Real-time control of a hearing instrument with EEG-based attention decoding

Jens Hjortkjær[1,2], Daniel D.E. Wong[3,4], Alessandro Catania[1], Jonatan Märcher-Rørsted[1], Enea Ceolini[5], Søren A. Fuglsang[1,2], Ilya Kiselev[5], Giovanni Di Liberto[6], Shih-Chii Liu[5], Torsten Dau[1], Malcolm Slaney[7] and Alain de Cheveigné[3,4]

[1] Hearing Systems Section, Department of Health Technology, Technical University of Denmark, Kgs. Lyngby, Denmark

[2] Danish Research Centre for Magnetic Resonance, Centre for Functional and Diagnostic Imaging and Research, Copenhagen University Hospital - Amager and Hvidovre, Copenhagen, Denmark

[3] Laboratoire des Systèmes Perceptifs, Paris, CNRS UMR 8248, France

[4] Département d'Etudes Cognitives, Ecole Normale Supérieure, Paris, PSL, France

[5] Institute of Neuroinformatics, University of Zurich and ETH Zurich, Zurich, Switzerland

[6] School of Computer Science and Statistics, Institute of Neuroscience, Trinity College, The University of Dublin, Dublin, Ireland

[7] Center for Computer Research in Music and Acoustics (CCRMA), Stanford University, Stanford, CA, United States of America

E-mail: `jhjort@dtu.dk`

Abstract. Restoring normal speech perception in everyday noisy acoustic environments remains an outstanding challenge for hearing aids. Speech separation technology is improving rapidly but hearing instrument technology cannot fully exploit this advance without knowing which sound sources the user wants to hear. Even with high-quality source separation, the hearing aid must know which speech streams to enhance and which to suppress. Advances in EEG-based decoding of auditory attention raise the potential of a neuro-steered hearing instrument that selectively enhances the sound sources that a hearing-impaired listener is focusing their attention on. Here, we present a real-time brain-computer interface (BCI) system implementing this concept. Our system combines a stimulus-response model based on canonical correlation analysis (CCA) for real-time EEG attention decoding with a multi-microphone hardware platform enabling low-latency real-time speech separation through spatial beamforming. In this paper, we provide an overview of the system and its various components and discuss prospects and limitations of the technology. We illustrate its application with case studies of listeners steering acoustic feedback of competing speech streams via real-time attention decoding. A software implementation code of the system is publicly available for further research and explorations.

1. INTRODUCTION

Listeners with normal hearing can effortlessly follow speech in 'cocktail party' scenarios with competing speech sources by selectively attending to relevant speakers (Cherry, 1953; Yost, 1997). Hearing loss greatly reduces this ability, even when the sound input to the ear is amplified by a hearing aid (Bronkhorst, 2000). Consequently, difficulties with communication in everyday noisy situations continue to be the most common complaint from hearing-aid users (Kochkin, 2010; Lesica, 2018). Potential solutions to this challenge are limited by the very design of hearing aids as sound processing devices. Acoustic amplification or noise suppression in hearing aids can restore some degree of audibility, but the hearing aid cannot help users follow individual speech sources in the presence of others without knowing which speakers the user is attending. Even if the hearing instrument could perform high-quality acoustic scene analysis and separate all the individual sound sources in a complex acoustic environment, it would still need to know which sound sources to enhance and which to suppress. This requires a user input to the hearing aid.

Consequently, the idea of a *cognitively controlled hearing aid* has been proposed to overcome this basic challenge (cocoha.org, Dau et al., 2018; Slaney et al., 2020; Geirnaert et al., 2021). The notion is that if a user's selective auditory attention can be measured from brain signals, then the brain-decoded attention signal can be used as input to a speech separation process in a hearing instrument. The hearing aid would then use the attention signal to control the relative levels of competing speakers in a real-time speech separation system. Such a hearing instrument is a form of neural prosthesis that assists the hearing-impaired listener by automatically enhancing relevant sound sources in a complex acoustic scene.

The vision of such neuro-steered hearing technology has been fueled by successes in auditory attention decoding from EEG or other neurophysiological signals (O'Sullivan et al., 2015; Mirkovic et al., 2015; Van Eyndhoven et al., 2016; Aroudi et al., 2016; Akram et al., 2016; Miran et al., 2018; Ciccarelli et al., 2019). In offline attention decoding studies, a main direction has been to examine the correlation between continuous EEG responses to speech mixtures and acoustic stimulus features of competing speech streams (Ding and Simon, 2012a; O'Sullivan et al., 2015). The main observation is that low-frequency cortical activity (<10 Hz) synchronize with slow fluctuations in the speech envelope (<10 Hz) and that this synchronization is stronger for attended speech streams compared to ignored ones (Mesgarani and Chang, 2012; Power et al., 2012; Ding and Simon, 2012b; Golumbic et al., 2013). In seminal work, O'Sullivan et al. (2015) used a linear regression model to reconstruct speech envelopes from continuous multi-channel EEG responses to natural speech. When selectively attending to one speaker in a two-talker mixture, the speech envelopes reconstructed from the EEG responses were more correlated with the envelope of the attended speech stream compared to the ignored one. This, in turn, allowed listeners' attention to be predicted from single-trial (1 minute) EEG data (O'Sullivan et al., 2015). Later studies have demonstrated that EEG-based

attention decoding with similar models can also be achieved in more realistic acoustic scenarios (Fuglsang et al., 2017; Aroudi et al., 2019), with shorter EEG data segments (de Cheveigné et al., 2021; Ciccarelli et al., 2019), and in older listeners with hearing loss (Fuglsang et al., 2020).

These successes in offline attention decoding raise hope for a real-time implementation where the brain-decoded signal is used to steer the acoustic feedback to the listener (e.g., in a hearing aid). In a real-time system, EEG decoding would be coupled with an acoustic speech separation system that separates different voices in the listening space. The decoded attention signal can then control the relative gains of the separated speech sources in order to enhance attended ones and suppress ignored ones.

