Simon Geirnaert, Servaas Vandecappelle, Emina Alickovic, Alain de Cheveigné, Edmund Lalor, Bernd T. Meyer, Sina Miran, Tom Francart, and Alexander Bertrand

Electroencephalography-Based Auditory Attention Decoding

Toward neurosteered hearing devices

©SHUTTERSTOCK.COM/TEX VECTOR

Digital Object Identifier 10.1109/MSP.2021.3075932
Date of current version: 28 June 2021

People suffering from hearing impairment often have difficulties participating in conversations in so-called cocktail party scenarios where multiple individuals are simultaneously talking. Although advanced algorithms exist to suppress background noise in these situations, a hearing device also needs information about which speaker a user actually aims to attend to. The voice of the correct (attended) speaker can then be enhanced through this information, and all other speakers can be treated as background noise. Recent neuroscientific advances have shown that it is possible to determine the focus of auditory attention through noninvasive neurorecording techniques, such as electroencephalography (EEG). Based on these insights, a multitude of auditory attention decoding (AAD) algorithms has been proposed, which could, combined with appropriate speaker separation algorithms and miniaturized EEG sensors, lead to so-called neurosteered hearing devices. In this article, we provide a broad review and a statistically grounded comparative study of EEG-based AAD algorithms and address the main signal processing challenges in this field.

Introduction

State-of-the-art hearing devices, such as hearing aids and cochlear implants, contain advanced signal processing algorithms to suppress acoustic background noise and thus assist the constantly expanding group of people suffering from hearing impairment. However, situations where multiple people are simultaneously speaking (dubbed the *cocktail party problem*) still cause major difficulties for a hearing device user, often leading to social isolation and a decreased quality of life. Beamforming algorithms that use microphone array signals to suppress acoustic background noise and extract a single speaker from a mixture of voices lack a fundamental piece of information to assist a hearing device user in cocktail party scenarios: which speaker should be treated as the attended one (i.e., the person whom the user/listener wants to hear) and which other speakers should be regarded as interfering noise sources? This issue is often addressed by using simple

heuristics, for example, by selecting the loudest speaker or by assuming that the attended speaker is in front of the listener. However, in many practical situations, these heuristics will select and enhance a speaker to whom a user is not listening. For instance, when listening to a passenger while driving a car or when listening to a public address system, a selection based on a look direction will fail.

Although advanced algorithms exist to suppress background noise in these situations, a hearing device also needs information about which speaker a user actually aims to attend to.

Recent neuroscientific insights into how the brain synchronizes with the attended speech envelope [1], [2] have laid the groundwork for a new strategy to tackle this problem: extracting attention-related information directly from the origin, i.e., the brain. This is generally referred to as the *AAD problem*. In the past 10 years, following these groundbreaking advances in the field of auditory neuroscience and neural engineering, the topic of AAD has gained traction in the biomedical signal processing community. AAD can be performed through several neurorecording modalities, such as EEG [3], electrocorticography (ECoG) [1], and magnetoencephalography (MEG) [2]. However, the invasiveness of ECoG and the high cost and lack of wearability of MEG limit their applicability in practical hearing devices for daily use. On the other hand, EEG is considered a good candidate to be integrated with hearing devices, as it is a noninvasive, wearable, and relatively cheap neurorecording technique.

In [3], the first successful speech-based AAD algorithm based on unaveraged single-trial EEG data was proposed. The main idea in [3] is to decode the attended speech envelope from a multichannel EEG recording by using a neural decoder and correlating its output with the speech envelope of each speaker. Following this seminal work, many AAD algorithms have been developed [4]–[10]. In combination with effective speaker separation algorithms [11]–[15], and relying on rapidly evolving improvements in the miniaturization and wearability of EEG sensors [16]–[19], these advances could lead to a new assistive solution for the hearing impaired: a neurosteered hearing device. Figure 1 provides a conceptual overview of a neurosteered hearing device in a situation where there are two competing speakers. The AAD block contains an algorithm that determines the attended speaker by integrating the demixed speech envelopes and the EEG.

Despite the large variety of AAD algorithms, an objective and transparent comparative study has not been performed, making it hard to identify which strategies are most successful. We will briefly review different types of AAD algorithms and their most common instances and provide an objective and quantitative comparative survey using two independent, publicly available data sets [20], [21]. This study has been reviewed and endorsed by the authors of the papers in which the algorithms were proposed to ensure fairness and correctness. While the article's main focus is on this AAD block, we also discuss other practical challenges along the road ahead, such as evaluations in more realistic listening scenarios, the interaction of AAD with speech demixing and beamforming algorithms, and challenges related to EEG sensor miniaturization.

Review of AAD algorithms

In this section, we provide a brief overview of various AAD algorithms. Our study includes only papers published before 2020, when this article was conceptualized. However, since this field is quickly progressing and because several papers have appeared during the past year, the reader is encouraged to look up new AAD algorithms (and extensions thereof) and compare them with the presented methods.

For the sake of easy exposition, we assume that there are only two speakers (one attended and one unattended), although all the algorithms can be generalized to more than two. In the remainder of the article, we also make an abstraction of the speaker separation and denoising block in Figure 1 and assume that the AAD block has direct access to the envelopes of the original unmixed speech sources, as is often done in the AAD literature. However, we will briefly return to the combination of both blocks in the "Open Challenges and Outlook" section.

Most AAD algorithms adopt a stimulus reconstruction (SR) approach (also known as *backward modeling* and *decoding*). In this strategy, a multiple-input, single-output (MISO) neural decoder is applied to all EEG channels to reconstruct the attended speech envelope. This neural decoder is pretrained to optimally reconstruct the attended speech envelope from the EEG data while blocking other (unrelated) neural activity. It is in this training procedure that most AAD algorithms differ. The reconstructed speech envelope is afterward

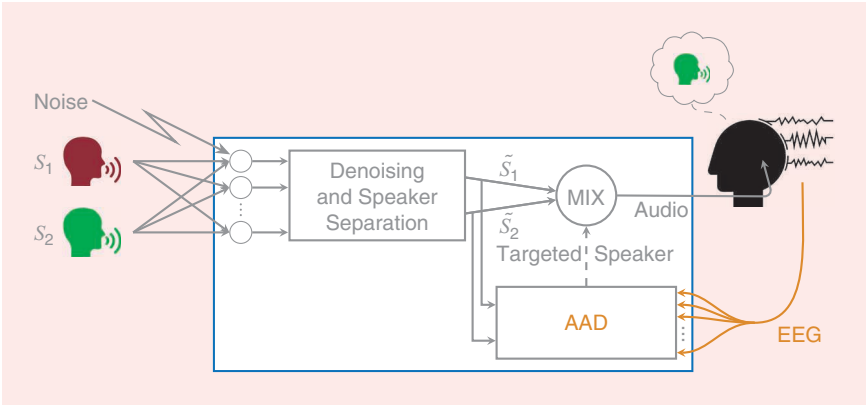

FIGURE 1. A conceptual overview of a neurosteered hearing device when there are two competing speakers. The green speaker (S_2) corresponds to the attended one, while the red speaker (S_1) corresponds to the unattended one.

correlated with the speech envelopes of all speakers, after which the one with the highest Pearson correlation coefficient is identified as the attended speaker. This correlation coefficient is estimated across a window of τ seconds, which is referred to as the *decision window length*, corresponding to the amount of EEG data used in each decision about the attention. Typically, the AAD accuracy strongly depends on the decision window length because the Pearson correlation estimates are very noisy due to the low signal-to-noise ratio of the output signal of the neural decoder.

Alternatively, the neural response in each EEG channel can be predicted from the speech envelopes via an encoder (also known as *forward modeling* and *encoding*) and can then be correlated with the measured EEG [5], [22]. When the encoder is linear, this corresponds to estimating impulse responses (that is, temporal response functions) between the speech envelopes and the recorded EEG signals. For AAD, backward MISO decoding models have been demonstrated to outperform forward encoding models [5], [22], as the former can exploit the spatial coherence across the different EEG channels at its input. In this study, we focus only on backward AAD models, except for the canonical correlation analysis (CCA) algorithm, in the "CCA" section, which combines forward and backward approaches. Due to the emergence of deep learning methods, a third approach has become popular: direct classification [9], [10]. In it, attention is directly predicted in an end-to-end fashion, without explicitly reconstructing the speech envelope.

State-of-the-art hearing devices, such as hearing aids and cochlear implants, contain advanced signal processing algorithms to suppress acoustic background noise.

The decoder models are typically trained in a supervised fashion, which means that the attended speaker must be known for each data point in the training set. This requires collecting "ground-truth" EEG data during a dedicated experiment in which a subject is asked to pay attention to a predefined speaker in a speech mixture. The models can be trained either in a subject-specific fashion (based on EEG data from the actual subject under test) or in a subject-independent approach (based on EEG data from subjects other than the one under test). The latter leads to a universal (subject-independent) decoder, which has the advantage that it can be applied to new subjects without the need to go through such a tedious ground-truth EEG data collection for every new subject. However, since each person's brain responses are different, the accuracy achieved by universal decoders is typically lower [3]. In this article, we consider only subject-specific decoders, which achieve better accuracy, as they are tailored to the EEG of a specific end user. Transfer learning techniques, which are becoming popular in the field of brain–computer interfaces [23], may close the gap between subject-specific and subject-independent models, although this remains to be researched in the context of AAD. Figure 2 reviews and classifies the algorithms included in our study, discriminated based on their fundamental properties. In the following sections, we distinguish between linear and nonlinear algorithms.

