*algorithms*

MDPI

*Article*

# Closed-Loop Cognitive-Driven Gain Control of Competing Sounds Using Auditory Attention Decoding

**Ali Aroudi [1,2,*], Eghart Fischer [1], Maja Serman [1], Henning Puder [1] and Simon Doclo [2]**

[1] WS Audiology, 91058 Erlangen, Germany; eghart.fischer@sivantos.com (E.F.); maja.serman@sivantos.com (M.S.); henning.puder@wsa.com (H.P.)
[2] Department of Medical Physics and Acoustics and Cluster of Excellence Hearing4all, University of Oldenburg, 26122 Oldenburg, Germany; simon.doclo@uol.de
* Correspondence: ali.aroudi@uni-oldenburg.de

**Abstract:** Recent advances have shown that it is possible to identify the target speaker which a listener is attending to using single-trial EEG-based auditory attention decoding (AAD). Most AAD methods have been investigated for an open-loop scenario, where AAD is performed in an offline fashion without presenting online feedback to the listener. In this work, we aim at developing a closed-loop AAD system that allows to enhance a target speaker, suppress an interfering speaker and switch attention between both speakers. To this end, we propose a cognitive-driven adaptive gain controller (AGC) based on real-time AAD. Using the EEG responses of the listener and the speech signals of both speakers, the real-time AAD generates probabilistic attention measures, based on which the attended and the unattended speaker are identified. The AGC then amplifies the identified attended speaker and attenuates the identified unattended speaker, which are presented to the listener via loudspeakers. We investigate the performance of the proposed system in terms of the decoding performance and the signal-to-interference ratio (SIR) improvement. The experimental results show that, although there is a significant delay to detect attention switches, the proposed system is able to improve the SIR between the attended and the unattended speaker. In addition, no significant difference in decoding performance is observed between closed-loop AAD and open-loop AAD. The subjective evaluation results show that the proposed closed-loop cognitive-driven system demands a similar level of cognitive effort to follow the attended speaker, to ignore the unattended speaker and to switch attention between both speakers compared to using open-loop AAD. Closed-loop AAD in an online fashion is feasible and enables the listener to interact with the AGC.

**Keywords:** auditory attention decoding; adaptive gain control; speech enhancement; EEG; brain computer interface



**Citation:** Aroudi, A.; Fischer, E.; Serman, M.; Puder, H.; Doclo, S. Closed-Loop Cognitive-Driven Gain Control of Competing Sounds Using Auditory Attention Decoding. *Algorithms* **2021**, *14*, 287. https://doi.org/10.3390/a14100287

Academic Editor: Maryam Ravan

Received: 9 August 2021
Accepted: 29 September 2021
Published: 30 September 2021



## 1. Introduction

Hearing aids aim at restoring the normal hearing abilities by several processing steps including speech enhancement. The main objective of speech enhancement is to improve the intelligibility of the recorded microphone signals, which are often corrupted by various noise sources [1,2]. In a scenario with multiple competing speakers, the performance of many speech enhancement algorithms, for example, beamforming and blind source separation, depends on correctly identifying the target speaker, i.e., the speaker which the listener is attending to.

Recent advances in electroencephalography (EEG) have shown that it is possible to identify the target speaker from single-trial EEG recordings [3–18], which are non-invasive and have appropriate temporal resolution for auditory stimuli. Several single-trial EEG-based auditory attention decoding (AAD) methods have been proposed to identify the speaker which the listener is attending to, aiming to be incorporated in a real-world applications, e.g., to control a hearing aid. AAD methods aim at identifying the attended speaker by relating the EEG responses of the listener to speech signals of speakers. These

methods are based on, for example, a least-squares cost function [3–7,11,14,16], canonical correlation analysis [8,13], a state-space model [9,18], and neural networks [12,15,17]. The least-squares-based AAD method used in Reference [3,5–7,11,14] aims at reconstructing the attended speech envelope from the EEG responses of the listener using a trained spatio-temporal envelope estimator. To identify the attended speaker, the reconstructed speech envelope is compared with the speech envelopes of the competing speakers using correlation coefficients. Since these correlation coefficients are typically highly fluctuating, a large correlation window of about 30 s is typically required to obtain a reliable decoding performance, which causes a large processing delay [9,18,19].

The possibility of decoding auditory attention from EEG recordings has led to an increasing research interest in the topic of incorporating AAD in a brain-computer interface for real-world applications, for example, to cognitively drive speech enhancement algorithms [19–25]. Cognitive-driven speech enhancement algorithms potentially provide the listener with the ability to selectively attend to a specific speaker. It should, however, be noted that the performance of most aforementioned AAD methods and cognitive-driven speech enhancement algorithms has been investigated for an open-loop scenario, where AAD is performed in an offline fashion without presenting online feedback to the listener. In addition, scenarios with no attention switch between speakers have typically been investigated, which is unrealistic in practice.

To investigate the performance of AAD for real-world applications, closing the loop by presenting feedback according to the AAD results in an online fashion is of crucial importance. Feedback presentation may influence the subsequent intent of the listener and the brain signals that encode that intent. In Reference [5], the feasibility of a closed-loop system based on least-squares-based AAD has been shown by presenting the AAD results as visual feedback, i.e., using different colors or a sphere with different radii. However, the feasibility of closed-loop AAD enabling the listener to interact with speech enhancement in an online fashion and allowing the listener to switch attention between speakers remains to be investigated.

In this paper, we aim at developing a closed-loop AAD system that allows to enhance a target speaker, suppress an interfering speaker and switch attention between both speakers. Specifically, we propose a cognitive-driven adaptive gain controller (AGC), which is based on real-time AAD (RAAD). The RAAD first generates correlation coefficients for both speakers from the EEG responses of the listener and the speech signals of both speakers. To this end, we adopted the least-squares-based AAD method from Reference [3], either using a small correlation window of length 0.25 s or a large correlation window of length 15 s. The fluctuating correlation coefficients are then translated into more reliable probabilistic attention measures, based on which the attended and the unattended speaker are identified. To this end, we propose an AAD algorithm either using a generalized linear model (GLM) or using a state-space model (SSM), similarly to Reference [9,18]. The AGC as an ideal speech enhancement algorithm then amplifies the identified attended speaker and attenuates the identified unattended speaker, where the gains for both speakers are based on the probabilistic attention measures. Finally, the loop of cognitive-driven gain control is closed by presenting the amplified attended speaker and the attenuated unattended speaker to the listener via loudspeakers, enabling the listener to interact with the AGC in an online fashion and switch attention between speakers. For an acoustic scenario comprising two competing speakers where one speaker is located on the left side and the other speaker is located on the right side, we investigate the decoding performance and the speech enhancement performance of the proposed closed-loop cognitive-driven gain controller system with 10 participants based on objective and subjective evaluations. In addition, we provide a detailed analysis and experimental comparison between the open-loop and the closed-loop AAD system using either the GLM or the SSM.

The paper is organized as follows. In Section 2, we introduce the experiment protocol used to calibrate and evaluate the proposed cognitive-driven gain controller system, describe the stimuli and the data acquisition used for the experiments, and present the

proposed cognitive-driven gain controller system. In Section 3, we evaluate the decoding performance and the speech enhancement performance of the proposed cognitive-driven gain controller system. In Section 4, we discuss the experimental results in more detail, summarize the main contributions, and suggest possible topics for further research.

## 2. Methods

### *2.1. Experiment Protocol*

In this section, we present the experiment protocol used to calibrate and evaluate the cognitive-driven gain controller system. The experiment protocol consists of a calibration phase, an open-loop AAD phase, and a closed-loop AAD phase (see Figure 1).

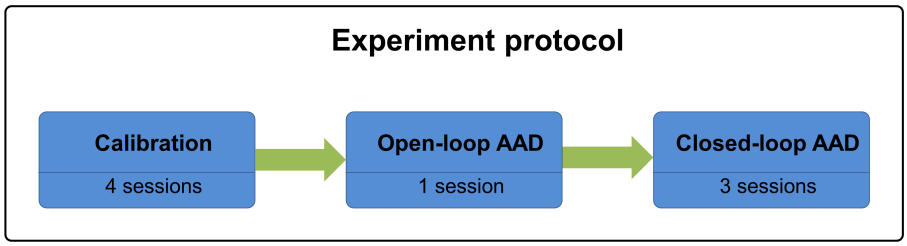



**Figure 1.** Experiment protocol used to calibrate and evaluate the cognitive-driven gain controller system. The experiment protocol consists of a calibration phase with four sessions, an open-loop AAD phase with one session and a closed-loop AAD phase with three sessions.

#### 2.1.1. Calibration Phase

In the calibration phase, the cognitive-driven gain controller system was individually calibrated for each participant using the EEG responses for a scenario with two competing speakers (see Figure 2). Participants were cued by an arrow on a screen to listen attentively to one of the speakers while recording the ongoing EEG responses. Participants were also instructed to minimize eye movement and blinking, which may cause EEG artifacts. The EEG responses were recorded during four sessions, lasting 30 min in total. The first and the second session each lasted 10 min, while the third and the fourth session each lasted 5 min. For the first and the third session, the participants were cued to attend to the left speaker, whereas, for the second and the fourth session, the participants were cued to attend to the right speaker. Following each session, the participants were asked to fill out a questionnaire consisting of multiple-choice questions about the stories uttered by the speakers. There was one question per minute of each story. The questionnaire was used to check whether the participants attended to the cued speaker. After the fourth session, there was a short break. During this break, the recorded EEG responses were used to calibrate the cognitive-driven gain controller system (see Section 2.5), individualized per participant.

#### 2.1.2. Open-Loop AAD Phase

In the open-loop AAD phase, the calibrated AAD algorithms were used to identify the attended and the unattended speaker without presenting feedback to the participants. The open-loop AAD phase consisted of one session lasting 10 min. During this session, participants were cued by an arrow on a screen every minute to switch attention between the competing speakers while recording the ongoing EEG responses. Afterwards, participants were asked to rate how much effort it took to follow the attended speaker, to ignore the unattended speaker, and to switch attention to the cued speaker on a scale from 0 to 10, with 0 being least effort and 10 being most effort. In addition, participants were asked to rate how well they understood the attended story on a scale from 0 to 10, with 0 being nothing understood and 10 being everything understood. Furthermore, the participants were asked to fill out a questionnaire consisting of multiple-choice questions about the stories uttered by the speakers. While participants were rating and answering the questionnaire, the decoding performance of several AAD algorithms (see Section 2.5.1)

was evaluated using the recorded EEG responses. The following AAD algorithms were considered:

- LW–GLM: AAD algorithm using a generalized linear model (GLM) with a large correlation window (LW) of 15 s.
- LW–SSM: AAD algorithm using a state-space model (SSM) with a large correlation window (LW) of 15 s.
- SW–SSM: AAD algorithm using a state-space model with a small correlation window (SW) of 0.25 s. Using a small correlation window was motivated by the results in Reference [9], where it was shown that the state-space model is able to translate highly fluctuating coefficients of the spatio-temporal envelope estimators into reliable probabilistic attention measures.