Such a BCI works by selectively amplifying those speech streams that are more strongly correlated with EEG responses. This implies that the system alters the speech mixture perceived by the user. If the relative level of the perceived speech sources influences the neural response, and hence the stimulus-response correlation (Ding and Simon, 2012a; Das et al., 2018), this creates a closed-loop feedback situation (see Fig. 1). Such a closed-loop system immediately raises several challenges to be considered. For instance, if an ignored speech stream is attenuated by the BCI, can the user then switch attention to the ignored speaker? How audible do different speakers in a scenario have to be to remain decodable? Is there an influence of visual feedback on closed-loop acoustic steering? How fast and how accurate must an attention decoder be to be usable for acoustic steering in real-life communication? How should the acoustic feedback be rendered to the user? Should the BCI seek to enhance attended sources or suppress ignored ones? Can users, typically elderly hearing-impaired people, adapt to the device and learn to harness its power to their benefit?

A real-time auditory attention decoding system based on stimulus-response correlation is unique in a BCI context. Most existing BCI systems classify brain signals to generate some form of system command (e.g., steering some application), i.e., the neural data are the only input to the BCI. The neuro-steered hearing instrument discussed here, on the other hand, takes both EEG and audio signals as inputs. Classification now relies on attention-related changes in the *relation* between EEG and speech stimulus features. This imposes a strong prior on the classification problem: the BCI can act specifically on the part of the neural response that is related to the stimulus. This can potentially help remove noise and task-unrelated changes in the EEG, which may in turn help the BCI's ability to generalize beyond its training conditions.

Here we present a real-time implementation of such a closed-loop auditory attention decoding system. Our system addresses each requirement of a real device (hardware, software, usability) as explored within in the COCOHA project (cocoha.org). It consists first of a wireless multi-microphone hardware platform that performs online low-latency acoustic speech separation via spatial beamforming. Next, an attention decoding module performs online analysis of the correlation between speech stimulus features of the separated speech streams and continuous EEG signals. By correlating the EEG to envelope features of each speech stream in a mixture, the decoder identifies the attended

stream as the most correlated one. The decoded attention signal is then used to steer the relative gains of the separated speech streams to enhance the attended speaker in the acoustic feedback to the listener.

The structure of this paper is as follows. We first present an overview of the real-time system and its various components. Next, we demonstrate its functionality in real-time closed-loop demonstrations performed with human subjects, both with normal and clinically impaired hearing. In a first demonstration, we investigate the ability of older hearing-impaired listeners to perform real-time attention steering with speech mixtures using our system. Next, we investigate how visual input of the speakers face might affect closed-loop attention decoding. The demonstrations serve as proof-of-principle case studies of real-time steering of acoustic feedback via attention. Details of the implementation and the demonstrations are given in the Methods section at the end.

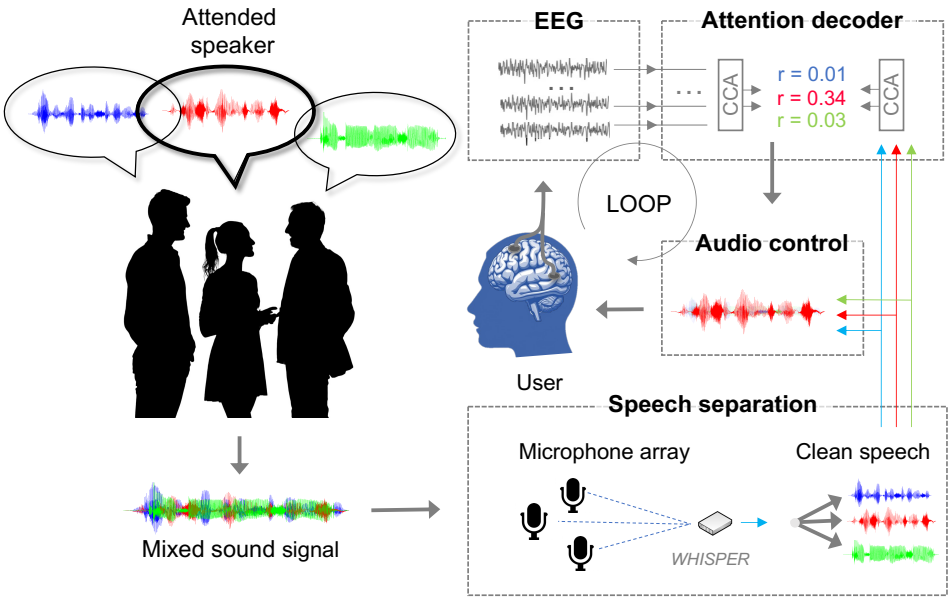

Figure 1: The cognitively controlled hearing aid in overview. Speech is picked up by an *ad hoc* array of microphones and processed to produce clean speech streams, one per speaker. Continuous EEG signals from the user are synchronized with the separated audio streams and transmitted to an attention-decoding module that classifies the attended speech stream based on correlations between audio and EEG features. The time-varying classification outputs are then used to control the relative gains of the separated audio streams.

2. IMPLEMENTATION

2.1. Overview of the system

Figure 1 shows an overview of the real-time system. First, mixed sound signals in a free-field acoustic environment are picked up by an *ad hoc* wireless microphone array placed in the listening space. A hardware platform synchronizes microphone signals between

each other and performs low-latency beamforming to output separate speech streams corresponding to N different speakers in the listening space. Concurrently with this audio processing, EEG signals are recorded from the listener's scalp and synchronized with the audio signals. Synchronized EEG and audio signals are then streamed to an attention decoding module. The attention decoding module estimates which stream is currently being attended to by the listener, and this knowledge is used to modulate the relative gains of the audio channels in the mixture. The re-rendered audio mixture with the attended speaker enhanced is then presented to the listener via an insert earpiece (e.g., a hearing aid speaker).

2.2. Microphone array platform and acoustic scene analysis

For acoustic separation of speech mixtures, our system uses a wireless acoustic sensor network called WHISPER, a custom microphone hardware platform developed for this project (see Kiselev et al. 2017 and Ceolini et al. 2020b for details). It consists of hardware modules that each support up to four microphones (see Fig. 2) which can be combined in wireless sensor networks of any arbitrary number. WHISPER synchronizes microphone signals between each other, and the array can then be used for low-latency, high-quality speech separation via beamforming computed locally on the platform. Here, we use WHISPER for beamforming with synchronized wireless microphone networks distributed in *ad hoc* arrays in the listening space. In an *ad hoc* array, the microphones occupy arbitrary positions, and the system adapts to this geometry (in contrast to preset array geometries commonly proposed in the literature). Details of the implementation can be found in Ceolini et al. (2020b).