Linear methods

All linear methods included in this study, which differ in the features shown in the linear branch of Figure 2, adopt the so-called SR framework [Figure 3(a)]. This boils down to applying a linear time-invariant spatiotemporal filter $d_c(l)$ on the C-channel EEG $x_c(t)$ to reconstruct the attended speech envelope $s_a(t)$:

$$\hat{s}_a(t) = \sum_{c=1}^{C} \sum_{l=0}^{L-1} d_c(l) x_c(t+l), \tag{1}$$

where c is the channel index, ranging from one to C, and l is the time lag index, ranging from zero to $L-1$, with L being the per-channel filter length. The corresponding MISO filter is anticausal as the brain responds to the stimulus, such that only future EEG time samples can be used to predict the current stimulus sample. Equation (1) can be rewritten as $\hat{s}_a(t) = \mathbf{d}^T \mathbf{x}(t)$ by using $\mathbf{d} \in \mathbb{R}^{LC\times 1}$, collecting all decoder coefficients for all time lags and channels, and using $\mathbf{x}(t) = [\underline{\mathbf{x}}_1(t)^{\mathrm{T}} \ \underline{\mathbf{x}}_2(t)^{\mathrm{T}} \ \cdots \ \underline{\mathbf{x}}_C(t)^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{LC\times 1}$, with $\underline{\mathbf{x}}_c(t) = [x_c(t) \ x_c(t+1) \ \cdots \ x_c(t+L-1)]^{\mathrm{T}}$ (the same indexing holds for the decoder $\mathbf{d}$). In the next three sections, we introduce the different linear methods included in this study. The approaches, which are all correlation based, can be extended to more than two competing speakers by simply correlating the reconstructed speech envelope with those of the competing speakers and taking the maximum.

Supervised minimum mean-square error backward modeling

The most basic way of training the decoder, first presented in the EEG-based AAD-context in [3], is by minimizing the mean-square error (MSE) between the actual attended envelope and the reconstructed one. In [4], it is shown that minimizing the MSE is equivalent to maximizing the Pearson correlation coefficient between the reconstructed and attended speech envelopes. Using sample estimates, assuming that there are T samples available, the minimum MSE (MMSE)-based formulation becomes equivalent to the least-squares (LS) formulation:

$$\hat{\mathbf{d}} = \underset{\mathbf{d}}{\operatorname{argmin}} \| \mathbf{s}_{\mathrm{a}} - \mathbf{X}\mathbf{d} \|_2^2, \tag{2}$$

with $\mathbf{X} = [\mathbf{x}(0) \cdots \mathbf{x}(T-1)]^{\mathrm{T}} \in \mathbb{R}^{T\times LC}$ and $\mathbf{s}_{\mathrm{a}} = [s_{\mathrm{a}}(0) \cdots s_{\mathrm{a}}(T-1)]^{\mathrm{T}} \in \mathbb{R}^{T\times 1}$. The normal equations lead to the solution $\hat{\mathbf{d}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{s}_{\mathrm{a}}$. The first factor corresponds to an estimation of the autocorrelation matrix $\hat{\mathbf{R}}_{\mathrm{xx}} = 1/T \times \Sigma_{t=0}^{T-1} \mathbf{x}(t)\mathbf{x}(t)^{\mathrm{T}} \in \mathbb{R}^{LC\times LC}$, while the second one corresponds to the cross-correlation vector $\hat{\mathbf{r}}_{xs_{\mathrm{a}}} = 1/T \Sigma_{t=0}^{T-1} \mathbf{x}(t) s_{\mathrm{a}}(t) \in \mathbb{R}^{LC\times 1}$.

To avoid overfitting, two types of regularization are used in the AAD literature: ridge regression/L_2-norm regularization and L_1-norm/sparse regularization, also known as the *least absolute shrinkage and selection operator* (*LASSO*).

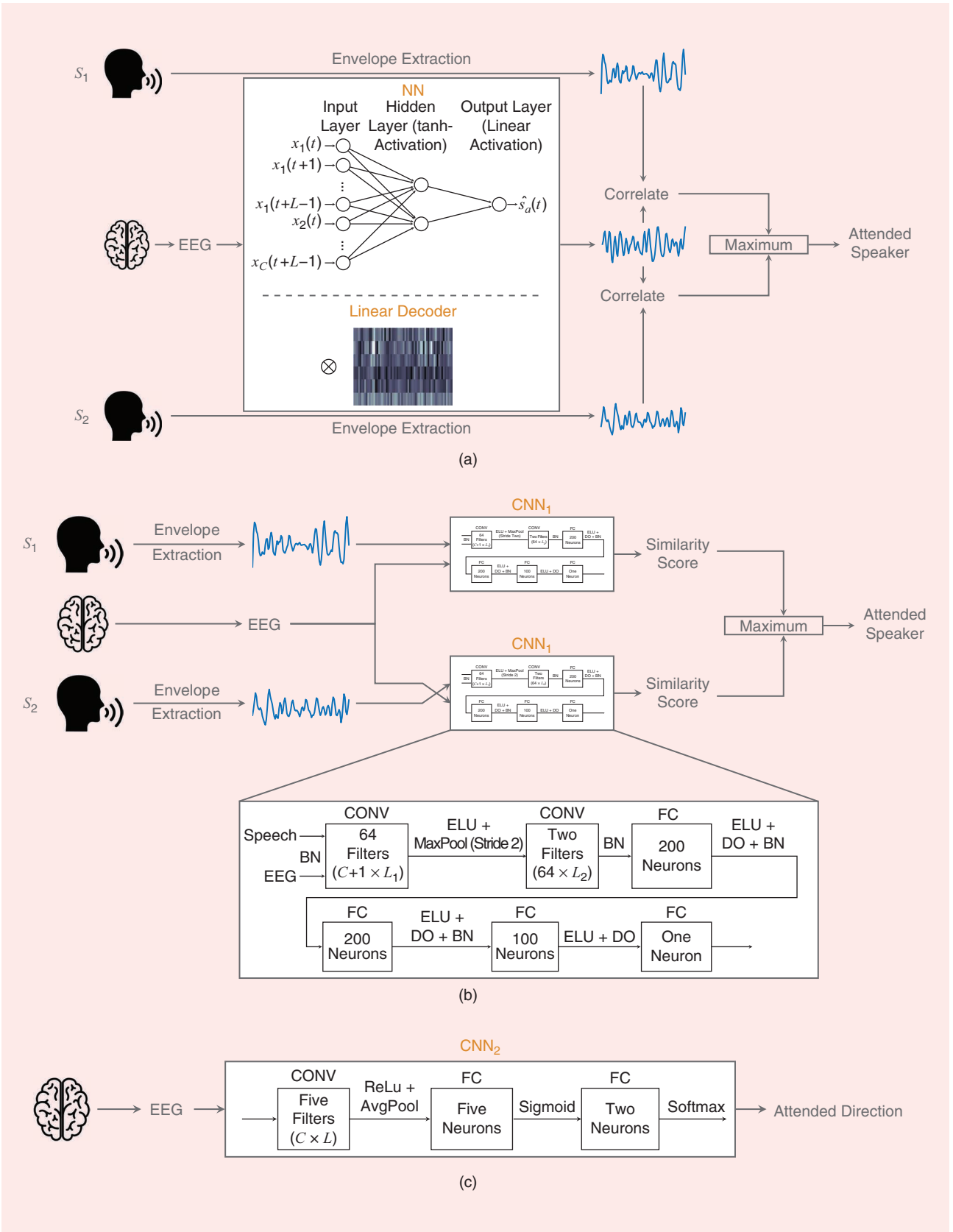

FIGURE 2. The AAD algorithms in this study (except the forward models; see the introduction to the "Review of AAD Algorithms" section) and the planned contrasts in the statistical analysis. An asterisk indicates a significant difference ($p < 0.05$), while (n.s.) designates a nonsignificant difference (see the "Statistical Analysis" section for more details). CNN: convolutional neural network; MMSE: minimum mean-square error; LASSO: least absolute shrinkage and selection operator; CCA: canonical correlation analysis; adap: adaptive; avgdec: averaging decoders; avgcorr: averaging correlation matrices.

The corresponding cost functions are shown in Table 1, where the regularization hyperparameter λ is defined relative to $z = \text{trace}(\mathbf{X}^T\mathbf{X})/LC$ (for ridge regression)/$q = ||\mathbf{X}^T\mathbf{s}_a||_\infty$ (for the LASSO). Similar to [5], we use the alternating direction method of multipliers to iteratively obtain the solution to the LASSO problem. The optimal value λ can be found by using a cross-validation scheme. Other regularization methods, such as Tikhonov regularization, have been proposed as well [22].

