



An Improved Stimulus Reconstruction Method for EEG-Based Short-Time Auditory Attention Detection

Kai Yang¹, Zhuo Zhang¹, Gaoyan Zhang¹(✉), Unoki Masashi², Jianwu Dang^{1,2}, and Longbiao Wang¹

¹ Tianjin Key Laboratory of Cognitive Computing and Application, College of Intelligence and Computing, Tianjin University, Tianjin, China {kai_y, zhanggaoyan}@tju.edu.cn, zhang-zhuo@ecc.u-tokyo.ac.jp
² Japan Advanced Institute of Science and Technology, Ishikawa, Japan

Abstract. Short-time auditory attention detection (AAD) based on electroencephalography (EEG) can be utilized to help hearing-impaired people improve their perception abilities in multi-speaker environments. However, the large individual differences and very low signal-to-noise ratio (SNR) of EEG signals may prevent the AAD from working effectively across subjects in a short time duration. To address the above issues, this paper firstly used a sparse autoencoder with the same trial constraint (SAE-T) method to extract common features across subjects from EEG signals in a 2-s time window. Then we use a CNN-based speech temporal amplitude envelopes (TAEs) reconstruction model for attention detection by comparing the reconstructed accuracy of attended with unattended speech, and the time delay and segmented SAE-T features were also considered in the model. Moreover, the dataset we used has no directional information of speech, which can train a more general model for practical application. Experimental results show that the proposed method can achieve AAD detection accuracy to 86.31%, higher than the method of removing time delay or segmented SAE-T features.

Keywords: Auditory Attention Detection · EEG · Speech Stimulus Reconstruction · Short-time · Cross-subject

1 Introduction

According to the results of the latest Global Burden of Disease (GBD) study, the burden of hearing loss due to aging has increased over time, and the demand for hearing aids has thus increased globally [1]. The general picture that has emerged is that hearing-impaired elderly people spend most of their time wearing hearing aids in favorable listening conditions, such as in quiet or speech-moderate environments, rather than in noise or speech with noise scenarios [2]. Traditional hearing aids have been improved somewhat by the use of a built-in microphone to reduce the noise and a beamforming method to enhance the speech of a specific

speaker [3]. However, this approach is not suitable for situations typified by competitive speech, because in such cases, it is impossible to distinguish which is the target speech and which is the background noise.

To resolve this issue, some studies have analyzed electrophysiological signals by using electroencephalography (EEG) or magnetoencephalography (MEG) to detect the auditory attention of listeners [4, 5], which is called auditory attention detection (AAD) [6]. These studies were based on previous findings that persistent neural excitability oscillations can modulate responses and affect perceptual, motoric, and cognitive processes [7], and intrinsic oscillations are entrained by external rhythms, which allows the brain to optimize the processing of predictable events, such as speech. In 2008, Aiken et al. showed that the human auditory cortex either directly follows the speech temporal amplitude envelopes (TAEs) or consistently responds to changes in these envelopes [8]. In the scenario of a cocktail party, selective attention has been found to enhance the cortical entrainment of the focused speech and inhibit synchronization with the ignored speech [6]. AAD can potentially be combined with speech separation for application to smart hearing aids in the brain-computer interface (BCI) field in the future.

Earlier research on the AAD method mainly utilized the multivariate temporal response function (mTRF) method to perform a linear mapping between EEG and attended TAE [9]. Through this linear regression model, EEG can be used to reconstruct TAE and detect attention by comparing the Pearson correlation between the reconstructed TAE and the two original TAEs. The classification accuracy can reach about 85% within a 60-s time window [5]. However, the detection accuracy dramatically declines when using a subject-independent AAD algorithm [10]. Deep learning technology has been increasingly used in the field of BCI. When applied in EEG signal processing, it has shown excellent automatic feature extraction ability and a competitive decoding performance [11]. Some studies have used deep neural network (DNN) models to reconstruct the TAE of the attended speech. The cross-subject detection accuracy in a 2-second time window is about 67.8% [12]. In addition to stimulus reconstruction, many studies also used direct classification of EEG signals by utilizing orientation information. Some researchers have used the convolutional neural network (CNN) to perform AAD within a smaller time window (1–2 s) and found that the within-subject decoding accuracy was increased to about 80% [13]. Compared with the above-mentioned within-subject AAD studies, a recent study performed cross-subject AAD in a 2-s time window using a multi-task learning model, in which the direct AAD classification task was assisted by the TAE reconstruction task, and the results showed an AAD accuracy of 82% [14].

