are denoted as $\mathbf{E} \in \mathbb{R}^{128\times64}$ with 128 samples and 64-channel.

In the SAnet or STAnet, the spatial attention mechanism comprises a convolution layer with the size of 128×1 and two fc layers (input: 64, hidden: 8, output: 64). The output of the spatial attention mechanism is $\mathbf{E}' \in \mathbb{R}^{128\times64}$, that has the same dimension as the input EEG signals $\mathbf{E}$.

In all models, we employ a convolutional layer with a kernel of 5×64, and a max pooling size of 4×1, as shown in the Conv block in Fig. 1. With or without the spatial attention mechanism, the Conv block produces an EEG feature representation $\mathbf{E_s} \in \mathbb{R}^{32\times5}$.

The TAnet or STAnet involves a temporal attention mechanism. We set both Q and K to 32×50, and V to be of the same size as the input $\mathbf{E_s}$. Therefore, the size of the temporal mask is $M_t \in \mathbb{R}^{32\times32}$ and the output is $\mathbf{E_t} \in \mathbb{R}^{32\times5}$.

We set both Q and K to 32×50, and V the same as the input $\mathbf{E_s}$. Therefore, the size of the temporal mask is $M_t \in \mathbb{R}^{32\times32}$ and the output is $\mathbf{E_t} \in \mathbb{R}^{32\times5}$.

For classification decision, we employ a fc layer of 160×2. To prevent overfitting, we apply dropout and batch normalization. The Adaptive Moment Estimation (Adam) optimizer [52] is employed to minimize the cross-entropy loss function with the learning rate of 10^{-3}. All hyperparameters are chosen empirically with a grid search approach through a 5-fold CV experiment.

IV. RESULTS

A. Ablation Analysis

We conduct ablation analysis using 1-second decision window as a case study. The ASAD accuracy of the four models in Table II are reported across all subjects on KUL dataset in Fig. 2. For 1-second decision window, CNN model attains ASAD accuracy of 84.1%, with a standard deviation (SD) of 10.16%. SAnet outperforms CNN model with an average improvement of 3.8% (87.9%, SD: 10.10%). Similarly, TAnet outperforms CNN model with an average improvement of 4.2% (88.3%, SD: 9.39%). The proposed STAnet outperforms all others, with an average accuracy of 90.1% (SD: 8.95%).

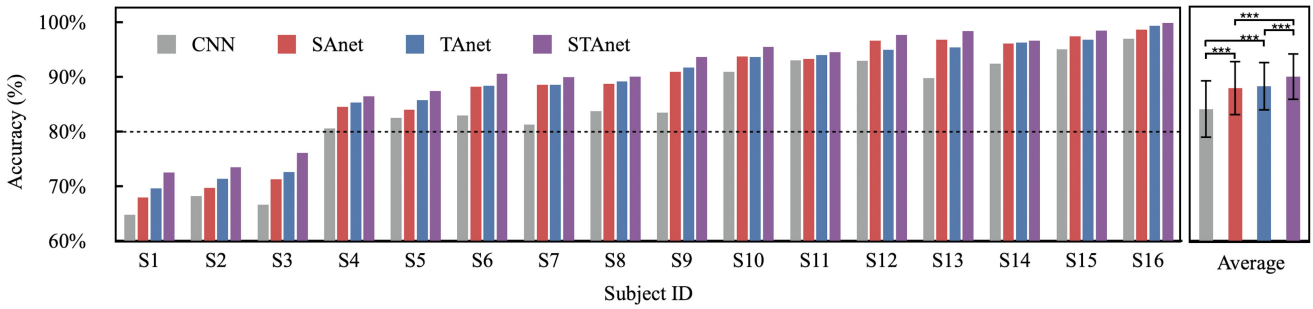

Fig. 2. Accuracy of baseline model and attention-based models for decoding auditory spatial attention among all subjects in KUL dataset with 1-second decision window. These subjects are ranked according to the accuracy of the STAnet. The horizontal dotted line shows a reference accuracy at 80%. Statistically significant difference: $^{***}p < 0.001$.