Assume a given training set consisting of K data segments of a specific length T. These segments can either be artificially constructed by segmenting a continuous recording (usually for the sake of cross validation) or they can correspond to different experimental trials (potentially from different subjects, e.g., when training a subject-independent decoder). There exist various methods of combining these segments in the process of training a decoder. As suggested in the seminal paper [3], decoders $\mathbf{d}_k$ can be trained per segment k, after which all decoders are averaged to obtain a single, final decoder $\mathbf{d}$. In [4] (also adopted in, e.g., [11], [15], [19], and [24]–[26]), an alternative scheme is proposed, where, instead of separately estimating one decoder per segment, the loss function (2) (with regularization) is minimized across all K segments at once. As evident from the solution in Table 1, this is equivalent to first estimating the autocorrelation matrix and the cross-correlation vector via averaging the sample estimates per segment, whereafter one decoder is computed. It is easy to see that this is mathematically equivalent to concatenating all the data in one big matrix $\mathbf{X} \in \mathbb{R}^{KT\times LC}$ and vector $\mathbf{s}_a \in \mathbb{R}^{KT\times 1}$

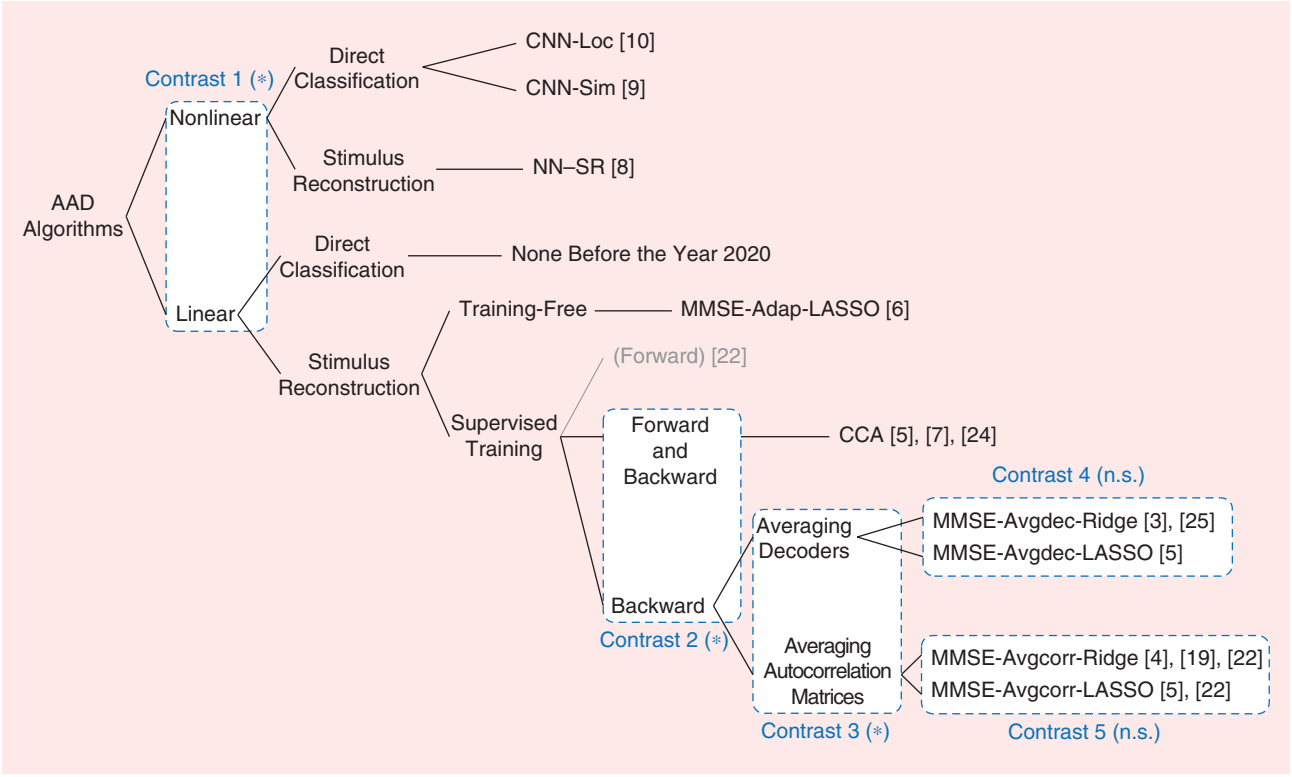

FIGURE 3. A conceptual overview of the different AAD algorithms and the different network topologies of (a) a linear SR decoder and neural network (NN)-SR, (b) convolutional NN (CNN)-sim, and (c) CNN-loc. CONV = convolutional layer, FC: fully connected; BN: batch normalization; ELU: exponential linear unit; ReLu: rectified linear unit; DO: dropout: MaxPool: maximum pooling; AvgPool: average pooling.

Table 1. The supervised backward MMSE decoder and its different varieties.

Method	Cost Function	Solution
Ridge regression + averaging of decoders [3] (MMSE-avgdec-ridge)	$\hat{\mathbf{d}}_k = \underset{\mathbf{d}}{\text{argmin}} \|\mathbf{s}_{a_k} - \mathbf{X}_k\mathbf{d}\|_2^2 + \lambda z_k \|\mathbf{d}\|_2^2$	$\hat{\mathbf{d}}_k = (\mathbf{X}_k^T\mathbf{X}_k + \lambda z_k \mathbf{I})^{-1}\mathbf{X}_k^T\mathbf{s}_{a_k}$ and $\hat{\mathbf{d}} = \frac{1}{K}\sum_{k=1}^{K}\hat{\mathbf{d}}_k$
LASSO + averaging of decoders [5] (MMSE-avgdec-LASSO)	$\hat{\mathbf{d}}_k = \underset{\mathbf{d}}{\text{argmin}} \|\mathbf{s}_{a_k} - \mathbf{X}_k\mathbf{d}\|_2^2 + \lambda q_k \|\mathbf{d}\|_1$	ADMM and $\hat{\mathbf{d}} = \frac{1}{K}\sum_{k=1}^{K}\hat{\mathbf{d}}_k$
Ridge regression + averaging of correlation matrices [4] (MMSE-avgcorr-ridge)	$\hat{\mathbf{d}} = \underset{\mathbf{d}}{\text{argmin}} \sum_{k=1}^{K}\|\mathbf{s}_{a_k} - \mathbf{X}_k\mathbf{d}\|_2^2 + \lambda z \|\mathbf{d}\|_2^2$	$\hat{\mathbf{d}} = \left(\sum_{k=1}^{K}\mathbf{X}_k^T\mathbf{X}_k + \lambda z\mathbf{I}\right)^{-1}\sum_{k=1}^{K}\mathbf{X}_k^T\mathbf{s}_{a_k}$
LASSO + averaging of correlation matrices [5] (MMSE-avgcorr-LASSO)	$\hat{\mathbf{d}} = \underset{\mathbf{d}}{\text{argmin}} \sum_{k=1}^{K}\|\mathbf{s}_{a_k} - \mathbf{X}_k\mathbf{d}\|_2^2 + \lambda q \|\mathbf{d}\|_1$	ADMM

ADMM: alternating direction method of multipliers.

and straightforwardly computing the decoder. Therefore, it is an example of the early integration paradigm, as opposed to late integration in the former case when averaging K separate decoders. Both versions are included in our study. Table 1 shows the four configurations of the MMSE/LS-based decoder that were proposed as different AAD algorithms in [3]–[5], adopting different regularization techniques (L_2/L_1-regularization) and ways to train the decoder (via averaging decoders and correlation matrices).

CCA

CCA to decode the auditory brain was proposed in [7] and [27]. It was applied to the AAD problem for the first time in [5]. CCA combines a spatiotemporal backward (decoding) model $\mathbf{w}_x \in \mathbb{R}^{LC\times 1}$ on the EEG and a temporal forward (encoding) model $\mathbf{w}_{s_a} \in \mathbb{R}^{L_a\times 1}$ on the speech envelope, with L_a being the number of filter taps of the encoding filter. In this sense, CCA differs from the previous approaches, which were different configurations of the same MMSE/LS-based decoder. In CCA, the forward and backward model are jointly estimated such that their outputs are maximally correlated:

$$\max_{\mathbf{w}_x,\mathbf{w}_{s_a}} \frac{\mathbb{E}\{(\mathbf{w}_x^{\mathrm{T}}\mathbf{x}(t))(\mathbf{w}_{s_a}^{\mathrm{T}}\mathbf{s}_a(t))\}}{\sqrt{\mathbb{E}\{(\mathbf{w}_x^{\mathrm{T}}\mathbf{x}(t))^2\}}\sqrt{\mathbb{E}\{(\mathbf{w}_{s_a}^{\mathrm{T}}\mathbf{s}_a(t))^2\}}} = \max_{\mathbf{w}_x,\mathbf{w}_{s_a}} \frac{\mathbf{w}_x^{\mathrm{T}}\mathbf{R}_{xs_a}\mathbf{w}_{s_a}}{\sqrt{\mathbf{w}_x^{\mathrm{T}}\mathbf{R}_{xx}\mathbf{w}_x}\sqrt{\mathbf{w}_{s_a}^{\mathrm{T}}\mathbf{R}_{s_as_a}\mathbf{w}_{s_a}}}, \tag{3}$$

where $\mathbf{s}_a(t) = [s_a(t)\ \ s_a(t-1)\ \ \cdots\ \ s_a(t-L_a+1)]^{\mathrm{T}} \in \mathbb{R}^{L_a\times 1}$. As opposed to the EEG filter $\mathbf{w}_x$, the audio filter $\mathbf{w}_{s_a}$ is a causal one, as the stimulus precedes the brain response. The solution of the optimization problem in (3) can be easily retrieved by solving a generalized eigenvalue decomposition (details are given in [4] and [5]).

