INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Auditory Attention Detection in Real-Life Scenarios Using Common Spatial Patterns from EEG

Kai Yang[1], *Zhuang Xie*[2], *Di Zhou*[3], *Longbiao Wang*[1], *Gaoyan Zhang*[1*]

[1]Tianjin key Laboratory of Cognitive Computing and Application, Tianjin University, China
[2]School of Software, Henan University, China
[3]Japan Advanced Institute of Science and Technology, Japan

kai_y@tju.edu.cn, zhuangxie_cs@163.com, zhoudi@jaist.ac.jp, zhanggaoyan@tju.edu.cn

Abstract

Auditory attention detection (AAD) methods based on electroencephalography (EEG) could be used in neuro-steered hearing devices to help hearing-loss people improve their hearing ability. However, previous studies have mostly obtained EEG data in laboratory settings which limits the practical application of neuro-steered hearing devices. In this study, we employ a common spatial pattern (CSP) algorithm to perform AAD using EEG signals collected by a wireless mobile EEG system, from real-life scenarios when people are walking and sitting. The results show that the CSP method can achieve AAD accuracy between 81.3% and 87.5% when using different decision windows (1 s- 30 s), which is better than previous methods based on linear mapping methods and convolutional neural networks (CNN). This proves that the CSP algorithm can decode people's attention efficiently even outside the laboratory. Analysis of EEG frequency bands shows that the δ and β bands have high activity in attention tasks.

Index Terms: auditory attention detection, electroencephalography, real-life scenarios, common spatial pattern, brain-computer interface

1. Introduction

The latest Global Burden of Disease (GBD) study shows that the burden of hearing loss due to aging is increasing over time, and the global demand for assistive listening devices is growing [1]. Assistive listening devices, such as hearing aids and cochlear implants, can restore hearing in hearing-loss patients. Although these hearing devices have been improved over the past few decades, including the use of more advanced speech enhancement, directional beamforming, and noise suppression technologies [2], the most advanced hearing devices still do not work well in "cocktail party" [3] scenarios when multiple people are speaking at the same time. In such a scenario, normal-hearing people can easily distinguish and track the sound source of interest, while ignoring other sources. However, people with hearing impairments often have difficulty participating in conversations. While advanced speech enhancement algorithms can suppress background noise and enhance a speaker from a mix of speech, they often do not know which speaker to enhance. Researchers proposed extracting attention-related information from the brain to determine the attended speaker [4, 5, 6]. This problem is commonly known as Auditory Attention Detection (AAD). AAD algorithms can be integrated with speech separation technology, miniature EEG sensors, and intelligent gain systems in neuro-steered hearing devices [7] to selectively amplify the attended source, which will improve the quality of life of hearing-loss people.

Previous AAD researches have been based on the physiological basis that the brain can track speech envelope [4, 5]. Envelope tracking is reflected as a phase-locking effect between the neural signal and the speech envelope. Moreover, in multi-speaker scenarios, envelope tracking is enhanced for attended envelopes than unattended envelopes [4, 5, 8, 9]. Based on these findings, some researchers perform AAD by stimulus reconstruction (SR) methods [6, 10]. The SR methods construct a linear decoder to reconstruct speech envelopes from the recorded brain signals, such as magnetoencephalography (MEG) or electroencephalography (EEG). By comparing the correlation between the reconstructed envelopes and the actual envelopes of different speakers, the speaker with a higher correlation is identified as the attended one. The detection accuracy of the linear SR method is in the range of 82%-89% for a 60 s decision window. Such a long decision window is not suitable for realistic applications in hearing devices. However, as the decision window length decreases, especially below 10 s, the detection accuracy of the SR method will drop sharply [11]. Alickovic et al. [12] used canonical correlation analysis (CCA) to improve the accuracy of AAD. CCA algorithm finds the optimal linear transform to apply to both the stimulus envelopes and neural signals to reveal correlations between them. Using the 60 s decision window, the detection accuracy reaches about 90% on different datasets. Similarly, the accuracy performs poorly on shorter decision windows, with accuracies of about 58% and 68% for 1 s and 5 s [13].