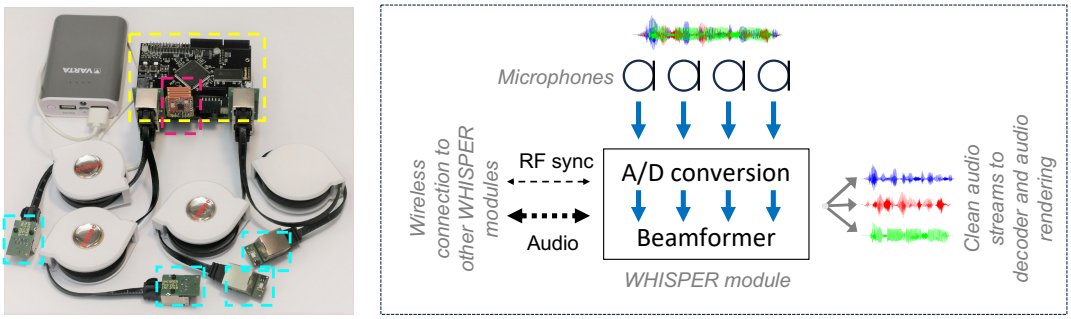

Figure 2: ***Left***: prototype WHISPER module (yellow box) equipped with four microphones (cyan boxes) and a wireless synchronizer (red box). ***Right***: a WHISPER module acquires audio from up to four microphones and exchanges sampled audio and synchronization signals with other WHISPER modules or a common laptop. Microphone streams are merged and processed by a beamformer to produce clean streams for the decoder and audio rendering modules.

Wireless *ad hoc* arrays as implemented with WHISPER offer several advantages as a solution for speech separation in hearing technologies. First, distributed microphones in the listening space allow good spatial coverage and high probability that relevant sources

are picked up by a close microphone. Second, with WHISPER, more microphones can readily be added due to its modular architecture. Theory allows for infinite interferer rejection as long as the number of microphones is at least equal to the number of sources (target and interferers). However, in the presence of noise and reverberation, or for certain degenerate configurations, a larger number of microphones may provide better performance, as observed empirically. It also offers leverage to minimize spectral distortion of the target as a result of propagation and spatial filtering. Third, a platform with distributed computing capabilities can offload the computational burden of potentially complex separation algorithms.

WHISPER can be used to deploy any separation or enhancement algorithm given its ability to run floating-point algebra. For our real-time BCI, we used the simple but powerful beamforming algorithm Minimum-Variance Distortionless Response (MVDR). In separate evaluations of speech separation quality with 2-4 speaker mixtures (with *a priori* SNR=0), MVDR on WHISPER obtains ∼6 dB SDR using 1 WHISPER node (4 microphones), and ∼9 dB SDR using 3 nodes (12 microphones, see Ceolini et al. 2020b). This corresponded to a short-time objective intelligibility (STOI) score of 0.6 and 0.9, respectively. We have also evaluated WHISPER running other similar beamforming algorithms, such as the speech distortion weighted multichannel Wiener filter (SDW-MWF), maximum SNR (MSNR), masked-based MVDR (MB-MVDR) or mask-based generalized eigenvalue (MB-GEV), with equivalent high performance (Ceolini et al., 2020b). Notably, WHISPER can also be used to deploy high-performing speech enhancement algorithms based on deep neural networks (see Ceolini and Liu, 2019).

Achieving low latency is critical for successful applications in hearing technology. The effective latency that can be achieved with beamforming on WHISPER depends on several factors, including the number of modules and microphones, the beamforming algorithm, and WiFi transmission latency (see Ceolini et al., 2020b). The BCI experiments reported here were conducted using a single WHISPER module (with 4 microphones) wired to a remote laptop running an MVDR beamformer. In this situation, the processing latency is essentially the algorithmic latency dictated by the 512-point window on a 24 kHz input signal and 75% window overlap required for frequency-domain processing, roughly ∼11ms. Beamforming on several WHISPER modules with off-the-shelf wireless protocols would entail considerably greater latency, to be optimized in future work.

2.3. Attention decoding

The real-time attention decoding pipeline is illustrated in Figure 3. The decoder receives the separated audio streams from the microphone platform and multichannel EEG. Before an experiment, EEG and audio streams are synchronized via an initial trigger pulse sent along both paths. The online processing pipeline then performs feature extraction for both EEG and audio streams (Fig. 3 blue), including multichannel envelope extraction for the audio signals, and standard filtering and denoising for the