Note that, in this paper, an AAD algorithm using a generalized linear model with a small correlation window was not considered, since initial experiments showed highly fluctuating correlation coefficients with unreliable probabilistic attention measures.

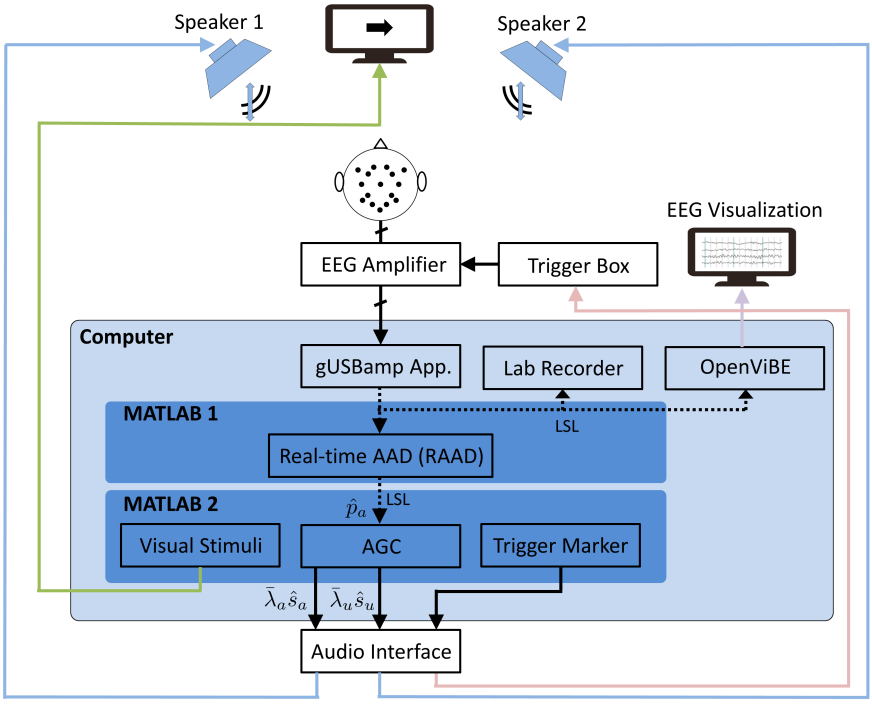



**Figure 2.** Overview of the proposed cognitive-driven gain controller system for a scenario with two competing speakers. The EEG responses were acquired using the EEG amplifier. The acquired EEG responses and EEG trigger markers were streamed using the gUSBamp application. Using the gUSBamp application and the LSL software package, the streamed EEG responses were forwarded to the real-time AAD (RAAD) for online decoding, to the Lab Recorder application for recording, and to the OpenViBE software for online EEG visualization. The RAAD was implemented and run using MATLAB (MATLAB 1). The RAAD identified the attended and the unattended speaker and generated their corresponding probabilistic attention measure. The generated probabilistic attention measure of the attended speaker ($\hat{p}_a$) was forwarded to the AGC using the LSL software package. Based on the probabilistic attention measure, the AGC amplified the attended speaker ($\bar{\lambda}_a \hat{s}_a$) and attenuated the unattended speaker ($\bar{\lambda}_u \hat{s}_u$) as acoustic stimuli. The AGC (together with trigger marker and visual stimuli) was implemented and run using MATLAB (MATLAB 2). The AAD loop was then closed by presenting the acoustic stimuli using the audio interface and two loudspeakers.

#### 2.1.3. Closed-Loop AAD Phase

In the closed-loop AAD phase, the calibrated cognitive-driven gain controller system was used to identify the attended and the unattended speaker and to close the loop

by presenting the amplified attended speaker and the attenuated unattended speaker in an online fashion via loudspeakers. The closed-loop AAD phase consisted of three sessions, each lasting 10 min. During each session, participants were cued by an arrow on a screen every minute to switch attention between the presented competing speakers while recording the ongoing EEG responses. To identify the attended and the unattended speaker, in each session a different AAD algorithm was used, i.e., LW–GLM, LW–SSM, and SW–SSM. These AAD algorithms were randomly assigned to the sessions for each participant. After each session, the participants were asked to fill out a questionnaire consisting of multiple-choice questions about the stories uttered by the speakers, similarly to the open-loop AAD phase. In addition, the participants were asked to rate how much effort it took to follow the attended speaker, to ignore the unattended speaker, to switch attention to the cued speaker, and how well they understood the attended story. For analyzing the experimental results, 24% of the results needed to be excluded, 5% due to a technical hardware problem when saving the results and 19% due to poor attentional performance reported by participants themselves after a few sessions (Some participants reported that they either completely lost concentration to attend to the cued speaker or they completely engaged with the story uttered by the non-cued speaker.).

### *2.2. Participants*

Ten native German-speaking participants (aged between 22 and 31 years; 6 male, 4 female) took part in this study (An EEG experiment participation was announced, to which ten persons responded.). Informed consent was received from all participants. All participants were self-reported normal hearing and reported no past or present neurological or psychiatric conditions. Informed consent was obtained from all subjects involved in the study. The study was carried out in accordance with the Declaration of Helsinki.

### *2.3. Stimuli*

Two German audio stories, uttered by two different male speakers, were used as the speech signals of the competing speakers. One story was from a German audio book website [26], and the other story was from a selection of audio books [27]. Before performing the experiment, participants reported no knowledge of the audio stories. Speech pauses from the audio stories that exceeded 0.5 s were shortened to 0.5 s. The audio stories were normalized to the same root-mean-square (RMS) value at a comfortable level which was individualized by each participant. The audio stories with no repetition were considered as the acoustic stimuli for the calibration phase, the open-loop AAD phase and the closed-loop AAD phase. The acoustic stimuli were presented at a sampling frequency of 44,100 Hz using MATLAB (MATLAB 2 in Figure 2), a Fireface UC audio interface system (provided by RME Audio, Germany) and two loudspeakers placed at the left side (with an azimuth of $-45°$) and the right side (with an azimuth of $45°$) and a distance of 1 m from the the participants. The visual stimuli consisting of an arrow for cueing were presented using a monitor in front of the participants. In addition, the EEG trigger markers synchronized with the acoustic and visual stimuli were generated using the Fireface UC audio interface system and a g.TRIGbox (provided by g.tec, Austria). The presentation of the acoustic and visual stimuli and the trigger marker generation were performed using the same computer employed for the cognitive-driven gain controller system (see Figure 2 and Table 1).

**Table 1.** Software and hardware used for the proposed cognitive-driven gain controller system.

| Software/Hardware | Description |
|---|---|
| Lab Streaming Layer (LSL) software 1.12 | Handles the gUSBamp application and the Lab Recorder program |
| gUSBamp application (part of LSL) | Streams EEG responses and EEG trigger markers |
| Lab Recorder program (part of LSL) | Record EEG responses |
| OpenViBE software 2.0 | Visualizes online EEG responses |
| MATLAB 11.2 | Runs RAAD algorithms, AGC algorithm, presents acoustic and visual stimuli via Fireface UC audio interface system, and generates EEG trigger marker |
| gUSBamp Research | Records EEG signals using 16 channel EEG amplifier |
| g.TRIGbox (multi-modal trigger box) | Generates EEG trigger markers synchronized with acoustic and visual stimuli |
| Fireface UC audio interface system | Presents acoustic stimuli |
| Computer | Runs MATLAB, LSL software, and OpenViBE software |

### *2.4. Data Acquisition*

Aiming at using a small number of electrodes for AAD, EEG responses were acquired using $C = 16$ electrodes. The electrodes were placed on the scalp area at F1, F2, FC3, FC4, FT7, FT8, Cz, C5, C6, P3, P4, P7, P8, Oz, PO3, and PO4 (see Figure 3). This electrode placement was inspired by the results in Reference [6,28], where it was shown that an electrode configuration covering the temporal, central, frontal, and parietal scalp areas yields a reliable decoding performance. The EEG responses were referenced to the P9 electrode (Since we did not observe a significant difference in AAD performance between referencing to the P9 electrode or to the nose electrode, we decided to use P9 as the fixed reference electrode for all phases.). The EEG responses were acquired using active (g.LADYbird) electrodes and a g.USBamp bio-signal amplifier (provided by g.tec, Austria). The acquired EEG responses and EEG trigger markers were streamed at a sampling frequency of 500 Hz using the gUSBamp application from the Lab Streaming Layer (LSL) software package (provided by Swartz Center for Computational Neuroscience, UCSD). Using the gUSBamp application, the streamed EEG responses were also forwarded to RAAD for online decoding, to the Lab Recorder application (provided by Swartz Center for Computational Neuroscience and Kothe) for recording, and to the OpenViBE software for online EEG visualization (see Table 1). The gUSBamp application, the OpenViBE software and the Lab Recorder application were run on the same computer employed for the cognitive-driven gain controller system (see Figure 2).

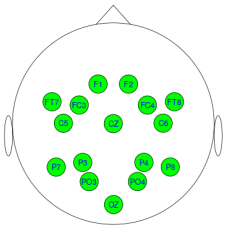


**Figure 3.** Scalp map of EEG electrodes.

### *2.5. Cognitive-Driven Gain Controller System*

In this section, we present the proposed cognitive-driven gain controller system consisting of RAAD and AGC (see Figure 4). Section 2.5.1 describes the RAAD, which generates probabilistic attention measures based on which the attended and the unattended speaker are identified. Section 2.5.2 describes the AGC, which amplifies the identified attended speaker and attenuates the identified unattended speaker based on the probabilistic attention measures.