However, these studies come with several problems and challenges. First, most of them had success primarily with the within-subject AAD performance, while the cross-subject accuracy remained unclear or not good because of the large inter-subject difference. Therefore, we need an effective method for extracting the common features among subjects in order to improve the cross-subject AAD accuracy. Second, most of these previous studies used data with orientation

information, such as the classic binaural listening experiment. In such cases, the AAD may be affected by not just the attention of the audio but also that of the direction. Therefore, in the present study, we use a data set without direction information and build a general algorithm that can achieve real-time attention decoding in cross-subject situations.

In this work, we firstly developed a sparse autoencoder with the same trial constraint (SAE-T) method to preprocess the data before AAD training. This method extracts the common features of EEG across subjects and reduces the dimensionality of input samples. Secondly, we developed a CNN-based segmented reconstruction model and reconstructed the attended TAE to detect the attention in a 2-s time window. The response delay and the two original TAEs were also considered to assist reconstruction of TAE. The segmented input makes the size of the model smaller, which improves the training efficiency.

2 Proposed Method

2.1 Sparse Autoencoder with the Same Trial Constraint (SAE-T) Method for Extracting Common Features

Due to the large individual difference of subjects and the low SNR of EEG signals, the cross-subject accuracy of AAD is low. Therefore, we propose a SAE-T method to extract common features between subjects and further reduce the noise and dimension of EEG signals. The autoencoder has the characteristics of good noise reduction and dimensionality reduction, so it has been increasingly applied to the EEG features and achieved good results [15]. The autoencoder is an unsupervised learning model, where the distribution of the number of neurons is symmetrical between layers, usually decreasing first and then increasing layer by layer, simulating the process of encoding and then decoding. The number of neurons in the final output layer of the autoencoder is the same as that of the input layer. Usually, the autoencoder uses the mean square error (MSE) of the output layer and the input layer as the cost function for training, and the output of the intermediate layer is used as the result of encoding. The sparse autoencoder (SAE) increases the sparsity constraint based on the autoencoder. The sparsity constraint makes the expressions passed at each layer as sparse as possible. The principle is similar to the propagation of neurons in the human brain, that is, certain stimuli will only activate some neurons, and most of the remaining neurons are inactivated, so the sparse expressions are usually more effective than other expressions.

The structure of the proposed SAE-T method is shown in Fig. 1. In order to extract the common features from different subjects, we added the average signals of different subjects under the same trial as a constraint when training the SAE because all subjects listened to the same speech stimuli. Under the training of the autoencoder, the reconstructed sample is not only close to itself but also close to the signals of other subjects in the same trial. The SAE-T is trained by minimizing the Pearson correlation between the input sample X and the reconstructed sample X' , the Pearson correlation between the reconstructed