Statistical analyses are performed using IBM SPSS statistics software at a significance level of 0.05. Descriptive statistics are used for means and standard deviations. The Kolmogorov-Smirnov test is used to confirm the normality of the distribution of the data, prior to selection of appropriate statistical tests. Paired *t*-tests are employed to compare the models to identify which one gives a significant improvement. We observe that SAnet attains a significantly higher average ASAD accuracy than CNN model ($p < 0.001$), and that TAnet attains a significantly higher performance than CNN model ($p < 0.001$). There is no statistically significant difference of ASAD accuracy between SAnet and TAnet ($p = 0.15$). Moreover, STAnet gains a significant increase of ASAD accuracy over SAnet (2.2%, $p < 0.001$) and TAnet (1.8%, $p < 0.001$).

The results on the DTU Dataset corroborate the findings on the KUL Dataset. Specifically, CNN model attains ASAD accuracy of 63.3% (SD: 5.96%). SAnet outperforms CNN model with an average improvement of 4.3% (67.6%, SD: 8.96%), and TAnet outperforms CNN model with an average improvement of 5.1% (68.4%, SD: 8.98%). STAnet achieves the best ASAD performance with an average accuracy of 71.9% (SD: 8.94%).

In sum, the spatial attention mechanism and the temporal attention mechanism are the contributing factors to the performance gains over the baseline CNN model. The fact that STAnet outperforms all competing models clearly confirms the advantage of the proposed spatiotemproal attention.

B. Effect of Decision Windows

We further compare the ASAD performance of the STAnet for different detection window sizes ranging from 0.1-second to 10-second, as shown in Fig. 3.

On the KUL dataset, the STAnet achieves an average decoding accuracy across all subjects of 90.1% (SD: 8.95%) for 1-second decision window, 91.4% (SD: 8.22%) for 2-second decision window, 92.6% (SD: 6.75%) for 5-second decision window, and 93.9% (SD: 6.54%) for 10-second decision window. In general, a larger decision window provides a better result, which corroborates with findings in previous studies [14], [15], [24]. It is worth noting that the STAnet is capable of decoding auditory spatial attention with a very short decision window (<1 s). For 0.5-second and 0.2-second decision windows, the STAnet obtains a high ASAD accuracy of 87.2% (SD: 9.77%) and 84.3%

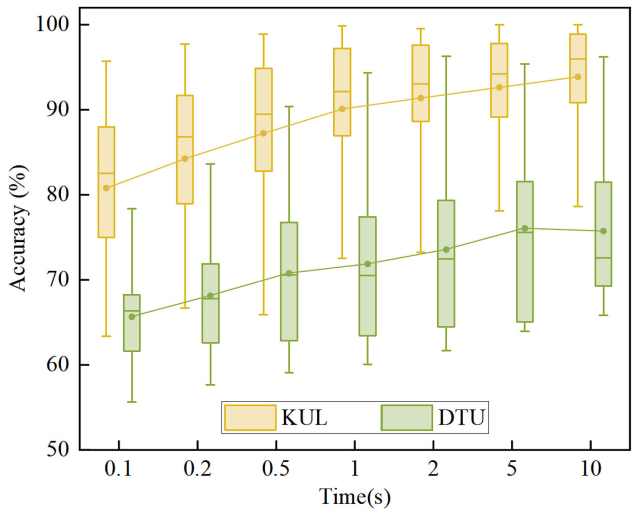

Fig. 3. Auditory spatial attention detection accuracy of the STAnet for seven decision window sizes across all subjects in KUL and DTU datasets, respectively.

(SD: 9.73%), respectively. Although the accuracy for 0.1-second decision window is lower than that for 1-second, the STAnet maintains a competitive ASAD accuracy (80.8%, SD: 9.87%).

On the DTU dataset, the STAnet obtains an average accuracy of 65.7% (SD: 5.50%) for 0.1-second, 68.1% (SD: 7.08%) for 0.2-second, 70.8% (SD: 8.04%) for 0.5-second, 71.9% (SD: 8.94%) for 1-second, 73.7% (SD: 9.59%) for 2-second, 76.1% (SD: 9.63%) for 5-second, and 75.8% (SD: 9.17%) for 10-second decision windows, respectively. The ASAD accuracy on the DTU dataset is significantly lower than that on the KUL dataset, which is consistent with the observations in [18], [27]. One of the possible reasons could be that the DTU dataset has two speech streams arriving 60° to the left and 60° to the right of the listening subjects [50], while the KUL dataset has the two speech streams arriving from ± 90° instead [49]. Therefore, the DTU dataset presents a more challenging task than the KUL dataset. Another major difference between the DTU and the KUL dataset is that the former's auditory stimuli are presented to the listeners with room reverberation at various levels, which might reduce the cortical speech tracking accuracy in human brain [53], hence decrease the differential responses between the attended and