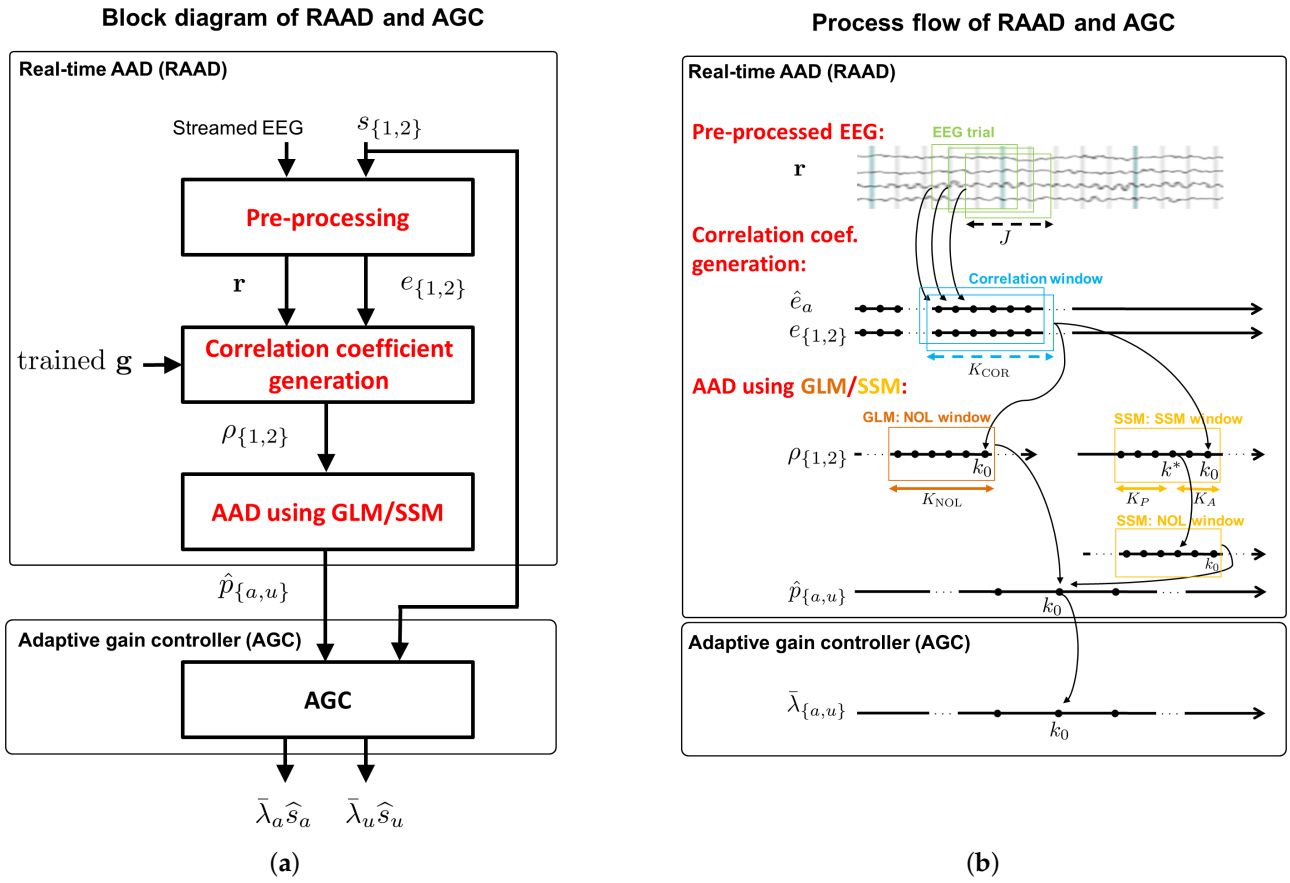



**Figure 4.** Real-time AAD (RAAD) and adaptive gain controller (AGC): (**a**) block diagram and (**b**) process flow.

#### 2.5.1. Real-Time Auditory Attention Decoding (RAAD)

The RAAD consists of three blocks (see Figure 4), i.e., pre-processing of the EEG responses and speech signals, correlation coefficient generation and AAD using either GLM or SSM.

A. *Pre-Processing:*
The streamed EEG responses from the gUSBamp application were re-referenced to a common average reference, band-pass filtered between 0.5 Hz and 9 Hz using a fourth-order Butterworth band-pass filter, and, subsequently, downsampled to 64 Hz in an online fashion. Contrary to the online EEG pre-processing, the speech pre-processing was performed in an offline fashion, since the speech signal $s_{1,t}$ of speaker 1 and the speech signal $s_{2,t}$ of speaker 2, with $t$ the discrete time index for $t = 1 \ldots T$, are available. The envelopes of both speech signals $e_{1,k}$ and $e_{2,k}$, with $k$ the sub-sampled time index for $k = 1 \ldots K$, were obtained using a Hilbert transform, followed by low-pass filtering at 9 Hz and downsampling to 64 Hz. The pre-processed EEG responses and the speech envelopes were then provided in an online fashion to the correlation coefficient generation block.

B. *Correlation Coefficient Generation:*
To generate the correlation coefficients of speaker 1 and speaker 2, we adopted the least-squares-based AAD method from Reference [3], which estimates the attended speech envelope from the EEG responses using a spatio-temporal envelope estimator trained during the calibration phase.

(1) *Training step (calibration phase):* In the training step, the attended speaker is assumed to be known. The attended speech envelope is then estimated from the

pre-processed EEG responses $r_{c,k}$, with $c$ the electrode index for $1 \dots C$, using a spatio-temporal envelope estimator $\mathbf{g}$ [3], i.e.,

$$\hat{e}_{a,k} = \mathbf{g}^T \mathbf{r}_k, \tag{1}$$

with

$$\mathbf{g} = \left[\mathbf{g}_1^T \, \mathbf{g}_2^T \cdots \, \mathbf{g}_C^T\right]^T, \tag{2}$$

$$\mathbf{g}_c = \left[g_{c,0} \;\; g_{c,1} \cdots \; g_{c,J-1}\right]^T, \tag{3}$$

$$\mathbf{r}_k = \left[\mathbf{r}_{1,k}^T \, \mathbf{r}_{2,k}^T \; \cdots \; \mathbf{r}_{C,k}^T\right]^T, \tag{4}$$

$$\mathbf{r}_{c,k} = \left[r_{c,k} \, r_{c,k+1} \; \cdots \; r_{c,k+J-1}\right]^T, \tag{5}$$

where $J$ denotes the number of envelope estimator coefficients per electrode. The trained envelope estimator $\mathbf{g}$ is obtained by minimizing the least-squares error between the (known) attended speech envelope $e_{a,k}$ and the reconstructed envelope $\hat{e}_{a,k}$, regularized with the squared norm of the derivative of the envelope estimator coefficients to avoid over-fitting [3,14,29], i.e.,

$$\min_{\mathbf{g}} \frac{1}{K} \sum_{k=1}^{K} (e_{a,k} - \underbrace{\mathbf{g}^T \mathbf{r}_k}_{\hat{e}_{a,k}})^2 + \beta \mathbf{g}^T \mathbf{\Lambda} \mathbf{g}, \tag{6}$$

with $\mathbf{\Lambda}$ denoting the derivative matrix [14] and $\beta$ denoting a regularization parameter. The solution of (6) is equal to

$$\mathbf{g} = (\mathbf{Q} + \beta \mathbf{\Lambda})^{-1} \mathbf{q}, \tag{7}$$

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

$$\mathbf{Q} = \frac{1}{K} \sum_{k=1}^{K} \mathbf{r}_k \mathbf{r}_k^T, \;\; \mathbf{q} = \frac{1}{K} \sum_{k=1}^{K} \mathbf{r}_k e_{a,k}. \tag{8}$$

(2) *Correlation coefficient generation step (open-loop and closed-loop AAD phase):* To generate the correlation coefficients of speaker 1 and speaker 2, we compute the Pearson correlation coefficients between the estimated attended envelope $\hat{e}_{a,k}$ in (1) and the speech envelopes $e_{1,k}$ and $e_{2,k}$, i.e.,

$$\rho_{1,k} = \rho(\mathbf{e}_{1,k}, \; \hat{\mathbf{e}}_{a,k}), \;\; \rho_{2,k} = \rho(\mathbf{e}_{2,k}, \; \hat{\mathbf{e}}_{a,k}), \tag{9}$$

where $\hat{\mathbf{e}}_{a,k}$ denotes the stacked vector of estimated attended envelopes corresponding to a correlation window of length $K_{COR}$, i.e.,

$$\hat{\mathbf{e}}_{a,k} = \left[\hat{e}_{a,k-K_{\mathrm{COR}}+1} \, \hat{e}_{a,k-K_{\mathrm{COR}}+2} \; \cdots \; \hat{e}_{a,k}\right]^T, \tag{10}$$

and $\mathbf{e}_{1,k}$ and $\mathbf{e}_{2,k}$ are defined similarly as in (10).

In the training step, the pre-processed EEG responses obtained from the calibration phase were segmented into trials of length 15 s, shifted by 1 sample (corresponding to $\frac{1}{64}$ s). The parameters $J$ and $\beta$ of the envelope estimator in (3) and (7) were determined for each participant using a leave-one-trial-out cross-validation approach, similarly as in Reference [3,14]. Using these parameters, a trained spatio-temporal envelope estimator $\mathbf{g}$ in (7) was then computed for each participant using all trials from the calibration phase.

In the correlation coefficient generation step, the pre-processed EEG responses were segmented in the same way as in the training step. The correlation coefficients $\rho_{1,k}$

and $\rho_{2,k}$ in (9) were computed either using a large correlation window of length $K_{\text{COR}} = 960$ samples (corresponding to 15 s) with an overlap of 959 samples or using a small correlation window of length $K_{\text{COR}} = 16$ samples (corresponding to 0.25 s) with no overlap. In Reference [4,7,9,18], it has been shown that the performance of AAD algorithms is affected by fluctuations of the correlation coefficients. In this paper, we propose two methods (GLM and SSM) to translate the fluctuating correlation coefficients into more reliable probabilistic attention measures.

C. *Auditory Attention Decoding Using Generalized Linear Model:*
The AAD algorithm using the GLM consists of a training and a decoding step. The training step takes place during the calibration phase, whereas the decoding step takes place during the open-loop and the closed-loop AAD phase.

(1) *Training step:* The correlation coefficients of speaker 1 and speaker 2 in (9) are first segmented into non-overlapping (NOL) windows of length $K_{\text{NOL}}$, i.e.,

$$\boldsymbol{\rho}_{1,i} = \left[\rho_{1,(i-1)K_{\text{NOL}}+1}\ \rho_{1,(i-1)K_{\text{NOL}}+2}\ \cdots\ \rho_{1,iK_{\text{NOL}}}\right]^T, \tag{11}$$

$$\boldsymbol{\rho}_{2,i} = \left[\rho_{2,(i-1)K_{\text{NOL}}+1}\ \rho_{2,(i-1)K_{\text{NOL}}+2}\ \cdots\ \rho_{2,iK_{\text{NOL}}}\right]^T, \tag{12}$$

with $i$ the window index for $i = 1 \ldots I$. The mean differential correlation coefficient between speaker 1 and speaker 2 in window $i$ is computed as

$$\bar{\Delta}\rho_i = \frac{1}{K_{\text{NOL}}} \sum_{n=1}^{K_{\text{NOL}}} (\rho_{1,(i-1)K_{\text{NOL}}+n} - \rho_{2,(i-1)K_{\text{NOL}}+n}). \tag{13}$$

We model the attention state $\bar{d}_i$ in window $i$ as a binary random variable [30], i.e.,

$$\begin{cases} \bar{d}_i = 1, & \text{attending to speaker 1 in window } i \\ \bar{d}_i = 2, & \text{attending to speaker 2 in window } i \end{cases}, \tag{14}$$

which is assumed to follow a Bernoulli distribution with probability $\bar{p}_i$, i.e.,

$$P(\bar{d}_i) = \bar{p}_i{}^{2-\bar{d}_i}(1-\bar{p}_i)^{\bar{d}_i-1} = \begin{cases} \bar{p}_i, & \text{if } \bar{d}_i{=}1 \\ 1-\bar{p}_i, & \text{if } \bar{d}_i{=}2 \end{cases}. \tag{15}$$

