Impact of Different Acoustic Components on EEG-Based Auditory Attention Decoding in Noisy and Reverberant Conditions

Ali Aroudi, *Student Member, IEEE*, Bojana Mirkovic, Maarten De Vos,
and Simon Doclo, *Senior Member, IEEE*

***Abstract*—Identifying the target speaker in hearing aid applications is an essential ingredient to improve speech intelligibility. Recently, a least-squares-based method has been proposed to identify the attended speaker from single-trial EEG recordings for an acoustic scenario with two competing speakers. This least-squares-based auditory attention decoding (AAD) method aims at decoding auditory attention by reconstructing the attended speech envelope from the EEG recordings using a trained spatio-temporal filter. While the performance of this AAD method has been mainly studied for noiseless and anechoic acoustic conditions, it is important to fully understand its performance in realistic noisy and reverberant acoustic conditions. In this paper, we investigate AAD using EEG recordings for different acoustic conditions (anechoic, reverberant, noisy, and reverberant-noisy). In particular, we investigate the impact of different acoustic conditions for AAD filter training and for decoding. In addition, we investigate the influence on the decoding performance of the different acoustic components (i.e., reverberation, background noise, and interfering speaker) in the reference signals used for decoding and the training signals used for computing the filters. First, we found that for all considered acoustic conditions it is possible to decode auditory attention with a considerably large decoding performance. In particular, even when the acoustic conditions for AAD filter training and for decoding are different, the decoding performance is still comparably large. Second, when using speech signals affected by either reverberation and/or background noise there is no significant difference in decoding performance ($p > 0.05$) compared to when using clean speech signals as reference signals. In contrast, when using reference signals affected by the interfering speaker, the decoding performance significantly decreases. Third, the experimental results indicate that it is even feasible to use training signals affected by reverberation, background noise and/or the interfering speaker for computing the filters.**

***Index Terms*—Auditory attention decoding, electroencephalography (EEG), background noise, reverberation, interfering speaker.**

I. Introduction

IN COMPLEX acoustic conditions the human auditory system has a remarkable ability to segregate a speaker of interest from a mixture of speakers and background noise [1], [2]. In contrast with normal-hearing persons, hearing-impaired persons typically have more difficulties with such auditory segregation, particularly in multi-talker scenarios [3]. Although many acoustic signal processing algorithms are available to reduce background noise or to perform source separation in multi-talker scenarios [4], [5], these algorithms typically need to rely on assumptions about the target speaker to be enhanced. For example, in hearing aid applications the target speaker is typically assumed to be located in front of the user or is assumed to be the loudest speaker. As in real-world conditions such assumptions are often violated, the performance of these algorithms may substantially decrease. Therefore, successfully identifying the target speaker in hearing aid applications is very important to improve speech intelligibility.

Recent studies have shown that auditory cortical responses are correlated with the envelope of the attended speech signal [6]–[8], based on which decoding and encoding properties of the speech signal, e.g. spectrotemporal features and perceptual unites, have been studied in the brain auditory pathway [9], [10]. Based on this finding, an auditory attention decoding (AAD) method has been proposed in [11] to identify the attended speaker from single-trial EEG recordings. This method aims at reconstructing the attended speech envelope from the EEG recordings using a trained spatio-temporal filter. In the *training step*, the clean speech signal of the attended speaker is used to train a spatio-temporal filter by minimizing the least-squares error between the attended speech envelope and the reconstructed envelope. In the *decoding step*, the clean speech signals of both the attended and the unattended speaker are used as reference signals. In [11] it has

Manuscript received May 24, 2018; revised October 3, 2018, December 7, 2018 and January 15, 2019; accepted February 9, 2019. Date of publication March 7, 2019; date of current version April 8, 2019. This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)-Cluster of Excellence 1077 Hearing4all. *(Corresponding author: Ali Aroudi.)*

A. Aroudi and S. Doclo are with the Department of Medical Physics and Acoustics, University of Oldenburg, 26111 Oldenburg, Germany (e-mail: ali.aroudi@uni-oldenburg.de; simon.doclo@uni-oldenburg.de).

B. Mirkovic is with the Department of Psychology, University of Oldenburg, 26129 Oldenburg, Germany (e-mail: bojana.mirkovic@uni-oldenburg.de).

M. De Vos is with the Department of Engineering, Institute of Biomedical Engineering, University of Oxford, Oxford OX3 7DQ, U.K. (e-mail: maarten.devos@eng.ox.ac.uk).

Digital Object Identifier 10.1109/TNSRE.2019.2903404

been shown that for high-density EEG recordings it is possible to decode auditory attention when presenting the clean speech signals of the different speakers to different ears of a listener (i.e. dichotic stimuli presentation). When presenting competing speech signals in a simulated anechoic condition including head filtering effects, it has been shown in [12] that a larger AAD performance can be obtained compared to dichotic presentation. Recently, a large research effort has focused on investigating how to use AAD as part of a brain-computer interface for real-world applications, e.g., to control a hearing aid [12]–[24], mainly however for anechoic conditions. Aiming at integrating a small-size EEG recording system in hearing aids, in [13]–[15] the reliability of AAD using a low number of EEG electrodes has been shown in an anechoic condition. Aiming at investigating the effect of neurofeedback, in [16] the feasibility of an online closed-loop system for AAD has been shown in an anechoic condition. Instead of using the clean speech signals of the attended and the unattended speaker as reference signals for decoding, in [17]–[21] the effect of different reference signals on the AAD performance has been investigated for an anechoic condition. Using simulated noisy reference signals for decoding, in [17] we have investigated the robustness of AAD to residual interference and background noise. In [18] and [19] a neuro-steered noise reduction algorithm has been proposed to suppress the unattended speaker based on the AAD decision for an anechoic condition. In [20] an AAD-based sound source separation algorithm using deep neural networks has been presented to suppress the unattended speaker. In [21], we have investigated steerable beamformers to generate reference signals for AAD in an anechoic condition.

While the performance of the aforementioned least-squares-based AAD method has been extensively investigated for noiseless and anechoic acoustic conditions, in practice also background noise and reverberation, i.e. acoustic reflections against walls and objects, are present. Reverberation is known to spectro-temporally distort speech signals, causing the binaural spatial cues and pitch to become less reliable for performing auditory attention tasks [25]–[28]. In addition, interfering speakers and background noise degrade the attended speech signal, possibly leading to a severe speech encoding degradation at the level of the auditory nerve and the brainstem [29], [30]. Since in noisy and reverberant conditions the available signals at the ears contain several acoustic components (i.e. reverberation, background noise and interfering speaker), fully understanding the impact of each acoustic component on AAD is of crucial importance, e.g., in order to generate appropriate reference signals for decoding from these signals. Recently, in [31] the performance of the least-squares-based AAD method was investigated for noisy and reverberant acoustic conditions. In [31] the same acoustic condition was used for AAD filter training and for decoding and the feasibility of using reverberant speech signals both as training and as reference signals was investigated. It was shown that in this way a comparable decoding performance for the reverberant condition as for the anechoic condition can be obtained. In this paper, we perform a more detailed analysis of the performance of the least-squares-based AAD method for an acoustic scenario comprising two competing speakers, background noise and reverberation. Compared to [31] we consider more acoustic conditions, especially with regard to background noise, and we specifically investigate the impact of different acoustic conditions for the training and the decoding steps. In addition, we investigate the influence on the decoding performance of the different acoustic components in the reference signals used for decoding and the training signals used for computing the filters. Some preliminary results were presented in [32], where we investigated the feasibility of using the (unprocessed) signals at the ears, containing reverberation, background noise and the interfering speaker, as reference and training signals.

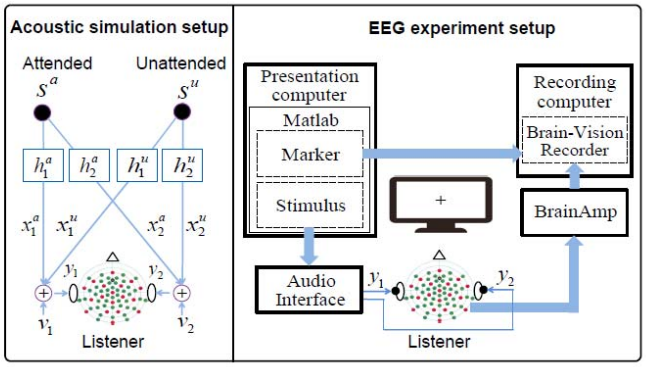


Fig. 1. Acoustic simulation setup and EEG experiment setup. The acoustic simulation setup was used for simulating the presented stimuli in different acoustic conditions. For the EEG experiment setup, MATLAB was used for sending the acoustic stimuli to the audio interface and the event markers to the Brain-Vison recorder software. The acoustic stimuli were presented to the participants via earphones using the audio interface. The EEG responses were amplified using BrainAmp and recorded together with the event markers using Brain-Vision.