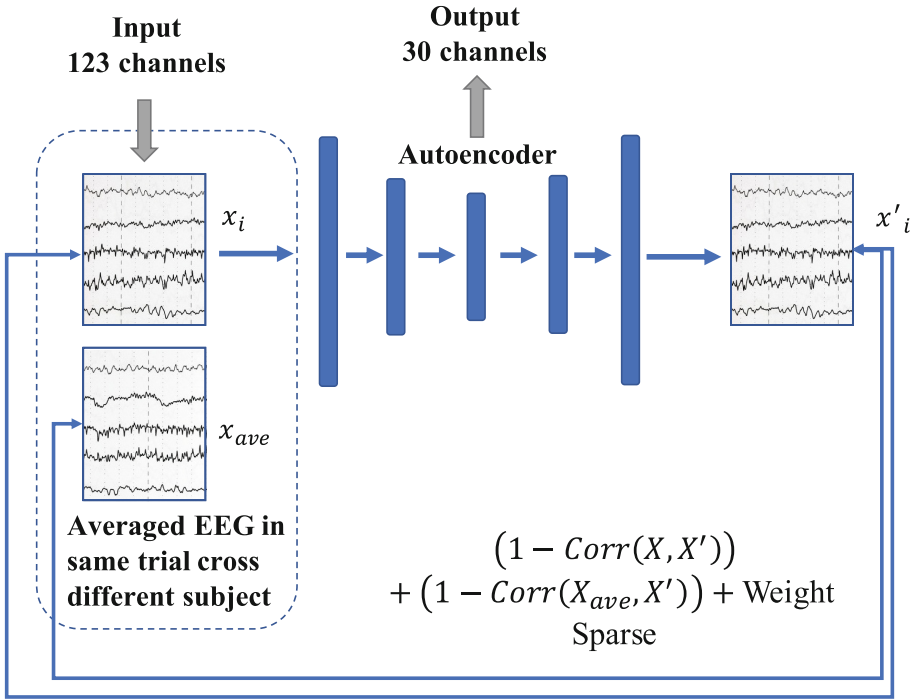


Fig. 1. Proposed SAE-T for reducing the impact of cross-subject EEG variations.

sample X' and the mean of other original samples X_{ave} in the same class, and the sparsity constraints. The cost function is shown in formula (1). $Corr()$ represents the Pearson correlation between two sets of signals. ρ is the sparse parameter, and $\hat{\rho}_j$ means the average activation of the hidden layer. On the basis of correlation as a loss function, we added regularization and added KL divergence as sparsity constraints.

$$\begin{aligned}
 C(X) &= (1 - Corr(X, X')) + (1 - Corr(X_{ave}, X')) \\
 &+ \frac{\lambda}{2} \sum_{l=1}^2 \sum_i^n \sum_j^m |W_{i,j}^{(l)}| + \beta \sum_{j=1}^m \left(\rho \log \frac{\rho}{\hat{\rho}_j} + (1 - \rho) \log \frac{1 - \rho}{1 - \hat{\rho}_j} \right) \quad (1)
 \end{aligned}$$

In order to reduce the complexity of the training model, we cut the samples by a sliding window across the time points and then input the window blocks. The size of the EEG sample is (channel \times timepoint), and each sample is divided into (channel \times 1) blocks along the timepoint through the sliding window. Each window block is used as input and obtained the (channel \times 1) reconstruction block, the reconstruction sample (channel \times timepoint) can finally be obtained.

The SAE-T model we used has four layers and the encoding process has two layers. The number of neurons in each encoding layer is 60 and 30, respectively. The number of neurons in the decoding layer is symmetrical: 30 and 60,

respectively. Therefore, after training through SAE-T, we obtain the ($30 \times$ time-points) data output by the encoding process, which means that we reduce the dimensionality of the EEG data from the number of original channels to 30. The optimizer we use is RMSProp, the learning rate is set to 1×10^{-3} , the decay rate of the learning rate is 0.99, and the sparsity weight is 0.03.

2.2 Proposed Segmented R Model with Added TAEs for AAD

In the AAD task, the two original speech signals, and the EEG data are known. We use the EEG signal to reconstruct the attended speech so we can detect which of the two originals is the attended speech by a higher correlation of reconstructed speech with the original one, which is called the reconstruction model (R model). This model is different from the direction binary classification model (D model) that skips the reconstruction step and performs classification directly.