Using a GLM, the probability of attending to speaker 1 is then given by [31]

$$\bar{p}_i = P(\bar{d}_i = 1) = 1 - P(\bar{d}_i = 2) = \frac{1}{1+e^{-\bar{z}_i}}, \tag{16}$$

with the linear predictor $\bar{z}_i$, i.e.,

$$\bar{z}_i = \mathbf{x}_i^T \boldsymbol{\alpha}, \tag{17}$$

$$\mathbf{x}_i = [1\ \ \bar{\Delta}\rho_i]^T, \tag{18}$$

$$\boldsymbol{\alpha} = [\alpha_0\ \ \alpha_1]^T, \tag{19}$$

where $\alpha_0$ and $\alpha_1$ denote the GLM parameters. Obviously, the probability of attending to speaker 1 monotonically increases from 0 to 1 for $\bar{z}_i \in (-\infty, \infty)$. The probability mass function in (15) can be written as an exponential distribution using the canonical link function $\theta_i = \text{logit}(\bar{p}_i) = \bar{z}_i$, with $\text{logit}(\bar{p}_i) = \log\left(\frac{\bar{p}_i}{1-\bar{p}_i}\right)$, i.e.,

$$P(\bar{d}_i) = \exp(\bar{d}_i\theta_i - b(\theta_i)), \tag{20}$$

with

$$b(\theta_i) = \log(1+\exp(\theta_i)). \tag{21}$$

The maximum likelihood (ML) estimate of the GLM parameters in (19) is then obtained by maximizing the log-likelihood function, i.e.,

$$\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}} \ell(\mathbf{x}_i, i = 1 : I|\boldsymbol{\alpha}) = \sum_{i=1}^{I} \bar{d}_i \theta_i - b(\theta_i). \tag{22}$$

This estimate can be computed, for example, by using an iteratively re-weighted least-squares algorithm and Newton–Raphson method [32,33], i.e.,

$$\hat{\boldsymbol{\alpha}}^{(r+1)} = (\mathbf{X}^T \mathbf{W}^{(r)} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{(r)} \mathbf{y}^{(r)}, \tag{23}$$

with $r$ the iteration index and

$$\mathbf{X} = [\mathbf{x}_1 \, \mathbf{x}_2 \ldots \, \mathbf{x}_I]^T, \;\; \mathbf{X} \in \mathbb{R}^{I \times 2}, \tag{24}$$

$$\mathbf{W}^{(r)} = \mathrm{diag}\left\{ \frac{1}{\bar{p}_i^{(r)} \left(1 - \bar{p}_i^{(r)}\right)} \right\}, \;\; \mathbf{W}^{(r)} \in \mathbb{R}^{I \times I}, \tag{25}$$

$$\bar{p}_i^{(r)} = \mathrm{logit}^{-1}(\bar{z}_i^{(r)}), \tag{26}$$

$$\bar{z}_i^{(r)} = \mathbf{x}_i^T \hat{\boldsymbol{\alpha}}^{(r)}, \tag{27}$$

$$\mathbf{y}^{(r)} = [y_1^{(r)} \, y_2^{(r)} \ldots \, y_I^{(r)}]^T, \;\; \mathbf{y}^{(r)} \in \mathbb{R}^{I \times 1}, \tag{28}$$

$$y_i^{(r)} = \bar{z}_i^{(r)} + (\bar{d}_i - \bar{p}_i^{(r)}) \mathrm{logit}'(\bar{p}_i^{(r)}), \tag{29}$$

where $(\cdot)'$ denotes the derivative operator. Algorithm 1 summarizes the GLM parameter estimation in the training step.

**Algorithm 1** : GLM Training

**input:** $\mathbf{x}_i$ and $\bar{d}_i$ for $i = 1 \ldots I$

1: initialization: $\bar{p}_i^{(0)} = \bar{d}_i$, $\bar{z}_i^{(0)} = \mathrm{logit}(\bar{p}_i^{(0)})$, calculate $\hat{\boldsymbol{\alpha}}^{(1)}$ using (23)
2: **for** $r = 1 \ldots R$ **do**
3: calculate $\bar{p}_i^{(r)}$ and $\bar{z}_i^{(r)}$ using (26) and (27), respectively
4: calculate $\mathbf{y}^{(r)}$ and $\mathbf{W}^{(r)}$ using (28), (29), and (25), respectively
5: update the GLM parameters $\hat{\boldsymbol{\alpha}}^{(r+1)}$ using (23)
6: **end for**

**output:** $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\alpha}}^{(R+1)}$

(2) *Decoding step:* To decode which speaker a participant is attending to in window $i$, the mean differential correlation coefficient $\bar{\Delta}\rho_i$ is computed using (13), based on which the linear predictor $\bar{z}_i$ is computed using the (trained) GLM parameters $\hat{\boldsymbol{\alpha}}$ in (17). The probability of attending to speaker 1, i.e., $P(\bar{d}_i = 1)$, and the probability of attending to speaker 2, i.e., $P(\bar{d}_i = 2)$, are then obtained using (16). Based on these probabilities, it is decided that the participant attended to speaker 1 if $P(\bar{d}_i = 1) > P(\bar{d}_i = 2)$, or attended to speaker 2 otherwise.
The probabilistic attention measure of the attended speaker $\hat{p}_{a,i}$ in window $i$ is, hence, determined as

$$\begin{cases} \hat{p}_{a,i} = P(\bar{d}_i = 1), & \text{if } P(\bar{d}_i = 1) > P(\bar{d}_i = 2) \\ \hat{p}_{a,i} = P(\bar{d}_i = 2), & \text{otherwise.} \end{cases} \tag{30}$$

Obviously, the probabilistic attention measure of the attended speaker $\hat{p}_{a,i}$ lies between 0.5 and 1. The probabilistic attention measure of the unattended speaker

$\hat{p}_{u,i}$ is determined as $\hat{p}_{u,i} = 1 - \hat{p}_{a,i}$. The process flow of AAD using the GLM is depicted in Figure 4.

The AAD algorithm using the GLM was implemented and run using MATLAB (MATLAB 1 of RAAD in Figure 2). For the training step, Algorithm 1 was executed with $R = 30$ iterations using the correlation coefficients obtained from the calibration phase. Both for the training and the decoding steps, the correlation coefficients were computed using the large correlation window (i.e., $K_{\text{COR}} = 960$ samples) and the mean differential correlation coefficient in (13) was computed using a window of length $K_{\text{NOL}} = 16$ samples (corresponding to 0.25 s). During the decoding step, the probabilistic attention measures $\hat{p}_{a,i}$ and $\hat{p}_{u,i}$ were forwarded to the AGC using the LSL software package (see Figure 2). Each participant's own data were used for training the GLM parameters and for decoding. To evaluate the performance of the proposed LW–GLM algorithm, the decoding performance for each participant was computed as the percentage of correctly decoded NOL windows. To evaluate the delay to detect a cued attention switch of the proposed LW–GLM algorithm, the delay was computed as the time takes for the LW–GLM algorithm to detect an attention switch after the moment the arrow on a screen cued to switch attention.

D. *Auditory Attention Decoding Using State-Space Model:*
As an alternative to the GLM, it has been proposed in Reference [9] to use a SSM to translate the absolute values of the coefficients of the spatio-temporal envelope estimator into probabilistic attention measures. Contrary to Reference [9], in this paper, we propose to use the absolute values of the correlation coefficients

$$\phi_{1,k} = |\rho_{1,k}|, \tag{31}$$

$$\phi_{2,k} = |\rho_{2,k}|, \tag{32}$$

instead of the coefficients of the spatio-temporal envelope estimator, which need to be obtained for both the attended and the unattended speaker.
Similarly to (14), we model the attention state $d_k$ at time instance $k$ as a binary random variable, i.e.,

$$\begin{cases} d_k = 1, & \text{attending to speaker 1 at time instance } k \\ d_k = 2, & \text{attending to speaker 2 at time instance } k \end{cases}, \tag{33}$$

which is assumed to follow a Bernoulli distribution with probability $p_k$. Similarly to (16), the probability of attending to speaker 1 is given by

$$p_k = P(d_k = 1) = 1 - P(d_k = 2) = \frac{1}{1 + e^{-z_k}}, \tag{34}$$

where the variable $z_k$ is now modeled as an autoregressive (AR) process, i.e.,

$$z_k = c_0 z_{k-1} + w_k. \tag{35}$$

The parameter $c_0$ is a hyperparameter ensuring stability of the AR process, and the noise process $w_k$ is assumed to follow a normal distribution with variance $\eta_k$, i.e.,

$$w_k \sim \mathcal{N}(0, \eta_k), \tag{36}$$

$$\eta_k \sim \text{Inverse-Gamma}(a_0, b_0), \tag{37}$$

where $a_0$ and $b_0$ are hyperparameters. The AR model in (35) implies that the variable $z_k$ at time instance $k$ is predicted from $z_{k-1}$ at the previous time instance with some uncertainty, which is modeled by the noise process $w_k$.
To relate the correlation coefficients $\rho_{1,k}$ and $\rho_{2,k}$ in (9) to the attention state $d_k$, we model the probability of the absolute values of the correlation coefficients, given

attention to speaker 1 or speaker 2, using a log-normal distribution (Please note that modeling the probabilities of the absolute values of the correlation coefficients with log-normal distributions allows for a closed-form iterative solution [9], compared to modeling the probabilities of the correlation coefficients either with normal or von Mises-Fisher distributions [30].), i.e.,

$$p(\phi_{l,k} \mid d_k = l) \sim \text{Log-Normal}(\delta_a, \mu_a), \quad l = 1, 2, \tag{38}$$

with

$$\delta_a \sim \text{Gamma}(\bar{\gamma}_a, \bar{\nu}_a), \;\; p(\mu_a \mid \delta_a) \sim \mathcal{N}(\bar{\mu}_a, \delta_a), \tag{39}$$

where $\bar{\gamma}_a$, $\bar{\nu}_a$ and $\bar{\mu}_a$ denote the hyperparameters of the attended log-normal distribution. Similarly, we model the probability of the absolute values of the correlation coefficients, given no attention to speaker 1 or speaker 2, as

$$p(\phi_{l,k} \mid d_k \neq l) \sim \text{Log-Normal}(\delta_u, \mu_u), \quad l = 1, 2, \tag{40}$$

with

$$\delta_u \sim \text{Gamma}(\bar{\gamma}_u, \bar{\nu}_u), \;\; p(\mu_u \mid \delta_u) \sim \mathcal{N}(\bar{\mu}_u, \delta_u), \tag{41}$$

where $\bar{\gamma}_u$, $\bar{\nu}_u$, and $\bar{\mu}_u$ denote the hyperparameters of the unattended log-normal distribution. Since a small overlap between the attended and the unattended log-normal distributions is desired for a reliable decoding performance, the hyperparameters $\bar{\gamma}_{\{a,u\}}$, $\bar{\nu}_{\{a,u\}}$, and $\bar{\mu}_{\{a,u\}}$ are tuned to minimize the overlap.

Aiming at estimating the probability of attending to speaker 1 and speaker 2 at time instance $k = k^*$ (see Figure 4), we now consider the absolute values of the correlation coefficients within a sliding window of length $K_{\text{SSM}} = K_P + K_A + 1$, with $K_P$ and $K_A$ denoting the number of correlation coefficients prior to and after $k^*$, respectively. The set of parameters to be estimated in this window is given by $\mathbf{\Omega} = \left\{z_{k^*-K_P:k^*+K_A}, \eta_{k^*-K_P:k^*+K_A}, \delta_a, \mu_a, \delta_u, \mu_u\right\}$. The maximum a posteriori (MAP) estimate is obtained by maximizing the log-posterior function, i.e.,

$$\hat{\mathbf{\Omega}} = \arg\max_{\mathbf{\Omega}} \ell(\mathbf{\Omega} | \phi_{1,k}, \phi_{2,k}, k = k^* - K_P : k^* + K_A), \tag{42}$$

which can be computed iteratively using the Expectation Maximization (EM) algorithm as in Reference [9,30].