The paper is organized as follows. In Section II the different acoustic conditions used for recording the EEG responses and the different acoustic signals used for the experimental analysis are introduced. In Section III the training and decoding steps of the least-squares-based AAD method are briefly reviewed. Section IV describes the acoustic and EEG measurement setup used for the experiments. In Section V the experimental results are presented and discussed, exploring the influence on the decoding performance of the different acoustic conditions and acoustic components.

II. Acoustic Conditions and Components

We consider an acoustic scenario comprising two competing speakers and background noise in a reverberant environment (see left part of Fig. 1). The clean speech signal of the attended speaker is denoted as $s^a[i]$, while the clean speech signal of the unattended speaker is denoted as $s^u[i]$, with i the discrete time index. The signals at the ears of the listener consist of a mixture of both speakers, including head filtering effects, reverberation and background noise. The signal $y_m[i]$ at the m-th ear, with $m = 1$ denoting the left ear and $m = 2$ denoting the right ear, can be written as

$$y_m[i] = \underbrace{h_m^a[i] * s^a[i]}_{x_m^a[i]} + \underbrace{h_m^u[i] * s^u[i]}_{x_m^u[i]} + v_m[i], \tag{1}$$

TABLE I
ACOUSTIC SIGNALS USED FOR EXPERIMENTAL ANALYSIS

Signal	Definition
s^a, s^u	clean speech signal
$x_m^{a,an}$, $x_m^{u,an}$	anechoic speech signal
x_m^a, x_m^u	reverberant speech signal
x_m^{an}	interfered speech signal
$x_m^{a,no}$, $x_m^{u,no}$	noisy speech signal
y_m	binaural speech signal

where $h_m^a[i]$ and $h_m^u[i]$ denote the (reverberant) acoustic impulse response between the m-th ear and the attended and the unattended speaker, respectively, $*$ denotes the convolution operation, and $v_m[i]$ denotes the background noise component at the m-th ear. The reverberant speech signal of the attended and the unattended speaker at the m-th ear is denoted as $x_m^a[i]$ and $x_m^u[i]$, respectively. These reverberant speech signals consist of an anechoic speech signal encompassing the (anechoic) head filtering effect, i.e. $x_m^{a,an}[i]$ and $x_m^{u,an}[i]$, and a reverberation component. For notational conciseness the index i will be omitted in the remainder of this paper, except where explicitly required.

For the EEG recordings we will consider four different acoustic conditions, i.e. anechoic, reverberant, noisy and reverberant-noisy. We refer to the EEG data recorded in a specific acoustic condition as the *EEG condition*. Depending on the acoustic condition, the stimuli presented at the ears of the listener obviously comprise different acoustic components:

- in the *anechoic* condition (an), the mixture of the anechoic speech signals of the attended and the unattended speaker is presented.
- in the *noisy* condition (no), the mixture of the anechoic speech signals of the attended and the unattended speaker and background noise is presented.
- in the *reverberant* condition (re), the mixture of the reverberant speech signals of the attended and the unattended speaker is presented.
- in the *reverberant-noisy* condition (rn), the mixture of the reverberant speech signals of the attended and the unattended speaker and background noise is presented.

To investigate the impact of the different acoustic components on the AAD performance, we will consider several acoustic signals (see Table I) to compute envelopes for filter training and evaluation:

- the clean speech signals s^a and s^u.
- the anechoic speech signals $x_m^{a,an}$ and $x_m^{u,an}$, i.e. the clean speech signals affected by head filtering effects.
- the reverberant speech signals x_m^a and x_m^u, i.e. the anechoic speech signals affected by reverberation.
- the interfered speech signals, i.e. the anechoic speech signals affected by an interfering speaker

$$x_m^{an} = x_m^{a,an} + x_m^{u,an}. \tag{2}$$

- the noisy speech signals, i.e. the anechoic speech signals affected by background noise

$$x_m^{a,no} = x_m^{a,an} + v_m, \quad x_m^{u,no} = x_m^{u,an} + v_m. \tag{3}$$

- the binaural speech signals y_m in (1), i.e. the anechoic speech signals affected by reverberation, background noise and an interfering speaker.

It should be noted that in the experiments (see Section IV-B) the positions of the attended and the unattended speaker are not always the same, i.e. for some participants the attended speaker is on the right side (and the unattended speaker on the left side), whereas for some participants the attended speaker is on the left side (and the unattended speaker on the right side). Due to the head filtering effect, the broadband energy ratio between the attended speech component and the unattended speech component in the signals at the ears is always smaller at the side of the unattended speaker than at the side of the attended speaker for the considered scenario. Therefore, the speech signals in Table I at the side of the attended speaker will be referred to as attended speech signals and the speech signals at the side of the unattended speaker as unattended speech signals.

III. AUDITORY ATTENTION DECODING METHOD

This section briefly reviews the least-squares-based AAD method proposed in [11]. This method aims at reconstructing the attended speech envelope from the EEG recordings using a trained spatio-temporal filter. Section III-A describes the training step, where the envelope of a training signal is used together with the EEG recordings to compute the filter. Section III-B describes the decoding step, where the envelopes of two reference signals (attended and unattended) are compared with an estimate of the attended speech envelope computed using the trained filter.

A. Training Step

In the training step, the attended speaker is assumed to be known and an attended speech signal (e.g., the clean speech signal of the attended speaker s^a) is used as training signal. From this signal the attended speech envelope $e^a[k]$, with $k = 1 \dots K$ the sub-sampled time index, is extracted, e.g., based on the Hilbert transform [33]. The attended speech envelope is then estimated from the EEG recordings $r_c[k]$, $c = 1 \dots C$, using a spatio-temporal filter as

$$\hat{e}^a[k] = \sum_{c=1}^{C} \sum_{l=0}^{L-1} g_{c,l} \; r_c[k + l + \Delta], \tag{4}$$

with $g_{c,l}$ the l-th filter coefficient in the c-th channel, L the number of filter coefficients per channel, and Δ modeling the latency of the attentional effect in the EEG responses to the speech stimuli. In vector notation, (4) can be written as

$$\hat{e}^a[k] = \mathbf{g}^T \mathbf{r}[k], \tag{5}$$

with

$$\mathbf{g} = \left[\mathbf{g}_1^T \, \mathbf{g}_2^T \dots \mathbf{g}_C^T\right]^T, \tag{6}$$

$$\mathbf{g}_c = \left[g_{c,0} \;\; g_{c,1} \dots \; g_{c,L-1}\right]^T, \tag{7}$$

$$\mathbf{r}[k] = \left[\mathbf{r}_1^T[k] \; \mathbf{r}_2^T[k] \dots \mathbf{r}_C^T[k]\right]^T, \tag{8}$$

$$\mathbf{r}_c[k] = \left[r_c[k + \Delta] \; r_c[k + 1 + \Delta] \dots r_c[k + L - 1 + \Delta]\right]^T, \tag{9}$$

with $(.)^T$ denoting the transpose operation. The spatio-temporal filter $\mathbf{g}$ is computed by minimizing the least-squares error between the attended speech envelope $e^a[k]$ and the reconstructed envelope $\hat{e}^a[k]$, regularized with the squared l_2-norm of the derivatives of the filter coefficients to avoid over-fitting [11], [13], [32], [33], i.e.

$$J(\mathbf{g}) = \frac{1}{K}\sum_{k=1}^{K}\left(e^a[k] - \mathbf{g}^T\mathbf{r}[k]\right)^2 + \beta\mathbf{g}^T\mathbf{D}\mathbf{g}, \quad (10)$$

with $\mathbf{D}$ denoting the derivative matrix [13] and β denoting a regularization parameter. The filter minimizing the regularized least-squares cost function in (10) is equal to

$$\mathbf{g} = (\mathbf{Q} + \beta\mathbf{D})^{-1}\mathbf{q}, \quad (11)$$

with the correlation matrix $\mathbf{Q}$ and the cross-correlation vector $\mathbf{q}$ given by

$$\mathbf{Q} = \frac{1}{K}\sum_{k=1}^{K}\left(\mathbf{r}[k]\mathbf{r}^T[k]\right), \quad \mathbf{q} = \frac{1}{K}\sum_{k=1}^{K}\left(\mathbf{r}[k]e^a[k]\right). \quad (12)$$

In this paper we will consider several *EEG training conditions* (tc) for computing the filter $\mathbf{g}$, i.e. $tc = an$ using EEG responses recorded in the anechoic condition, $tc = re$ using EEG responses recorded in the reverberant condition, $tc = no$ using EEG responses recorded in the noisy condition, and $tc = rn$ using EEG responses recorded in the reverberant-noisy condition. In addition, we will consider the EEG training condition $tc = ac$, in which EEG responses from all conditions are used for computing the filter.

