INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Low Latency Auditory Attention Detection with Common Spatial Pattern Analysis of EEG Signals

Siqi Cai[1,2], *Enze Su*[1], *Yonghao Song*[1], *Longhan Xie*[1], *Haizhou Li*[2,3]

[1]Shien-Ming Wu School of Intelligent Engineering, South China University of Technology, China
[2]Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[3]Machine Listening Lab, University of Bremen, Germany

siqi.cai@u.nus.edu, {Enze.Su,Yonghao.Song} @mail.scut.edu.cn, melhxie@scut.edu.cn, haizhou.li@nus.edu.sg

Abstract

A listener listens to one speech stream at a time in a multi-speaker scenario. EEG-based auditory attention detection (AAD) aims to identify to which speech stream the listener has attended using EEG signals. The performance of linear modeling approaches is limited due to the non-linear nature of the human auditory perception. Furthermore, the real-world applications call for low latency AAD solutions in noisy environments. In this paper, we propose to adopt common spatial pattern (CSP) analysis to enhance the discriminative ability of EEG signals. We study the use of convolutional neural network (CNN) as the non-linear solution. The experiments show that it is possible to decode auditory attention within 2 seconds, with a competitive accuracy of 80.2%, even in noisy acoustic environments. The results are encouraging for brain-computer interfaces, such as hearing aids, which require real-time responses, and robust AAD in complex acoustic environments.

Index Terms: Auditory attention detection (AAD), convolutional neural networks (CNN), electroencephalogram (EEG), common spatial pattern (CSP).

1. Introduction

Just like computers, humans also have limited bandwidth and processing power when listening. However, humans have developed the ability to pay selective attention to one of the speech streams in a multi-speaker environment, or "cocktail party scenario"[1]. This is still a highly non-trivial task for machines. For example, hearing aids are not able to follow a target speaker in the presence of noise and other competing speech streams [2].

To enhance the listening experience of hearing prostheses users, many previous studies focused on reducing background noise and increasing speech intelligibility [3, 4, 5, 6]. Even if a good speech separation is available, selection of the attended speaker is still a fundamental problem in a cocktail party environment, that motivates us to look into brain-computer interface. Can our brains inform the hearing aids which speech stream we would like to pay attention to?

Recent studies have demonstrated that selective attention in a cocktail party scenario can be decoded using recordings of brain activity, such as magnetoencephalography (MEG)[7], EEG [8, 9] and electrocorticographic (ECoG) [10]. The study of auditory attention detection (AAD) helps us understand the human auditory processing, and auditory attention detection can become an important function of hearing aids devices in the future[11].

Among different measures of cortical activity, EEG is a realistic option for BCI applications [12], because it's a cheaper, non-invasive solution, and it is easy to use. There have been successful prior studies on auditory attention detection using EEG data. The stimulus reconstruction [13, 14] is a typical technique. Studies based on linear stimulus reconstruction [8, 9, 15, 16] have verified the feasibility of reliable AAD using EEG signals. However, Faure et al. [17, 18] pointed out that the human auditory system is inherently non-linear, therefore, linear methods are not the best to model the complex and dynamic nature of the brain.

Another limitation of the stimulus reconstruction technique is that AAD performance depends on the duration of the evaluated trial. A longer trial duration leads to better performance. The trial duration is an absolute delay on top of the time required for computing [19, 20]. The temporal resolution of existing approaches for reliable attention decoding is in the order of 30 seconds without counting the computing time, while humans are able to switch attention from one speaker to another at a temporal resolution of around 1 second [21]. In the case of hearing aids, a longer trial duration means a longer delay occurs for the brain to inform the device about the switching of auditory attention. The real-world BCI systems, such as hearing aids, video conferencing systems, call for real-time implementation of auditory attention detection.