Generally, when two original speech signals have other features, such as different direction information, the D model is quite effective and easy to use [16]. However, for practical applications, if the azimuths of the two sound sources are very close, it may have a negative impact on the detection performance of the D model, but the detection accuracy of the R model will not be affected.

In order to ameliorate the R model to obtain better detection accuracy, we propose the segmented R model with added TAEs method, which is shown in Fig. 2. First, we add the TAE features of the two original speech signals to the EEG. Prior studies have speculated that unattended speech is also processed in the brain [17, 18], so we believe that adding the original attended and unattended TAEs to the training model will help to improve the reconstruction performance. Next, in each sample, we use a sliding window with a certain window length (e.g. 100 ms) and step length of one sampling point to intercept the samples and add them to the model in blocks. This significantly reduces the complexity of the model compared with putting the entire sample in. We use a CNN-based model for training. The learning rate is set to 5×10^{-5} and the Adam optimizer is used. We set up a total of four convolutional layers, and the size of the convolution kernel of each layer is 3×3 . Each convolutional layer is followed by a ReLU activation function and a pooling layer. After the feature is extracted by convolution, there are four fully connected layers, and the number of neurons is 6, 3, 2, and 1, respectively. Every time we gave an input block, it outputs a corresponding timepoint of the reconstruction data. As the sliding window traverses the timepoint of the entire sample, we obtain a reconstruction TAE of length ($1 \times$ timepoint). The cost function is the difference between the Pearson correlation of the attended TAE and the unattended TAE as shown in formula (2). Y' represents reconstructed TAE and the Y_a and Y_u represent original attended TAE and unattended TAE, respectively.

$$C(Y', Y_a, Y_u) = Corr(Y', Y_u) - Corr(Y', Y_a) \quad (2)$$

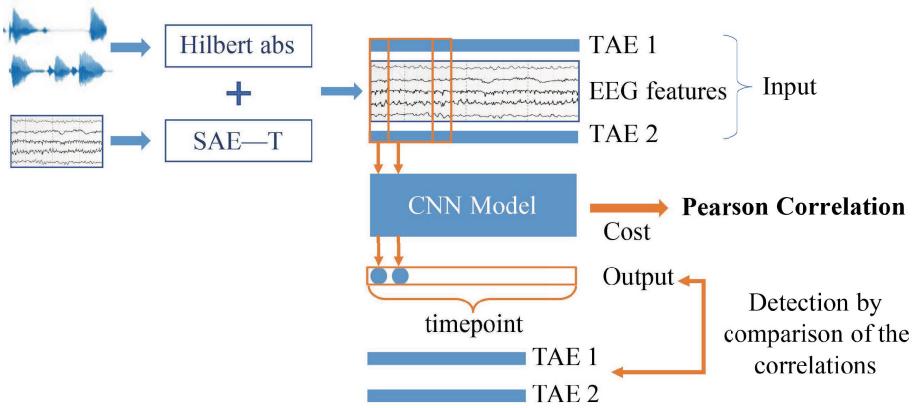


Fig. 2. Proposed Segmented R model with added TAEs method for AAD.

Using this method, we can train a model for reconstructing the attended TAE and then detect which speech the listener wants to focus on by comparing the correlation between the reconstructed TAE and the two original TAEs.

3 Experiments

3.1 Participants

The experimental dataset was collected by ourselves. A total of 21 participants (mean \pm standard deviation age, 21 ± 2.2 years; eleven women, ten men) took part in the study. All participants were undergraduate or graduate students, had normal or corrected to normal vision, normal hearing, and were native people. All were judged to be right-handed after applying the Edinburgh Handedness Inventory [19]. All subjects provided written informed consent to participate in the study and received a corresponding reward for their participation. This study was approved by the Institutional Review Board at Tianjin University before the experiment and was carried out in accordance with the principles of the Declaration of Helsinki [20].