Aiming at investigating the influence of each acoustic component, in this paper we will consider different attended speech signals (see Table I) as *training signals*, more in particular the clean attended speech signal s^a, the anechoic attended speech signal $x_m^{a,an}$, the reverberant attended speech signal x_m^a, the interfered attended speech signal x_m^{an}, the noisy attended speech signal $x_m^{a,no}$, and the binaural attended speech signal y_m.

B. Decoding Step

For each acoustic condition, the complete set of EEG responses is segmented into T trials (see Section IV-D for more details). To decode to which speaker a listener attended during trial t, first an estimate of the attended speech envelope $\hat{e}_t^a[k]$ is computed using the (trained) filter $\mathbf{g}_t$, i.e.

$$\hat{e}_t^a[k] = (\mathbf{g}_t)^T\mathbf{r}_t[k], \quad (13)$$

with $\mathbf{r}_t[k]$ denoting the EEG recordings of trial t. Next, the correlation coefficients between the estimated attended speech envelope $\hat{e}_t^a[k]$ and the envelope of two reference signals, i.e. namely the attended and the unattended reference signal, are computed as

$$\rho_t^a = \rho\left(e_t^a[k], \hat{e}_t^a[k]\right), \quad \rho_t^u = \rho\left(e_t^u[k], \hat{e}_t^a[k]\right), \quad (14)$$

where ρ_t^a and ρ_t^u denote the attended and the unattended correlation coefficient, respectively, and $e_t^a[k]$ and $e_t^u[k]$ denote the attended and the unattended speech envelope, respectively. When $\rho_t^a > \rho_t^u$, it is decided that auditory attention has been correctly decoded. Accordingly, a larger difference between the attended and the unattended correlation coefficient $\rho_t^a - \rho_t^u$ (referred to as correlation difference) is indicative of a more reliable AAD decision. The decoding performance P is defined as the percentage of correctly decoded trials over all considered trials and all participants. To compute the correlation coefficients in (14), EEG recordings in different acoustic conditions can be used for computing $\hat{e}_t^a[k]$. In addition, aiming at investigating the influence of each acoustic component on the decoding performance, different reference signals (see Table I) can be used for computing the attended and the unattended speech envelope $e_t^a[k]$ and $e_t^u[k]$, respectively.

In this paper we will investigate the decoding performance for several *EEG evaluation conditions* $ec \in \{an, re, no, rn, ac\}$, with P_{ec} denoting the decoding performance for a specific EEG evaluation condition. To decode trial t of an EEG evaluation condition using the filter trained in a specific EEG training condition which is not necessary the same as the EEG evaluation condition, the filter $\mathbf{g}_t$ is computed as follows:

- when the trial t to be decoded is part of the trials in the EEG training condition, the filter is computed using (11) as

$$\mathbf{g}_t = \left(\tilde{\mathbf{Q}}_t + \beta\mathbf{D}\right)^{-1}\tilde{\mathbf{q}}_t, \quad (15)$$

  with $\tilde{\mathbf{Q}}_t$ the average correlation matrix, computed by averaging all correlation matrices corresponding to trials in the EEG training condition *except* trial t, and $\tilde{\mathbf{q}}_t$ the average cross-correlation vector, computed by averaging all cross-correlation vectors corresponding to trials in the EEG training condition *except* trial t, i.e.

$$\tilde{\mathbf{Q}}_t = \frac{1}{T-1}\sum_{n=1,\, n\neq t}^{T}\mathbf{Q}_n, \quad \tilde{\mathbf{q}}_t = \frac{1}{T-1}\sum_{n=1,\, n\neq t}^{T}\mathbf{q}_n. \quad (16)$$

  This procedure corresponds to leave-one-out cross validation.
- when the trial t to be decoded is not part of the trials in the EEG training condition, the filter is computed using (11) as

$$\mathbf{g}_t = \left(\bar{\mathbf{Q}} + \beta\mathbf{D}\right)^{-1}\bar{\mathbf{q}}, \quad (17)$$

  with $\bar{\mathbf{Q}}$ the average correlation matrix, computed by averaging all correlation matrices corresponding to trials in the EEG training condition, and $\bar{\mathbf{q}}$ the average cross-correlation vector, computed by averaging all cross-correlation vectors corresponding to trials in the EEG training condition, i.e.

$$\bar{\mathbf{Q}} = \frac{1}{T}\sum_{n=1}^{T}\mathbf{Q}_n, \quad \bar{\mathbf{q}} = \frac{1}{T}\sum_{n=1}^{T}\mathbf{q}_n, \quad (18)$$

Since the number of trials across acoustic conditions is different (see Section IV-B), for $tc = ac$ the average correlation matrix and the average cross-correlation vector ($\tilde{\mathbf{Q}}_t$, $\bar{\mathbf{Q}}$, $\tilde{\mathbf{q}}_t$ and $\bar{\mathbf{q}}$) are computed in such a way that the contribution of trials from each acoustic condition is considered equally.

TABLE II

ACOUSTIC CONDITIONS USED FOR EXPERIMENTAL ANALYSIS AND STIMULI PRESENTATION

Experimental Analysis Condition	Stimuli Presentation	SNR[dB]	T_{60}[s]	Number of Trials
Anechoic (an)	Anechoic [34]	∞	<0.05	40
Reverberant (re)	Reverberant I [34]	∞	0.50	10
	Reverberant II [35], [36]	∞	1.00	10
Noisy (no)	Noisy I [34]	9.0	<0.05	10
	Noisy II [34]	4.0	<0.05	10
Reverberant-noisy (rn)	Reverberant-noisy I [34]	9.0	0.50	10
	Reverberant-noisy II [34]	4.0	0.50	10
	Reverberant-noisy III [35], [36]	9.0	1.00	10

In [17] it has been shown that the parameters involved in the filter design (Δ, L, β) play an important role in obtaining a good decoding performance. In order not to favour one specific EEG evaluation condition, the filter parameters have been determined to optimize the average decoding performance P_{ac} over all considered acoustic conditions. Please note that the filter parameters have been optimized per participant and for each EEG training condition (see Section IV-B and IV-D).

IV. ACOUSTIC AND EEG MEASUREMENT SETUP

In this section, we describe the acoustic and EEG measurement setup used for the experiments and information about the participants and the used paradigm.

A. Participants

Eighteen native German-speaking participants (right-handed and aged between 21 and 34 years) took part in this study. All participants were normal-hearing as was confirmed by pure tone audiometry. The participants reported no past or present neurological or psychiatric conditions. All participants signed an informed consent form and were paid for their participation. Two participants were excluded from the analysis, one participant due to poor attentional performance (as revealed by the questionnaire results) and the other participant due to a technical hardware problem.