In CCA, the backward model $\mathbf{w}_x$ and the forward model $\mathbf{w}_{s_a}$ are extended to a set of J filters $\mathbf{W}_x \in \mathbb{R}^{LC\times J}$ and $\mathbf{W}_{s_a} \in \mathbb{R}^{L_a\times J}$ for which the outputs are maximally correlated but mutually uncorrelated [the J outputs of $\mathbf{W}_x^{\mathrm{T}}\mathbf{x}(t)$ are uncorrelated to one another, and the J outputs of $\mathbf{W}_{s_a}^{\mathrm{T}}\mathbf{s}_a(t)$ are uncorrelated to one another]. There are now J Pearson correlation coefficients between the outputs of the J backward and forward filters (that is, canonical correlation coefficients), which are collected in the vector $\boldsymbol{\rho}_i \in \mathbb{R}^{J\times 1}$ for speaker i, whereas before, there was only one per speaker. Furthermore, because of the way CCA constructs the filters, it can be expected that the first components are more important than the later ones. To find the optimal way of combining the canonical correlation coefficients, a linear discriminant analysis (LDA) classifier can be trained, as proposed in [7]. To generalize the maximization of the correlation coefficients of the previous AAD algorithms (which is equivalent to taking the sign of the difference of the correlation coefficients of both speakers), we propose to construct a feature vector $\mathbf{f} \in \mathbb{R}^{J\times 1}$ by subtracting the canonical correlation vectors: $\mathbf{f} = \boldsymbol{\rho}_1 - \boldsymbol{\rho}_2$ and classifying $\mathbf{f}$ with an LDA classifier. As in [7], we use principle component analysis as a preprocessing step on the EEG to reduce the number of parameters. In fact, this is a way of regularizing CCA, and it can be viewed as an alternative to the regularization techniques proposed in other methods.

Training-free MMSE with LASSO (MMSE-adap-LASSO)

In [6], a fundamentally different AAD algorithm is proposed. In our study, all other AAD algorithms are supervised batch-trained algorithms, which have separate training and testing stages. First, the decoders need to be trained in a supervised manner by using a large amount of ground-truth data, after which they can be applied to new test data. In practice, this necessitates a (potentially cumbersome) a priori training stage, resulting in a fixed decoder, which does not adapt to nonstationary EEG signal characteristics, e.g., due to changing conditions and brain processes. The AAD algorithm in [6] aims to overcome these issues by adaptively estimating a decoder for each speaker and simultaneously using the outputs to decode the auditory attention. Therefore, this training-free AAD algorithm has the advantage of adapting the decoders to nonstationary signal characteristics, without requiring the amount of ground-truth data that the supervised AAD algorithms need.

In this study, we removed the state-space and dynamic decoder estimation modules to produce a single decision for each decision window, similar to the other AAD algorithms we review (a full description of the algorithm can be found in [6]). This leads to the following formulation:

$$\hat{\mathbf{d}}_{i,l} = \underset{\mathbf{d}}{\operatorname{argmin}} \left\| \mathbf{s}_{i,l} - \mathbf{X}_l\mathbf{d} \right\|_2^2 + \lambda q \left\| \mathbf{d} \right\|_1 \tag{4}$$

for the ith speaker in the lth decision window. In the context of AAD, for every new incoming window of τ seconds of EEG and audio data, two decoders are thus estimated (one for each speaker). As an attentional marker, these estimated decoders could be applied to the EEG data $\mathbf{X}_l$ of the lth decision window to compute the correlation with their corresponding stimuli envelopes. In addition, the authors of [6] propose to identify the attended speaker by selecting the one with the largest L_1-norm of its corresponding decoder $\hat{\mathbf{d}}_{i,l}$, as the attended decoder should exhibit more sparse, significant peaks, while the unattended decoder should have smaller, randomly distributed coefficients. The regularization parameter is again being cross validated and defined in the same way as for the MMSE-avgdec/corr-LASSO methods. To prevent overfitting by decreasing the number of parameters to be estimated, the authors of [6] proposed to a priori select a subset of EEG channels. In our study, we adopt this approach and select the same channels. While we do not adopt the extra postprocessing state-space modeling steps from [6] and [28] to focus on the core AAD algorithm, it is noted that such an extra smoothing step, which also takes previous and future decisions into account, can effectively enhance the performance of most AAD algorithms, albeit at the cost of a potential algorithmic delay in the detection of attention switches [6].

Nonlinear methods

Nonlinear methods based on (deep) neural networks (NNs) can adopt an SR approach [8] similar to the linear methods, or they can directly classify the attended speaker from the EEG and the audio (that is, direct classification) [9], [10]. However, they are more vulnerable to overfitting [10], particularly for the small data sets that are typically collected in AAD research. To appreciate the differences between current NN-based AAD approaches, Figure 3 presents a conceptual overview of the various strategies and network topologies of the presented nonlinear methods. We give a concise description of each architecture in the following and refer to the respective papers for further details.

Fully connected SR–NN (NN-SR)

In [8], the authors proposed a fully connected (FC) NN with a single hidden layer that reconstructs the envelope based on a segment of EEG data. As in Figure 3(a), the input layer consists of LC neurons (similar to a linear decoder), with L being the number of time lags and C the number of EEG channels. These neurons are connected to a hidden layer with two neurons and a `tanh` activation function. The neurons are then combined into a single output neuron that uses a linear activation function and outputs one sample of the reconstructed envelope. Thus, the network has $2 \times (LC + 1)$ (the hidden layer) $+ \ 2 + 1$ (the output layer) $\approx 3,446$ trainable parameters.

The network is trained to minimize $1 - \rho(\hat{\mathbf{s}}_\mathrm{a}, \mathbf{s}_\mathrm{a})$ across a segment of M training samples (within this segment, the NN coefficients are kept constant), with $\rho(\cdot)$ being the Pearson correlation coefficient and $\hat{\mathbf{s}}_\mathrm{a}, \mathbf{s}_\mathrm{a} \in \mathbb{R}^{M \times 1}$ the reconstructed and attended envelope, respectively. Minimizing this cost function is equivalent to maximizing the Pearson correlation coefficient between the reconstructed and attended speech envelopes, similar to linear SR approaches. The trained network is then used as a decoder, where the speech envelope showing the highest correlation with the decoder output is selected as the attended speaker. This algorithm can be extended to more than two competing speakers, similar to the other linear SR decoders.

Convolutional NN to compute similarity between EEG and stimulus (CNN-sim)

In [9], a convolutional NN (CNN) is proposed to directly compare a $C \times T$ EEG segment with a $1 \times T$ speech envelope. This network is trained to output a similarity score $\in [0, 1]$ (much like the correlation coefficient used in other approaches) between the EEG and the speech envelope by using a binary cross-entropy cost function. The speech envelope that, according to the trained CNN, is most similar to the EEG is then identified as the attended speaker. This technique can be easily extended to more than two speakers by computing a similarity score for each speaker and taking the maximum over all scores to identify the attended one. The network depicted in Figure 3(b) consists of two convolutional layers with maximum pooling (stride two) after the first convolutional layer and four FC layers. In total, this network has $64 \times (C + 1) \times L_1$ (the first convolutional layer) $+ \ 2 \times 64 \times L_2$ (the second convolutional layer) $+ 200 \times 3$ (the first FC layer) $+ 200 \times 201$ (the second FC layer) $+ 100 \times 201$ (the third FC layer) + 101 (the fourth FC layer) $\approx 69,070$ trainable parameters. An exponential linear unit is used as a nonlinear activation function. Furthermore, dropout is used as a regularization technique to prevent overfitting in the FC layers; batch normalization is used throughout the network. Details about the training can be found in [9].

CNN to determine spatial locus of attention (CNN-loc)

In [10], a CNN is proposed to determine the spatial locus of attention (i.e., the directional focus of attention, e.g., left or right), solely based on the EEG. This is a fundamentally different approach to tackle the AAD problem, which has the advantage of not requiring individual speech envelopes (see the "Open Challenges and Outlook" section). Furthermore, it avoids the requirement to estimate a correlation coefficient across a relatively long decision window, as in all aforementioned algorithms, thereby avoiding large algorithmic delays.

This CNN determines the spatial locus of attention, starting from a $C \times T$ EEG segment. As detailed in Figure 3(c), it consists of one convolutional layer and two FC layers. The convolutional layer consists of five spatiotemporal filters, with lags L similar to before, each outputting a 1D time series of length T on which a rectified linear unit activation function is applied. Afterward, an average pooling layer condenses each output series into a scalar, leading to a 5D vector, which is then used as an input for two FC layers, the first consisting of five neurons with a sigmoid activation function and the output layer consisting of two neurons and a softmax layer. In total, this network has $5 \times C \times L$ (the convolutional layer) $+ \ 5 \times 6$ (the first FC layer) $+ 2 \times 6$ (the second FC layer) $\approx 2,708$ trainable parameters. The CNN can be extended to more than two possible spatial locations (and, thus, competing speakers) by adding more output neurons to the network to generalize it to a multiclass problem in which each class corresponds to a location or zone where the attended speaker is believed to be positioned.

A cross-entropy cost function is minimized using minibatch gradient descent. Weight decay regularization is applied, as is a posttraining selection of the optimal model based on the validation loss. Furthermore, during training, data from the subject under test (as in all other methods) as well as data from other subjects are used, as the authors of [10] found that this prevents the model from overfitting on the training data in case only a limited amount of information about the subject under test is available. Therefore, the inclusion of data from other subjects can be seen as a type of regularization.

Comparative study of AAD algorithms

We compared the aforementioned AAD algorithms on two publicly available data sets [20], [21] in a subject-specific manner. Both data sets were collected for AAD by using a competing-talker setup in which two stories are simultaneously narrated. Details about the data sets and the preprocessing of the

Experiment Details

Data

The characteristics of both data sets are summarized in Table S1.

Speech envelope extraction

Individual speech signals are passed through a gammatone filter bank, which roughly approximates the spectral decomposition as performed by the human auditory system. Per subband, the audio envelopes are extracted, and their dynamic range is compressed using a power law operation with exponent 0.6, after which the subband envelopes are summed into a single broadband envelope [24].

Table S1. The data set characteristics.