3.2 Experiment and Hardware

The content of the auditory stimulation was short stories by the Japanese writer Shinichi Hoshi translated into the mother language. Auditory stimuli consisted of pre-recorded natural speeches of a male and female broadcaster recorded in a soundproof room with an LCT450 professional microphone and iCon Ultra4 sound card. There were three types of auditory stimuli: forward sequence playback, reverse sequence playback, and mixed playback of two speeches. These three categories of stimuli were played randomly in the experiment, but the continuity and sequence of the story were guaranteed. We only used the mixed audio

sequence, which included 20 trials with a duration of around 60 s per trial. The remaining data were used for other studies [21].

All trials were set to equal root mean square intensities that were considered equally loud. All audios kept the mute segment within 500 ms so as to prevent the attention from other unimportant speech when the target speaker was paused between sentences. We mixed two speeches without applying head-related transfer function to ensure that any attentional effects observed were entirely due to top-down selective attention and not produced by a more general allocation of spatial attention.

The subjects were asked to pay attention to a specific speaker (male or female) during the mixed audio experiment. After each trial, participants answered questions about the content of the listening materials to make sure they were taking the experiment seriously. The number of the trials that participants were asked to attend to female or male speakers are equal. The experiment was carried out in an electromagnetic shielding and soundproof room. We used a 128-channel Quick Cap EEG cap and a Neuroscan system (Neuroscan, USA) to record EEG data at a sampling rate of 1000 Hz. The auditory stimulus was low-pass filtered with a cut-off frequency of 44,100 Hz and presented through Etymotic ER2 air conduction headphones with an electromagnetic shielding box and a JDS LABS power amplifier at 60 dBA, which can reduce electromagnetic interference and ensure that the sound quality is not damaged.

3.3 Data Preprocessing

In order to improve the SNR of EEG data, we preprocessed it through the following steps. We first removed unnecessary electrodes such as electrooculogram (EOG) and picked out the desired EEG channels, for a total of 122. The data was down-sampled 250 Hz and then passed through a 1-Hz highpass filter and a 40-Hz low-pass filter. We then used the Artifact Subspace Reconstruction (ASR) component in EEGLAB toolbox [22] to remove any bad channels of the EEG data and performed an interpolation calculation through the electrode signals around it to obtain the replacement channel. Then, the whole brain signals were averaged as a re-reference. We repeated this process of removing bad channels and replacing them with re-references a total of three times to obtain the best preprocessing effect. Because we added the reference electrode in the preprocessing, after that, we obtained a total of 123 channels of electrode signals. Finally, we performed independent component analysis (ICA) on the EEG data to remove interfering signals such as electromyography.

We downsampled the preprocessed EEG data and stimuli data 100 Hz to make the sample sizes uniform. To achieve real-time auditory attention detection, we cut the preprocessed EEG data into 2-s data samples through a sliding window with an overlap rate of 50%. At the same time, we used the Hilbert method to extract TAE features for the original two audios and used the same cutting method to cut them into corresponding 2-s data. We standardize all data including EEG and TAEs through Z-score.

In total, we used data from 21 participants. The data of three subjects were randomly selected as the test set, and that of the remaining 18 subjects were used as the training set. After data cutting and shuffling, we had a total of 21,834 training set samples and 3639 test set samples. In the training set, the number of samples attending to male speeches is 10,530 and the number of samples attending to female speeches is 11,304. In the test set, the number of samples attending to male and female speeches are 1755 and 1884, respectively.

4 Results and Discussions

We compared the performance of different models on our dataset. We used the CNN-based R model with segmented input as the baseline model. All models have iterated 150 epochs. The EEG sample attended to male speech was assumed to be a positive example, and the EEG sample attended to female speech was assumed to be a negative example. We used accuracy as the evaluation index. Because there were differences between the quantity of the positive and negative samples, we also calculated the true positive rate (TPR), true negative rate (TNR), and F1 score as evaluations by obtaining the confusion matrix of different models. The results are shown in Table 1. First, the segmented R model with added TAEs method was effective. Using the simplest CNN-based R model, or the segmented R model with segmented input, the detection accuracy was very low, close to the chance level. After adding TAEs to the input, the classification accuracy increased from 52.02% to 71.94%. This demonstrates that adding the original TAEs plays a very important role in the reconstruction of the attended TAE.