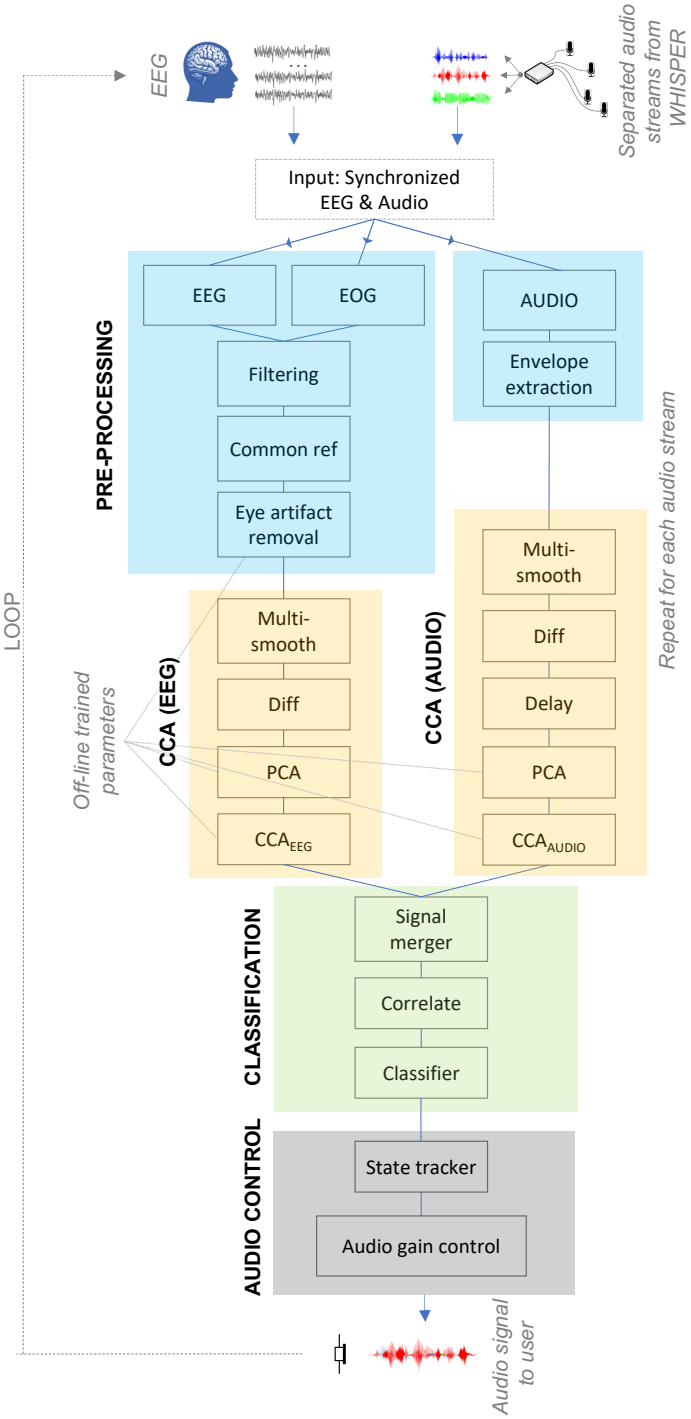

Figure 3: Schematic of the real-time attention decoding pipeline. Each box represents a processing object implemented in OpenVibe (Renard et al., 2010). Inputs are synchronized EEG and separated audio streams. The output is a gain control signal for each speech stream. The processing can be divided into pre-processing (blue), CCA transformation of EEG and audio (yellow), classification of attention (green), and audio control (grey).

EEG (see Methods for additional details).

The core of the attention classification pipeline is a real-time implementation of a linear decoder based on canonical correlation analysis (CCA) (see de Cheveigné et al. 2018 for additional details). This approach generalizes linear regression models to predict synchronized stimulus-response activity (de Cheveigné et al., 2018, 2021). Based on training data, a CCA model is trained to find linear transforms that are then applied to both EEG and audio features from each audio stream at test time (Fig. 3, yellow). As inputs to the CCA model, the EEG and audio streams are passed through a multichannel filterbank designed to extract stimulus-response correlations at multiple time scales. The filterbanks are identical on the EEG and audio side to minimize spectral mismatch between the signals to be correlated. On the EEG side, the CCA performs spatiotemporal filtering to capture different neural sources correlated with stimulus activity (de Cheveigné et al., 2018). On the audio side, the CCA combines envelope subbands to perform bandpass filtering that capture envelope fluctuations at different time scales (Dau et al., 1997). Once the CCA weight matrices are obtained on training data, the linear transformation of the data simply consists of parallel matrix multiplications at test time (one for the EEG data, one for each separated audio channel), implying low computational cost.

The CCA-transformed EEG and audio streams produce sets of pairwise correlations coefficients between EEG and audio features. For each pair, the correlation scores are continuously computed within a predefined decoding window (Fig. 3, green). The duration of the decoding window represents a trade-off between decoding speed and accuracy (Wong et al., 2018; de Cheveigné et al., 2021). In offline studies, we found that our CCA-pipeline can classify attention with an accuracy of around 70-80% with decoding windows around 6-8 seconds (see Fig. 6C below for an example of offline decoding accuracies). The windowed correlation component scores are passed to a linear classifier, here implemented as a two-class support vector machine to decide which of two streams is attended. The choice of classifier is not critical and can be replaced to accommodate an arbitrary number of audio streams. In the simplest case, the correlation component scores can simply be summed and compared between N streams.

2.4. Audio control

The output of the attention classifier is used to control the gain of the different audio streams in the speech mixture (Fig. 3 grey). We have explored different audio control strategies. A simple first choice was to map time-varying classifier probabilities to an instantaneous audio gain on each speech stream via a non-linear mapping function. For this, we used a point non-linearity that pulls low classification probabilities towards zero, and high classification probabilities towards a maximum applied gain (see Methods for details). The rationale is to leave the gain unchanged (i.e., no BCI feedback) unless the classifier has reached some level of certainty about the subject's attention.

A drawback of a simple instantaneous coupling between classification scores and gain is the potential of erratic and fast gain changes. Even if misclassification occurs

infrequently, such acoustic artefacts may be perceived as highly disturbing by the user. As an alternative strategy, we implemented an attention state space model to control the audio levels (see Methods). In this, the rate of gain change varies with the certainty of the classifier. In each classification window, the user's attention state is a function of the previous state and the current information transfer rate estimated from the classifier probabilities. In effect, the state tracker accumulates information over time to estimate the user's current attention state and the relative gains change based on this state. This makes the system less prone to detrimental sudden shifts in gain upon misclassifications while maintaining the possibility of fast state changes when the current information transfer rate of the system is high.

3. DEMONSTRATIONS

In this section, we report different real-time experiments performed to demonstrate that the system is functional and usable by a real subject. They also serve to gain insight as to the best choice of parameters (e.g., decoding window length) and potential audio control strategies. The demonstrations serve as proof-of-principle use-cases rather than quantitative evaluations of performance with end users.

3.1. Demonstration with real talkers

A video illustrating the real-time system is shown below (Fig. 4). A listener wearing a wireless EEG system focuses attention on either the male or female speaker. The two speakers alternately raise their hands to indicate that the listener should attend to her or him. The WHISPER microphone platform is placed on the table between the talkers and the listener in an *ad hoc* arrangement. The microphone signals are synchronized on the platform and streamed to the speech separation module that performs online beamforming. The separated speech streams are transmitted to the attention decoding module running on a remote laptop. The decoding module receives and synchronizes EEG and audio streams and performs CCA-based decoding of attention, as described above. The classification output controls the relative gain of the two speech streams in the acoustic feedback presented to the headphones of the subject. The audio track of the left video contains the recorded room audio (the mixed speech). The audio track of the right video contains the beamformed audio presented to the subject, i.e., emulating the hearing aid output. More video demonstrations are available at cocoha.org/2018/12/31/videos/.