SR and CCA are both linear mapping methods. To improve detection accuracy, some researchers have proposed constructing deep neural network (DNN) models that can extract nonlinear features to perform AAD. de Taillez et al. [14] used a simple Fully Connected Network (FCN) model to reconstruct the speech envelope from EEG signals and decode the attended speaker by correlation analysis. The accuracy of AAD is about 96.7% and 67.8% for decision windows of 60 s and 2 s. In [15], the authors proposed a convolutional neural network (CNN) model that uses EEG data and speech envelope features as input and implicitly computes the similarity between the EEG signals and the corresponding speech envelopes. The AAD accuracy is about 81% under a 10 s decision window. Cai et al. [16, 17] used the attention mechanism in neural networks to construct classification models with the accuracy of about 80%-88% for 2 s decision window and about 79%-84% for 1 s decision window. All the above studies employ clean speech envelopes as the model input, but this is not reasonable for practical applications because only mixed speech signals are available. Therefore, Vandecappelle et al. [18] used a CNN model to decode the locus of auditory attention (left/right) without knowledge of the speech envelopes, and the results show that the accuracy

Corresponding author:Gaoyan Zhang

10.21437/Interspeech.2023-1377

is about 80.8% for decision window of 1 s. A study used a multi-task learning model, in which the direct AAD classification task was assisted by the envelope reconstruction task to perform AAD, and the results show an AAD accuracy of 82% of 2 s decision window [19]. However, due to the small amount of subjects' data, deep learning methods have a high risk of overfitting and the results may be poorer when changing different subjects or datasets. On the contrary, the linear method has higher robustness and stability and is computationally cheaper [7].

What should be noted is that most of the current studies have collected EEG data in the laboratory, which limits the further application of AAD in daily environments. To address the above-mentioned issues, we focus on the linear filtering method common spatial patterns (CSP) to decode the directional focus of attention and we use EEG data collected in real-life outdoor scenarios instead of laboratory settings. The CSP method extracts spatial features based on transient lateralization effects in the brain, which does not require envelope features and avoids the need to compute correlations over longer time windows. It is important to shorten the time of AAD for further application in neuro-steered hearing devices. In addition, we conducted experiments in different frequency bands to explore the contribution of EEG frequency bands in AAD.

2. Methods

2.1. CSP filtering and feature extraction

CSP filtering is a spatial feature extraction method widely used in the field of brain-computer interface (BCI), such as in motor imagery [20, 21], emotion recognition [22], etc. In this study, the CSP algorithm is used to perform AAD. As a binary classification algorithm, the principle of CSP is to solve a set of optimal spatial filters by diagonalizing matrices. The original signals are projected through filters into a lower-dimensional subspace so that the variance difference between the two classes of signals is the largest. Suppose $x(t) \in R^{N\times 1}$ denotes the EEG signal of N channels at sampling point $t = 1 \dots T$. The $x(t)$ belongs to one of the two EEG classes C_1 and C_2 (e.g., attending left speaker or right speaker). The goal of CSP is to design M spatial filters $W \in R^{N\times M}$ with the first $M/2$ filters maximizing the output energy of class C_1 and minimizing the output energy of Class C_2. The other $M/2$ filters maximize the output energy of class C_2 and minimize the output energy of class C_1. For the first filter w_1, the objective function is shown below:

$$w_1 = \underset{w}{argmax} \frac{w^{\mathrm{T}} R_{C_1} w}{w^{\mathrm{T}} R_{C_2} w}, \tag{1}$$

where R_{C_1} and R_{C_2} are the sample covariance matrices of class C_1 and C_2, as shown in equation (2) and (3).

$$R_{C_1} = \frac{1}{|C_1|} \sum_{t\in C_1} x(t) x^{\mathrm{T}}(t) \tag{2}$$

$$R_{C_2} = \frac{1}{|C_2|} \sum_{t\in C_2} x(t) x^{\mathrm{T}}(t) \tag{3}$$

$|C_1|$ and $|C_2|$ are the number of time points of classed C_1 and C_2. By constraining $w^T R_{C_2} w = 1$ in equation (1), the problem can be transformed into an extreme value problem in Equation (4) by the method of Lagrange multipliers.

$$L(\lambda, w) = w^{\mathrm{T}} R_{C_1} w - \lambda (w^{\mathrm{T}} R_{C_2} w - 1) \tag{4}$$

Further transforming Equation (4) into:

$$R_{C_2}^{-1} R_{C_1} w = \lambda w, \tag{5}$$

which corresponds to an eigenvalue problem. The filters w_1 and w_M can be obtained for the eigenvectors corresponding to the maximum and minimum eigenvalues. The remaining filters can be obtained for the eigenvectors corresponding to the subsequent largest and smallest eigenvalues.