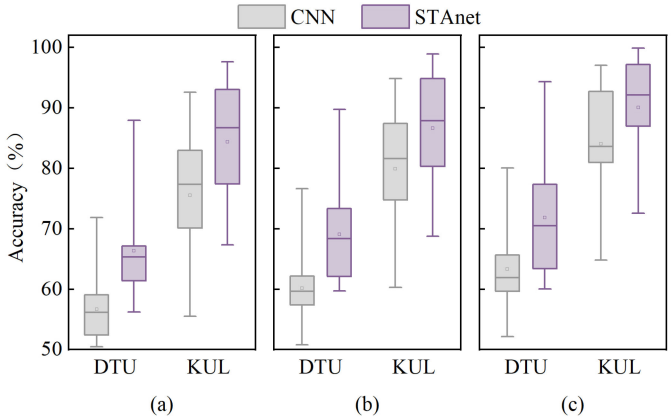

Fig. 4. Auditory spatial attention detection accuracy of the STAnet and CNN model for all subjects in the DTU and KUL datasets. (a) 16-channel EEG, (b) 32-channel EEG, and (c) 64-channel EEG.

unattended speakers [54]. However, the latter's auditory stimuli are presented to the listeners in a quiet acoustic environment.

We confirm that the STAnet performs reasonably well at a temporal resolution of 1 s, comparable to the time required for humans to switch their attention from one speaker to another. It is also encouraging to see that the STAnet still decodes well at a resolution of 100 milliseconds. We are not aware of other models that perform similarly in such low latency settings. With our results, we consider that real-time decoding of auditory attention is within reach.

C. Low-Density EEG Signals

This work is motivated to achieve real-time EEG-based ASAD in daily-life applications. High-density EEG signals involve more channels, therefore provide fine-grained spatial sampling attention detection. However, more channels also mean an increased setup time and effort [7], [41]. It is therefore desirable to reduce the number of channels required for an ASAD system. We further investigate how the STAnet performs in relatively low-density EEG systems.

Both KUL and DTU datasets are recorded with 64-channel with the BioSemi ActiveTwo system. We obtain 32-channel and 16 channel EEG following the electrode locations of the international 10/20 system [55]. Fig. 4 depicts the ASAD performance of the STAnet with 1-second decision windows based on 16-channel, 32-channel, and 64-channel EEG signals over all subjects in the KUL and DTU datasets. In general, more channels lead to better performance.

On the KUL dataset, the average accuracy of CNN model degrades from 64-channel (84.1%, SD: 10.16%) to 32-channel (79.9%, SD: 10.46%), and further to 16-channel (75.4%, SD: 11.01%). We observe an accuracy drop from 64-channel to 32-channel by 4.2%, and from 64-channel to 16-channel by 8.7%. We observe a modest accuracy drop of 3.4% and 5.7% for 32-channel and 16-channel over 64-channel EEG, respectively. However, the average accuracy for the STAnet remains competitive (16-channel, 84.4%, SD: 9.94%; 32-channel, 86.7%, SD: 9.85%), which significantly outperforms the CNN model (paired t-test: $p < 0.001$).

On the DTU dataset, the average accuracy of CNN model decreases significantly from 64-channel EEG (63.3%, SD: 5.96%) to 32-channel (60.2%, SD: 5.84%), and further to 16-channel (56.7%, SD: 5.18%). Relatively, the STAnet clearly reduces the performance gap over the CNN model between the performance of 64-channel (71.9%, SD: 8.94%) and 32-channel (69.1%, SD: 8.24%), as well as 16-channel (66.4%, SD: 7.66%).

To conclude, the STAnet is more robust than the CNN model with low-density EEG signals. We believe that the improved robustness comes from the spatiotemproal attention mechanisms.

D. STAnet vs Linear Decoder

We further compare the proposed STAnet with other competing models in the literature. We start by comparing the STAnet with the CCA model [9], which is considered to be one of the best linear AAD decoders to date [18]. It is noted that the CCA model requires the speech stimuli as the reference, that the STAnet doesn't require. In other words, the STAnet decodes the human attention purely from the brain signals themselves.

We re-implement the CCA model on both DTU and KUL datasets and report the performance for different decision windows, as summarized in Table III. The CCA model obtains an accuracy of 75.9% on the KUL dataset and 70.1% on the DTU dataset with 10-second decision window, and 60.2% and 53.4% on the two datasets respectively with 1-second decision window. With decision windows of less than 1-second, the accuracy of the CCA model further degrades, and drops below the chance level on the DTU dataset.