3.2. Demonstrations with normal-hearing and hearing-impaired listeners

We first tested the real-time system with the clinical target group: older listeners with hearing impairment. Hearing-impaired (HI) listeners with moderate-to-severe hearing loss and normal-hearing (NH) listeners were presented with two competing speech streams in an attention-switching paradigm. To be able to compare results across

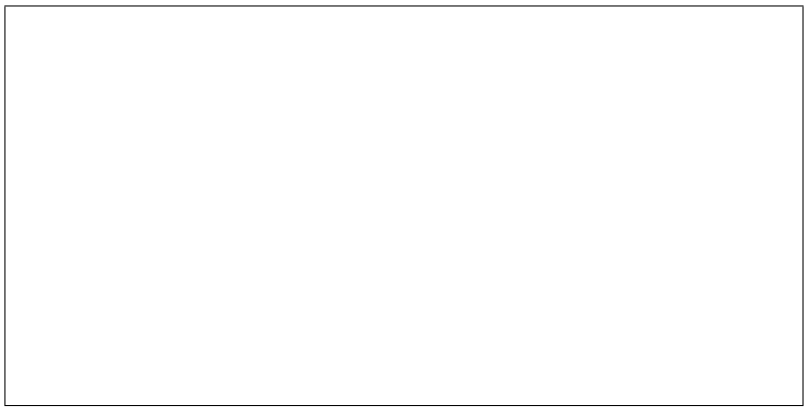

Figure 4: [Videos removed in this preprint. Please find the videos at cocoha.org/2018/12/31/videos/]. Video demonstration of real-time attention decoding. Left: the mixed speech audio recorded in the room. Right: the separated speech audio received by the subject. The subject (on the right) is wearing a wireless EEG system and headphones receiving the acoustic feedback from the BCI. Speakers (on the left) raise their hand alternately to indicate that the listener should attend to them. The WHISPER microphone platform is placed on the table between the listener and the speakers

sessions and listeners, pre-recorded speech (audiobooks) was presented via loudspeakers. Training data were obtained with subjects attending one of two simultaneous speakers on separate 50 sec long trials for ∼30 min (as in offline studies). After training the attention decoder, subjects performed closed-loop steering on 90 sec-long trials. Subjects were prompted to attend to one speaker at the beginning of the trial and to switch attention to the other speaker after 45 sec. The gain was controlled as an instantaneous function of classifier output, and a maximum gain difference of 10 dB was used.

Data from these real-time experiments are shown in Figure 5. The average gain applied to the two speech streams as a result of real-time decoding are shown in Fig. 5C and were similar for the HI (blue) and NH (red) listeners. At the onset of trials, the two speakers are presented at the same baseline level (0 dB target-to-masker ratio, TMR). As expected from offline decoding results, the system can correctly classify the attended speaker and the acoustic gain changes accordingly at the onset of the trial. When cued to shift attention, the subject attends to the other speaker. At that point, the target speaker has a negative target-to-masker ratio, here the maximum gain difference of -10 dB TMR (supposing that the system classified the speaker correctly on the first half of the trial). As can be seen, the gain successfully switches to enhance speaker 2. The average switch time, computed as the duration from the switch cue to the zero-crossing point in the classifier output, was found to be 4.7 sec (min=2.9 sec, max=6.3 sec).

After the experiments, participants were debriefed and interviewed about their experience of using the neurofeedback system. All participants reported a subjective sense of ability to steer the acoustic feedback that improved during the experiment. This could indicate that closed-loop training with the BCI may be important for good performance. Participants also indicated that occasional misclassification by the system

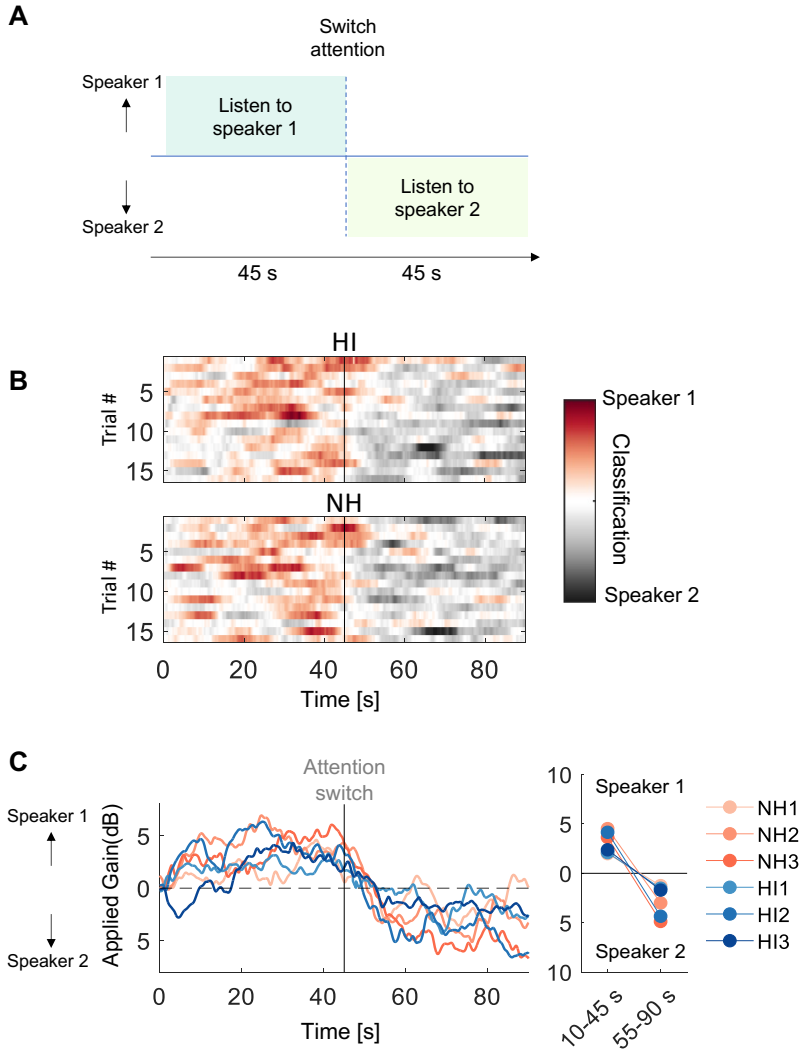

Figure 5: Real-time attention steering of acoustic feedback in hearing-impaired (HI) and normal-hearing (NH) listeners. (**A**) Paradigm: subjects were presented with two simultaneous speech streams and cued to switch attention to speaker 2 after listening to speaker 1 for 45 secs. (**B**) Outputs of the attention classifier as a function of time. The line indicates the point of the attention switch. (**C**) Applied gain (in dB) to speaker 1 and speaker 2 as a function of time. Red traces indicate NH listeners, blue traces indicate HI listeners. The panel on the right summarizes the average gain applied across trials 0-35 sec before and 10-35 sec after the attention switch, i.e., omitting the 10 sec following the switch.