By using the CSP filter, the original EEG signal with T sampling points $X \in R^{N\times T}$ is first converted to $Y \in R^{M\times T}$ in the lower dimensional space. Typically, the log-energy of each decision window is then calculated as feature input to the classifier [13, 23]. The decision window size determines the amount of EEG data used for AAD. Assuming that the decision window contains T_d sampling points, the feature vector f can be expressed as:

$$f = \begin{bmatrix} \log(\sum_{t=1}^{T_d} y_1(t)^2) \\ \vdots \\ \log(\sum_{t=1}^{T_d} y_M(t)^2) \end{bmatrix}, \tag{6}$$

where $y_i(t)$, $i = 1, \dots, M$ denotes the value of Y in the i-th row and t-th column. Therefore, the feature vector f of each decision window is an M-dimensional vector.

2.2. Classification

Feature vectors f can be classified with classifiers. In this study, we use two classifiers, linear discriminant analysis (LDA) and support vector machine (SVM), to detect the direction of attention and compare the performance between them. LDA is usually used in combination with CSP [23], and its classification principle is very simple. It finds a projection vector v such that similar samples are projected to be as close as possible and dissimilar samples are as far away as possible [24]. The algorithm finds the optimal vector v by the following solution:

$$v = S_W^{-1}(\mu_2 - \mu_1), \tag{7}$$

where S_W is the within-class scatter matrix of feature f as shown in Equation (8). μ_1 and μ_2 are the class feature means.

$$S_W = \sum_{f\in C_1} (f - \mu_1)(f - \mu_1)^{\mathrm{T}} + \sum_{f\in C_2} (f - \mu_2)(f - \mu_2)^{\mathrm{T}} \tag{8}$$

After getting v, we can get the projection distance $v^{\mathrm{T}} f$ based on the input feature f. Then we can choose a bias to classify the projection distance of the samples. The commonly used bias is the average of the LDA projected class means:

$$b = -\frac{1}{2} v^{\mathrm{T}} (\mu_1 + \mu_2). \tag{9}$$

Therefore, the discriminant model is:

$$D(f) = v^{\mathrm{T}} f + b. \tag{10}$$

f is classified into class C_1 if $D(f) > 0$ and into class C_2 if $D(f) < 0$.

SVM has good performance in binary classification problems with small sample data. With the given training data, the SVM will obtain a hyperplane as a decision surface that maximizes the interval between the two classes of samples. The principle can be found in [25]. This study also uses SVM classifier to classify the extracted log-energy features. Here, two different kernels in SVM are used: linear kernel and radial basis function (RBF) kernel.

2.3. Baseline AAD models

In this study, we compare three different AAD methods, including the classical SR method [6, 26], CCA method [12, 27], and a binary classification method [18] using advanced CNN. The SR method reconstructs the speech envelopes from the EEG signals by constructing linear decoders. The correlations between the reconstructed envelopes and the actual envelopes are calculated to detect the subjects' attention.

CCA combines a spatio-temporal backward model (i.e. SR, mapping the EEG to the envelopes) and a temporal forward model (mapping the envelopes to the EEG) to make their output maximally correlated. CCA finds the best transformation matrices W_1 and W_2 for EEG signals X_1 and speech envelopes X_2. The columns of X_1W_1 are mutually uncorrelated, as are the columns of X_2W_2, while pairs of columns taken from both ("canonical correlate pairs") are maximally correlated [27]. The Pearson correlation coefficients between these pairs are defined as the canonical correlation coefficients. The attended speaker is identified by classifying the difference between the canonical correlation coefficients of the competing speakers using a classifier.

The CNN binary classification method uses EEG as input to predict the directional focus of subjects' attention. It consists of a convolutional layer with 5 kernels (kernel size: channels $\times$ 17) and two fully connected (FC) layers. The convolutional layer uses the rectified linear unit (ReLU) activation function and average pooling. The two FC layers have 5 and 2 units, respectively, followed by sigmoid activation. Loss is calculated using the cross-entropy loss function.