B. Acoustic Stimuli

Two German audio stories, uttered by two different male speakers, were used as the clean speech signals (sampling frequency of 16 kHz). One story was from the German audio book website [37] and the other story was from a selection of audio books [38]. Speech pauses that exceeded 0.5 s were shortened to 0.5 s. Before performing the experiment, the participants reported no, or very limited, knowledge of the audio stories. The acoustic stimuli were simulated by convolving the clean speech signals (i.e. the audio stories) with non-individualized binaural acoustic impulse responses, either from [34], [35], or [36], and by adding diffuse babble noise, generated according to [39]. The competing speakers were simulated at $-45°$ (left) and $45°$ (right). Eight different acoustic conditions were considered for the stimuli (see Table II): anechoic, reverberant with a moderate and a large reverberation time ($T_{60} = 0.5$ s, $T_{60} = 1$ s), noisy with two different broadband signal-to-noise ratios (SNR $= 9.0$ dB, SNR $= 4.0$ dB), and three combinations of reverberation and noise. The SNR is defined as the broadband energy ratio between the reverberant speech signal of the attended and the unattended speaker at the ears and the background noise component at the ears, i.e.

$$\mathrm{SNR} = 10\log_{10}\frac{\sum_i |x_1^a[i]|^2 + |x_1^u[i]|^2 + |x_2^a[i]|^2 + |x_2^u[i]|^2}{\sum_i |v_1[i]|^2 + |v_2[i]|^2}. \tag{19}$$

For the experimental analysis, the acoustic conditions were grouped based on acoustic similarity as shown in Table II, resulting in four experimental analysis conditions, i.e. anechoic, reverberant, noisy, and reverberant-noisy. The acoustic stimuli were presented to the participants via insert earphones (E-A-RTONE 3A) using an RME HDSP 9632 PCI Audio Interface, Tucker Davis Technologies programmable attenuators, and MATLAB, which was also used for generating the EEG marker stream (see Fig. 1).

C. Paradigm

The stimuli were presented in 11 sessions, each of length 10 minutes, interrupted by short breaks. Among all participants, 8 participants were instructed to attend to the left speaker, while 10 participants were instructed to attend to the right speaker. The participants were also instructed to look at a fixation cross on a screen and minimize eye blinking. For each participant, the anechoic condition was always assigned to the first session and subsequently to every other third session (i.e. session 4, 7, and 10). Aiming at minimizing the influence of the speech material on AAD, the acoustic conditions (except for the anechoic condition) were randomly assigned to the other sessions. Following each session, the participants were asked to fill out a questionnaire consisting of 10 multiple-choice questions related to each story. The questionnaire was aimed to indicate whether the participants attended to the instructed speaker and whether the audio story was intelligible in the different acoustic conditions. The experiment for each participant took place on two different days.

D. EEG Setup and Signal Pre-Processing

The EEG responses were recorded using a BrainAmp system, provided by BrainProducts GmbH, Germany, and

$C = 64$ channels, provided by Easycap GmbH, Germany, with a sampling frequency of 500 Hz (see EEG experiment setup in Fig. 1). The EEG responses were referenced to the nose electrode and recorded using the Brain-Vision recorder software. The EEG recordings were re-referenced offline to a common average reference, band-pass filtered between 2 Hz and 8 Hz using a third-order Butterworth band-pass filter (as in [11], [13], [14]), and subsequently downsampled to $f_s = 64$ Hz. The envelopes of all considered 16 kHz speech signals were obtained using a Hilbert transform [33], followed by low-pass filtering at 8 Hz and downsampling to $f_s = 64$ Hz. For the training and decoding steps (see Section III), the EEG recordings of each session were split into 10 trials, each of length 60 seconds (see Table II). For filter training, the filter was computed using all considered trials based on (15) and (17), as proposed in [32] and [33], instead of computing a filter per trial and averaging per-trial filters as proposed in [11]. For filter training and evaluation, each participant's own data were used. The decoding performance was computed by averaging the percentage of correctly decoded trials over all considered trials and all participants.

V. Results and Discussion

In this section, the decoding performance of the least-squares-based AAD method is investigated for different acoustic conditions (see Table II) using the experimental setup discussed in the previous section. Section V-A discusses the results of the questionnaire. In Section V-B the impact of different acoustic conditions for the training and decoding steps is investigated. In Section V-C the impact of the head filtering effect is explored by comparing the decoding performance using either the clean or the anechoic speech signals. Finally, in Section V-D the influence of each acoustic component is investigated by comparing the decoding performance using reference and training signals affected by background noise, reverberation, and/or interfering speaker.

A. Questionnaire Analysis

For all considered acoustic conditions, Fig. 2 presents the correct answer scores related to the attended story, averaged across all participants. The highest score is obtained for the anechoic condition, while the lowest score is obtained for the reverberant-noisy condition. The statistical multiple comparison test (Kruskal-Wallis test followed by the post-hoc Dunn and Sidak test [40]) showed a significant difference (Kruskal-Wallis test: $\chi^2 = 19.0$, $p = 0.002$) in terms of the correct answer score between the anechoic condition and either the noisy or the reverberant-noisy condition (post-hoc Dunn and Sidak test: $p = 0.022$ and $p = 0.000$, respectively) and between the reverberant condition and the reverberant-noisy condition (post-hoc Dunn and Sidak test: $p = 0.013$), implying that – as expected – the noisy and the reverberant-noisy condition are more challenging.

B. Impact of Acoustic Conditions

For all considered EEG evaluation conditions, Fig. 3 presents the decoding performance for different EEG training conditions when the clean speech signals are used as reference and training signals.

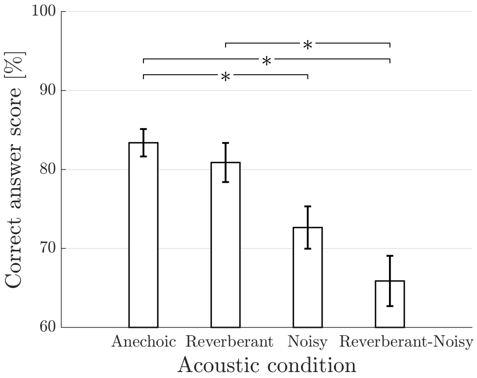


Fig. 2. The correct answer scores related to the attended story, averaged across all participants, for different acoustic conditions. Error bars represent one standard error around the mean and $*$ indicates a significant difference ($p < 0.05$) between acoustic conditions, based on the Kruskal-Wallis test followed by the post-hoc Dunn and Sidak test [40].

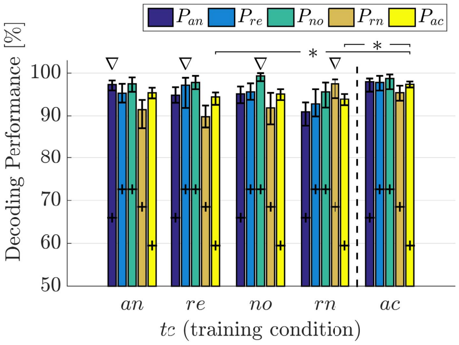


Fig. 3. The decoding performance for different EEG training and evaluation conditions when using the clean speech signals. The plus signs represent the upper boundary of the confidence interval corresponding to chance level, based on a binomial test at the 5% significance level, the error bars represent the bootstrap confidence interval at the 5% significance level, ∇ indicates the decoding performance when the EEG training and evaluation conditions are equal, and the dashed line separates the decoding performance obtained with filters trained in a specific acoustic condition and with filters trained in all *acoustic* conditions. A multiple comparison test (the Kruskal-Wallis test followed by the post-hoc Dunn and Sidak test) was performed across P_{ac} where $*$ indicates a significant difference ($p < 0.05$).