(the speaker to-be-ignored being amplified) was experienced as highly detrimental for listening, making them aware of the BCI and ways to steer it correctly.

3.3. Real-time experiments with audio-visual speech

Auditory attention can be decoded while listeners only hear the speaker they are attending, but listeners typically look at the speaker they are listening to in real-world situations. It is unclear whether a BCI trained only on auditory inputs might also

work successfully in natural audiovisual conditions. The system could potentially fail to generalize to this mismatching stimulus condition. Alternatively, the visual face may be correlated with audio envelope information (O'Sullivan et al., 2017a; Pedersen et al., 2022), and it may be easier to focus auditory attention on a speaker that can be seen, which may in turn improve decoding (O'Sullivan et al., 2013). Potential visual benefits are important to investigate, for instance in the case were the user wants to switch attention to a previously ignored speaker. In this case, visual information correlated with the speech signal may help 'recover' a target with low audibility.

We tested these questions in a larger experiment. Training data were obtained from normal-hearing participants ($N = 20$) listening to two-talker speech mixtures with or without watching the visual input of the speakers' moving faces (Fig. 6). The listeners were then tested using the closed-loop system either with audiovisual stimuli or with the audio alone. The stimuli consisted of two separate pre-recorded speech videos presented to all subjects, i.e. the microphone platform was not used for this experiment. As a control condition, we included trials without BCI control, i.e. trials where the audio gains were left unchanged, referred to as 'open-loop' trials. In test trials of 50 seconds, subjects were again prompted to follow one speaker and then cued to switch attention midway through the trial (i.e., after 25 sec). In closed-loop trials, the audio gains were controlled by a state tracker with a sliding decoding window of 6.5 sec and a classification rate of 4 Hz. The same speech envelope decoding pipeline as described above was used in both audiovisual and audio-only conditions (i.e. no visual features were used in the decoder).

Figure 6 shows data from the real-time experiments with audiovisual speech. Participants were similarly able to steer the BCI with and without seeing the speakers face. A modest improvement with visual feedback was seen in the offline training data (Fig. 6C), but not in the real-time data. Fig. 6B shows the output of the attention state tracker in the different BCI conditions (closed-loop audio-only, closed-loop audiovisual, open-loop) and the different training conditions (audio training, audiovisual training). The average switching time of state tracker, again computed as the duration from the switch cue to the zero-crossing point, was found to be 12.6 secs., with no difference found between BCI conditions. As a summary measure of BCI tracking performance, Fig. 6D shows the average accuracy of the differential output of the state tracker across subjects, i.e., whether the state tracker steps in direction of the target speaker in each classification window. No differences between closed-loop vs open-loop steering were observed, indicating that subjects can switch attention and 'recover' an attenuated (previously ignored) speaker. While the real-system accurately tracks attention for most subjects, outlier subjects performing at chance-level can be observed (Fig. 6D).

4. DISCUSSION

Helping hearing-impaired listeners follow speech in noisy everyday environments remains a fundamental challenge for hearing technologies. Here we present, as a proof-of-

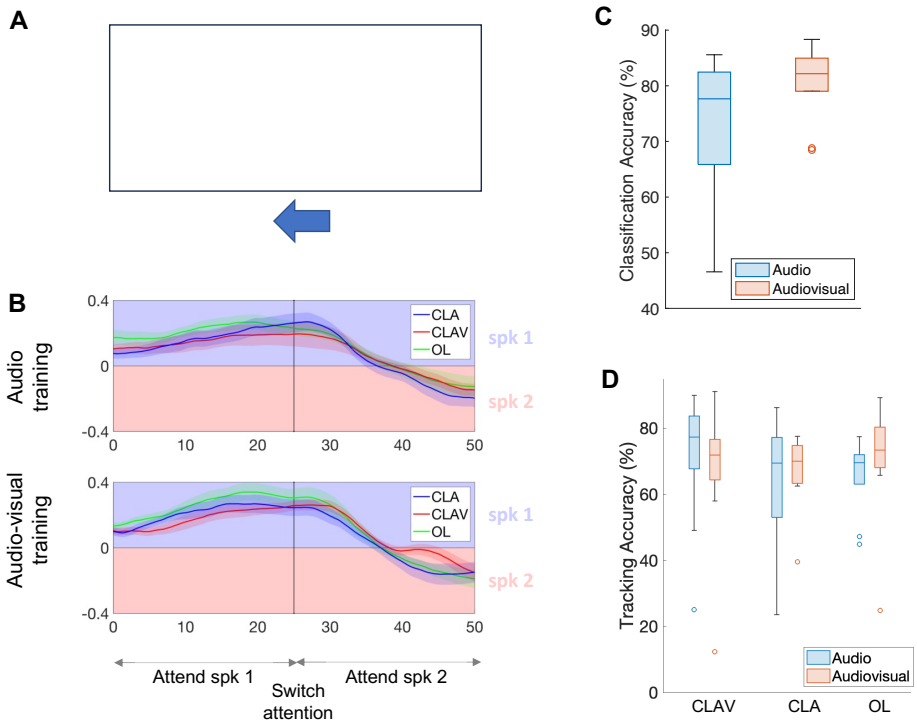

Figure 6: Experiment II: real-time attention decoding with and without visual input. (**A**) Listeners were cued to attend a male or female speaker in audiovisual or audio-only conditions. [Photographs removed in preprint]. (**B**) Real-time outputs of the attention state tracker for listeners trained with audio-only (above) or audio-visual (below) stimuli. Line colors indicate the different BCI stimulus conditions. Black: closed-loop audio-only (CLA), blue: closed-loop audio-visual (CLAV), red: open-loop. (**C**) Offline average attention decoding accuracies based on analysis of the training data in segments of 6.5 sec. for listeners trained with audio-only stimuli (blue) or audio-visual stimuli (red). (**D**) Online accuracies of the attention state tracker.

principle, a real-time system for acoustic speech separation controlled by attention. Our initial closed-loop demonstrations indicate that listeners, including elderly people with hearing loss, can use such a system to enhance the speech sources they are focusing attention on. If the signal-to-noise ratio (SNR) of speech sources can be directly controlled and improved in this way, a gain in speech understanding follows inherently. Yet, numerous challenges in bringing such a technology to real-life use obviously remain. Here we discuss some of the critical prospects and limitations.