Using the estimated variable $z_k$, the probability $p_k = P(d_k = 1)$ of attending to speaker 1 at time instance $k$ is obtained using (34). These probabilities are segmented into non-overlapping windows of length $K_{\text{NOL}}$, i.e.,

$$\mathbf{p}_i = \left[p_{(i-1)K_{\text{NOL}}+1} \; p_{(i-1)K_{\text{NOL}}+2} \; \cdots \; p_{iK_{\text{NOL}}}\right]^T, \tag{43}$$

and the probability of attending to speaker 1 in window $i$ is then computed as the mean of the probabilities, i.e.,

$$P\left(\hat{d}_i = 1\right) = \frac{1}{K_{\text{NOL}}} \sum_{n=1}^{K_{\text{NOL}}} p_{(i-1)K_{\text{NOL}}+n}, \tag{44}$$

with $\hat{d}_i$ the attention state in window $i$. The probability of attending to speaker 2 in window $i$ is computed as

$$P\left(\hat{d}_i = 2\right) = 1 - P\left(\hat{d}_i = 1\right). \tag{45}$$

Based on these probabilities, it is decided that the participant attended to speaker 1 if $P\left(\hat{d}_i = 1\right) > P\left(\hat{d}_i = 2\right)$, or attended to speaker 2 otherwise.

The probabilistic attention measure of the attended speaker $\hat{p}_{a,i}$ in window $i$ is, hence, determined as

$$\begin{cases} \hat{p}_{a,i} = P\left(\hat{d}_i = 1\right), & \text{if } P\left(\hat{d}_i = 1\right) > P\left(\hat{d}_i = 2\right) \\ \hat{p}_{a,i} = P\left(\hat{d}_i = 2\right), & \text{otherwise.} \end{cases} \tag{46}$$

The probabilistic attention measure of the unattended speaker $\hat{p}_{u,i}$ is determined as $\hat{p}_{u,i} = 1 - \hat{p}_{a,i}$. The process flow of AAD using the SSM is depicted in Figure 4.
The AAD algorithm using the SSM was implemented and run using MATLAB (MATLAB 1 of RAAD in Figure 2). The hyperparameters in (35) and (37) were set to $c_0 = 1$, $a_0 = 2.008$ and $b_0 = 0.2016$, similarly as in Reference [9]. The hyperparameters $\bar{\gamma}_a$, $\bar{\nu}_a$ and $\bar{\mu}_a$ in (39) were set by fitting a gamma and a normal distribution to the absolute values of the correlation coefficients of the (oracle) attended speaker obtained from the calibration phase. Similarly, the hyperparameters $\bar{\gamma}_u$, $\bar{\nu}_u$, and $\bar{\mu}_u$ in (41) were set by fitting a gamma and a normal distribution to the absolute values of the correlation coefficients of the (oracle) unattended speaker obtained from the calibration phase. The SSM parameter set $\boldsymbol{\Omega}$ was estimated using the EM algorithm as in Reference [9] with 20 iterations. On the one hand, for the LW–SSM algorithm using a large overlapping correlation window (i.e., $K_{\text{COR}} = 960$ samples, 1 sample shift), a small SSM window of length $K_{\text{SSM}} = 1$ sample (corresponding to $\frac{1}{64}$ s) with $K_P = 0$ and $K_A = 0$ was used. On the other hand, for the SW–SSM algorithm using a small non-overlapping correlation window (i.e., $K_{\text{COR}} = 16$ samples), a large SSM window of length $K_{\text{SSM}} = 60$ samples (corresponding to 15 s) with $K_P = 53$ (corresponding to 13.25 s) and $K_A = 6$ (corresponding to 1.50 s) was used as in Reference [9]. The length of the window $K_{\text{NOL}}$ in (43) was set such that both algorithms generated the probabilistic attention measure of the attended speaker $\hat{p}_{a,i}$ in (46) every 0.25 s. This means that for the LW–SSM algorithm a window of length $K_{\text{NOL}} = 16$ samples was used, while, for the SW–SSM algorithm, a window of length $K_{\text{NOL}} = 1$ sample was used. Each participant's own data were used for hyperparameter and parameter setting, as well as for decoding. To evaluate the performance of the proposed LW–SSM and SW–SSM algorithms, the decoding performance for each participant was computed as the percentage of correctly decoded NOL windows. To evaluate the delay to detect a cued attention switch of the proposed LW–SSM and SW–SSM algorithms, the delay was computed as the time takes for the LW–SSM and SW–SSM algorithms to detect an attention switch after the moment the arrow on a screen cued to switch attention.

#### 2.5.2. Adaptive Gain Controller (AGC)

The probabilistic attention measure of the attended speaker $\hat{p}_{a,i}$ in window $i$, either obtained using the GLM in (30) or using the SSM in (46), is then used to drive the AGC (see Figure 2).

The speech signal $s_{1,t}$ of speaker 1 and the speech signal $s_{2,t}$ of speaker 2 are first segmented into non-overlapping windows of length $K_{\text{AGC}}$, i.e., for window $i$

$$\mathbf{s}_{1,i} = \begin{bmatrix} s_{1,(i-1)K_{\text{AGC}}+1} & s_{1,(i-1)K_{\text{AGC}}+2} & \cdots & s_{1,iK_{\text{AGC}}} \end{bmatrix}^T, \tag{47}$$

$$\mathbf{s}_{2,i} = \begin{bmatrix} s_{2,(i-1)K_{\text{AGC}}+1} & s_{2,(i-1)K_{\text{AGC}}+2} & \cdots & s_{2,iK_{\text{AGC}}} \end{bmatrix}^T. \tag{48}$$

Based on the AAD result for window $i$, the attended speech vector $\hat{\mathbf{s}}_{a,i}$ and the unattended speech vector $\hat{\mathbf{s}}_{u,i}$ are determined as

$$\begin{cases} \hat{\mathbf{s}}_{a,i} = \mathbf{s}_{1,i}, \ \hat{\mathbf{s}}_{u,i} = \mathbf{s}_{2,i} & \text{if the identified attended speaker is speaker 1} \\ \hat{\mathbf{s}}_{a,i} = \mathbf{s}_{2,i}, \ \hat{\mathbf{s}}_{u,i} = \mathbf{s}_{1,i} & \text{otherwise.} \end{cases} \tag{49}$$

By multiplying the attended speech vector $\hat{\mathbf{s}}_{a,i}$ with the gain $\lambda_{a,i}$ and multiplying the unattended speech vector $\hat{\mathbf{s}}_{u,i}$ with the gain $\lambda_{u,i}$, the objective of the AGC is to achieve a desired signal-to-interference-ratio (SIR) between the identified attended and unattended speakers in window $i$. The desired SIR in window $i$ is defined as a linear function of the probabilistic attention measure $\hat{p}_{a,i}$, i.e.,

$$\mathrm{SIR}_i^{des} = 2\mathrm{SIR}_{max}\hat{p}_{a,i} - \mathrm{SIR}_{max}, \tag{50}$$

such that $\hat{p}_{a,i} = 1$ corresponds to $\mathrm{SIR}_{max}$, i.e., the maximum desired SIR, and $\hat{p}_{a,i} = 0.5$ corresponds to SIR $= 0$ dB.

The SIR in window $i$ at the output of the AGC is equal to

$$\mathrm{SIR}_i = 10\log_{10}\left(\frac{\lambda_{a,i}^2\varphi_{a,i}}{\lambda_{u,i}^2\varphi_{u,i}}\right), \tag{51}$$

with the energy of the attended and unattended speech vector in window $i$ given by

$$\varphi_{a,i} = \hat{\mathbf{s}}_{a,i}^T\hat{\mathbf{s}}_{a,i}, \quad \varphi_{u,i} = \hat{\mathbf{s}}_{u,i}^T\hat{\mathbf{s}}_{u,i}. \tag{52}$$

By setting (51) equal to the desired SIR in (50) and constraining the overall energy at the output of the AGC to be equal to the overall input energy, i.e.,

$$\lambda_{a,i}^2\varphi_{a,i} + \lambda_{u,i}^2\varphi_{u,i} = \varphi_{a,i} + \varphi_{u,i}, \tag{53}$$

the gains $\lambda_{u,i}$ and $\lambda_{a,i}$ can be computed as

$$\lambda_{u,i}^2 = \frac{1 + \frac{\varphi_{a,i}}{\varphi_{u,i}}}{1 + 10^{\frac{\mathrm{SIR}_i^{des}}{10}}}, \tag{54}$$

$$\lambda_{a,i}^2 = 10^{\frac{\mathrm{SIR}_i^{des}}{10}}\frac{\varphi_{u,i}}{\varphi_{a,i}}\lambda_{u,i}^2. \tag{55}$$

To avoid annoying artefacts due to highly time-varying gains, the gains $\lambda_{u,i}$ in (54) and $\lambda_{a,i}$ in (55) are averaged over four windows, i.e.,

$$\bar{\lambda}_{u,i} = \frac{1}{4}\sum_{n=i-3}^{i}\lambda_{u,n}, \quad \bar{\lambda}_{a,i} = \frac{1}{4}\sum_{n=i-3}^{i}\lambda_{a,n}. \tag{56}$$

The amplified attended speech vector $\bar{\mathbf{s}}_{a,i}$ and the attenuated unattended speech vector $\bar{\mathbf{s}}_{u,i}$ in window $i$ are finally obtained as

$$\bar{\mathbf{s}}_{a,i} = \bar{\lambda}_{a,i}\hat{\mathbf{s}}_{a,i}, \tag{57}$$

$$\bar{\mathbf{s}}_{u,i} = \bar{\lambda}_{u,i}\hat{\mathbf{s}}_{u,i}. \tag{58}$$

These signals are then presented to the participant via two loudspeakers. The AGC was implemented and run using MATLAB (MATLAB 2 in Figure 2). The sampling frequency of the speech signals of both speakers was equal to 44,100 Hz. The maximum desired SIR in (50) was set to 7 dB.