It is observed in Table III that the STAnet clearly outperforms the CCA model by a large margin (>20%) across all decision window sizes. On the KUL dataset, the STAnet achieves a robust performance of above 80% accuracy even for 0.1-second decision window.

E. STAnet vs CNN Decoder

We also compare the STAnet with CNN model by Vandecappelle *et al.* [24], that decodes the direction of auditory attention and achieves impressive results. For a fair comparison, we re-implement the CNN model with our experiment setup, and evaluate the ASAD accuracy for various decision windows.

As shown in Table III, the CNN model offers a significantly better accuracy than the CCA model on both KUL and DTU datasets. At the same time, the STAnet consistently outperforms the CNN model by average 6.1% on the KUL dataset and 8.8% on the DTU dataset respectively in terms of accuracy over various decision window sizes.

To summarize, the STAnet consistently outperforms the state-of-the-art linear and non-linear models on two publicly available datasets. These results confirm the effectiveness of the spatiotemporal attention network.

V. Discussions

Multi-channel EEG signals are collected from multiple sites of the scalp. The signals acquired from various electrode positions are not equally informative as far as auditory attention detection is concerned [41]. To gain insight into how the spatially

TABLE III

AUDITORY SPATIAL ATTENTION DETECTION ACCURACY (%) COMPARISON OF DIFFERENT MODELS ON KUL DATASET [49] AND DTU DATASET [50] FOR SEVEN DIFFERENT DECISION WINDOW LENGTHS. LINEAR MODEL DENOTES THE SETTING IN [9], WHILE CNN MODEL DENOTES THE SETTING IN [24]. NOTE THAT THE ACCURACY OF THE PROPOSED STANET SIGNIFICANTLY OUTPERFORMS BOTH THE LINEAR MODEL (P <0.001) AND NON-LINEAR MODEL (P <0.001)

Dataset	Model	Decision window (second)						
		0.1	0.2	0.5	1	2	5	10
KUL [49]	linear (CCA) [9]	50.9	53.6	55.7	60.2	63.5	69.4	75.9
	non-linear (CNN) [24]	74.3	78.2	80.6	84.1	85.7	86.9	87.9
	STAnet	**80.8**	**84.3**	**87.2**	**90.1**	**91.4**	**92.6**	**93.9**
DTU [50]	linear (CCA) [9]	-	-	-	53.4	57.7	61.9	70.1
	non-linear (CNN) [24]	56.7	58.4	61.7	63.3	65.2	67.4	67.8
	STAnet	**65.7**	**68.1**	**70.8**	**71.9**	**73.7**	**76.1**	**75.8**

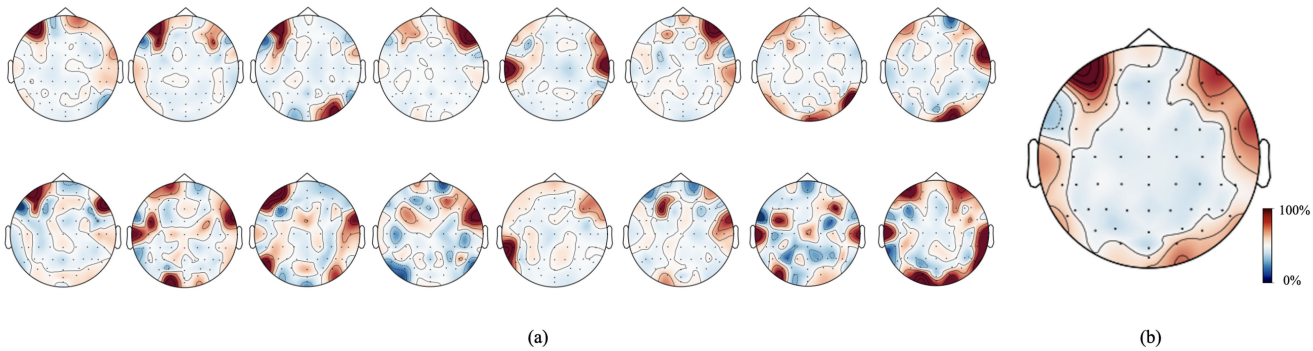

Fig. 5. Topography maps of the decoder weights associated with the EEG electrodes on KUL dataset. We aggregate spatial attention weights for all 1-second decision windows and plot the average. (a) The decoder activation patterns for all individual subjects. (b) The decoder activation pattern averaged over all 16 subjects. Black dots represent all 64 EEG electrodes. The attention weights are denoted by colors, with red color corresponding to a higher weightage.