First, we investigate the feasibility of decoding EEG responses in different acoustic conditions $ec \in \{an, re, no, rn, ac\}$ when using filters trained using EEG responses in a *specific* acoustic condition $tc \in \{an, re, no, rn\}$ (i.e. left part of Fig. 3, separated by dashed line). When the EEG evaluation and training conditions are equal (indicated by ∇), it can be observed that a very good decoding performance (>96%) is obtained for all EEG evaluation conditions. These results are consistent with previous findings for the anechoic condition [12], [14], [16]–[19] as well as with recent findings for the reverberant and the reverberant-noisy conditions [32]. For each EEG training condition $tc \in \{an, re, no, rn\}$, it can be observed that the decoding performance when the EEG evaluation and training conditions are equal (indicated by ∇) is among the highest decoding performances for all EEG evaluation conditions. When the EEG evaluation and

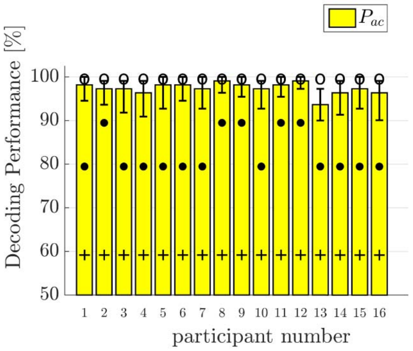


Fig. 4. The decoding performance P_{ac} per participant obtained with filters trained in all acoustic conditions and using clean speech signals. The plus signs represent the upper boundary of the confidence interval corresponding to chance level, based on a binomial test at the 5% significance level, the error bars represent the bootstrap confidence interval at the 5% significance level, the solid circles represent the minimum decoding performance and the void circles represent the maximum decoding performance.

training conditions are not equal, typically a lower decoding performance is obtained (except in some cases for the anechoic and the reverberant EEG training conditions). For example, for the reverberant-noisy EEG training condition the highest decoding performance is obtained for the reverberant-noisy EEG evaluation condition (>97%), while a lower decoding performance is obtained for the anechoic, reverberant, and noisy EEG evaluation conditions (>90%). In addition, for all EEG training conditions $tc \in \{an, re, no, rn\}$ it can be observed that the average decoding performance for all conditions P_{ac} is considerably high (>93%).

Secondly, we investigate the feasibility of decoding EEG responses in different acoustic conditions $ec \in \{an, re, no, rn, ac\}$ when using filters trained using EEG responses in *all acoustic conditions* $tc = ac$ (i.e. right part of Fig. 3, separated by dashed line). It can be observed that a very good decoding performance (>95%) is obtained for all EEG evaluation conditions and that the decoding performance across EEG evaluation conditions is more consistent compared to when using filters trained in a specific acoustic condition. In addition, the average decoding performance for all conditions P_{ac} obtained with filters trained in all conditions is occasionally significantly larger than with filters trained in a specific acoustic condition.[1] For example, the decoding performance P_{ac} obtained with filters trained in all conditions ($tc = ac$) is significantly larger than with filters trained either in the reverberant condition ($tc = re$) or in the reverberant-noisy condition ($tc = rn$) (Kruskal-Wallis test: $\chi^2 = 16.5$, $p = 0.002$; post-hoc Dunn and Sidak test comparisons of $tc = ac$ with $tc = re$ and $tc = rn$: $p = 0.020$, $p = 0.001$, respectively). To investigate how much the average decoding performance for all conditions P_{ac} varies across participants, Fig. 4 presents P_{ac} per participant, obtained with filters trained in all conditions. It can be observed that the decoding performance per participant ranges between 80% and 100%.

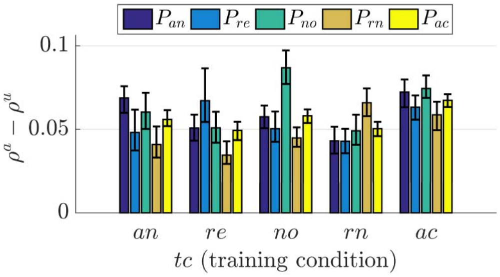


Fig. 5. Average correlation differences for different EEG training and evaluation conditions when using the clean speech signals. The error bars represent the bootstrap confidence interval at the 5% significance level.

The feasibility of using either filters trained in a specific acoustic condition or filters trained in all acoustic conditions to perform AAD in different acoustic conditions may be explained by considering the robust neural responses to degraded – but still intelligible – speech signals. Several studies have shown that auditory cortical responses resemble the clean attended speech signal more than the speech signal degraded by different acoustic components (e.g., background noise, interfering speaker), suggesting a robust neural representation of the clean attended speech signal [6], [7], [30], [31], [41]. To decode auditory attention, the trained filters aim at reconstructing the clean attended speech envelope from EEG responses that are largely invariant to degradations. Hence, the reconstructed attended envelope is expected to be more correlated to the clean attended speech envelope than to the clean unattended speech envelope, i.e. the correlation difference $(\rho^a - \rho^u)$ is expected to be larger than zero. For all considered EEG evaluation conditions, Fig. 5 presents the correlation difference for different EEG training conditions, averaged across all considered trials and participants (note that these average correlation coefficients are not directly used for decoding). It can be observed that a correlation difference significantly larger than zero is obtained for all considered acoustic conditions, which is consistent with a robust neural representation of the clean attended speech signal.

Finally, we investigate the parameters involved in the filter design (Δ, L, β) across EEG training conditions. Fig. 6 depicts the optimal parameter values (see Section III-B), averaged across all considered trials and all participants. It can be observed that the optimal value for Δ varies only slightly between 93.8 ms to 101.6 ms, while the optimal value for L varies more substantially between 109.3 ms to 128.9 ms. Accordingly, the EEG responses contributing most to the AAD performance are those with latencies between 93.8 ms and 230.5 ms, consistent with previous findings in [13], [17], and [30]. In addition, the optimal value for the regularization parameter β varies between 10^{-1} to 10^{2}. It can be observed that the optimal regularization parameter is smaller when using filters trained in all conditions than when using filters trained in a specific acoustic condition. A possible explanation may be that training in all conditions can by itself be considered as some form of regularization, consistent with previous finding in [33].

[1] We also performed the AAD experiment using filters trained with 40 trials that were arbitrarily selected from different acoustic conditions and observed similar findings.

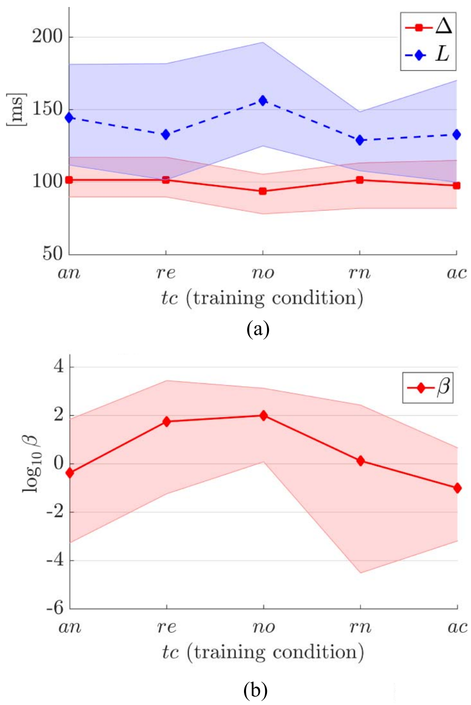


Fig. 6. The optimal values for the filter parameters (a) Δ and L, and (b) the regularization parameter β, averaged across all trials and all participants when using the clean speech signal. The shaded area indicates the bootstrap confidence interval at the 5% significance level.

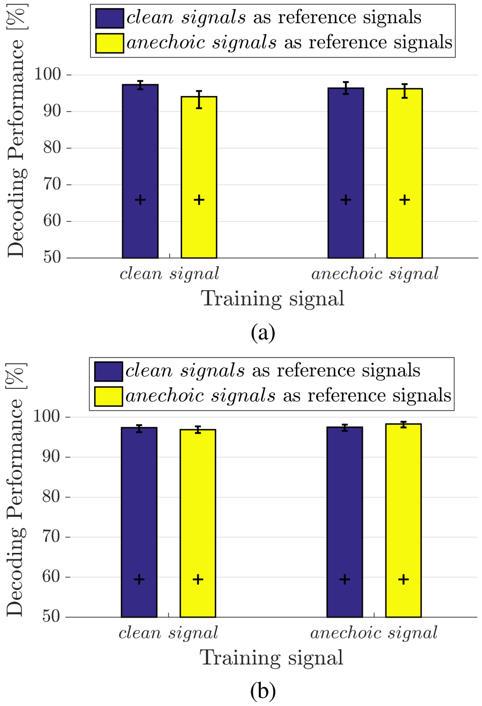


Fig. 7. Influence of head filtering effect on AAD. Comparison of decoding performance using either the clean or the anechoic speech signals when the EEG evaluation and training conditions are equal to (a) the anechoic condition or (b) all conditions. The plus signs represent the upper boundary of the confidence interval corresponding to chance level based on a binomial test at the 5% significance level, the error bars represent the bootstrap confidence interval at the 5% significance level.

In summary, the results in this section show the feasibility of using either filters trained in a specific acoustic condition or filters trained in all conditions to perform AAD in different acoustic conditions.[2] While these results were obtained using the clean speech signals as training and reference signals, in the next sections we will investigate in more detail the influence of the different acoustic components (head filtering effect, reverberation, background noise, interfering speaker) in the training and reference signals.