With the advent of deep learning in computer vision and speech processing, neural networks provide us an effective way to understand the complex and highly non-linear nature of auditory processes in human brain. Taillez et al. [22] firstly investigated whether machine learning methods can improve the performance of AAD. Recently, CNN-based models have been applied in AAD [23, 24, 25, 26] and demonstrated from EEG signals collected in acoustically controlled environments with two-talkers and without background noise. However, it is unclear how this performs in more realistic environments, i.e., with background noise.

This paper aims to investigate whether it is possible to realize low latency AAD in noisy environments. We propose the use of common spatial pattern (CSP) for EEG signal enhancement under different auditory attention, and a convolutional neural network as a classifier. Common spatial pattern [27] is a specifically designed spatial filter that constructs very few new time-series whose variances contain the most discriminative information. Considering that CSP is an effective method for decoding oscillatory EEG data and shows good performances in some BCI systems [28, 29], we would like to study the effect of CSP in AAD task.

The rest of the paper is organized as follows. Section 2 presents the CSP algorithm for EEG classification and the CNN

http://dx.doi.org/10.21437/Interspeech.2020-2496

model for AAD. Experimental setup and the results are introduced in Section 3. Conclusions are drawn in Section 4.

2. Low Latency Auditory Attention Detection

Auditory attention detection is usually formulated as a binary classification problem in a two-speaker scenario [24, 25, 26]. Given a multi-channel EEG signal, and two single-speaker speech streams, we hope to detect which speech stream the EEG signal is associated with.

A typical EEG-based AAD system consists of a signal processing front-end, that is followed by a backend classifier. Stimulus reconstruction is the basic theory for EEG-based AAD, in which cortical responses are used to approximate the envelope of the speech stream heard by the participant, that is then compared with the original speech stimulus to reveal the attended or unattended speaker in a cocktail party scenario.

Previous studies have demonstrated that both the temporal resolution [19, 20, 30] and the acoustic scenes [23] have an impact on the AAD accuracy. Specifically, AAD accuracy depends on the length of the decision window, which means how much EEG data are needed to make a decision. On the other hand, noise has adverse affects on the attended speech representation in the neural responses, which leads to significant decline in AAD performance. The previous studies prompt us to study how to improve ADD in noisy acoustic environments and with low latency.

To address the research problems, we propose to apply common spatial pattern analysis to perform spatial enhancement of original EEG signals, which is of low signal-to-noise ratio. At the same time, we apply an auditory-inspired linear filter bank and power-law compression to improve the speech envelope extraction process. Finally, a CNN-based decoder is developed as the binary classifier, as shown in Figure 1.

2.1. CSP for EEG Enhancement

CSP is a spatial feature enhancement algorithm for binary classification problems, which can be employed to extract spatial distribution components of two classes [27, 28, 29]. In our AAD method, we attempt to find an optimal spatial filter for each subject with diagonalization calculation to project EEG signals into a new feature space and maximize the variance between the classes. Then the features with higher discrimination are obtained.

Suppose that we have two EEG signals for two opposite classes of auditory attention, G_1 and G_2. Then G_1 and G_2 are multi-channel evoked response matrix with $N \times T$ dimension, where N is the number of channels and T is the number of samples collected from each channel. The mixed covariance matrix with eigenvalue decomposition of the two classes is:

$$\begin{aligned} C &= C_1 + C_2 \\ &= \frac{G_1 G_1^T}{tr(G_1 G_1^T)} + \frac{G_2 G_2^T}{tr(G_2 G_2^T)} \\ &= U \lambda U^T \end{aligned} \tag{1}$$

where C_1 and C_2 is the covariance matrices of G_1 and G_2, $tr(\cdot)$ is sum of elements on the main diagonal of a matrix as the trace of the matrix, λ is the a diagonal matrix of eigenvalues and U is the corresponding eigenvector. $P = \sqrt{\lambda^{-1}} U^T$ is used for transformation to obtain the decomposition:

$$TM_1 = PC_1P^T = E_1 \lambda_1 E_1^T \tag{2}$$

$$TM_2 = PC_2P^T = E_2 \lambda_2 E_2^T \tag{3}$$

where the eigenvectors E_1 and E_2 of TM_1 and TM_2 are equal with $\lambda_1 + \lambda_2 = I$. Then we get the projection matrix $W = E^T P$ with the eigenvectors from the decomposition. The feature after spatial filtering can be expressed as:

$$F = W \times G \tag{4}$$

In this way, we increase the separation between different classes of EEG data, that is expected to improve the classification.