Using attention for seamless BCI control

Using attention to steer acoustic feedback has inherent benefits in the context of a neural prosthetics. BCIs often require the user to direct attention to the BCI device itself (e.g., a P3-speller). The attention-controlled hearing instrument, on the other hand, decodes and 'amplifies' a natural listening process. Listeners naturally orient attention towards relevant sources in real-world listening situations, allowing them to

better perceive relevant sources and suppress background noise. Our BCI assists this natural behavior without interrupting it. If the system correctly decodes attention, then the interaction is seamless. An alternative solution might be to have the user control the speech separation system via an overt device (e.g., selecting which speakers to enhance on a mobile phone). However, such a solution has the detrimental effect that it diverts the user's attention towards steering the device rather than facilitating it. With the proposed BCI, listeners simply attend to speech as they do in everyday situations to better perceive the relevant source.

BCI based on stimulus-response correlation

Most existing BCI systems rely on classification of brain signals related to a behavioural state to form a control signal. The BCI proposed here relies instead on behavior-related changes in the stimulus-response relation for classification, i.e., the system uses both brain and audio signals as inputs. This imposes an important constraint on the classification problem: the BCI can act specifically on the part of the neural response that is related to the stimulus and not on global changes in brain states. This may make the BCI less susceptible to random EEG noise or task-unrelated neural signals.

An alternative solution is to classify a listener's spatial attention based on changes in EEG alpha power (Kelly et al., 2005; Geirnaert et al., 2020; Popov et al., 2023), i.e., classifying spatial attention based only on EEG features. Yet, this solution suffers the inherent problem that changes in lateralised EEG alpha power coincides with many task-unrelated changes in the cognitive state of the subject. In pilot studies using this approach, we found that high attention classification accuracies ($> 90\%$) could be achieved with short EEG segments ($<$1 sec), but the system failed to work robustly in closed-loop. A BCI based on stimulus-response metrics may help the critical generalization to real-world conditions beyond those used for training the system.

Yet, a stimulus-response-based BCI may also imply a potential weakness: if the stimulus-response relation is absent, then the BCI cannot classify attention. If a speech stream has been suppressed by the BCI, the listener may still *intend* to listen to the speech but may not be able to attend to it. Our initial attention switching experiments indicate that it is possible to 'recover' previously ignored sources attenuated by up to $10dB$ relative to the previous target source, but the boundary conditions for this remain to be explored (Das et al., 2018). Offline experiments with audiovisual speech tracking indicate that 'envelope' speech tracking may be possible in silent lipreading conditions where the speaker can only be seen (O'Sullivan et al., 2017a). Visual feedback of the speaker may thus help a BCI track a speech source even when it cannot be heard.

What is accurate decoding?

Attention classification implies a trade-off between speed and accuracy. A critical limitation of the technology is a very low tolerance to misclassification. In experiments with hearing-impaired listeners, subjective reports indicated that misclassifications were

experienced as highly disturbing. With decoding windows below 10 sec., the accuracy is currently below 100% correct classification for most subjects. This implies that the system occasionally amplifies the unattended source and suppresses the attended target, i.e., the opposite of the intended effect. Even if misclassification occurs infrequently (e.g., 10% of time time), the effect may be disastrous. While the BCI is seamless when classification is correct (the subject simply listens to speech), incorrect classification immediately makes the listener aware of the presence of the BCI. Many test participants reported that this meant switching strategy into actively steering the device, a process reported as effortful. In consequence, the opposite effect of the intended purpose of the BCI.

Better attention decoding?

Progress towards a successful technology may therefore be to improve the accuracy of the attention decoding. Our current implementation relies on linear mappings (CCA) between audio and EEG features. Non-linear extensions have been proposed in offline attention decoding studies (de Taillez et al., 2020). An obvious extension is to consider deep learning-based models (Ciccarelli et al., 2019; Vandecappelle et al., 2021; Thornton et al., 2022). However, a number of pitfalls, often deceitfully hidden, should be pointed out. One concern is the limited availability of large amounts of unbiased training data. In typical attention decoding experiments, listeners are instructed to attend to one speaker in one ear and ignore a different speaker in the other ear. Linear methods based on envelope-EEG correlation can generalize to new speakers because they explicitly model this correlation. However, a non-linear neural network trained to classify attention may simply ignore the EEG and classify acoustic differences between the speakers. To avoid this, experiments that generate training data should be carefully counterbalanced, e.g., attending equally to both speakers, equally attending left/right, etc. Even here, subtle differences may leak into the training data. For instance, speakers may slightly shift voice pitch when speaking over longer durations, and conventional cross-validation procedures (such as leave-one-trial-out) are prone to bias from such details. Reported results must be taken with a grain of salt.

What is attention?

Another critical challenge for obtaining high decoding accuracy is the lack of ground truth measures of attention. Decoding models, as the ones presented here, are trained on data obtained with listeners cued to attend a given target during training trials. The subjects' attention state is assumed to remain stationary throughout a trial, but the attention state cannot be monitored in any clear way. In natural listening situations, attention fluctuates with the listener's endogenous state and with external salient events (Esterman et al., 2013; Chun et al., 2011). When instructed to attend a target, a listener's attention may still occasionally switch to the non-target, or attention may simply drift from the task altogether (Smallwood and Schooler, 2015; Weissman et al.,

2006). The lack of measures of such attention dynamics puts decoding models on uncertain ground. An alternative approach may be to dispose of attention labels and focus instead on modelling the stimulus-response relation with simpler stimuli, e.g., listening to a single speaker. For instance, de Cheveigné et al. (2021) proposed to train stimulus-response models on whether or not a segment of EEG matches the auditory stimulus that evoked it. This approach allows models to be trained in a self-supervised manner. If attention modulates the stimulus-response mapping, then robust stimulus-response models may be key to progress without relying on attention labels.