3. Experiments

3.1. AAD Dataset

The dataset used in this study was collected by Straetmans et al [26]. It contains EEG data of 20 subjects (24.2 ± 2.8 years; 16 female, 4 male). A two-competitive speaker paradigm was employed for the experiment. Subjects were asked to focus on one continuous speech stream while ignoring the other simultaneously presented speech stream. The stimuli were presented after head-related transfer function filtering to simulate speech from 45° to the left and 45° to the right of subjects. Speech stimuli were presented in six approximately five-minute-long trials. Each stimulus consisted of a coherent short story from an audio storybook narrated in German by a male speaker. Audio stimuli were presented through in-ear headphones. The side of the to-be-attended stream alternated across subjects. Stimuli were presented to subjects in random order.

The experiment was conducted in a public cafeteria. In three of the trials, subjects were asked to walk at a comfortable speed along a predetermined route. In the remaining three trials, subjects sat on chairs in front of a white wall. EEG was recorded by a wireless 24-channel electrode cap and connected to a direct current amplifier. EEG data were recorded at a sampling rate of 250 Hz and channel Fz was used as the reference electrode. The EEG signals were wirelessly transmitted to a smartphone via Bluetooth. More dataset details can be found in [26].

3.2. Data Preprocessing

The original EEG data were preprocessed using the EEGLAB toolbox [28]. The data were firstly downsampled to 128 Hz and band-pass filtered between 1 and 30 Hz. The artifact subspace reconstruction method [29] was applied to detect and remove high-amplitude non-brain activity (produced by eye blinks, muscle activity, sensor motion, etc.). The bad channels with a correlation less than 0.8 with surrounding channels were replaced by an estimate calculated using spherical spline interpolation [30] and all EEG channels were re-referenced to the average reference. Finally, we performed independent component analysis (ICA) on the EEG data to remove artifacts. As for the speech envelopes needed in the SR and CCA methods, we directly used the ones provided in the dataset without additional processing.

3.3. Experiment Setup

For the training of the different AAD methods, for each trial of all subjects, we randomly selected a continuous 20% length of data as the test set and the rest data as the training set. There is no overlap between them. We sliced different length samples (1 s, 2 s, 5 s, 10 s, 20 s, and 30 s) in each set, and also no overlap between samples.

For the CSP method, to avoid overfitting, the sample covariance matrices in Equation (2) and Equation (3) are regularized using ridge regression. Using the method in [31] which is the recommended state-of-the-art covariance matrix estimator, the regularization parameters are analytically determined. We chose $M = 6$ filters as in [13], the first three filters maximizing the output energy of class C_1 and minimizing the energy of class C_2, and the last three filters doing the opposite.

For the SR method, we used EEG data with time lags ranging from 0 ms to 250 ms [6] to reconstruct the stimulus envelopes. In CCA, a forward lag of -250 ms was used on the speech envelopes and a backward lag of 250 ms was used on the EEG [12]. The differences between the first ten canonical correlation coefficients of the competing speakers were selected as the classifier's inputs. Here, we used the LDA classifier to decode the attended speaker. In both methods, we downsampled the EEG signals to the same 64 Hz as the envelopes provided by the dataset when preprocessing the EEG. The EEG signals and speech envelopes were again filtered between 1 and 8 Hz which was determined to be optimal for linear mapping methods [6, 32].

We constructed the same CNN model as in [18]. During the training, the batch size was set to 64 and we used the stochastic gradient descent (SGD) optimizer with an initial learning rate of 0.1 and a momentum of 0.9. The learning rate decayed to 0.5 times the original every five epochs. Regularization consisted of weight decay with a value of 0.01. The early stopping strategy was also employed if the loss on the validation set did not decrease for ten consecutive epochs. In all methods, we performed a 5-fold cross-validation on the training set to select the optimal models and hyperparameters, then tested them on the test set.