Table 1. Attention detection performance using different models.

Models	Accuracy	TPR	TNR	F1 score
Segmented R model+CNN	0.5202	0.5231	0.5175	0.5126
Segmented R model with added TAEs+CNN	0.7194	0.5162	0.9087	0.6396
Segmented R model with added TAEs+CNN+SAE-T	0.7593	0.6434	0.8447	0.7109
Segmented R model with added TAEs+CNN+delay	0.9692	0.9452	0.9915	0.9673
Segmented R model with added TAEs+CNN+SAE-T+delay	0.9708	0.9459	0.9942	0.9691
Segmented R model with added TAEs+CNN+SAE-T+delay*	0.8631	0.8541	0.8715	0.8576

* denotes that the model uses a random TAE order for the experiments. In addition, all other models use a fixed female-male TAE order for the experiments.

Second, SAE-T also played an important role in cross-subject detection. It showed a good performance in the CNN model which increased the AAD accuracy from 71.94% to 75.93%. We also performed MSE and correlation analysis

between the reconstructed EEG signal by SAE-T and the original signal and found that SAE-T reduced the MSE of the EEG across different subjects from 1.6215×10^{-2} to 1.0285×10^{-3} . Meanwhile, the Pearson correlation was slightly improved, as shown in Table 2. These results demonstrate that SAE-T helps reduce the difference among subjects.

Table 2. Comparison of MSE and Pearson correlation between reconstructed speech and original ones without or with SAE-T processing.

	Mean square error	Pearson correlation
	(AVE/STD)	(AVE/STD)
No SAE-T	0.01621/0.00010	0.00283/0.00681
SAE-T	0.00103/0.00002	0.00308/0.00743

Third, prior studies have shown that the EEG signal has a time delay of 180 ms when tracking the attended TAE [9]. We adjusted the original unit length of the sliding window from 10 ms to 180 ms to take into account the time delay. After taking into account the time delay, the model detection accuracy increased from 75.93% to 97.08%, which is a significant improvement.

Finally, to make our model more convenient for practical application, we examined the detection accuracy of the cases in which the male and female TAEs were put into a specific order and in random order also as shown in Table 1. In the former case, the female and male TAEs were always placed in the first and last rows of input data. However, it has a problem in that it is necessary to classify the separated TAEs by gender in advance, which increases the complexity of AAD. Therefore, we tried to randomly add these two TAEs to the first and last rows, respectively, when assuming that the gender features of the two TAEs were not known. Although the detection accuracy was reduced to 86.31% when using random order of TAEs, it is still high and more suitable for practical applications.

5 Conclusions and Future Work

In this paper, we proposed a novel approach for short-time auditory attention detection based on EEG signals. First, we proposed the SAE-T method to extract the common feature to reduce the impact of inter-subject differences and compress data dimensions. Then we proposed an AAD model, which segments the data into a short time window, and added speech TAEs and EEG-speech delay into a CNN model. Results showed an accuracy of 86.31%. Finally, in contrast to the datasets in previous studies, we used a dataset without information about the spatial locations of speakers, which helps to train a more general model for daily life.

Our study has two shortcomings. First, we did not examine whether our method would also have a good detection performance on datasets with directional information. Second, we only used the mixed speech of one male and one female for the experiment and did not consider the mixed speech from the same gender or more than two speakers. These issues will be resolved in future work.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (No. 61876126 and 61503278).