Attribute	Das-2015 [20]	Fuglsang-2018 [21]
Number of subjects	16	18
Amount of data (per subject)	72 min	50 min
EEG system	64-channel Biosemi (wet EEG)	64-channel Biosemi (wet EEG)
Speakers	Male and male	Male and female
Azimuth direction sources	±90°	±60°
Acoustic room condition	Dichotic and HRTF-filtered in anechoic room	HRTF-filtered in anechoic, mildly, and highly reverberant room

EEG: electroencephalography; head-related transfer function.

Frequency range

For computational efficiency, the speech envelopes as well as the electroencephalography signals are downsampled to $f_s = 64\,\text{Hz}$ and bandpass filtered between 1 and 32 Hz [8]–[10]. For the linear algorithms, this was further reduced to $f_s = 20\,\text{Hz}$ and 1–9 Hz to be able to reduce the number of parameters in the spatiotemporal decoders; linear stimulus reconstruction (SR) methods have been demonstrated not to exploit information above 9 Hz [24].

Hyperparameter settings

The decoder lengths and convolutional neural network (CNN) kernel lengths are set as in the original papers. For all linear methods, this is $L = 250\,\text{ms}$ for NN–SR, $L = 420\,\text{ms}$ for CNN-loc, $L = 130\,\text{ms}$ and for CNN-sim, $L_1 = 30\,\text{ms}$ (the first layer), and $L_2 = 10\,\text{ms}$ (the second layer). For canonical correlation analysis (CCA), 1.25 s is chosen as the encoder length. The full set of 64 channels is used in all algorithms, except for the minimum mean-square error-adap-least absolute shrinkage and selection operator, where the 28 channels in [6] are chosen to reduce the number of parameters (since the decoder is estimated on much less data). The regularization parameters are cross validated using 10 values in the range $[10^{-6}, 0]$. For CCA, it turned out that retaining all principle component analysis components for both data sets is optimal.

EEG and audio data are described in "Experiment Details." All algorithms, including the deep learning methods, are separately retrained from scratch on each data set.

Given a decision window length τ, the performance of each algorithm is evaluated via the accuracy $p \in [0, 100]\%$, defined as the percentage of correctly classified decision windows. Since an EEG is the superimposed activity of many different (neural) processes, the correlation ρ between the reconstructed and attended envelopes is typically quite low (on the order of 0.05–0.2). Therefore, it is important to use a sufficiently long decision window such that the decision process is less affected by estimation noise in ρ due to the finite sample size. As a result, the accuracy p generally increases for longer decision windows τ, leading to a so-called "$p(\tau)$–performance curve." These accuracies are obtained using the cross-validation procedure described in "Details of the Cross-Validation Procedure."

The $p(\tau)$–performance curve thus presents a tradeoff between accuracy and decision delays in the AAD system (a long decision length implies a slower reaction time to a switch in attention). In [26], the minimal expected switch duration (MESD) metric was proposed to resolve this tradeoff to compare AAD algorithms more easily. The MESD determines the most optimal point on the $p(\tau)$–performance curve in the context of attention-steered gain control by minimizing the expected time it takes to switch the gain between two speakers in an optimized, robust gain control system. Therefore, it outputs a single-number time metric (the MESD, in seconds) for a $p(\tau)$–performance curve and removes the loss of statistical power due to multiple-comparison corrections in statistical hypothesis evaluation (due to testing for multiple decision window lengths). Furthermore, the MESD ensures that the statistical comparison is automatically focused on the most practically relevant points on the $p(\tau)$–performance curve, which typically turn out to be the ones corresponding to short decision window lengths $\tau < 10\,\text{s}$ [26]. A higher MESD corresponds to worse AAD performance and vice versa. This MESD is a theoretical metric that is not based on actual attention switches in the data, which are also not present in the data sets that are used. It is merely employed here as a comparative measure, which does not necessarily reflect the true switching time, as it relies on independence assumptions in the underlying Markov model, which can be violated in practice.

Statistical analysis

To statistically compare the included AAD algorithms, we adopt a linear mixed-effects model (LMM) on the MESD values, with the AAD algorithm as a fixed effect and with subjects as a repeated-measure random effect. Five contrasts of interest were set a priori according to the binary tree structure in Figure 2. Algorithms that were not competitive and that did not perform significantly better than chance are excluded from the statistical analysis, which is why some are not present in the contrasts (see the "Performance Curves" section). The planned contrasts reflect the most important different features between AAD algorithms, as in Figure 2, motivating how they are set. The significance level is set at $\alpha = 0.05$.

Results

Performance curves

Figure 4 gives the $p(\tau)$–performance curves of the different AAD algorithms on both data sets. For the MMSE-based decoders, it is observed that there is barely an effect on the type of regularization and that averaging correlation matrices (early integration) consistently outperform averaging decoders (late integration). Furthermore, CCA outperforms all other linear algorithms. Finally, on the Das-2015 data set, it is clear that decoding the spatial locus of attention using the CNN-loc method substantially outperforms the SR methods for short decision windows ($< 10s$), where the CNN-loc method appears to be less affected by the decision window length. However,

Details of the cross-validation procedure

Two-stage cross validation

The different algorithms are evaluated via a two-stage cross-validation (CV) procedure applied per subject and decision window length. The auditory attention decoding (AAD) accuracy is determined via an outer leave-one-segment-out CV (LOSO-CV) loop. Per outer fold, the optimal hyperparameter is determined via an inner 10-fold CV loop on the training set of the outer loop. The length of each left-out segment in the outer loop is chosen equal to 60 s, which is split into smaller disjointed decision windows. For example, for a decision window length of 30 s, each left-out segment results in two decisions. Additional details per AAD algorithm are provided in Table S2 (the standard CV corresponds to training on all but one segment and testing on the left-out segment).

Leave-one-speaker-out CV

When using the LOSO-CV method, the test set always contains a speaker that is also present in the training set. To avoid potential overfitting to speakers in the training set for the convolutional-neural-network-loc algorithm, we use the leave-one-speaker-out (LOSpO)-CV method for this algorithm, as proposed and explained in [10]. For linear methods, there is no difference between the LOSO-CV and LOSpO-CV. This is validated by performing 100 runs/subject, with another random CV split in each run (using the same number of folds as for LOSpO-CV). We then tested whether the LOSpO-CV performance significantly differed from the median of this empirical distribution (i.e., the median across all random splits) across all subjects. For the canonical correlation analysis (CCA) method, which has the most degrees of freedom to overfit, the difference between the LOSpO-CV and median random-CV accuracy is less than 1% on 20-s decision windows, and a paired Wilcoxon signed-rank test (across subjects) shows no significant difference ($W = 85, n = 16, p = 0.38$).

Table S2. Additional AAD algorithm details.

Method	Outer LOSO-CV Loop	Inner 10-CV Loop
MMSE-avgcorr-ridge/LASSO	Standard	Optimization of λ (independent of τ, tuned based on largest value of τ)
MMSE-avgdec-ridge/LASSO	The training data of each fold are split into windows of the same size as τ. A different decoder is estimated in each subwindow, and the decoders are averaged across all training folds (similar to [3]).	Optimization of λ (reoptimized for τ, due to the dependency of the training procedure on τ)
CCA	Standard, additional LOSO-CV loop to train and test LDA classifier	Optimization of the number of canonical correlation coefficients J as input for LDA (reoptimized for each τ)
MMSE-adap-LASSO	Optimization of λ per τ and fold by taking the hyperparameter with highest accuracy on training fold	—
NN–SR	Standard	—
CNN-loc	LOSpO-CV instead of LOSO-CV, training and testing redone for τ	—
CNN-sim	10-fold CV instead of LOSO-CV (due to computation time), training and testing redone for τ	—

MMSE: minimum mean-square error; LASSO: least absolute shrinkage and selection operator; CCA: canonical correlation analysis; LDA: linear discriminant analysis; NN: neural network; SR: stimulus reconstruction; CNN: convolutional NN; LOSpO: leave one speaker out.

the standard error on the mean is much higher for the CNN-loc algorithm than for the other methods, indicating higher inter-subject variability.

The performance of MMSE-adap-LASSO, CNN-sim, and NN–SR methods is not shown in Figure 4, as it did not exceed the significance level or was not competitive on either of the two data sets. For a decision window length of 10 s, the MMSE-adap-LASSO algorithm achieves an average accuracy of 52.9%, with a standard deviation of 4.3%, on the Das-2015 data set and 49.8%, with a standard deviation of 5.9%, on the Fuglsang-2018 data set. The CNN-sim algorithm achieves 51.7%, on average, with a standard deviation of 2.3%, on Das-2015 (where there was no convergence for five subjects) and 58.1%, with a standard deviation of 9.2%, on Fuglsang-2018. Finally, the NN–SR algorithm achieves, on average, only 52.1% (with a standard deviation of 4.4%) on Das-2015 and 52.3% (with a standard deviation of 3.6%) on Fuglsang-2018. Since these algorithms did not significantly outperform a random classifier or were not competitive, they were excluded from the statistical analysis. Furthermore, the CNN-loc method did not perform well on Fuglsang-2018 (i.e., 56.3%, with a standard deviation of 4.5%, on 10-s decision windows). Therefore, planned contrast 1 was excluded from the analysis for that data set.