4. Results and Discussion

4.1. AAD performance of different methods

After applying CSP filtering, we used different classifiers to detect the locus of auditory attention, including SVM with an RBF kernel (CSP-SVM$_1$), SVM with a linear kernel (CSP-SVM$_2$), and LDA classifier (CSP-LDA). We show the AAD accuracy of the different methods in Table 1. The accuracy of SR and CCA methods using linear mapping is low, especially on short decision windows, below 60% on the 1 s and 2 s decision windows. The methods using CSP filtering and

Table 1: *AAD performance using different methods. Boldface indicated the best result.*

Method	Window(s)					
	1	**2**	**5**	**10**	**20**	**30**
SR [6]	54.9	56.9	59.8	62.7	68.4	74.4
CCA [12]	55.3	58.7	62.8	65.5	71.8	77.5
CNN [18]	80.9	82.3	83.1	83.2	83.6	84.6
CSP-SVM$_1$	80.1	81.5	82.0	84.6	85.0	84.6
CSP-SVM$_2$	**81.3**	83.1	84.8	87.5	**87.8**	**87.5**
CSP-LDA	80.8	**83.5**	**86.1**	**87.6**	87.2	**87.5**

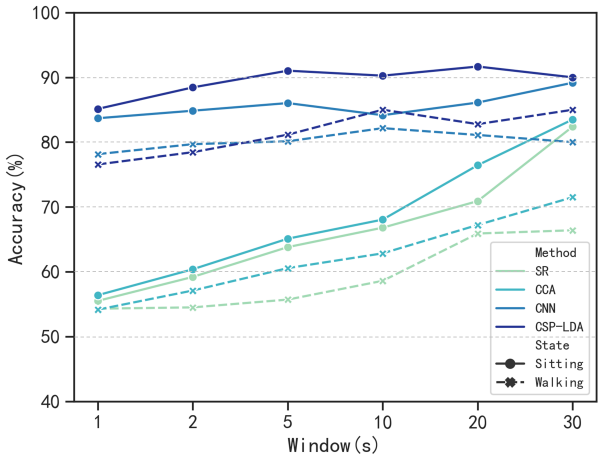

Figure 1: *AAD performance of different methods in sitting and walking states.*

CNN achieve accuracies of over 80% on 1 s window and the CSP achieves the highest accuracy on all decision windows. Paired t-tests were conducted to compare the performance of different methods. The results of the significance test show that the CSP-SVM$_2$ and CSP-LDA significantly outperform the CNN method ($t = 3.471, p<0.05; t = 3.702, p<0.05$), and there is no significant difference between CSP-SVM$_1$ and CNN ($t = 0.036, p = 0.973$). This indicates that the linear CSP filtering method meets or even exceeds the method using advanced neural networks. In addition, the accuracy of CSP-SVM$_2$ with the linear kernel is significantly better than that of CSP-SVM$_1$ with RBF kernel ($t = 7.621, p<0.01$). There is also no significant difference between CSP-SVM$_2$ and CSP-LDA ($t = 0.413, p = 0.696$). However, LDA is faster to compute and requires fewer computational resources compared to SVM. Therefore, we used the LDA classifier in the subsequent analysis.

We further tested the AAD accuracy of the four methods in different states (sitting and walking), and the results are shown in Figure 1. It is found that the AAD accuracy in the sitting state is significantly higher than the accuracy in the walking state among all methods. The CSP method reached 85.13% and 76.56% accuracies for the 1 s decision window in the sitting and walking state, which shows that the CSP method can decode subjects' attention stably and accurately in real-life scenarios.

4.2. The impact of different frequency bands on AAD

To investigate the contribution of different frequency bands in AAD, we further filtered the EEG data into different sub-bands,

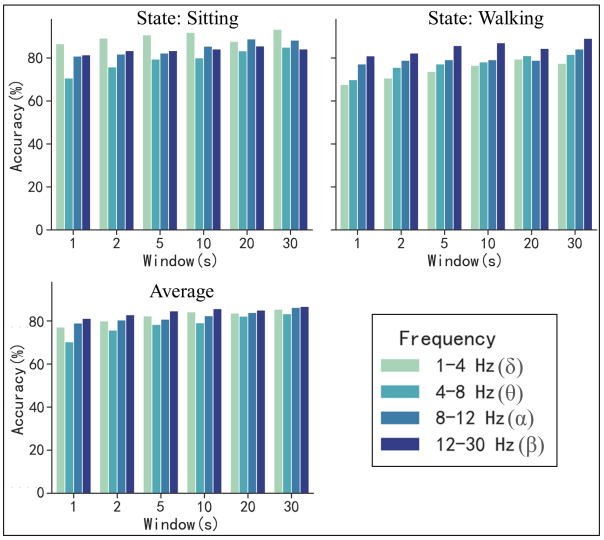

Figure 2: *Comparison of the contribution of EEG frequency bands in different states.*