References

1. Haile, L., Orji, A., Briant, P., Adelson, J., Davis, A., Vos, T.: Updates on hearing from the global burden of disease study. *Innov. Aging* **4**(Suppl 1), 808 (2020)
2. Humes, L.E., Rogers, S.E., Main, A.K., Kinney, D.L.: The acoustic environments in which older adults wear their hearing aids: insights from datalogging sound environment classification. *Am. J. Audiol.* **27**(4), 594–603 (2018)
3. Haykin, S., Liu, K.R.: *Handbook on Array Processing and Sensor Networks*. John Wiley & Sons, Hoboken (2010)
4. Kurmanavičiūtė, D., Rantala, A., Jas, M., Väililä, A., Parkkonen, L.: Target of selective auditory attention can be robustly followed with meg. *bioRxiv* p. 588491 (2019)
5. O’sullivan, J.A., et al.: Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cereb. Cortex* **25**(7), 1697–1706 (2015)
6. Ding, N., Simon, J.Z.: Emergence of neural encoding of auditory objects while listening to competing speakers. *Proc. Natl. Acad. Sci.* **109**(29), 11854–11859 (2012)
7. Thut, G., Miniussi, C., Gross, J.: The functional importance of rhythmic activity in the brain. *Curr. Biol.* **22**(16), R658–R663 (2012)
8. Aiken, S.J., Picton, T.W.: Human cortical responses to the speech envelope. *Ear Hear.* **29**(2), 139–157 (2008)
9. Crosse, M.J., Di Liberto, G.M., Bednar, A., Lalor, E.C.: The multivariate temporal response function (mTRF) toolbox: a MATLAB toolbox for relating neural signals to continuous stimuli. *Front. Hum. Neurosci.* **10**, 604 (2016)
10. Geirnaert, S., et al.: Neuro-steered hearing devices: decoding auditory attention from the brain (2021)
11. Gu, X., et al.: EEG-based brain-computer interfaces (BCIs): a survey of recent studies on signal sensing technologies and computational intelligence approaches and their applications. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **18**, 1645–1666 (2021)
12. de Taillez, T., Kollmeier, B., Meyer, B.T.: Machine learning for decoding listeners’ attention from electroencephalography evoked by continuous speech. *Eur. J. Neurosci.* **51**(5), 1234–1241 (2020)
13. Vandecappelle, S., Deckers, L., Das, N., Ansari, A.H., Bertrand, A., Francart, T.: EEG-based detection of the locus of auditory attention with convolutional neural networks. *Elife* **10**, e56481 (2021)
14. Zhang, Z., Zhang, G., Dang, J., Wu, S., Zhou, D., Wang, L.: EEG-based short-time auditory attention detection using multi-task deep learning. In: *INTERSPEECH*, pp. 2517–2521 (2020)

15. Yao, Y., Plested, J., Gedeon, T.: Deep feature learning and visualization for EEG recording using autoencoders. In: Cheng, L., Leung, A.C.S., Ozawa, S. (eds.) ICONIP 2018. LNCS, vol. 11307, pp. 554–566. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-04239-4_50
16. Deckers, L., Das, N., Ansari, A., Bertrand, A., Francart, T.: EEG-based detection of the attended speaker and the locus of auditory attention with convolutional neural networks. *bioRxiv*. 475673 (2018)
17. Lewis, J.L.: Semantic processing of unattended messages using dichotic listening. *J. Exp. Psychol.* **85**(2), 225 (1970)
18. Vanthornhout, J., Decruy, L., Francart, T.: Effect of task and attention on neural tracking of speech. *Front. Neurosci.* **13**, 977 (2019)
19. Oldfield, R.C.: The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia* **9**(1), 97–113 (1971)
20. Association, W.M., et al.: World medical association declaration of Helsinki. Ethical principles for medical research involving human subjects. *Bull. World Health Organ.* **79**(4), 373 (2001)
21. Zhou, D., Zhang, G., Dang, J., Wu, S., Zhang, Z.: A multi-subject temporal-spatial hyper-alignment method for EEG-based neural entrainment to speech. In: 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 881–887. IEEE (2020)
22. Delorme, A., Makeig, S.: EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* **134**(1), 9–21 (2004)