Subject-specific MESD performance

A visual analysis of the per-subject MESD values (Figure 5) confirms the trends based on the performance curves. These trends are also verified by a statistical analysis using the LMM (the two outlying subjects of the CNN-loc algorithm were removed in all comparisons on the Das-2015 data set). Indeed, there is a significant improvement when decoding the spatial locus of attention via a nonlinear method versus the linear SR methods [$p < 0.001$ (Das-2015)]. Furthermore, CCA significantly outperforms all backward SR decoders [$p < 0.001$ (Das-2015); $p < 0.001$ (Fuglsang-2018)], while there is also a significant improvement when averaging correlation matrices compared to averaging decoders [$p = 0.0028$ (Das-2015); $p < 0.001$ (Fuglsang-2018)]. There is no notable effect from the specific regularization technique [$p = 0.79$ (Das-2015) and $p = 0.30$ (Fuglsang-2018) in the averaging correlation matrices; $p = 0.57$ (Das-2015) and $p = 0.91$ (Fuglsang-2018) in the averaging decoders].

Discussion

From the results and statistical analysis, it is clear that CCA [7], which adopts a joint forward and backward model, outperforms the other SR methods. Furthermore, the CNN-loc method [10], which decodes the spatial locus of attention based on the EEG alone (i.e., without using speech stimuli), substantially outperforms all SR methods on the Das-2015 data set at short decision window lengths, leading to substantially lower MESDs. The relatively high performance on short decision windows is attributed to the fact that this approach avoids correlating the decoded EEG with the speech envelope, thereby not suffering from the noise-susceptible correlation estimation. However, the nonsignificant performance of the CNN-loc method on the Fuglsang-2018 data set implies that alternative algorithms for decoding the spatial locus of attention might be required to improve robustness and generalization to different conditions.

Remarkably, while the traditional linear SR methods are found to perform well across data sets, none of the tested nonlinear (NN) methods achieve a competitive performance on both benchmark data sets, even though significant performance was obtained on the respective data sets used in [8]–[10]. This shows that these architectures do not always generalize well, even after retraining them on a new data set (the original authors validated the implementations in our study to rule out potential discrepancies in the implementation). Due to the black-box nature of these methods, it remains unclear what causes success on one data set and failure on another. One possible explanation is that the design process that eventually led to the reported network architecture was too tailored to a

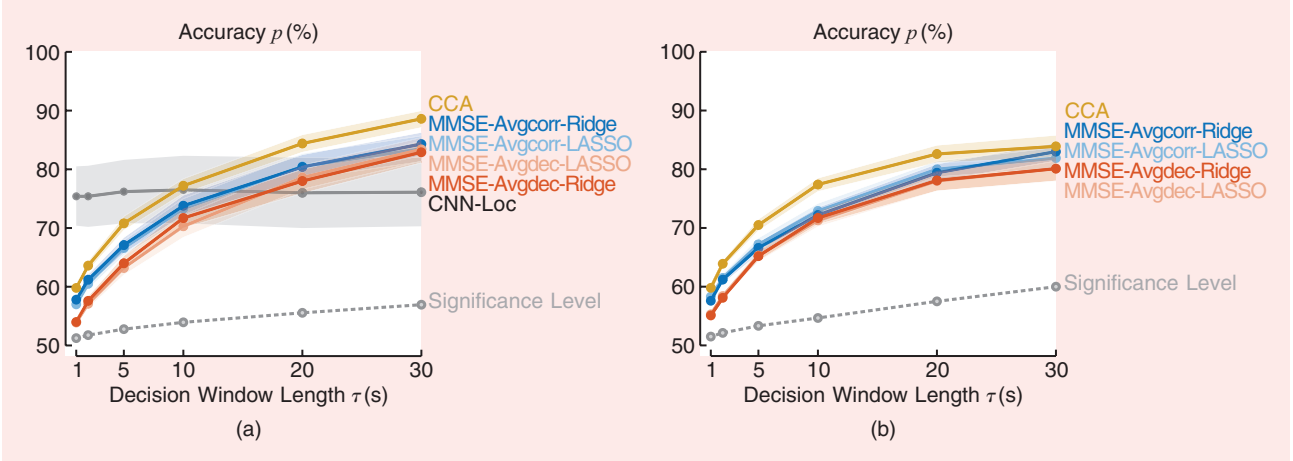

FIGURE 4. The accuracy p (the mean error ± the standard error on the mean across subjects) as a function of the decision window length τ for (a) the Das-2015 data set and (b) the Fuglsang-2018 data set. The MMSE-adap-LASSO, CNN-sim, and NN–SR methods did not perform significantly better than a random classifier and are not depicted. The CNN-loc method achieved competitive results only on Das-2015.

particular data set (and its size) despite proper cross validation. Furthermore, (deep) NNs may potentially pick up subtle patterns that may change or become absent in different experimental setups, due to differences in equipment, speech stimuli, and experiment protocols. Although this lack of reproducibility across data sets seems to undermine the practical usage of the presented nonlinear AAD methods, the current benchmark data sets are possibly too small for these techniques to facilitate drawing firm conclusions. AAD based on (deep) NNs may become more robust when larger data sets become available, with more subjects, EEG information per subject, and variation in experimental conditions. Nevertheless, the results of this study point out the risks of overfitting and overdesigning these architectures, thereby emphasizing the importance of extensive validation with multiple independent data sets.

Open challenges and outlook

Validation in realistic listening scenarios

In this study, we investigated and compared different AAD algorithms on data that were collected in a very controlled environment, with only two competing speakers, without much background noise and heavy reverberation, with well-separated competing speakers, and without switches in attention. Many of these AAD algorithms need to be further validated in more complex listening scenarios.

While we tested the algorithms on data with only two competing speakers, the algorithm in [3] was extended to four in [29], with limited performance loss. Thus, it is hoped that all other configurations of this decoder, including the CCA and MMSE-adap-lasso extensions and the NN–SR and CNN-sim models, which are based on the same principles, similarly generalize to multiple speakers. However, the effect of an increasing number of competing speakers and speaker locations on the CNN-loc algorithm is not immediately clear due to the fundamentally different decoding strategy. Decoding the spatial locus of attention may become much harder when there are more than two speaker locations. To what extent this affects performance remains to be investigated.

The impact of background noise (such as babble) and reverberation on the AAD performance of SR decoders has been extensively investigated in [21], [25], and [30]. For example, in [30], it was shown that the AAD accuracy increases when there is moderate background noise compared to no noise. Similarly, in [25], the AAD performance was comparable across different noisy and reverberant conditions. Moreover, even when training decoders with data collected in different acoustic conditions (noise and reverberation) than the test conditions, good

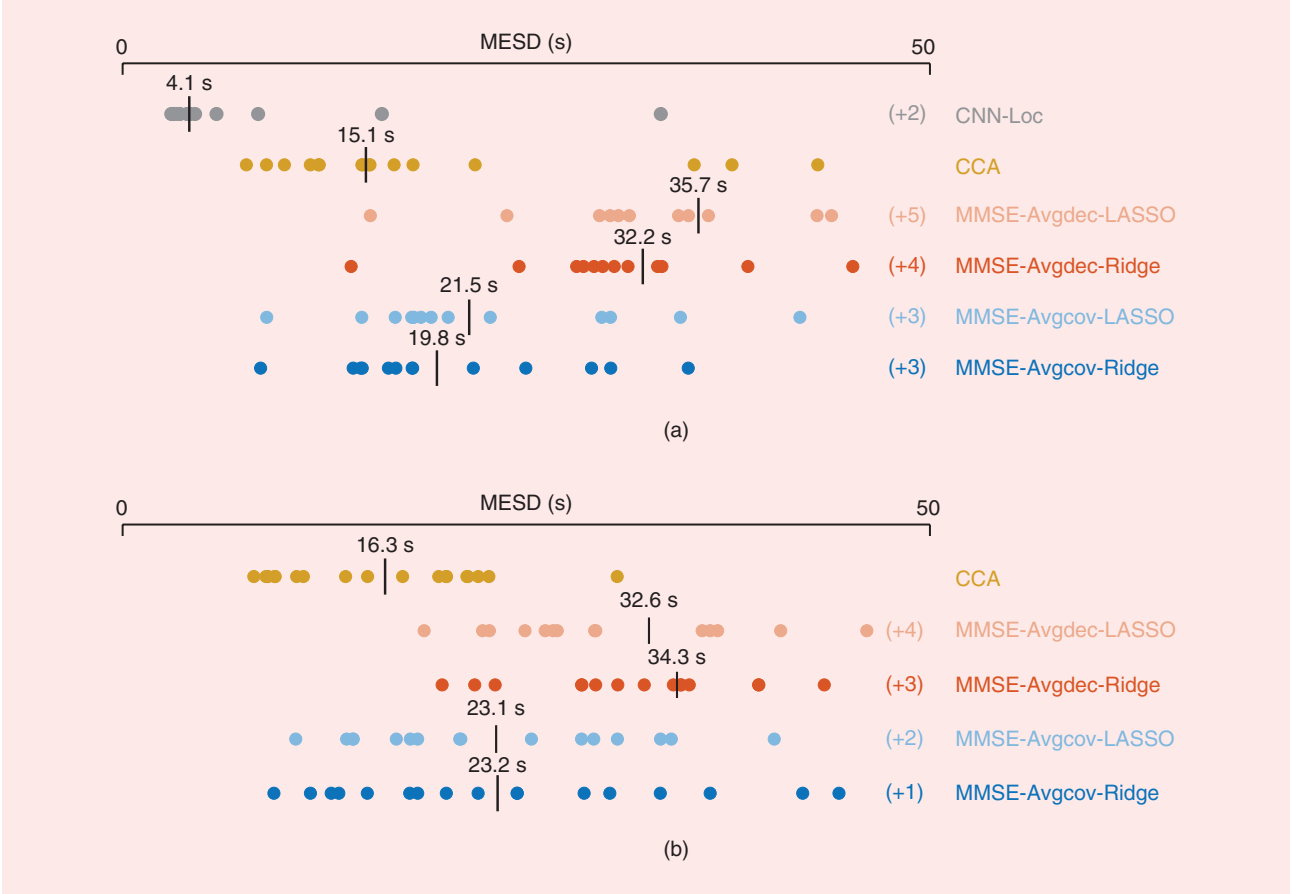

FIGURE 5. The per-subject MESD values, with the median indicated by a bar, for (a) Das-2015 and (b) Fuglsang-2018. The number of data points with an MESD of >50 s is indicated as (+x). However, the points were included in the computation of the medians.