C. Influence of Head Filtering Effect

In this section, we investigate the influence of the head filtering effect by comparing the decoding performance when using clean or anechoic speech signals either as training or as reference signals. Fig. 7a presents the decoding performance for the anechoic condition ($ec = an$) when using filters trained in the anechoic condition ($tc = an$). Fig. 7b presents the average decoding performance for all conditions ($ec = ac$) when using filters trained in all conditions ($tc = ac$). A paired Wilcoxon signed rank test revealed no significant difference ($p > 0.05$) between using either the clean speech signals or the anechoic speech signals as training or as reference signals. These results indicate that for all considered acoustic conditions head filtering effects have no significant influence on the decoding performance.

D. Influence of Background Noise, Reverberation and Interfering Speaker

To investigate the influence of each acoustic component on AAD, Fig. 8 presents the decoding performance for all considered acoustic conditions (anechoic, reverberant, noisy, reverberant-noisy) using the following signals as training signals or as reference signals:

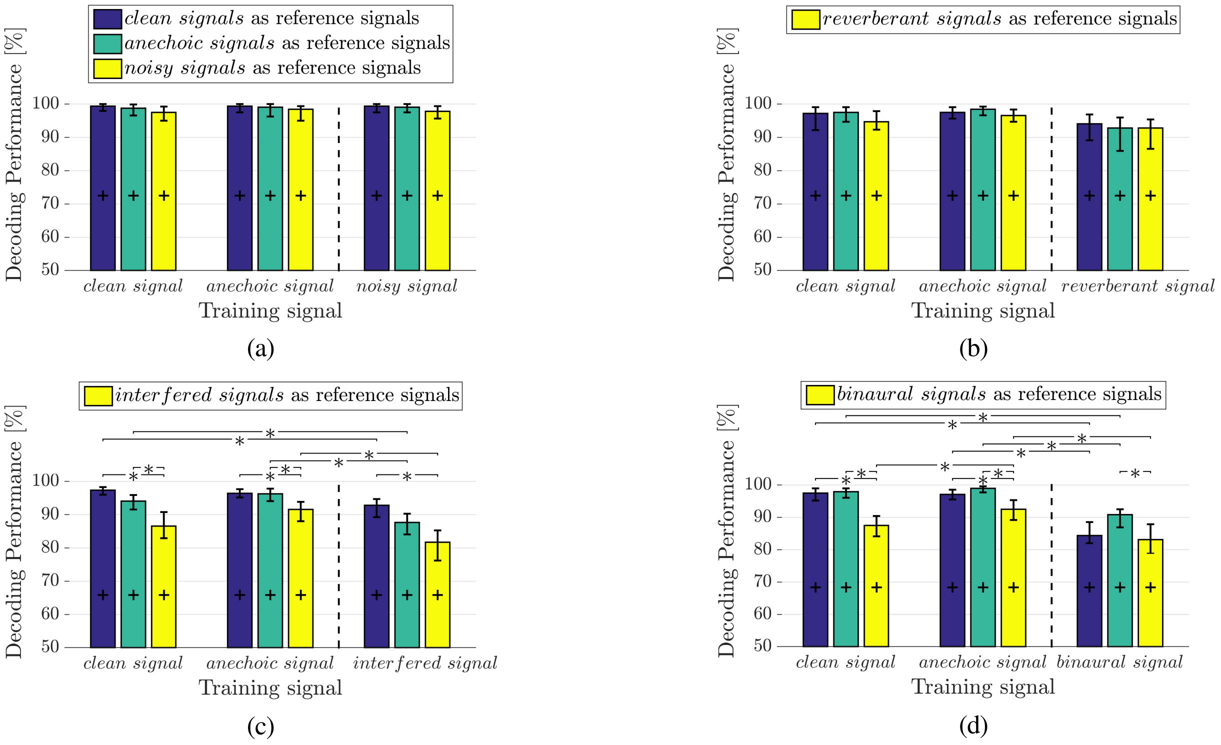
- the clean speech signals s^a and s^u.
- the anechoic speech signals $x_m^{a,an}$ and $x_m^{u,an}$.
- the anechoic speech signals affected by different acoustic components, i.e. the noisy speech signals $x_m^{a,no}$ and $x_m^{u,no}$ in (3) for the noisy condition, the reverberant speech signals x_m^a and x_m^u in (1) for the reverberant condition, the interfered speech signal x_m^{an} (attended and unattended side) in (2) for the anechoic condition,[3] and the binaural speech signals y_m (attended and unattended side) in (1) for the reverberant-noisy condition.

Similarly, Fig. 9 presents the correlation difference ($\rho^a - \rho^u$), averaged across all considered trials and participants (note that these average correlation coefficients are not directly used for decoding).

First, we investigate the case where the clean or the anechoic attended speech signal is used as training signal (i.e. left part of Fig. 8 and 9, separated by dashed line). When using the clean or anechoic speech signals as reference signals,

[2] We also performed the AAD experiment using trial lengths of 30 seconds and observed similar findings.

[3] The interfered speech signal is used in the anechoic condition to exclude the influence of other acoustic components (background noise and reverberation) on the analysis.



(a) (b) (c) (d)

Fig. 8. Influence of different acoustic components (background noise, reverberation and interfering speaker) on AAD. Comparison of decoding performance when using (a) the noisy speech signals in the noisy condition, (b) the reverberant speech signals in the reverberant condition, (c) the interfered speech signals in the anechoic condition, (d) the binaural speech signals in the reverberant-noisy condition, either as training signal or as reference signals. The plus signs represent the upper boundary of the confidence interval corresponding to chance level based on a binomial test at the 5% significance level, and the error bars represent the bootstrap confidence interval at the 5% significance level. The dashed line separates the case where the clean or the anechoic attended speech signals are used as training signals and the case where the attended speech signals affected by different acoustic components are used as training signals. A paired Wilcoxon signed rank test was performed between the decoding performance using the clean or the anechoic speech signals as reference signals and using the anechoic speech signals affected by different acoustic components as reference signals. In addition, a paired Wilcoxon signed rank test was performed between the decoding performance using the clean or the anechoic speech signals as training signals and using the anechoic speech signals affected by different acoustic components as training signals. $*$ indicates a significant difference ($p < 0.05$) based on the paired Wilcoxon signed rank test.

a very good decoding performance (>94%) is obtained for all acoustic conditions, as already shown in Fig. 7. When using the noisy speech signals (in the noisy condition, Fig. 8a) or the reverberant speech signals (in the reverberant condition, Fig. 8b) as reference signals, there is no significant difference in decoding performance ($p > 0.05$) compared to when using the clean or anechoic speech signals as reference signals. On the other hand, when using the interfered speech signals (in the anechoic condition, Fig. 8c) or the binaural speech signals (in the reverberant-noisy condition, Fig. 8d) as reference signals, the decoding performance is significantly lower ($p < 0.05$) than when using the clean or anechoic speech signals as reference signals, although the decoding performance is still considerably large (>87%). The feasibility of using either the interfered speech signals or the binaural speech signals as reference signals for AAD can be explained by considering the broadband energy ratio between the attended and unattended speech components in the signals at the ears. As already mentioned in Section II, due to the head filtering effect this broadband energy ratio is smaller at the side of the unattended speaker than at the side of the attended speaker. In summary, the results in Fig. 8 (left side) show that when using reference signals affected by reverberation or background noise, a comparable decoding performance can be obtained as when using clean or anechoic speech signals, whereas when using reference signals affected by the interfering speaker the decoding performance significantly decreases. This also suggests that in order to generate appropriate reference signals, it is more important to reduce the interfering speaker than to reduce background noise or reverberation.