2.2. CNN for Auditory Attention Detection

Convolutional neural networks (CNN) make use of 'convolution' and 'pooling' techniques to reduce a large amount of input data into their essential features, and uses those features for classification. In auditory attention detection, we build a CNN classifier, that takes the envelopes of two speech streams and the EEG features as the input, and decides which speech stream is associated with the EEG features in a binary decision.

In this study, we have the 2 speech streams coming from a male and a female speaker. The envelopes of stimulus were sorted by gender. The male speaker's speech envelope is always at the top row of the input matrix, while the female stimulus envelope is at the bottom row. As a result, the input matrix has 66 rows (64 EEG channels and 2 stimulus envelopes), as shown in Figure 1. The EEG data was presented in black and the envelopes of stimulus in blue. Speaker A and Speaker B are male and female speaker, respectively. 'H-LP' represents speech processing, which is described in detail in Section 3.2.

We study two contrastive implementation, one is with CSP analysis, that is called CSP+CNN, and another is without CSP, that is called CNN. During training, the network is optimized to predict the correct label, i.e. 0 or 1, that represents the attended speaker.

We adopt the same CNN architecture for both CSP+CNN and CNN systems. The CNN architecture includes a convolution layer $[66 \times 9]$, an average pooling and two fully-connected layers (Input:10, hidden:10, output:2). The activate function is rectifying linear unit (ReLu) and the loss function is the weighted cross-entropy. Considering that mapping coefficients comprise salient peaks at a particular lag to the stimulus, a lag is added between the stimulus envelopes and EEG data. According to Ding et al.[7], effects of auditory attention are most distinguishable in the M/EEG signals starting 100 post-stimulus. Therefore, stimulus envelopes were shifted in time 7 samples with respect to the EEG data, corresponding to a 100 ms time-lag at the 70 Hz sampling rate. Finally, We use the SGD optimization with a learning rate of 0.1 to train the networks.

3. Experiments and Results

3.1. Experimental Setup

We evaluated the methods on the 'EEG and audio dataset for auditory attention decoding' dataset [31, 32], which contains EEG signals from 18 normal-hearing subjects listening to one of two competing speakers. 64-channel EEG was recorded using a BioSemi ActiveTwo system (Biosemi, Amsterdam, The Netherlands) at a sampling rate of 512Hz. The electrodes were placed