i.e. δ band (1-4 Hz), θ band (4-8 Hz), α band (8-12 Hz), and β band (12-30 Hz) to train the CSP filters and decode the directional focus of attention. The results are shown in Figure 2, and it can be found that the δ band has the highest decoding accuracy in the sitting state, but performs the worst in the walking state. The β band contributes the most in the walking state, and the decoding accuracy even exceeds the sitting state in some decision windows. We conjecture that selective attention during sitting would work by increasing the gain of the low-frequency EEG signal for the attended speech [8, 6]. In contrast, during walking, subjects need to focus their attention more, the δ band is suppressed and the β band dominates. Averaging the results of sitting and walking reveals that the β band continues to contribute the most in the AAD task, consistent with the results obtained in previous studies [33, 34].

5. Conclusions

In this study, we used a CSP linear filtering method to detect the directional focus of attention in real-life scenarios and achieved a decoding accuracy of 81.3% on a 1 s decision window. We distinguished the AAD accuracy of subjects in different behavioral states (sitting and walking). Although the accuracy is slightly lower in the walking state, it still shows potential for application. Experiments with sub-bands reveal significant contributions from δ and β bands. Compared with traditional linear mapping methods and advanced CNN models, the CSP method is more accurate, responsive, robust, and requires fewer computational resources, which lays the foundation for realistic applications of neuro-steered hearing aids. Future research can be extended to extract EEG signals with a higher signal-to-noise ratio in living environments to improve AAD accuracy and explore AAD in scenes with more speakers.

6. Acknowledgements

This work was supported by the National Natural Science Foundation of China (NO.61876126) and the Innovation Fund of Tianjin University (No. 2023XSU-0026).

7. References

[1] L. Haile, A. Orji, P. Briant, J. Adelson, A. Davis, and T. Vos, "Updates on hearing from the global burden of disease study," *Innovation in Aging*, vol. 4, no. Suppl 1, p. 808, 2020.

[2] S. Haykin and K. R. Liu, *Handbook on array processing and sensor networks*. John Wiley & Sons, 2010.

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

[4] N. Ding and J. Z. Simon, "Emergence of neural encoding of auditory objects while listening to competing speakers," *Proceedings of the National Academy of Sciences*, vol. 109, no. 29, pp. 11 854–11 859, 2012.

[5] 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.

[6] 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.

[7] S. Geirnaert, S. Vandecappelle, E. Alickovic, A. de Cheveigne, E. Lalor, B. T. Meyer, S. Miran, T. Francart, and A. Bertrand, "Electroencephalography-based auditory attention decoding: Toward neurosteered hearing devices," *IEEE Signal Processing Magazine*, vol. 38, no. 4, pp. 89–102, 2021.

[8] N. Ding and J. Z. Simon, "Neural coding of continuous speech in auditory cortex during monaural and dichotic listening," *Journal of neurophysiology*, vol. 107, no. 1, pp. 78–89, 2012.

[9] E. M. Z. Golumbic, N. Ding, S. Bickel, P. Lakatos, C. A. Schevon, G. M. McKhann, R. R. Goodman, R. Emerson, A. D. Mehta, J. Z. Simon *et al.*, "Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party"," *Neuron*, vol. 77, no. 5, pp. 980–991, 2013.

[10] M. J. Crosse, G. M. Di Liberto, A. Bednar, and E. C. Lalor, "The multivariate temporal response function (mtrf) toolbox: a matlab toolbox for relating neural signals to continuous stimuli," *Frontiers in human neuroscience*, vol. 10, p. 604, 2016.

[11] W. Biesmans, N. Das, T. Francart, and A. Bertrand, "Auditory-inspired speech envelope extraction methods for improved eeg-based auditory attention detection in a cocktail party scenario," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 5, pp. 402–412, 2016.

[12] E. Alickovic, T. Lunner, F. Gustafsson, and L. Ljung, "A tutorial on auditory attention identification methods," *Frontiers in neuroscience*, p. 153, 2019.