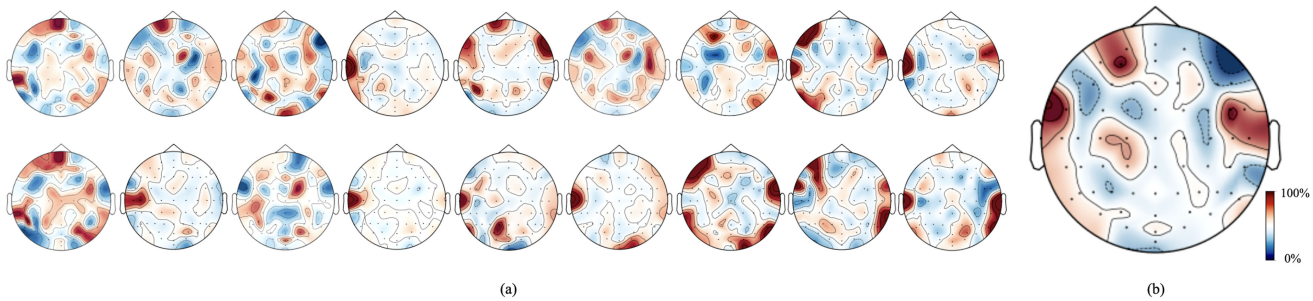

Fig. 6. Topography maps of the decoder weights associated with the EEG electrodes on DTU dataset. (a) The decoder activation patterns for all individual subjects. (b) The decoder activation pattern averaged over all 18 subjects.

differentiated weights on channels contribute to the performance gain, we aggregate and visualize the masking weights generated by the spatial attention mechanism over all 1-second decision windows for individual subjects in Fig. 5(a), and overall average in Fig. 5(b) on the KUL dataset, and in Fig. 6(a) and Fig. 6(b) respectively on the DTU dataset.

A. Subject-Independent Spatial Attention

It is expected that the locations indicative of neural activity contributing to speech processing have higher weights [56]. As illustrated in Fig. 5(b) and Fig. 6(b), in the topography with weights assigned by spatial attention mechanism, we see higher weights at electrodes placed over the frontal and temporal regions than elsewhere. These results are consistent with the findings by others that activations are observed prominently in the fronto-temporal cortex [15], [17], [24]. In addition, Fig. 5(b) and Fig. 6(b) also suggest that attentional modulation of speech tracking is mainly manifested in the auditory-specific temporal regions, that corroborates previous studies [4], [7], [22], [38], [51].

B. Subject Individuality

Fig. 5(a) and Fig. 6(a) illustrate how the differentiated weights generated by the spatial attention mechanism are distributed across the scalp of different individuals. Without surprise, it is observed that the weights, that reflect an individuals attentional focus, vary across the subjects. Additionally, it is worth noting that the decoder activation pattern may vary across subjects as well, therefore, the spatial patterns of brain activity related to auditory attention differ considerably across the subjects. These

findings support the claim that EEG signals exhibit subject-specific patterns due to the physiological and psychological individuality [28], [57], [58].

Considering the individuality in the modulation of auditory attention, the pre-defined handcrafted EEG features find it hard to have one size fit all. In contrast, the data-driven end-to-end learning, and the spatial attention mechanism in this study learn the ability to assign differentiated weights dynamically to EEG channels subject by subject, that effectively addresses subject individuality issue.

C. Data-Driven vs Handcrafted Features

Generally, most traditional auditory spatial attention decoding techniques involve a frontend process to remove artifacts, and to extract handcrafted features from EEG signals. In this study, we propose a data-driven solution to ASAD in an end-to-end manner. The raw EEG data are directly taken by the STAnet without any manipulation. The STAnet is capable of learning what is important for feature representation by itself as far as ASAD is concerned. As the end-to-end process is simple and straightforward, it facilitates the implementation in low-resource devices, such as neuro-steered hearing aids.

We consider that the data-driven approach is simple and effective, as can be seen in Table III where both STAnet and CNN [24] models clearly outperforms CCA [9]. The STAnet and CNN models employ data-driven frontend, while CCA involves handcrafted features. Besides the network architecture, we consider that the feature representation techniques deserve further exploration. For instance, the EEG characteristics in the frequency-domain may contribute to further enhancement of the ASAD performance [22], [23], [28], [38]. The fact that the time-domain frontend learns from the training data makes it less generalizable than other frequency-domain frontends across different recording conditions.