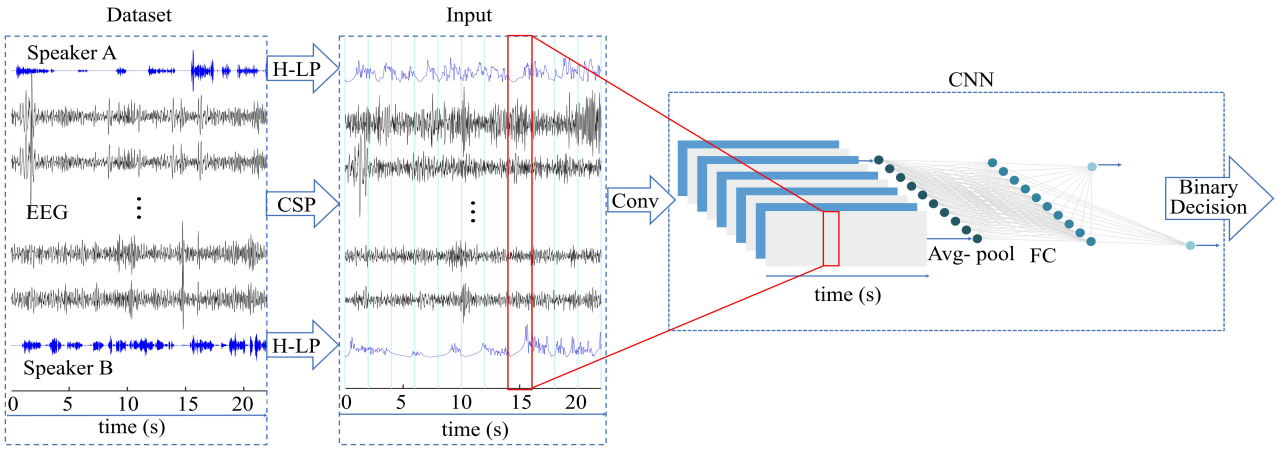

Figure 1: *The proposed CSP+CNN network for auditory attention detection. The convolution is highlighted in red. The network is trained to output two values, i.e. 0 and 1, to indicate the attended speaker.*

on the head according to the international 10/20-standard. Two additional electrodes were placed on the mastoids as physiological reference signals.

Each subject listened to sixty trials in which they were presented the 50 seconds of speech mixtures and instructed to attend to one particular speaker. Although silent gaps exceeding 0.5 second were truncated in some studies [8, 9, 15, 16, 23] to minimize the fluctuation of the subjects' attention, this dataset did not shorten silent periods and presented more realistic speech stimuli.

The auditory scenes comprises a male and a female simultaneously speaking in simulated rooms with different degrees of reverberation. Specifically, three different types of acoustic condition, including no noise, mild reverberation and high reverberation, were independently randomized across trials for each subject. In all trials, two concurrent speech streams were mixed with equal root-mean-square values of sound amplitude, presented roughly at a 65 dB sound pressure level (SPL). Besides these two target speakers, 6 additional speakers (3 male, 3 female) were simulated in the reverberant scenarios. The clarity, defined as the ratio of the direct 80-ms sound energy to the remaining energy, according to Fuglsang et al. [31], mild reverberation ranges between $C_{80,63Hz} = 5.7dB$ and $C_{80,63Hz} = 7.4dB$, and high reverberation ranges between $C_{80,63Hz} = 6.7dB$ and $C_{80,63Hz} = 9.7dB$.

3.2. Data Processing

The first step of data processing was to filter out 50 Hz line noise and harmonics in EEG data[33] and remove the eye artifacts using joint decorrelation framework [34]. Then, all EEG data was re-referenced to the average response of the mastoid electrodes and subsequently filtered offline with a band-pass filter between 2 and 32 Hz. The envelopes of the speech stimuli were passed through a gammatone filterbank with a range of 150 to 4,000 Hz and all of the sub-bands were power-law compression with 0.6 [9]. The speech envelope was then transformed into its respective absolute envelope by a Hilbert transformation, low-pass filtered with 32 Hz and downsampled from 512 Hz to 70 Hz to match the EEG data [22, 23], denoted as H-LP in Figure 1. Finally, envelopes of speech stimuli were normalized. EEG data was also normalized for each trial and spatial filtered by the CSP algorithm in which the training set of each subject was used to obtain a projection matrix.

The data set was randomly split into a training set (80%) and a validation set (10%), and a test set (10%). For all trails, decision windows of 1s, 2s, and 5s (separate experiments) with 50% overlap were used to cut the EEG data and the envelope of the speech stimuli into several segments. It is noted that all the repetitions windows were discarded to keep training, validation, and test set independent.

3.3. Experiment Results

The networks were trained in three different scenarios, no noise, mild reverberation and high reverberation. We report the performance of AAD for 3 different window sizes: 1 second, 2 seconds and 5 seconds in terms of the percentage of correctly classified decision windows.

Generally, CSP+CNN model outperforms CNN system with an average improvement of 9% in AAD accuracy in all the testing scenarios. As shown in Figure 2, accuracy for 2-second decision window was significantly different between the CSP+CNN and CNN models (paired t-test: no noise, $P < 0.001$; mild reverberation, $P = 0.013$ and high reverberation, $P = 0.002$).