AAD performance can be achieved. In [30], the effect of different speaker positions on the SR decoder was investigated, producing better performance with increasing speaker separation and acceptable accuracies for closely positioned competing speakers.

Finally, the effect of switches in auditory attention on the operation of several AAD algorithms is still unclear. While a theoretical analysis of the performance of AAD algorithms on attention switches was performed in [26], and some preliminary results on artificial attention switches were analyzed in [12], the performance of AAD algorithms on natural attention switches largely remains to be investigated.

Effects of speaker separation and denoising algorithms

As explained in the "Review of AAD Algorithms" section, most AAD algorithms require access to the speech envelopes of individual speakers. Although it is also possible to apply the SR decoders for AAD on unprocessed microphone signals, as shown in [11] and [25], the performance highly depends on a favorable relative position of the speakers and microphones. Thus, in the context of neurosteered hearing devices, the extraction of the per-speaker envelopes from a hearing aid's microphone recordings is generally required. It is expected that the performed speaker separation is not perfect, impacting the quality of the speech envelopes, and thus affecting the AAD algorithms that use these envelopes. Correspondingly, AAD algorithms that do not rely on this speaker separation step, such as decoding the spatial locus of attention [10], have an inherent advantage. In any case, a speech enhancement algorithm is required to eventually extract the attended speaker, for which advanced and well-performing signal processing algorithms exist (e.g., [31]).

A few studies have already combined AAD with speaker separation and denoising algorithms, using traditional beamforming approaches [11], [14], [15], [32] and deep NNs for speaker separation [12], [13], [32]. Remarkably, many of these studies show minor or very few effects on AAD performance when using demixed speech signals, even in challenging noisy conditions and despite significant distortions on the envelopes [15], [32]. These positive results are paramount for the practical applicability of neurosteered hearing devices. Finally, instead of treating speaker extraction and AAD as separate problems (as is the case in all aforementioned studies), one could aim to solve both problems simultaneously. In [14], speaker extraction and AAD are coupled in a joint optimization problem, where the beamformer is enforced to generate an output signal that is correlated to the output of a backward MMSE neural decoder, showing promising results.

EEG miniaturization and wearability effects

The data used in this article were acquired using expensive, heavy, bulky, and wet EEG recording systems. The realization of neurosteered hearing devices requires a wearable, concealable EEG monitoring system, for which the research is very active, resulting in novel miniature devices to acquire an EEG, for example, in the ear (e.g., [16]) or around the ear (e.g., [17]). However, such wearable, concealable EEG systems, also called *miniature EEG sensor devices*, provide only a limited number of EEG channels, which record brain activity within a small area. A first analysis using such an around-the-ear EEG system in the context of AAD showed potential, albeit with a significant decrease in performance [18].

In another (top-down) approach, it was shown that using data-driven selection of the best 10 EEG channels of a standard 64-channel EEG cap does not reduce the AAD performance of the linear SR decoder [19]. Similarly, in [9], the number of channels was reduced from 64 to 18, without any negative impact on the performance of the AAD system. Moreover, in [19], it was demonstrated that using EEG measured with strategically positioned electrode pairs with a <5-cm interelectrode distance results in AAD performance similar to standard long-distance montages. This is important for EEG miniaturization, where only a small number of electrodes within a confined area are available per device.

As mentioned, the data used here are collected using a wet EEG system, which requires a trained professional to apply the electrode gel and mount the system [33]. This seriously hampers practical applicability. Alternatively, dry EEG systems, which are easier to apply and thus more user friendly and suitable for long-term recording [33], are being developed (e.g., [16]). Although [33] shows that dry EEG systems can be used to record EEGs with quality similar to wet EEG systems, and while [9] briefly showed that a dry EEG system could achieve similar AAD performance, more extensive experimentation with dry EEG systems in the context of AAD is required, particularly in combination with miniaturization strategies [16]. While these results indicate that AAD is possible with fewer EEG electrodes and with dry and miniaturized EEG systems, the development of unobtrusive and wearable EEG systems for AAD remains an important hurdle to user-friendly and practical neurosteered hearing devices.

The realization of neurosteered hearing devices requires a wearable, concealable EEG monitoring system.

Outlook

Several studies have demonstrated that it is possible to decode auditory attention from a noninvasive neurorecording technique, such as EEG. In our study, we showed that most of these results are reproducible on different data sets. However, even for the best linear (SR) method (CCA), the accuracy with short decision windows is still too low, potentially leading to unacceptably slow system reactions to shifts in auditory attention, as indicated by a median MESD of 15 s. The results of this study demonstrated that an alternative strategy, such as decoding the spatial locus of attention, could significantly improve these short decision window lengths. Although nonlinear (deep learning) methods are believed to be able to improve AAD performance substantially, our study has demonstrated that the reported results obtained through these methods are hard to replicate on multiple independent AAD data sets. A major

future challenge for AAD research is the design of an algorithm or NN architecture that reliably improves short decision windows and is reproducible on different independent data sets.

Furthermore, most of the presented AAD algorithms require supervised training and are fixed during operation. To avoid cumbersome a priori training sessions for each individual user, as well as to adapt to the time-varying statistics of an EEG (e.g., in different listening scenarios), training-free and unsupervised adaptive AAD algorithms should be developed. While several steps have been made in that direction [6], the results of this study show that we are still far from a practical solution. Moreover, such online adaptive AAD algorithms are paramount to the development of closed-loop systems for neurosteered hearing devices, in which end users can react to and interact with the AAD algorithm and speech enhancement system. The interplay between the algorithmic processes in a hearing device and an end user could enable neurofeedback effects that significantly improve the performance of a hearing device [34].

Finally, these AAD algorithms need to be further evaluated in real-life situations, taking various realistic listening scenarios into account, as well as on potential hearing device users [35]. The individual building blocks of a neurosteered hearing device (Figure 1) need to be integrated such that an AAD algorithm is combined with a reliable and low-latency speaker separation algorithm, a miniaturized EEG sensor system, and a smart gain control system. Despite the many challenges ahead, the application of neurosteered hearing devices as a neurorehabilitative assistive device has been shown to be within reach, with the potential to substantially improve the functionality and user acceptance of future generations of hearing devices.

Acknowledgments

This research was funded by an Aspirant Grant, for Simon Geirnaert, from the Research Foundation–Flanders (grant 1136219N); the KU Leuven Special Research Fund (grant C14/16/057); Research Foundation–Flanders project G0A4918N; the European Research Council, through the European Union's Horizon 2020 research and innovation program (grants 802895 and 637424); and the Flemish Government, under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" program. The scientific responsibility is assumed by its authors. The first two authors implemented all the algorithms of the comparative study to ensure uniformity. All implementations were checked and approved by at least one of the authors of the original paper in which the method was presented.

Authors

Simon Geirnaert (simon.geirnaert@esat.kuleuven.be) received his M.S. degree (summa cum laude) in mathematical engineering from KU Leuven, Belgium, in 2018. He is currently a Ph.D. researcher in the Department of Electrical Engineering, STADIUS Center for Dynamical Systems, Signal Processing, and Data Analytics, KU Leuven, Leuven, 3000, Belgium, and in the Experimental Otorhinolaryngology Research Group, Department of Neurosciences, KU Leuven, Leuven, 3001, Belgium. His research interests include (biomedical) signal processing algorithm design. In 2018, he was nominated for the Agoriaprijs, and he won the 2019 Best Paper Award at the 40th WIC/IEEE Symposium on Information Theory and Signal Processing.

Servaas Vandecappelle (servaas.vandecappelle@kuleuven.be) received his M.S. degree in computer sciences from KU Leuven, Belgium, in 2017. From 2018 to 2020, he researched electroencephalography-based auditory attention decoding in the Department of Neurosciences and the Department of Electrical Engineering, KU Leuven, Leuven, 3001, Belgium.

Emina Alickovic (eali@eriksholm.com) received her Ph.D. degree in electrical and electronics engineering in 2015 from International Burch University, Bosnia and Herzegovina. She is an adjunct associate professor in the Automatic Control Group, Linköping University, Linköping, 581 83, Sweden. Her research interests include statistical and adaptive signal processing and machine learning, particularly with applications in auditory neuroscience.

Alain de Cheveigné (alain.de.cheveigne@ens.fr) received his Habilitation in neurosciences from Université Pierre et Marie Curie. He is a senior scientist at the Centre National de la Recherche Scientifique, Ecole Normale Supérieure, Paris, 75005, France, and an honorary professor at University College London, London, WC1E 6BT, U.K. His research interests include auditory psychophysics and modeling, audio signal processing, and electrophysiological and imaging data analysis.

Edmund Lalor (edmund_lalor@urmc.rochester.edu) received his Ph.D. degree in biomedical engineering from University College Dublin in 2006. He is an associate professor in the Department of Biomedical Engineering and the Department of Neuroscience, University of Rochester, Rochester, New York, 14627, USA. His research focuses on exploring how humans perceive and attend to the kinds of stimuli we encounter in daily life, such as speech and music, and how the differences in processing of such stimuli might inform our understanding of psychiatric and neurodevelopmental disorders.