[13] S. Geirnaert, T. Francart, and A. Bertrand, "Fast eeg-based decoding of the directional focus of auditory attention using common spatial patterns," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 5, pp. 1557–1568, 2020.

[14] T. de Taillez, B. Kollmeier, and B. T. Meyer, "Machine learning for decoding listeners' attention from electroencephalography evoked by continuous speech," *European Journal of Neuroscience*, vol. 51, no. 5, pp. 1234–1241, 2020.

[15] G. Ciccarelli, M. Nolan, J. Perricone, P. T. Calamia, S. Haro, J. O'sullivan, N. Mesgarani, T. F. Quatieri, and C. J. Smalt, "Comparison of two-talker attention decoding from eeg with nonlinear neural networks and linear methods," *Scientific reports*, vol. 9, no. 1, p. 11538, 2019.

[16] S. Cai, P. Li, E. Su, and L. Xie, "Auditory attention detection via cross-modal attention," *Frontiers in Neuroscience*, vol. 15, p. 652058, 2021.

[17] 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. 52, no. 2, pp. 256–266, 2021.

[18] 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.

[19] Z. Zhang, G. Zhang, J. Dang, S. Wu, D. Zhou, and L. Wang, "Eeg-based short-time auditory attention detection using multi-task deep learning." in *INTERSPEECH*, 2020, pp. 2517–2521.

[20] U. Talukdar, S. M. Hazarika, and J. Q. Gan, "Adaptation of common spatial patterns based on mental fatigue for motor-imagery bci," *Biomedical Signal Processing and Control*, vol. 58, p. 101829, 2020.

[21] C. Zhang and A. Eskandarian, "A computationally efficient multi-class time-frequency common spatial pattern analysis on eeg motor imagery," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2020, pp. 514–518.

[22] M. Yan, Z. Lv, W. Sun, and N. Bi, "An improved common spatial pattern combined with channel-selection strategy for electroencephalography-based emotion recognition," *Medical Engineering & Physics*, vol. 83, pp. 130–141, 2020.

[23] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Muller, "Optimizing spatial filters for robust eeg single-trial analysis," *IEEE Signal processing magazine*, vol. 25, no. 1, pp. 41–56, 2007.

[24] M. Kołodziej, A. Majkowski, and R. J. Rak, "Linear discriminant analysis as eeg features reduction technique for brain-computer interfaces," *Przeglad Elektrotechniczny*, vol. 88, no. 3, pp. 28–30, 2012.

[25] L. Wang, *Support vector machines: theory and applications*. Springer Science & Business Media, 2005, vol. 177.

[26] L. Straetmans, B. Holtze, S. Debener, M. Jaeger, and B. Mirkovic, "Neural tracking to go: auditory attention decoding and saliency detection with mobile eeg," *Journal of Neural Engineering*, vol. 18, no. 6, p. 066054, 2022.

[27] A. de Cheveigné, D. D. Wong, G. M. Di Liberto, J. Hjortkjær, M. Slaney, and E. Lalor, "Decoding the auditory brain with canonical component analysis," *NeuroImage*, vol. 172, pp. 206–216, 2018.

[28] A. Delorme and S. Makeig, "Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis," *Journal of neuroscience methods*, vol. 134, no. 1, pp. 9–21, 2004.

[29] M. Plechawska-Wojcik, M. Kaczorowska, and D. Zapala, "The artifact subspace reconstruction (asr) for eeg signal correction. a comparative study," in *Information systems architecture and technology: proceedings of 39th international conference on information systems architecture and technology–ISAT 2018: part II*. Springer, 2019, pp. 125–135.

[30] T. C. Ferree, "Spherical splines and average referencing in scalp electroencephalography," *Brain topography*, vol. 19, pp. 43–52, 2006.

[31] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of multivariate analysis*, vol. 88, no. 2, pp. 365–411, 2004.

[32] N. Das, J. Zegers, T. Francart, A. Bertrand *et al.*, "Linear versus deep learning methods for noisy speech separation for eeg-informed attention decoding," *Journal of Neural Engineering*, vol. 17, no. 4, p. 046039, 2020.

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

[34] A.-L. Giraud and D. Poeppel, "Cortical oscillations and speech processing: emerging computational principles and operations," *Nature neuroscience*, vol. 15, no. 4, pp. 511–517, 2012.