The speech enhancement performance of the AGC was evaluated in terms of the SIR improvement ΔSIR, i.e.,

$$\Delta\mathrm{SIR} = \mathrm{SIR}_{out} - \mathrm{SIR}_{in}, \tag{59}$$

with

$$\mathrm{SIR}_{in} = 10\log_{10}\left(\frac{\sum_{i=1}^{I}\mathbf{s}_{a,i}^{T}\mathbf{s}_{a,i}}{\sum_{i=1}^{I}\mathbf{s}_{u,i}^{T}\mathbf{s}_{u,i}}\right), \tag{60}$$

$$\mathrm{SIR}_{out} = 10\log_{10}\left(\frac{\sum_{i=1}^{I}\bar{\mathbf{s}}_{a,i}^{T}\bar{\mathbf{s}}_{a,i}}{\sum_{i=1}^{I}\bar{\mathbf{s}}_{u,i}^{T}\bar{\mathbf{s}}_{u,i}}\right), \tag{61}$$

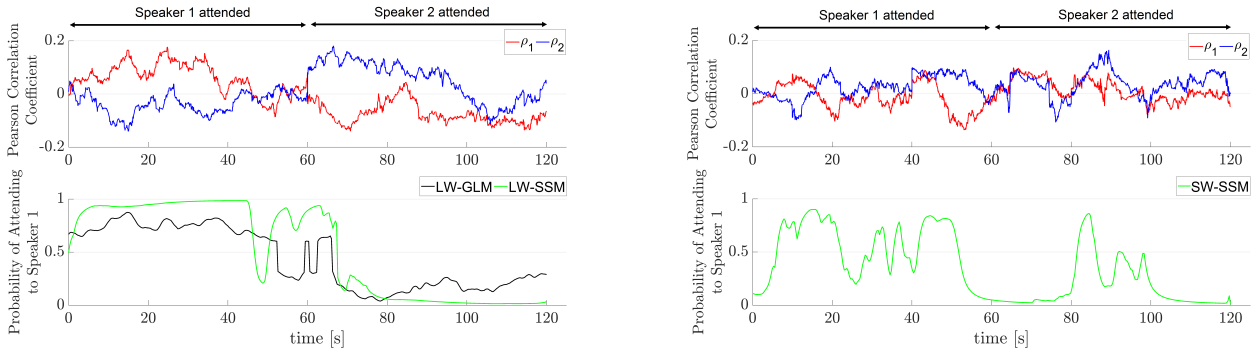
where $\mathbf{s}_{a,i}$ and $\mathbf{s}_{u,i}$ denote the (oracle) attended and unattended speech vectors, defined similarly as in (47).

## 3. Results

In this section, we evaluate the decoding performance and the speech enhancement performance of the proposed cognitive-driven gain controller system described in the previous section. In Section 3.1, we evaluate the decoding performance of the proposed AAD algorithms for the open-loop and the closed-loop AAD phase. In Section 3.2, we evaluate the speech enhancement performance of the AGC for the closed-loop AAD phase. Finally, in Section 3.3, we compare the subjective evaluation between the open-loop and the closed-loop AAD phase.

### *3.1. Auditory Attention Decoding Performance*

For all considered AAD algorithms (LW–GLM, LW–SSM, and SW–SSM), Figure 5 depicts the correlation coefficients $\rho_{1,k}$ and $\rho_{2,k}$ of speaker 1 and speaker 2 and the probability of attending to speaker 1, i.e., $P\left(\hat{d}_i = 1\right)$ or $P\left(\hat{d}_i = 2\right)$, for an exemplary session from the open-loop AAD phase. It can be observed that all AAD algorithms translate the fluctuating correlation coefficients into smooth probabilistic attention measures. When using the large correlation window, i.e., LW–GLM and LW–SSM, the correlation coefficients are more discriminative and the probabilistic attention measures are more reliable with a lower variability compared to using the small correlation window, i.e., SW–SSM. This can mainly be explained by the fact that the large correlation window provides a larger amount of data from the reconstructed attended envelope and the envelopes of the speech signals compared to the small correlation window. A large discriminability and reliability of the correlation coefficients and the probabilistic attention measures are obviously essential to obtain a large decoding performance.

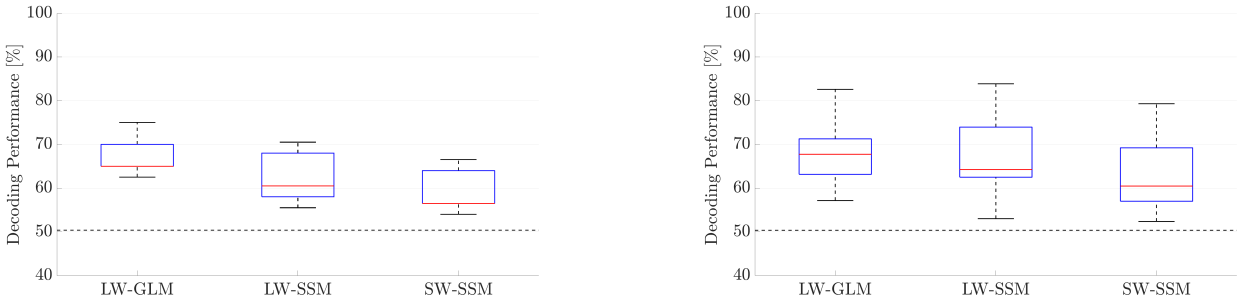



(**a**) Large correlation window ($K_{\mathrm{COR}} = 960$ samples, 1 sample shift)

(**b**) Small correlation window ($K_{\mathrm{COR}} = 16$ samples, no overlap)

**Figure 5.** Exemplary correlation coefficients of speaker 1 and speaker 2 and probability of attending to speaker 1 from the open-loop AAD phase when using AAD algorithms employing (**a**) a large correlation window (LW–GLM, LW–SSM) and (**b**) a small correlation window (SW–SSM).

For the considered AAD algorithms, Figure 6 depicts the decoding performance for the open-loop and the closed-loop AAD phase. It can be observed that all AAD algorithms yield a median decoding performance that is larger than chance level (50%). For the open-loop AAD phase, the LW–GLM, LW–SSM, and SW–SSM algorithms yield a median decoding performance of 65.0%, 60.5%, and 56.5%, respectively. For the closed-loop AAD phase, the LW–GLM, LW–SSM, and SW–SSM algorithms yield a median decoding performance of 67.7%, 64.2%, and 60.4%, respectively. The larger median decoding performance obtained by the LW–GLM and LW–SSM algorithms is consistent with the probabilistic attention measures in Figure 5, where due to the large correlation window more reliable probabilistic attention measures are obtained compared to the SW–SSM algorithm. A statistical multiple comparison test (Kruskal–Wallis test followed by post-hoc Dunn and Sidak test [34]) revealed no significant difference ($p > 0.05$) in decoding performance between the open-loop and the closed-loop AAD phase nor between the considered AAD algorithms.




(**a**) open-loop AAD (**b**) closed-loop AAD

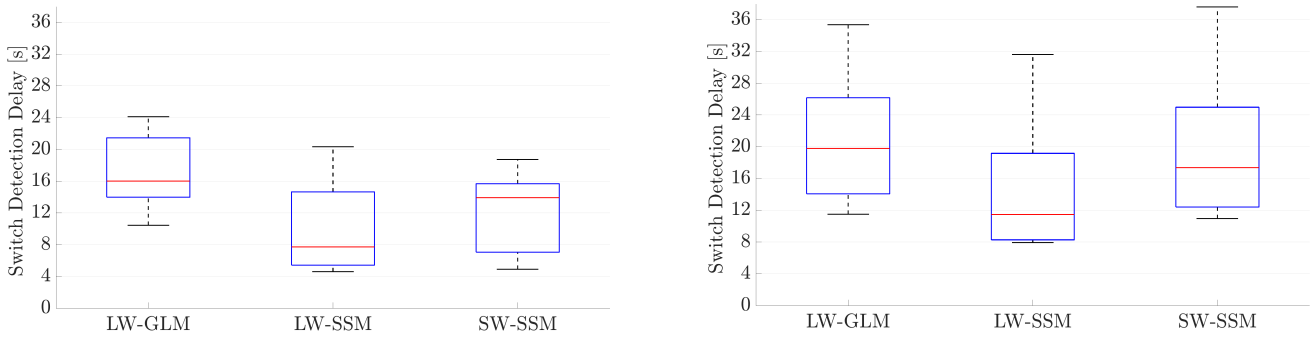
**Figure 6.** Decoding performance for (**a**) the open-loop AAD phase and (**b**) the closed-loop AAD phase when using the LW–GLM, LW–SSM, and SW–SSM algorithms. The boxplots display the median, the first quartile, the third quartile, the minimum, and the maximum of the decoding performance across participants. The dashed-line represents the upper boundary of the confidence interval corresponding to chance level based on a binomial test at the 5% significance level.

To further investigate the performance of the proposed AAD algorithms, Figure 7 depicts the delay to detect a cued attention switch for the open-loop and the closed-loop AAD phase. For the open-loop AAD phase, the LW–GLM, LW–SSM, and SW–SSM algorithms yield a median delay of 16.0 s, 7.7 s, and 13.9 s, respectively. For the closed-loop AAD phase, the LW–GLM, LW–SSM, and SW–SSM algorithms yield a median delay of 19.8 s, 11.5 s, and 17.4 s, respectively. A statistical multiple comparison test (Kruskal–Wallis test followed by post-hoc Dunn and Sidak test) revealed no significant difference ($p > 0.05$) between the open-loop and the closed-loop AAD phase nor between the considered AAD algorithms.

### *3.2. Signal-to-Interference Reduction of Adaptive Gain Controller*

For the considered AAD algorithms, Figure 8 depicts the SIR improvement for the closed-loop AAD phase. It can be observed that the LW–GLM, LW–SSM, and SW–SSM algorithms yield a median SIR improvement of 1.1 dB, 1.7 dB, and 0.5 dB, respectively. The larger SIR improvement obtained by the LW–GLM and LW–SSM algorithms can be explained by the larger decoding performance compared to the SW–SSM algorithm. The larger decoding performance leads to a larger number of windows during which the attended speaker is correctly amplified and the unattended speaker is correctly attenuated. In addition, it can be observed that the SW–SSM algorithm yields an SIR improvement with a larger variability (−2.7–3.0 dB) than the LW–GLM algorithm (0.6–2.1 dB) and the LW–SSM algorithm (0.7–3.8 dB). This can be explained by the larger variability of the probabilistic attention measures obtained by the SW–SSM algorithm (see Figure 5). Due to the linear role of the probabilistic attention measure in the AGC for determining the desired

SIR between the attended and the unattended speaker, as shown in (50), probabilistic attention measures with a large variability lead to SIRs with a large variability.




(**a**) open-loop AAD (**b**) closed-loop AAD

**Figure 7.** Delay to detect a cued attention switch for (**a**) the open-loop AAD phase and (**b**) the closed-loop AAD phase when using the LW–GLM, LW–SSM, and SW–SSM algorithms. The boxplots display the median, the first quartile, the third quartile, the minimum, and the maximum of the switch detection delay across participants.

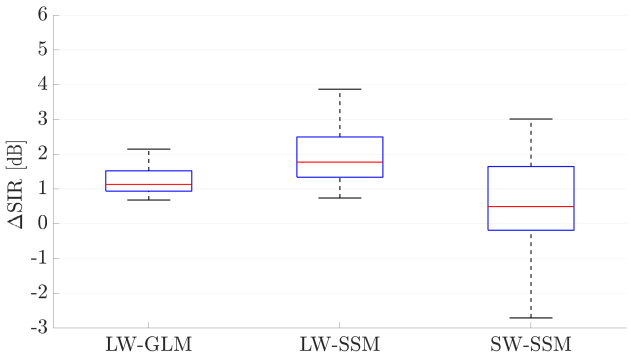



**Figure 8.** SIR improvement of the proposed cognitive-driven gain controller system when using the LW–GLM, LW–SSM, and SW–SSM algorithms. The boxplots display the median, the first quartile, the third quartile, the minimum, and the maximum of the SIR improvement across participants.