The decoding performance results in Fig. 8 can be further explained by considering the influence of each acoustic component on the correlation difference in Fig. 9. For the noisy condition (Fig. 9a), there are no significant differences between the considered reference signals, which corresponds to the decoding performance results in Fig. 8a. For the reverberant condition (Fig. 9b), it can be observed that the correlation differences significantly decrease ($\rho^a - \rho^u < 0.04$) when using the reverberant speech signals as reference signals, but only when using the clean attended speech signal as training signal. Nevertheless, this lower correlation difference does not result in a significantly lower decoding performance in Fig. 8b. For the anechoic condition (Fig. 9c) and the reverberant-noisy condition (Fig. 9d), it can be observed that the correlation differences significantly decrease when using the interfered

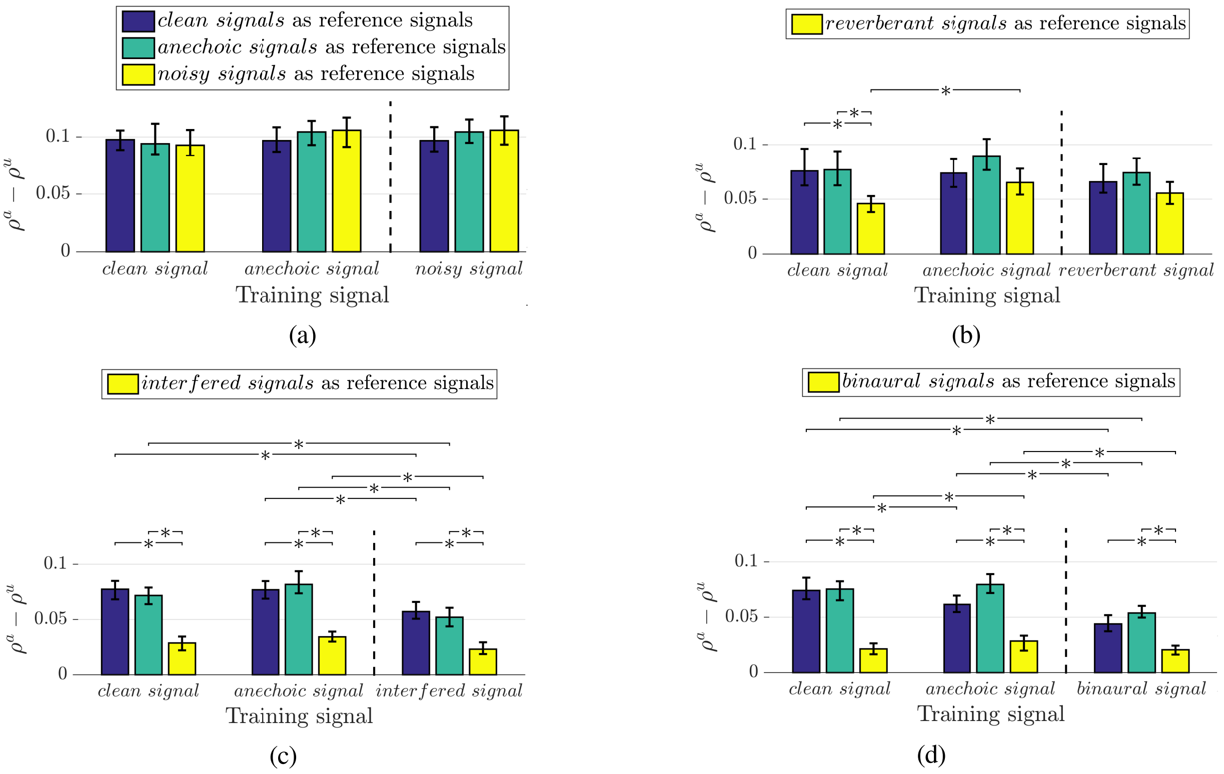


Fig. 9. Influence of different acoustic components (background noise, reverberation and interfering speaker) on AAD. Comparison of correlation difference when using (a) the noisy speech signals in the noisy condition, (b) the reverberant speech signals in the reverberant condition, (c) the interfered speech signals in the anechoic condition, (d) the binaural speech signals in the reverberant-noisy condition, either as training signal or as reference signals. The error bars represent the bootstrap confidence interval at the 5% significance level, and $*$ indicates a significant difference ($p < 0.05$) based on the paired Wilcoxon signed rank test.

speech signals ($\rho^a - \rho^u < 0.03$) or the binaural speech signals ($\rho^a - \rho^u < 0.02$) as reference signals. These lower correlation differences are also reflected by significantly lower corresponding decoding performances in Fig. 8c and 8d.

Secondly, we explore the potential of using the attended speech signal affected by different acoustic components as training signal (i.e. right part of Fig. 8 and 9, separated by dashed line). On the one hand, when using the noisy attended speech signal (in the noisy condition, Fig. 8a) or the reverberant attended speech signal (in the reverberant condition, Fig. 8b) as training signal, there is no significant difference in decoding performance ($p > 0.05$) compared to when using the clean or the anechoic attended speech signal as training signal (for all considered reference signals). On the other hand, when using the interfered attended speech signal (in the anechoic condition, Fig. 8c) or the binaural attended speech signal (in the reverberant-noisy condition, Fig. 8d) as training signal, the decoding performance is significantly lower compared to when using either the clean or the anechoic attended speech signal as training signal (for all considered reference signals). Nevertheless, even when using the binaural attended speech signal as training signal in the reverberant-noisy condition, it is still feasible to perform AAD with a decoding performance larger than 82%. The decoding performance results in Fig. 8 when using attended speech signals affected by different acoustic components as training signal are mostly consistent with the correlation differences in Fig. 9.

TABLE III
ACOUSTIC SIGNAL AS TRAINING OR REFERENCE SIGNALS USING WHICH THE LARGEST DECODING PERFORMANCE FOR A SPECIFIC EEG (TRAINING AND EVALUATION) CONDITION IS OBTAINED

EEG Condition	Acoustic Signal
Noisy	Clean, anechoic and noisy signals
Reverberant	Clean, anechoic and reverberant signals
Anechoic	Clean and anechoic signals
Reverberant-noisy	Clean and anechoic signals

In summary, the results in this section show that using speech signals affected by background noise and reverberation as training or reference signals results in a decoding performance that is comparable to using the clean or anechoic speech signals as training or reference signals. On the contrary, using speech signals affected by the interfering speaker as training or reference signals typically results in a significantly lower decoding performance. Table III presents which training/reference signals lead to the largest decoding performance for a specific EEG training/evaluation condition.

VI. CONCLUSIONS

In this paper, we investigated the performance of the least-squares-based AAD method for different acoustic conditions (anechoic, reverberant, noisy, and reverberant-noisy), both in the training step as well as in the decoding step. The experimental results showed that for all considered acoustic conditions it is possible to decode auditory attention with a considerably large decoding performance, even when the acoustic conditions for training and decoding are different. In addition, for most acoustic conditions there is no significant difference in decoding performance when using filters trained in all conditions or filters trained in a specific condition. This suggests that for an unseen realistic acoustic condition AAD can be performed using filters trained in, e.g., a laboratory acoustic condition.

Furthermore, we investigated the influence of the head filtering effect and of acoustic components (reverberation, background noise and interfering speaker) on the decoding performance. The experimental results showed that for all considered acoustic conditions the head filtering effect has no significant impact on the decoding performance. Moreover, when using speech signals affected by either reverberation or background noise as reference signals, a comparable decoding performance is obtained as when using clean speech signals as reference signals. On the contrary, when using speech signals affected by the interfering speaker as reference signals, the decoding performance significantly decreases. This suggests that for generating appropriate reference signals, e.g., using acoustic signal pre-processing algorithms, it is more important to reduce the interfering speaker than to reduce background noise or reverberation. Furthermore, when using the binaural speech signals as reference signals for decoding, a relatively large decoding performance can be obtained. This implies that decoding is feasible for the considered scenario even based on the unprocessed noisy and reverberant signals.

Finally, we explored the potential of using the attended speech signal affected by different acoustic components as training signal for computing the filter. When using attended speech signals affected by either reverberation or by background noise as training signal, a comparable decoding performance is obtained as when using the clean attended speech signal as training signal. However, when using attended speech signals affected by the interfering speaker as training signal, the decoding performance may significantly decrease. Nevertheless, even when using the binaural attended speech signal as training signal, it is still feasible to achieve a large decoding performance.

While the discussion in this paper has been limited to the least-squares-based AAD method, in which auditory attention is decoded using an envelope reconstruction model, AAD approaches based on empirical mode decomposition [42], [43], inherent fuzzy entropy [44] or a neural encoding model [45], [46] have not been investigated in this paper. Further work could therefore include a study on how reverberation and noise influence these AAD approaches.