We present AAD performance of CSP+CNN model among 18 subjects for 2-second decision window in three different acoustic conditions, as shown in Figure 3. High detection accuracy was obtained in no noise scenario (mean:81.2%, SD:7.8%), followed by high reverberant scenario (mean:80.8%, SD:9.3%) and mild reverberant scenario (mean:77.6%, SD:9.5%). Differences in the AAD accuracy of different acoustic conditions were tested for statistical significance using a paired t-test. Statistical analysis was performed using IBM SPSS statistics software (ver. 24.0, IBM Corp., Armonk, NY, USA) and a level of significance of 0.05 was selected. It is noted that there were no statistical differences between the different acoustic conditions, which is consistent with previous study[31]. This results verify that our method can detect the attended speech stimuli accu-

rately and remained robust in realistic acoustic environments.

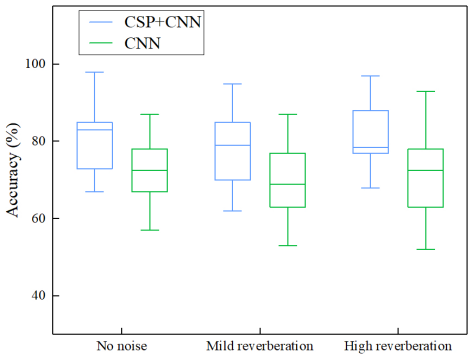

Figure 2: *AAD accuracy of the proposed CNN model and CSP+CNN networks with 2-second decision window.*

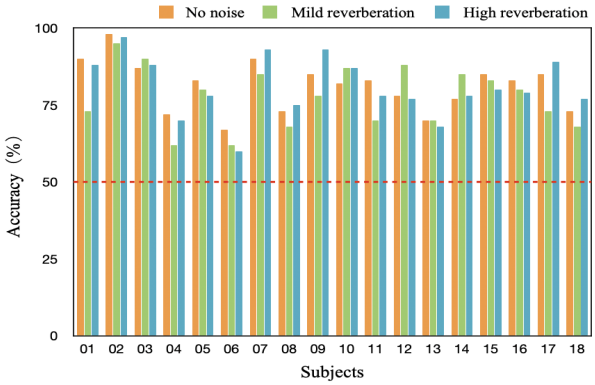

Figure 3: *AAD accuracy for 2-second decision window among all subjects under different acoustic conditions.*

Fuglsang et al. [31] and Wong et al. [32] reported results on the same 'EEG and audio dataset' with linear stimulus reconstruction model, that we use in Table 1 as a benchmarking reference. In [31], neural response was linearly mapped to the speech features based on temporal response functions. In [32], the regularization technique was used to optimize the generalizability of the linear mapping model. With 30-second of decision window, the linear model achieves the accuracy of 81% and 83% for two test conditions, while CNN+CSP achieves a higher AAD accuracy of 86.5%. We only use the 30-second case as a point of reference for comparison as this paper is focused on low latency AAD.

Our CSP+CNN approach achieves significant improvements over the linear models in low latency AAD, which makes our model more promising for BCI systems, such as hearing aids. Due to different experimental setup, we can't directly compare with the state-of-the-art CNN models [23, 24, 25, 26]. Therefore, we re-implement the CNN with our experiment setup and present the AAD accuracy for various decision windows, from 1-second to 5-second, in high reverberant environment, as shown in Figure 4. Accuracy of both CSP+CNN and CNN models drops as detection window size decreases, which was observed in some previous studies [23, 25]. However, for 10 out of 18 subjects, AAD accuracy was above 80%, even for 1-second of decision window using CSP+CNN model. We also

Table 1: *Auditory attention detection accuracy (%) in a comparative study of different models on the same 'EEG and audio dataset for auditory attention decoding' dataset. Linear Model (O) denotes the setting in [31], while Linear Model (R) denotes the use of regularization [32].*

Model	Decision window			
	1s	2s	5s	30s
Linear model (O)[31]	52	56	65	81
Linear model (R)[32]	55	61	70	83
CNN	69.2	71.2	71.9	—
CSP+CNN	78.6	80.2	82.1	86.5

note that the accuracy of CSP+CNN is significantly higher than CNN (paired t-test: 1-second of decision window, P = 0.001; 2-second of decision window, P = 0.001 and 5-second of decision window, P = 0.012).