D. Effect of EEG Signal Recording Conditions

Consistent with the results of previous ASAD studies [17], [18], [24], the ASAD accuracy varies across subjects, which reflects the differences in recording conditions [49], [50], [59], as well as the physiological and psychological characteristics of the individuals [60], [61]. In general, the content of the speech stimuli, the spatial origin of the brain responses, and the physical layout of the listening experiments, among others, all contribute to the variation of EEG signals. For example, as discussed in Section VI-B, the differences between the KUL and DTU datasets in terms of the physical layouts of the listening experiments possibly lead to the large performance difference in the ASAD experiments. Furthermore, we hypothesize that the variation across subjects observed in our study is partially due to the small size of the dataset. A sufficiently large dataset related to the selective auditory attention task will benefit for the non-linear ASAD decoders.

VI. Conclusion

In this paper, we propose the STAnet, which incorporates two attention components into an end-to-end deep learning architecture. Our model infers attention maps along two separate dimensions, i.e., spatial and temporal, then the attention maps are multiplied to EEG signals feature map for adaptive feature refinement. This spatiotemporal encoding enables a high density of information, hence with high ASAD performance. Results indicate that the STAnet significantly outperformed conventional linear as well as the current state-of-the-art non-linear approaches in two publicly available datasets. As it does not require clean speech envelopes, the STAnet has the potential to enhance the signal processing in realistic hearing aids and other BCIs by incorporating information about the attention of the user.

Acknowledgment

The authors would like to thank Peiwen Li for useful discussions and technical assistance.

References

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *J. Acoustical Soc. Amer.*, vol. 25, no. 5, pp. 975–979, 1953.
[2] N. Mesgarani and E. F. Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature*, vol. 485, no. 7397, pp. 233–236, 2012.
[3] N. Ding and J. Z. Simon, "Emergence of neural encoding of auditory objects while listening to competing speakers," *Proc. Nat. Acad. Sci.*, vol. 109, no. 29, pp. 11854–11859, 2012.
[4] J. A. O'Sullivan *et al.*, "Attentional selection in a cocktail party environment can be decoded from single-trial EEG," *Cereb. Cortex*, vol. 25, no. 7, pp. 1697–1706, 2015.
[5] S. Van Eyndhoven, T. Francart, and A. Bertrand, "EEG-informed attended speaker extraction from recorded speech mixtures with application in neuro-steered hearing prostheses," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 5, pp. 1045–1056, May 2017.
[6] E. Ceolini *et al.*, "Brain-informed speech separation (BISS) for enhancement of target speaker in multitalker speech perception," *NeuroImage*, vol. 223, 2020, Art. no. 117282.
[7] B. Mirkovic *et al.*, "Decoding the attended speech stream with multi-channel EEG: Implications for online, daily-life applications," *J. Neural Eng.*, vol. 12, no. 4, 2015, Art. no. 046007.
[8] W. Biesmans *et al.*, "Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 5, pp. 402–412, May 2017.
[9] A. de Cheveigné *et al.*, "Decoding the auditory brain with canonical component analysis," *NeuroImage*, vol. 172, pp. 206–216, 2018.
[10] W. Nogueira *et al.*, "Toward decoding selective attention from single-trial EEG data in cochlear implant users," *IEEE Trans. Biomed. Eng.*, vol. 67, no. 1, pp. 38–49, Jan. 2020.
[11] I. Kuruvila *et al.*, "Inference of the selective auditory attention using sequential LMMSE estimation," *IEEE Trans. Biomed. Eng.*, vol. 68, no. 12, pp. 3501–3512, Dec. 2021.
[12] P. Faure and H. Korn, "Is there chaos in the brain?I. Concepts of nonlinear dynamics and methods of investigation," *Comptes Rendus de l'Académie des Sci.-Ser. III-Sci. de la Vie*, vol. 324, no. 9, pp. 773–793, 2001.
[13] H. Korn and P. Faure, "Is there chaos in the brain? II. Experimental evidence and related models," *Comptes Rendus Biol.*, vol. 326, no. 9, pp. 787–840, 2003.
[14] T. de Taillez, B. Kollmeier, and B. T. Meyer, "Machine learning for decoding listeners' attention from electroencephalography evoked by continuous speech," *Eur. J. Neurosci.*, vol. 51, no. 5, pp. 1234–1241, 2020.
[15] G. Ciccarelli *et al.*, "Comparison of two-talker attention decoding from EEG with nonlinear neural networks and linear methods," *Sci. Rep.*, vol. 9, no. 1, pp. 1–10, 2019.
[16] S. Cai *et al.*, "Low latency auditory attention detection with common spatial pattern analysis of EEG signals," in *Proc. Interspeech*, 2020, pp. 2772–2776.
[17] S. Geirnaert, T. Francart, and A. Bertrand, "Fast EEG-based decoding of the directional focus of auditory attention using common spatial patterns," *IEEE Trans. Biomed. Eng.*, vol. 68, no. 5, pp. 1557–1568, May 2021.