References

[1] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. Cambridge, MA, USA: MIT Press, 1997.

[2] J. C. Middlebrooks, J. Z. Simon, A. N. Popper, and R. R. Fay, *The Auditory System at the Cocktail Party*. New York, NY, USA: Springer, 2017.

[3] B. G. Shinn-Cunningham and V. Best, “Selective attention in normal and impaired hearing,” *Trends Amplification*, vol. 12, no. 4, pp. 283–299, 2008.

[4] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, “Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones,” *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 18–30, Mar. 2015.

[5] S. Gannot *et al.*, “A consolidated perspective on multimicrophone speech enhancement and source separation,” *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[6] 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, May 2012.

[7] N. Ding and J. Z. Simon, “Emergence of neural encoding of auditory objects while listening to competing speakers,” *Proc. Nat. Acad. Sci.*, vol. 109, no. 29, pp. 11854–11859, 2012.

[8] E. B. Petersen and M. Wöstmann, J. Obleser, and T. Lunner, “Neural tracking of attended versus ignored speech is differentially affected by hearing loss,” *J. Neurophysiol.*, vol. 117, no. 1, pp. 18–27, 2017.

[9] N. Ding, M. Chatterjee, and J. Z. Simon, “Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure,” *Neuroimage*, vol. 88, pp. 41–46, Mar. 2014.

[10] P. Patel, L. K. Long, J. L. Herrero, A. D. Mehta, and N. Mesgarani, “Joint representation of spatial and phonetic features in the human core auditory cortex,” *Cell Rep.*, vol. 24, no. 8, pp. 2051–2062, 2018.

[11] J. A. O’Sullivan *et al.*, “Attentional selection in a cocktail party environment can be decoded from single-trial EEG,” *Cerebral Cortex*, vol. 25, no. 7, pp. 1697–1706, 2014.

[12] 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, 2016, Art. no. 056014.

[13] B. Mirkovic, S. Debener, M. Jaeger, and M. De Vos, “Decoding the attended speech stream with multi-channel EEG: Implications for online, daily-life applications,” *J. Neural Eng.*, vol. 12, no. 4, p. 46007, 2015.

[14] B. Mirkovic, M. G. Bleichner, M. De Vos, and S. Debener, “Target speaker detection with concealed EEG around the ear,” *Frontiers Neurosci.*, vol. 10, p. 349, Jul. 2016.

[15] L. Fiedler, M. Wöestmann, C. Graversen, A. Brandmeyer, T. Lunner, and J. Obleser, “Single-channel in-Ear-EEG detects the focus of auditory attention to concurrent tone streams and mixed speech,” *J. Neural Eng.*, vol. 14, no. 3, 2017, Art. no. 036020.

[16] 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*, Nov. 2017.

[17] A. Aroudi, B. Mirkovic, M. De Vos, and S. Doclo, “Auditory attention decoding with EEG recordings using noisy acoustic reference signals,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 694–698.

[18] S. Van Eyndhoven, T. Francart, and A. Bertrand, “EEG-informed attended speaker extraction from recorded speech mixtures with application in neuro-steered hearing prostheses,” *IEEE Trans. Biomed. Eng.*, vol. 64, no. 5, pp. 1045–1056, May 2017.

[19] N. Das, S. Van Eyndhoven, T. Francart, and A. Bertrand, “Adaptive attention-driven speech enhancement for EEG-informed hearing prostheses,” in *Proc. 38th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Orlando, FL, USA, Aug. 2016, pp. 77–80.

[20] J. O’Sullivan *et al.*, “Neural decoding of attentional selection in multi-speaker environments without access to clean sources,” *J. Neural Eng.*, vol. 14, no. 5, 2017, Art. no. 056001.

[21] A. Aroudi, D. Marquardt, and S. Doclo, “EEG-based auditory attention decoding using steerable binaural superdirective beamformer,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 851–855.

[22] B. Ekin, L. Atlas, M. Mirbagheri, and A. K. C. Lee, “An alternative approach for auditory attention tracking using single-trial EEG,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 729–733.

[23] 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.

[24] 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, 2018, Art. no. 066017.

[25] C. J. Darwin and R. W. Hukin, “Effects of reverberation on spatial, prosodic, and vocal-tract size cues to selective attention,” *J. Acoust. Soc. Amer.*, vol. 108, no. 1, pp. 335–342, 2000.

[26] J. F. Culling, K. I. Hodder, and C. Y. Toh, "Effects of reverberation on perceptual segregation of competing voices," *J. Acoust. Soc. Amer.*, vol. 114, no. 5, pp. 2871–2876, 2003.

[27] P. M. Zurek, R. L. Freyman, and U. Balakrishnan, "Auditory target detection in reverberation," *J. Acoust. Soc. Amer.*, vol. 115, no. 4, pp. 1609–1620, 2004.

[28] D. Ruggles and B. Shinn-Cunningham, "Spatial selective auditory attention in the presence of reverberant energy: Individual differences in normal-hearing listeners," *J. Assoc. Res. Otolaryngol.*, vol. 12, no. 3, pp. 395–405, 2011.

[29] S. Anderson, E. Skoe, B. Chandrasekaran, and N. Kraus, "Neural timing is linked to speech perception in noise," *J. Neurosci.*, vol. 30, no. 14, pp. 4922–4926, 2010.

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

[31] S. A. Fuglsang, T. Dau, and J. Hjortkjær, "Noise-robust cortical tracking of attended speech in real-world acoustic scenes," *NeuroImage*, vol. 156, pp. 435–444, Apr. 2017.

[32] A. Aroudi and S. Doclo, "EEG-based auditory attention decoding using unprocessed binaural signals in reverberant and noisy conditions?" in *Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jeju Province, South Korea, Jul. 2017, pp. 484–488.

[33] 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, May 2017.

[34] H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses," *EURASIP J. Adv. Signal Process.*, vol. 2009, p. 6, Jan. 2009.

[35] M. Jeub and M. Schäfer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proc. 16th Int. Conf. Digit. Signal Process. (DSP)*, Santorini-Hellas, Greece, Jul. 2009, pp. 1–5.

[36] J. Thiemann and S. Van De Par, "Multiple model high-spatial resolution HRTF measurements," in *Proc. DAGA*, 2015, pp. 1–2.

[37] H. Ohrka. (2012). *Ohrka.de-Kostenlose Hörabenteuer Für Kinderohren*. Accessed: Apr. 30, 2015. [Online]. Available: http://www.ohrka.de

[38] E. Hering, *Kostbarkeiten Aus Dem Deutschen Märchenschatz*. Audiopool Hörbuchverlag MP3 CD, 2011.

[39] E. A. P. Habets, I. Cohen, and S. Gannot, "Generating nonstationary multisensor signals under a spatial coherence constraint," *J. Acoust. Soc. Amer.*, vol. 124, no. 5, pp. 2911–2917, Nov. 2008.

[40] Y. Hochberg and A. C. Tamhane, *Multiple Comparison Procedures*. Hoboken, NJ, USA: Wiley, 1987.

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

[42] D. Looney, C. Park, Y. Xia, P. Kidmose, M. Ungstrup, and D. P. Mandic, "Towards estimating selective auditory attention from EEG using a novel time-frequency-synchronisation framework," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2010, pp. 1–5.

[43] O. Etard, M. Kegler, C. Braiman, A. E. Forte, and T. Reichenbach, "Real-time decoding of selective attention from the human auditory brainstem response to continuous speech," *bioRxiv*, Feb. 2018. [Online]. Available: https://www.biorxiv.org/content/early/2018/02/05/259853

[44] Z. Cao and C.-T. Lin, "Inherent fuzzy entropy for the improvement of EEG complexity evaluation," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 2, pp. 1032–1035, Apr. 2018.

[45] S. Akram, J. Z. Simon, and B. Babadi, "Dynamic estimation of the auditory temporal response function from MEG in competing-speaker environments," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 8, pp. 1896–1905, Aug. 2017.

[46] S. Miran, S. Akram, A. Sheikhattar, J. Z. Simon, T. Zhang, and B. Babadi, "Real-time tracking of selective auditory attention from M/EEG: A Bayesian filtering approach," *Frontiers Neurosci.*, vol. 12, p. 262, May 2018. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnins.2018.00262