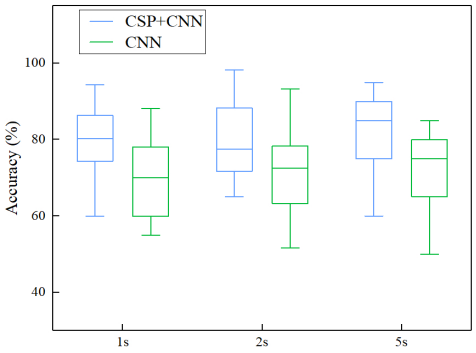

Figure 4: *AAD accuracy for different decision windows among all subjects under high reverberant condition.*

4. Conclusions

In this paper, we propose a model with combination of CSP and CNN for auditory attention detection in real-life acoustic environments. Experiments show that our proposed method not only achieves the better performance than the conventional linear model, but also outperforms the current state-of-the-art CNN models. Moreover, we also find the the proposed network performs well in low latency settings when operating in noisy environments. Given that more than two target speakers might be encountered in a cocktail party, we will explore the feasibility of our proposed model in a multi-speaker scenario.

5. Acknowledgements

This work is supported by Human-Robot Interaction Phase 1 (Grant No.192 25 00054), National Research Foundation (NRF) Singapore under the National Robotics Programme; AI Speech Lab (Award No. AISG-100E-2018-006), NRF Singapore under the AI Singapore Programme; Human Robot Collaborative AI for AME (Grant No. A18A2b0046), NRF Singapore. The work by H. Li is also partly funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (University Allowance, EXC 2077, University of Bremen, Germany).

6. References

[1] 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.

[2] K. Chung, "Challenges and recent developments in hearing aids: Part i. speech understanding in noise, microphone technologies and noise reduction algorithms," *Trends in Amplification*, vol. 8, no. 3, pp. 83–124, 2004.

[3] D. Wang, "Deep learning reinvents the hearing aid," *IEEE spectrum*, vol. 54, no. 3, pp. 32–37, 2017.

[4] C. Xu, W. Rao, X. Xiao, E. S. Chng, and H. Li, "Single channel speech separation with constrained utterance level permutation invariant training using grid lstm," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6–10.

[5] M. Zhang, J. Wu, Y. Chua, X. Luo, Z. Pan, D. Liu, and H. Li, "Mpd-al: an efficient membrane potential driven aggregate-label learning algorithm for spiking neurons," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 1327–1334.

[6] M. Zhang, X. Luo, Y. Chen, J. Wu, A. Belatreche, Z. Pan, H. Qu, and H. Li, "An efficient threshold-driven aggregate-label learning algorithm for multimodal information processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 592–602, 2020.

[7] 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.

[8] 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.

[9] 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.

[10] K. Dijkstra, P. Brunner, A. Gunduz, W. Coon, A. Ritaccio, J. Farquhar, and G. Schalk, "Identifying the attended speaker using electrocorticographic (ecog) signals," *Brain-Computer Interfaces*, vol. 2, no. 4, pp. 161–173, 2015.

[11] B. Shinn-Cunningham, V. Best, A. Kingstone, J. Fawcett, and E. Risko, "Auditory selective attention," *The handbook of attention*, vol. 99, 2015.

[12] A. K. Lee, E. Larson, R. K. Maddox, and B. G. Shinn-Cunningham, "Using neuroimaging to understand the cortical mechanisms of auditory selective attention," *Hearing research*, vol. 307, pp. 111–120, 2014.

[13] W. Bialek, F. Rieke, R. D. R. Van Steveninck, and D. Warland, "Reading a neural code," *Science*, vol. 252, no. 5014, pp. 1854–1857, 1991.