### *3.3. Subjective Evaluation of Open-Loop and Closed-Loop AAD*

For the open-loop and the closed-loop AAD phase, Figure 9 presents the perceived effort to follow the attended speaker, to ignore the unattended speaker, to switch attention between both speakers, and the level of story understanding.

In terms of the perceived effort to follow the attended speaker and to ignore the unattended speaker (Figure 9a,b), it can be observed that the lowest median effort is obtained for the open-loop AAD, while a higher median effort is required for the closed-loop AAD, especially when using the SW–SSM algorithm. This can be attributed to the negative SIR improvement in some windows (see Figure 8), where the attended speaker is wrongly attenuated and the unattended speaker is wrongly amplified. Nevertheless, a statistical multiple comparison test (Kruskal–Wallis test followed by post-hoc Dunn and Sidak test) revealed no significant difference ($p > 0.05$) between all considered open-loop and closed-loop AAD cases. Similarly, in terms of the effort to switch attention between both speakers (Figure 9c), a statistical multiple comparison test revealed no significant difference ($p > 0.05$) between all considered open-loop and closed-loop AAD cases. These results show that the proposed closed-loop cognitive-driven gain controller system demands a similar perceived effort to follow the attended speaker, to ignore the unattended speaker and to switch attention compared to the open-loop AAD system. In terms of the level of story understanding (Figure 9d), the highest median understanding level is obtained for the open-loop AAD, while a lower median understanding level is

obtained for the closed-loop AAD. This is consistent with the perceived cognitive effort (Figure 9a,b,c), where the open-loop AAD demands the lowest effort, possibly resulting in more cognitive resources available for story understanding compared to the closed-loop AAD. Nevertheless, a statistical multiple comparison test revealed no significant difference ($p > 0.05$) between all considered open-loop and closed-loop AAD cases.

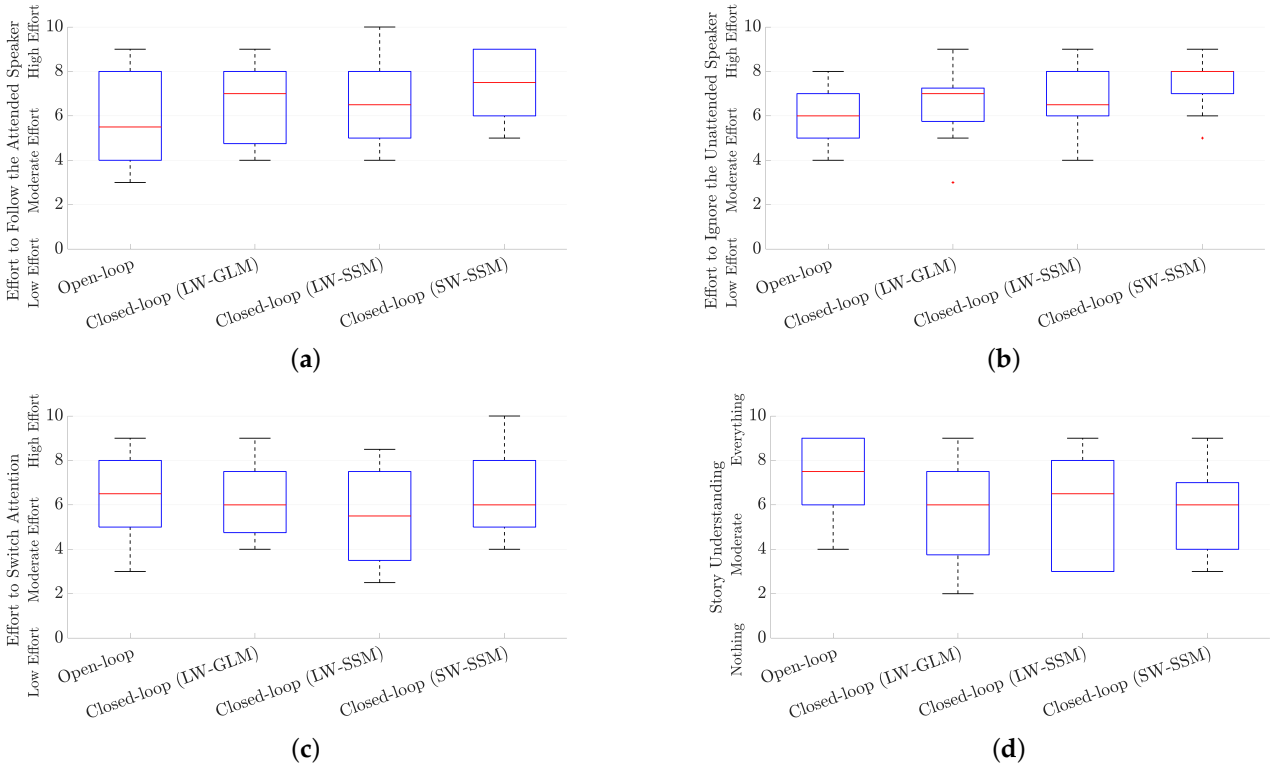



**Figure 9.** Subjective evaluation results of the open-loop and the closed-loop AAD phase using the LW–GLM, LW–SSM, and SW–SSM algorithms in terms of perceived effort to (**a**) follow the attended speaker, (**b**) ignore the unattended speaker, (**c**) switch attention between both speakers, and (**d**) understand the story.

Finally, Figure 10 presents the level of improvement in system usage achieved by the participants throughout the sessions of the closed-loop AAD phase. It can be observed for all considered AAD algorithms that a significant improvement in system usage is obtained.

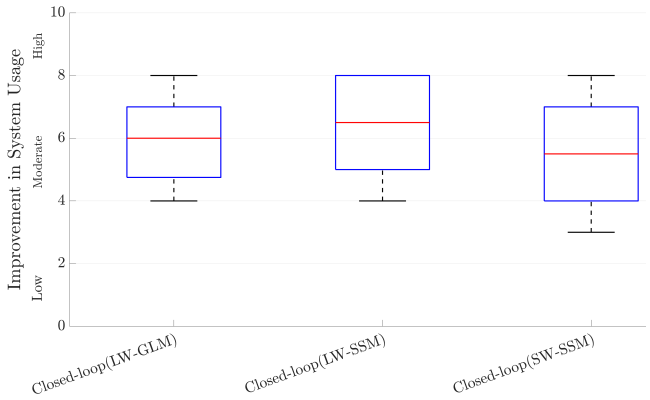



**Figure 10.** Subjective improvement in system usage for the closed-loop AAD system when using the LW–GLM, LW–SSM, and SW–SSM algorithms.

## 4. Discussion

The experimental results for the open-loop AAD system show that the largest median decoding performance is obtained by the LW–GLM algorithm (65%). This is in accordance with the experimental results in Reference [6], where it has been shown that open-loop AAD using a low number of electrodes with a correlation window smaller than 15 s results in a decoding performance lower than 75%. It should, however, be noted that the decoding performance in Reference [6] was obtained based on an optimal EEG electrode configuration, whereas the decoding performance reported in this paper was obtained based on a fixed EEG electrode configuration. In addition, the experimental results show that there is no significant difference in decoding performance between the open-loop and the closed-loop AAD system using the proposed AAD algorithms. This is consistent with the experimental results in Reference [5], where no significant difference in decoding performance between an open-loop and a closed-loop AAD system using visual feedback has been observed.

The experimental results show that the LW–GLM and LW–SSM algorithms using the large correlation window yield a larger median decoding performance compared to the SW–SSM algorithm using the small correlation window. The large correlation window provides a larger amount of data from the reconstructed attended envelope and the envelopes of the speech signals compared to the small correlation window, resulting in more discriminative correlation coefficients, more reliable probabilistic attention measures and a larger decoding performance. This is in accordance with the experimental results in Reference [4,6], where it has been shown that a larger correlation window results in a larger decoding performance. In addition, the experimental results show that the LW–GLM algorithm yields a larger median decoding performance than the LW–SSM algorithm. This may be explained by the fact that the LW–GLM algorithm infers the probabilistic attention measures based on the mean differential correlation coefficients rather than the absolute value of the correlation coefficients, hence providing a larger dynamic range including positive and negative values.

In conclusion, the results demonstrate the feasibility of closed-loop AAD in an online fashion, enabling the listener to interact with an adaptive gain controller (as an ideal speech enhancement algorithm) for a scenario with two competing speakers. On the one hand, the closed-loop cognitive-driven gain controller system improves the SIR between the attended and the unattended speaker. This may make it easier to follow the attended speaker, ignore the unattended speaker and switch attention between both speakers, resulting in a lower cognitive effort compared to open-loop AAD. On the other hand, the closed-loop cognitive-driven gain controller system introduces a significant delay to detect attention switches, which causes the attended speaker to be wrongly attenuated and the unattended speaker to be wrongly amplified for sometime. This may make it more difficult to follow the attended speaker and ignore the unattended speaker, resulting in a higher cognitive effort compared to open-loop AAD. Nevertheless, the subjective evaluation results indicate that overall the closed-loop cognitive-driven gain controller system demands a similar effort as open-loop AAD.

A delay to detect attention switches significantly influences the performance of the closed-loop cognitive-driven gain controller system. Recently, methods that are able to decode auditory attention with low delay have been proposed, e.g., based on a state-space model [9,18], neural networks [12,18,35], and common spatial patterns [36]. Therefore, investigating the potential of fast AAD methods for a closed-loop cognitive-driven gain controller system to detect attention switches could be interesting as future work.

While the closed-loop AAD experiments were performed without incorporating a practicing phase for the participants, the subjective evaluation results suggest that a significant improvement in system usage was obtained throughout the closed-loop AAD experiment. Future work could, therefore, investigate the impact of incorporating a practicing phase on the decoding and the speech enhancement performance of the cognitive-driven gain controller system. This practicing phase could simply be an extended version of the closed-

loop phase with many sessions where the participants can gather enough experience to fully master (i.e. find an intelligent way to control) the cognitive-driven gain controller system.

Although the application of the proposed cognitive-driven gain controller system was limited to acoustic scenarios with two competing speakers, it was shown in Reference [10] that open-loop AAD is feasible for an acoustic scenario with four competing speakers when using perfectly separated clean speech signals for decoding. In addition, the evaluation of the proposed system was limited to acoustic scenarios with non-moving speakers and with speakers located on opposite sides of the listener, whereas in real-world conditions speakers may move and may be located on the same side of the listener. Therefore, it would certainly be interesting as future work to investigate the performance of (an extension of) the proposed cognitive-driven speech enhancement system for more realistic acoustic scenarios.

The application of the proposed cognitive-driven gain controller system is obviously not limited to hearing devices. The system could also be used, e.g., for virtual reality (VR) that simulates a remote environment, e.g., for entertainment, training and medicine. It could be used to adapt the simulated world based on the auditory attention of the VR user.