Bernd T. Meyer (bernd.meyer@uol.de) received his Ph.D. degree from Carl von Ossietzky Universität Oldenburg, Oldenburg, 26129, Germany, in 2009, where he is a professor of communication acoustics. His research interests include the relation of speech and hearing, with a special interest in models of human speech perception, automatic speech processing, and neurophysiological data.

Sina Miran (sina_miran@starkey.com) received his Ph.D. degree in electrical and computer engineering from the University of Maryland, College Park in 2019. He is a research engineer at Starkey Hearing Technologies, Eden Prairie, Minnesota, 55344, USA. His research interests include the development of statistical and adaptive signal processing and machine learning algorithms for the analysis of neuroimaging data. He was a finalist for the Best Student Paper Award at the 2018 Annual International Conference of

the IEEE Engineering in Medicine and Biology Society and ICASSP 2017.

Tom Francart (tom.francart@med.kuleuven.be) received his Ph.D. degree in engineering from KU Leuven, Belgium, in 2008. He is an associate professor with the research group ExpORL in the Department of Neurosciences, KU Leuven, Leuven, 3001, Belgium. His research interests include sound processing for auditory prostheses, binaural hearing, and objective measures of hearing.

Alexander Bertrand (alexander.bertrand@esat.kuleuven.be) received his Ph.D. degree in engineering sciences from KU Leuven, Belgium, in 2011. Currently, he is an associate professor in the Department of Electrical Engineering, KU Leuven, Leuven, 3001, Belgium. His research interests include signal processing algorithm design, with a focus on biomedical sensor arrays and distributed algorithms. He is an associate editor of *IEEE Transactions on Signal Processing* and an elected member of the IEEE Signal Processing Society Sensor Array and Multichannel Technical Committee.

References

[1] 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. doi: 10.1038/nature11020.

[2] 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. 11,854–11,859, 2012. doi: 10.1073/pnas.1205381109.

[3] J. 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, 2014.

[4] 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 Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 5, pp. 402–412, 2017. doi: 10.1109/TNSRE.2016.2571900.

[5] E. Alickovic, T. Lunner, F. Gustafsson, and L. Ljung, "A tutorial on auditory attention identification methods," *Front. Neurosci.*, vol. 13, p. 153, Mar. 2019. doi: 10.3389/fnins.2019.00153.

[6] S. Miran, S. Akram, A. Sheikhattar, J. Z. Simon, T. Zhang, and B. Babdi, "Real-time tracking of selective auditory attention from M/EEG: A Bayesian filtering approach," *Front. Neurosci.*, vol. 12, p. 262, 2018. doi: 10.3389/fnins.2018.00262.

[7] A. de Cheveigné, D. D. E. Wong, G. M. Di Liberto, J. Hjortkjaer, M. Slaney, and E. Lalor, "Decoding the auditory brain with canonical component analysis," *NeuroImage*, vol. 172, pp. 206–216, May 2018. doi: 10.1016/j.neuroimage.2018.01.033.

[8] 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, 2017.

[9] 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, p. 11,538, 2019. doi: 10.1038/s41598-019-47795-0.

[10] S. Vandecappelle, L. Deckers, N. Das, A. H. Ansar, A. Bertrand, and T. Francart, "EEG-based detection of the locus of auditory attention with convolutional neural networks," *bioRxiv*, 2020.

[11] 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, 2017. doi: 10.1109/TBME.2016.2587382.

[12] J. O'Sullivan, Z. Chen, J. Herrero, G. M. McKhann, S. A. Sheth, A. D. Mehta, and N. Mesgarani, "Neural decoding of attentional selection in multi-speaker environments without access to clean sources," *J. Neural Eng.*, vol. 14, no. 5, p. 056001, 2017. doi: 10.1088/1741-2552/aa7ab4.

[13] C. Han, J. O'Sullivan, Z. Chen, Y. Luo, J. Herrero, A. D. Mehta, and N. Mesgarani, "Speaker-independent auditory attention decoding without access to clean speech sources," *Sci. Adv.*, vol. 5, no. 5, pp. 1–12, 2019. doi: 10.1126/sciadv.aav6134.

[14] W. Pu, J. Xiao, T. Zhang, and Z-Q Luo, "A joint auditory attention decoding and adaptive binaural beamforming algorithm for hearing devices," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2019, pp. 311–315.

[15] A. Aroudi and S. Doclo, "Cognitive-driven binaural beamforming using EEG-based auditory attention decoding," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 862–875, Jan. 2020. doi: 10.1109/TASLP.2020.2969779.

[16] S. L. Kappel, M. L. Rank, H. O. Toft, M. Andersen, and P. Kidmose, "Dry-contact electrode ear-EEG," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 1, pp. 150–158, 2019. doi: 10.1109/TBME.2018.2835778.

[17] S. Debener, R. Emkes, M. De Vos, and S. Debener, "Unobtrusive ambulatory EEG using a smartphone and flexible printed electrodes around the ear," *Sci Rep.*, vol. 5, no. 1, p. 16,743, July 2015. doi: 10.1038/srep16743.

[18] B. Mirkovic, M. G. Bleichner, M. De Vos, and S. Debener, "Target speaker detection with concealed EEG around the ear," *Front. Neurosci.*, vol. 10, p. 349, July 2016. doi: 10.3389/fnins.2016.00349.

[19] 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, 2020. doi: 10.1109/TBME.2019.2911728.

[20] N. Das, T. Francart, and A. Bertrand, "Auditory attention detection dataset KULeuven, Version 1.1.0," Zenodo, 2019. https://zenodo.org/record/3997352 (accessed May, 11, 2021).

[21] S. A. Fuglsang, D. D. E. Wong, and J. Hjortkjaer, "EEG and audio dataset for auditory attention decoding," Zenodo, 2018. 10.5281/zenodo.1199011 https://zenodo.org/record/1199011 (accessed May 11, 2021).

[22] D. D. E. Wong, S. A. Fuglsang, J. Hjortjaer, E. Ceolini, M. Slaney, and A. de Cheveigné, "A comparison of regularization methods in forward and backward models for auditory attention decoding," *Front. Neurosci.*, vol. 12, p. 531, Aug. 2018. doi: 10.3389/fnins.2018.00531.

[23] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomanonjy, and F. Yeger, "A review of classification algorithms for EEG-based brain–computer interfaces: A 10-year update," *J. Neural Eng.*, vol. 15, no. 3, p. 031005, 2018. doi: 10.1088/1741-2552/aab2f2.

[24] N. Das, W. Biesmans, A. Bertrand, and T. Francart, "The effect of head-related filtering and ear-specific decoding bias on auditory attention detection," *J. Neural Eng.*, vol. 13, no. 5, p. 056014, 2016. doi: 10.1088/1741-2560/13/5/056014.

[25] A. Aroudi, B. Mirkovic, M. De Vos, and S. Doclo, "Impact of different acoustic components on EEG-based auditory attention decoding in noisy and reverberant conditions," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 4, pp. 652–663, 2019. doi: 10.1109/TNSRE.2019.2903404.

[26] 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 Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 1, pp. 307–317, 2020. doi: 10.1109/TNSRE.2019.2952724.

[27] J. P. Dmochowski, J. J. Ki, P. DeGuzman, P. Sajda, and L. C. Parra, "Extracting multidimensional stimulus-response correlations using hybrid encoding-decoding of neural activity," *NeuroImage*, vol. 180, pp. 134–146, May 2018. doi: 10.1016/j.neuroimage.2017.05.037.

[28] A. Aroudi, T. de Tailez, and S. Doclo, "Improving auditory attention decoding performance of linear and non-linear methods using state-space model," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process (ICASSP)*, 2020, pp. 8703–8707.

[29] P. J. Schäfer, F. I. Corona-Strauss, R. Hannemann, S. A. Hillyard, and D. J. Strauss, "Testing the limits of the stimulus reconstruction approach: Auditory attention decoding in a four-speaker free field environment," *Trends Hearing*, vol. 22, pp. 1–12, Oct. 2018. doi: 10.1177/2331216518816600.

[30] N. Das, A. Bertrand, and T. Francart, "EEG-based auditory attention detection: Boundary conditions for background noise and speaker positions," *J. Neural Eng.*, vol. 15, no. 6, p. 066017, 2018. doi: 10.1088/1741-2552/aae0a6.

[31] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 8, pp. 1256–1266, 2019. doi: 10.1109/TASLP.2019.2915167.

[32] N. Das, J. Zegers, H. Van hamme, T. Francart, and A. Bertrand, "Linear versus deep learning methods for noisy speech separation for EEG-informed attention decoding," *J. Neural Eng.*, vol. 17, no. 4, p. 046039, 2020. doi: 10.1088/1741-2552/aba6f8.

[33] J. W. Kam, S. Griffin, A. Shen, S. Patel, H. Hinrichs, H-J Heinze, L. Y. Deouell, and R. T. Knight, "Systematic comparison between a wireless EEG system with dry electrodes and a wired EEG system with wet electrodes," *NeuroImage*, vol. 184, pp. 119–129, Sept. 2019. doi: 10.1016/j.neuroimage.2018.09.012.

[34] 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*, 2017.

[35] S. A. Fuglsang, J. Märcher-Rørsted, T. Dau, and J. Hjortkjaer, "Effects of sensorineural hearing loss on cortical synchronization to competing speech during selective attention," *J. Neurosci.*, vol. 40, no. 12, pp. 2562–2572, 2020. doi: 10.1523/JNEUROSCI.1936-19.2020.

SP