[14] G. B. Stanley, F. F. Li, and Y. Dan, "Reconstruction of natural scenes from ensemble responses in the lateral geniculate nucleus," *Journal of Neuroscience*, vol. 19, no. 18, pp. 8036–8042, 1999.

[15] B. Mirkovic, S. Debener, M. Jaeger, and M. De Vos, "Decoding the attended speech stream with multi-channel eeg: implications for online, daily-life applications," *Journal of neural engineering*, vol. 12, no. 4, p. 046007, 2015.

[16] 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 Transactions on Biomedical Engineering*, vol. 64, no. 5, pp. 1045–1056, 2016.

[17] 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 Sciences-Series III-Sciences de la Vie*, vol. 324, no. 9, pp. 773–793, 2001.

[18] H. Korn and P. Faure, "Is there chaos in the brain? ii. experimental evidence and related models," *Comptes rendus biologies*, vol. 326, no. 9, pp. 787–840, 2003.

[19] R. Zink, S. Proesmans, A. Bertrand, S. Van Huffel, and M. De Vos, "Online detection of auditory attention with mobile eeg: closing the loop with neurofeedback," *BioRxiv*, p. 218727, 2017.

[20] S. Geirnaert, T. Francart, and A. Bertrand, "An interpretable performance metric for auditory attention decoding algorithms in a context of neuro-steered gain control," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2019.

[21] R. Zink, A. Baptist, A. Bertrand, S. Van Huffel, and M. De Vos, "On-line detection of auditory attention in a neurofeedback application," in *Proc. 8th International Workshop on Biosignal Interpretation*, 2016, pp. 1–4.

[22] 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*, 2017.

[23] N. Das, A. Bertrand, and T. Francart, "Eeg-based auditory attention detection: boundary conditions for background noise and speaker positions," *Journal of neural engineering*, vol. 15, no. 6, p. 066017, 2018.

[24] L. Deckers, N. Das, A. H. Ansari, A. Bertrand, and T. Francart, "Eeg-based detection of the attended speaker and the locus of auditory attention with convolutional neural networks," *bioRxiv*, p. 475673, 2018.

[25] 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, pp. 1–10, 2019.

[26] 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," *bioRxiv*, p. 475673, 2020.

[27] H. Ramoser, J. Muller-Gerking, and G. Pfurtscheller, "Optimal spatial filtering of single trial eeg during imagined hand movement," *IEEE transactions on rehabilitation engineering*, vol. 8, no. 4, pp. 441–446, 2000.

[28] H. Lu, H.-L. Eng, C. Guan, K. N. Plataniotis, and A. N. Venetsanopoulos, "Regularized common spatial pattern with aggregation for eeg classification in small-sample setting," *IEEE transactions on Biomedical Engineering*, vol. 57, no. 12, pp. 2936–2946, 2010.

[29] D. Wu, J. T. King, C. H. Chuang, C. T. Lin, and T. P. Jung, "Spatial filtering for eeg-based regression problems in brain-computer interface (bci)," *IEEE Transactions on Fuzzy Systems*, vol. 26, no. 2, pp. 771–781, 2017.

[30] S. Miran, S. Akram, A. Sheikhattar, J. Z. Simon, T. Zhang, and B. Babadi, "Real-time tracking of selective auditory attention from m/eeg: A bayesian filtering approach," *Frontiers in neuroscience*, vol. 12, p. 262, 2018.

[31] 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.

[32] D. D. Wong, S. Fuglsang, J. Hjortkjær, E. Ceolini, M. Slaney, and A. de Cheveigné, "A comparison of temporal response function estimation methods for auditory attention decoding," 2018.

[33] A. de Cheveigné and D. Arzounian, "Robust detrending, rereferencing, outlier detection, and inpainting for multichannel data," *NeuroImage*, vol. 172, pp. 903–912, 2018.

[34] A. de Cheveigné and L. C. Parra, "Joint decorrelation, a versatile tool for multichannel data analysis," *Neuroimage*, vol. 98, pp. 487–505, 2014.