## 5. Conclusions

In this paper, we proposed a closed-loop gain system which cognitively steers an adaptive gain controller based on real-time AAD for a scenario with two competing speakers. The real-time AAD infers the probabilistic attention measures of the attended and the unattended speaker from EEG recordings of the listener and the speech signals of both speakers. Based on these probabilistic attention measures, the adaptive gain controller amplifies the identified attended speaker and attenuates the identified unattended speaker. The loop of cognitive-driven gain control is then closed by presenting the amplified attended speaker and the attenuated unattended speaker via loudspeakers. The experimental results demonstrate the feasibility of the proposed closed-loop cognitive-driven gain controller system (both using AAD algorithms based on GLM and SSM), enabling the listener to interact with the system in real-time. Although there is a significant delay to detect attention switches, which causes the attended speaker to be wrongly attenuated and the unattended speaker to be wrongly amplified for some time, the proposed closed-loop system is able to improve the SIR between the attended and the unattended speaker. Moreover, the subjective evaluation results show that the proposed closed-loop cognitive-driven system demands a similar perceived level of cognitive effort to follow the attended speaker, to ignore the unattended speaker, and to switch attention between both speakers compared to open-loop AAD. With this work, an attempt was made to bring closed-loop cognitive-driven speech enhancement closer to real-world applications.

**Author Contributions:** Conceptualization, A.A. and E.F.; Data curation, A.A.; Formal analysis, A.A.; Funding acquisition, E.F., M.S. and H.P.; Investigation, A.A., E.F., M.S. and H.P.; Methodology, A.A., E.F. and M.S.; Project administration, H.P.; Software, A.A.; Supervision, A.A., E.F., M.S. and H.P.; Validation, A.A.; Visualization, A.A.; Writing—original draft, A.A.; Writing—review & editing, A.A. and S.D. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by WS Audiology and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC 2177/1-Project ID 390895286.

**Institutional Review Board Statement:** The study was carried out in accordance with the Declaration of Helsinki.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Acknowledgments:** The authors would like to thank Ulrich Hoppe and Ronny Hannemann for their fruitful discussions concerning AAD and Stefan Handrick and Ivine Kuruvila for their assistance in EEG data collection.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Doclo, S.; Kellermann, W.; Makino, S.; Nordholm, S.E. Multichannel signal enhancement algorithms for assisted listening devices. *IEEE Signal Process. Mag.* **2015**, *32*, 18–30. [CrossRef]
2. Gannot, S.; Vincent, E.; Markovich-Golan, S.; Ozerov, A. A Consolidated Perspective on Multimicrophone Speech Enhancement and Source Separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 692–730. [CrossRef]
3. O'Sullivan, J.A.; Power, A.J.; Mesgarani, N.; Rajaram, S.; Foxe, J.J.; Shinn-Cunningham, B.G.; Slaney, M.; Shamma, S.A.; Lalor, E.C. Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cereb. Cortex* **2014**, 25, 1697–1706. [CrossRef] [PubMed]
4. Mirkovic, B.; Debener, S.; Jaeger, M.; De Vos, M. Decoding the attended speech stream with multi-channel EEG: Implications for online, daily-life applications. *J. Neural Eng.* **2015**, *12*, 46007. [CrossRef]
5. Zink, R.; Baptist, A.; Bertrand, A.; Van Huffel, S.; De Vos, M. Online detection of auditory attention in a neurofeedback application. In Proceedings of the 8th International Workshop on Biosignal Interpretation, Osaka, Japan, 1–3 November 2016; pp. 1–4.
6. Fuglsang, S.A.; Dau, T.; Hjortkjær, J. Noise-robust cortical tracking of attended speech in real-world acoustic scenes. *NeuroImage* **2017**, *156*, 435–444. [CrossRef] [PubMed]
7. Wong, D.D.; Fuglsang, S.A.; Hjortkjær, J.; Ceolini, E.; Slaney, M.; de Cheveigné, A. A Comparison of Regularization Methods in Forward and Backward Models for Auditory Attention Decoding. *Front. Neurosci.* **2018**, *12*, 531. [CrossRef]
8. de Cheveigné, A.; Wong, D.D.; Liberto, G.M.D.; Hjortkjær, J.; Slaney, M.; Lalor, E. Decoding the auditory brain with canonical component analysis. *NeuroImage* **2018**, *172*, 206–216. [CrossRef]
9. Miran, S.; Akram, S.; Sheikhattar, A.; Simon, J.Z.; Zhang, T.; Babadi, B. Real-Time Tracking of Selective Auditory Attention From M/EEG: A Bayesian Filtering Approach. *Front. Neurosci.* **2018**, *12*, 262. [CrossRef]
10. Schäfer, P.J.; Corona-Strauss, F.I.; Hannemann, R.; Hillyard, S.A.; Strauss, D.J. Testing the Limits of the Stimulus Reconstruction Approach: Auditory Attention Decoding in a Four-Speaker Free Field Environment. *Trends Hear.* **2018**, *22*, 1–12. [CrossRef]
11. Das, N.; Bertrand, A.; Francart, T. EEG-based auditory attention detection: Boundary conditions for background noise and speaker positions. *J. Neural Eng.* **2018**, *15*, 66017. [CrossRef]
12. de Taillez, T.; Kollmeier, B.; Meyer, B.T. Machine learning for decoding listeners' attention from electroencephalography evoked by continuous speech. *Eur. J. Neurosci.* **2018**, *51*, 1234–1241. [CrossRef] [PubMed]
13. Alickovic, E.; Lunner, T.; Gustafsson, F.; Ljung, L. A Tutorial on Auditory Attention Identification Methods. *Front. Neurosci.* **2019**, *13*, 153. [CrossRef] [PubMed]
14. Aroudi, A.; Mirkovic, B.; De Vos, M.; Doclo, S. Impact of Different Acoustic Components on EEG-based Auditory Attention Decoding in Noisy and Reverberant Conditions. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2019**, *27*, 652–663. [CrossRef]
15. Ciccarelli, G.; Nolan, M.; Perricone, J.; Calamia, P.T.; Haro, S.; O'Sullivan, J.; Mesgarani, N.; Quatieri, T.F.; Smalt, C.J. Comparison of Two-Talker Attention Decoding from EEG with Nonlinear Neural Networks and Linear Methods. *Sci. Rep. Nat.* **2019**, *9*, 1–10. [CrossRef]
16. Teoh, E.S.; Lalor, E.C. EEG decoding of the target speaker in a cocktail party scenario: Considerations regarding dynamic switching of talker location. *J. Neural Eng.* **2019**, *16*, 036017. [CrossRef]
17. Tian, Y.; Ma, L. Auditory attention tracking states in a cocktail party environment can be decoded by deep convolutional neural networks. *J. Neural Eng.* **2020**, *17*, 036013. [CrossRef]
18. Aroudi, A.; de Taillez, T.; Doclo, S. Improving Auditory Attention Decoding Performance of Linear and Non-Linear Methods using State-Space Model. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 8703–8707.
19. Geirnaert, S.; Vandecappelle, S.; Alickovic, E.; de Cheveigne, A.; Lalor, E.; Meyer, B.T.; Miran, S.; Francart, T.; Bertrand, A. Electroencephalography-Based Auditory Attention Decoding: Toward Neurosteered Hearing Devices. *IEEE Signal Process. Mag.* **2021**, *38*, 89–102. [CrossRef]
20. Van Eyndhoven, S.; Francart, T.; Bertrand, A. EEG-informed attended speaker extraction from recorded speech mixtures with application in neuro-steered hearing prostheses. *IEEE Trans. Biomed. Eng.* **2017**, *64*, 1045–1056. [CrossRef]
21. Dau, T.; Maercher Roersted, J.; Fuglsang, S.; Hjortkjær, J. Towards cognitive control of hearing instruments using EEG measures of selective attention. *J. Acoust. Soc. Am.* **2018**, *143*, 1744. [CrossRef]
22. Han, C.; O'Sullivan, J.; Luo, Y.; Herrero, J.; Mehta, A.D.; Mesgarani, N. Speaker-independent auditory attention decoding without access to clean speech sources. *Sci. Adv.* **2019**, *5*, eaav6134. [CrossRef]
23. Pu, W.; Xiao, J.; Zhang, T.; Luo, Z. A Joint Auditory Attention Decoding and Adaptive Binaural Beamforming Algorithm for Hearing Devices. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 311–315.
24. Aroudi, A.; Doclo, S. Cognitive-driven binaural beamforming using EEG-based auditory attention decoding. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* **2020**, *28*, 862–875. [CrossRef]

25. Mesgarani, N. Brain-Controlled Hearing Aids for Better Speech Perception in Noisy Settings. *Hear. J.* **2019**, *72*, 10–12. [CrossRef]
26. Hering, E. Kostbarkeiten aus dem Deutschen Märchenschatz. In *Audiopool Hörbuchverlag MP3 CD*; BUCHFUNK Verlag: Trier, Germany, 2012; ISBN 9783868471175.
27. Ohrka, H. Ohrka.de-Kostenlose Hörabenteuer für Kinderohren. *Accessed Apr* **2012**, *30*, 2015.
28. Mirkovic, B.; Bleichner, M.G.; De Vos, M.; Debener, S. Target Speaker Detection with Concealed EEG Around the Ear. *Front. Neurosci.* **2016**, *10*, 349. [CrossRef] [PubMed]
29. Biesmans, W.; Das, N.; Francart, T.; Bertrand, A. Auditory-Inspired Speech Envelope Extraction Methods for Improved EEG-Based Auditory Attention Detection in a Cocktail Party Scenario. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2017**, *25*, 402–412. [CrossRef] [PubMed]
30. Akram, S.; Simon, J.Z.; Babadi, B. Dynamic Estimation of the Auditory Temporal Response Function from MEG in Competing-Speaker Environments. *IEEE Trans. Biomed. Eng.* **2017**, *64*, 1896–1905. [CrossRef]
31. Collett, D. *Modeling Binary Data*; Chapman and Hall: New York, NY, USA, 2002.
32. Nelder, J.A.; Wedderburn, R.W.M. Generalized Linear Models. *J. R. Stat. Soc.* **1972**, *135*, 370–384. [CrossRef]
33. Train, K.E. *Discrete Choice Methods with Simulation*; Cambridge University Press: Cambridge, UK, 2009.
34. Hochberg, Y.; Tamhane, A.C. *Multiple Comparison Procedures*; John Wiley and Sons: Hoboken, NJ, USA, 1987.
35. Lu, Y.; Wang, M.; Yao, L.; Shen, H.; Wu, W.; Zhang, Q.; Zhang, L.; Chen, M.; Liu, H.; Peng, R.; et al. Auditory attention decoding from electroencephalography based on long short-term memory networks. *Biomed. Signal Process. Control* **2021**, *70*, 102966. [CrossRef]
36. Geirnaert, S.; Francart, T.; Bertrand, A. Fast EEG-Based Decoding Of The Directional Focus Of Auditory Attention Using Common Spatial Patterns. *IEEE Trans. Biomed. Eng.* **2021**, *68*, 1557–1568. [CrossRef]