[18] S. Geirnaert *et al.*, "Electroencephalography-based auditory attention decoding: Toward neurosteered hearing devices," *IEEE Signal Process. Mag.*, vol. 38, no. 4, pp. 89–102, Jul. 2021.

[19] N. Das *et al.*, "Linear versus deep learning methods for noisy speech separation for EEG-informed attention decoding," *J. Neural Eng.*, vol. 17, no. 4, 2020, Art. no. 046039.

[20] A. Aroudi and S. Doclo, "Cognitive-driven binaural beamforming using EEG-based auditory attention decoding," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 862–875, 2020.

[21] J. N. Frey *et al.*, "Selective modulation of auditory cortical alpha activity in an audiovisual spatial attention task," *J. Neurosci.*, vol. 34, no. 19, pp. 6634–6639, 2014.

[22] M. Wöstmann *et al.*, "Spatiotemporal dynamics of auditory attention synchronize with speech," *Proc. Nat. Acad. Sci.*, vol. 113, no. 14, pp. 3873–3878, 2016.

[23] A. Bednar and E. C. Lalor, "Where is the cocktail party? Decoding locations of attended and unattended moving sound sources using EEG," *NeuroImage*, vol. 205, 2020, Art. no. 116283.

[24] S. Vandecappelle *et al.*, "EEG-based detection of the locus of auditory attention with convolutional neural networks," *Elife*, vol. 10, 2021, Art. no. e56481.

[25] S. Geirnaert, T. Francart, and A. Bertrand, "Riemannian geometry-based decoding of the directional focus of auditory attention using EEG," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 1115–1119.

[26] S. Cai *et al.*, "Low-latency auditory spatial attention detection based on spectro-spatial features from EEG," 2021, *arXiv:2103.03621*.

[27] I. Kuruvila *et al.*, "Extracting the locus of attention at a cocktail party from single-trial EEG using a joint CNN-LSTM model," *Front. Physiol.*, vol. 12, 2021, Art. no. 1178.

[28] Y. Deng, I. Choi, and B. Shinn-Cunningham, "Topographic specificity of alpha power during auditory spatial attention," *NeuroImage*, vol. 207, 2020, Art. no. 116360.

[29] A. Bednar, F. M. Boland, and E. C. Lalor, "Different spatio-temporal electroencephalography features drive the successful decoding of binaural and monaural cues for sound localization," *Eur. J. Neurosci.*, vol. 45, no. 5, pp. 679–689, 2017.

[30] H. Ramoser, J. Muller-Gerking, and G. Pfurtscheller, "Optimal spatial filtering of single trial EEG during imagined hand movement," *IEEE Trans. Rehabil. Eng.*, vol. 8, no. 4, pp. 441–446, Dec. 2000.

[31] F. Lotte and C. Guan, "Regularizing common spatial patterns to improve BCI designs: Unified theory and new algorithms," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 2, pp. 355–362, Feb. 2011.

[32] N. Cheng *et al.*, "Brain-computer interface-based soft robotic glove rehabilitation for stroke," *IEEE Trans. Biomed. Eng.*, vol. 67, no. 12, pp. 3339–3351, Dec. 2020.

[33] J. Wang *et al.*, "Decoding single-hand and both-hand movement directions from noninvasive neural signals," *IEEE Trans. Biomed. Eng.*, vol. 68, no. 6, pp. 1932–1940, Jun. 2021.

[34] J. R. Kerlin, A. J. Shahin, and L. M. Miller, "Attentional gain control of ongoing cortical speech representations in a "cocktail party"," *J. Neurosci.*, vol. 30, no. 2, pp. 620–628, 2010.

[35] L. Meyer, "The neural oscillations of speech processing and language comprehension: State of the art and emerging mechanisms," *Eur. J. Neurosci.*, vol. 48, no. 7, pp. 2609–2621, 2018.

[36] D. S. Bassett and O. Sporns, "Network neuroscience," *Nature Neurosci.*, vol. 20, no. 3, pp. 353–364, 2017.

[37] C. Herff *et al.*, "Generating natural, intelligible speech from brain activity in motor, premotor, and inferior frontal cortices," *Front. Neurosci.*, vol. 13, 2019, Art. no. 1267.

[38] E. M. Z. Golumbic *et al.*, "Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party"," *Neuron*, vol. 77, no. 5, pp. 980–991, 2013.

[39] J. Samogin *et al.*, "Shared and connection-specific intrinsic interactions in the default mode network," *NeuroImage*, vol. 200, pp. 474–481, 2019.

[40] M. Arvaneh *et al.*, "Optimizing the channel selection and classification accuracy in EEG-based BCI," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 6, pp. 1865–1873, Jun. 2011.

[41] A. M. Narayanan and A. Bertrand, "Analysis of miniaturization effects and channel selection strategies for EEG sensor networks with application to auditory attention detection," *IEEE Trans. Biomed. Eng.*, vol. 67, no. 1, pp. 234–244, Jan. 2020.

[42] A. Vaswani *et al.*, "Attention is all you need," in *Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[43] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[44] S. Woo *et al.*, "Cbam: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[45] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and Accurate Deep Network Learning by Exponential Linear Units (Elus)," 2015, *arXiv:1511.07289*.

[46] M. R. Jones and M. Boltz, "Dynamic attending and responses to time," *Psychol. Rev.*, vol. 96, no. 3, pp. 459–491, 1989.

[47] M. R. Jones *et al.*, "Temporal aspects of stimulus-driven attending in dynamic arrays," *Psychol. Sci.*, vol. 13, no. 4, pp. 313–319, 2002.

[48] L.-V. Andreou, M. Kashino, and M. Chait, "The role of temporal regularity in auditory segregation," *Hear. Res.*, vol. 280, no. 1-2, pp. 228–235, 2011.

[49] N. Das, T. Francart, and A. Bertrand, "Auditory attention detection dataset KULeuven," Aug. 2020, Version 1.1.0. [Online]. Available: https://doi.org/10.5281/zenodo.3997352

[50] S. A. Fuglsang, D. D. Wong, and J. Hjortkjær, "EEG and audio dataset for auditory attention decoding," Mar. 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1199011

[51] 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.

[52] D. P. Kingma and J. Ba, "Adam: A. method for stochastic optimization," 2014, *arXiv:1412.6980*.

[53] N. Ding and J. Z. Simon, "Adaptive temporal encoding leads to a background-insensitive cortical representation of speech," *J. Neurosci.*, vol. 33, no. 13, pp. 5728–5735, 2013.

[54] J. M. Rimmele *et al.*, "The effects of selective attention and speech acoustics on neural speech-tracking in a multi-talker scene," *Cortex*, vol. 68, pp. 144–154, 2015.

[55] R. W. Homan, J. Herman, and P. Purdy, "Cerebral location of international 10-20 system electrode placement," *Electroencephalography Clin. Neuriophysiol.*, vol. 66, no. 4, pp. 376–382, 1987.

[56] A. de Cheveigné and J. Z. Simon, "Denoising based on spatial filtering," *J. Neurosci. Methods*, vol. 171, no. 2, pp. 331–339, 2008.

[57] I. Choi *et al.*, "Individual differences in attentional modulation of cortical responses correlate with selective attention performance," *Hear. Res.*, vol. 314, pp. 10–19, 2014.

[58] V. Viswanathan, H. M. Bharadwaj, and B. G. Shinn-Cunningham, "Electroencephalographic signatures of the neural representation of speech during selective attention," *Eneuro*, vol. 6, no. 5, pp. 1–14, 2019.

[59] N. Das *et al.*, "The effect of head-related filtering and ear-specific decoding bias on auditory attention detection," *J. Neural Eng.*, vol. 13, no. 5, 2016, Art. no. 056014.

[60] B. Blankertz *et al.*, "Optimizing spatial filters for robust EEG single-trial analysis," *IEEE Signal Process. Mag.*, vol. 25, no. 1, pp. 41–56, 2007.

[61] O.-Y. Kwon *et al.*, "Subject-independent brain-computer interfaces based on deep convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 10, pp. 3839–3852, Oct. 2020.