Need for speed?

Improved decoding accuracies can be obtained by integrating the classification over longer durations, i.e., by sacrificing classification speed. In dynamic communication situations involving many talkers, attention may switch rapidly and decoding windows of several seconds are already unfeasible. In other situations, however, longer time constants may be acceptable. In many communication situations the communication partners remain the same over longer periods of time – e.g., in a restaurant or at a meeting - where the BCI may slowly 'zoom in' on relevant sources. An alternative decoding strategy may be to identify such long-term stable communication contexts.

Alternative EEG solutions

A critical limitation of the technology is the quality of EEG signals recorded during natural behavior. Cortical signals recorded with scalp EEG are inherently weak and noisy. Our current demonstrations of a real-time BCI are still limited to stationary setups where the listener is sitting still and passively listening. Motion artifacts during natural behavior tend to dramatically surpass the brain signals of interest in scalp EEG. Ear-EEG (EEG recorded in the ear canal or behind the ear) offer feasible solutions for integration with a hearing aid (Looney et al., 2012; Bleichner and Debener, 2017), but does not appear to improve decoding performance (Fiedler et al., 2017). Invasive cortical recordings improve brain signal SNR considerably and O'Sullivan et al. (2017b) demonstrated fast and accurate decoding using electrocorticography (ECoG) recordings in neurological patients (O'Sullivan et al., 2017b). New multichannel implant solutions are emerging for BCI perspectives (Musk et al., 2019; Steinmetz et al., 2021). Apart from greatly improved SNR of intracranial signals, long-term recordings from the living brain propose new prospects for obtaining large amounts of data and developing large-scale attention decoding models. Less invasive solutions such as subcutaneous EEG are currently in clinical development (Viana et al., 2021) and offer interesting perspectives for decoding attention.

Acoustic speech separation solutions

Using a separate speech separation front-end based on distributed microphones was a design choice in our system. Distributed sensors on dedicated hardware offer high-quality speech separation at very low latencies, which is paramount for hearing instrument applications. One drawback with stationary microphones is the need for calibration for a given spatial configuration of sound sources. However, Voice Activity Detectors (VADs) could be used for dynamic calibration in non-stationary environments. More flexible extensions could be recently proposed 'acoustic swarms' of self-distributing microphones that dynamically locate and separate speech sources in changing environments (Itani et al., 2023). Other challenges arise with synchronization of wireless nodes based on radio transmission which is generally poorly suited for outdoor environments or rooms with plentiful reflective materials like metal surfaces.

An obvious drawback is the need for distributed microphones in the listening space. However, solutions with user-worn microphones (placed on the hearing aid) are also conceivable. Single-channel or two-channel solutions may exploit advances in speech separation within deep learning (Luo and Mesgarani, 2018). O'Sullivan et al. (2017b) already demonstrated that attention can be reliably decoded from the outputs of a single-channel speech separation system, i.e., without assuming access to the clean speech sources (O'Sullivan et al., 2017b). Successful attention decoding does not follow automatically from previous results with 'clean sources' since some degree of cross-talk between the separated streams must be expected even with high-performing speech separation systems (Subakan et al., 2021; Ravanelli et al., 2021).

In our system, the speech separation front-end is separate from the attention decoding process. The task of the decoder is then to decide which of N separated streams the listener is attending. An alternative framework is to combine the speech separation and decoding steps into one process (Ceolini et al., 2020a). Since a stimulus-response decoder estimates speech features (typically amplitude envelopes) of the attended stream, these can be used as a 'hint' of the clean audio source. Ceolini et al. (2020a) presented a neural network-based speech separation system trained on the neural data to extract the clean audio signal directly from the speech mixture. Using ECOG or EEG-reconstructed speech envelopes, the system directly separates the target from the mixture input. Higher envelope reconstruction accuracies were shown to translate into higher separation quality (e.g., an EEG-based reconstruction accuracy of 0.35 corresponds to around 4.2 dB SDR as reported in Ceolini et al. 2020a). This integrated approach could potentially be expanded to end-to-end neural networks that perform speech separation directly from the raw audio mixture and the continuous EEG (Hosseini et al., 2021). An advantage of such an approach is that the speech separation system does not have to separate all sources in an acoustic scene (and compare the brain response to each acoustic source) but can focus on separating only the relevant target source from the mixture. This may offer a considerable computational advantage in complex acoustic scenarios with many sources.

Audio control strategy

In our closed-loop experiments, we use the BCI to apply an audio gain to the attended speaker. Successful development of the technology is likely to require development of refined techniques to render the acoustic feedback in ways that are intuitive and comfortable to the user. Re-rendering the acoustic scene with spatially distributed sources may be critical for real-life usability. It should also be noted that amplification is not the only possible audio control strategy. Hearing aids can, for instance, benefit greatly from selectively applying different dynamic-range compression to the target and background signals (Overby et al., 2023). Selectively applying fast-acting compression to the attended foreground signal and slow compression to the ignored background avoids the problem of amplifying noise components along with the target speech, a recurrent problem faced by hearing aids today (May et al., 2020).

Other uses

Our BCI is presented as a technology for hearing restoration, but other uses can be envisioned. Closed-loop neurofeedback training of attention has been considered as a tool to treat attention deficit disorders, i.e., helping patients sustain focus over longer periods of time (Monastra et al., 2006; Arns et al., 2009). Closed-loop imaging designs have been proposed in which people monitor their own attention, for instance, via specially designed visual stimuli that change with an attention-driven neurofeedback (Debettencourt et al., 2015). Our closed-loop system decodes attention during natural listening and modifies the attended speech stimulus itself. This has the advantage of engaging a natural attention process where the 'reward' of successfully attending is inherent in a natural behavior (listening becomes easier when attending). Task difficulty can easily be titrated (e.g., via the amount of gain applied) making the system useable in attention training programs.

5. METHODS

EEG acquisition. Two different EEG acquisition systems were used. For demonstrations of the real-time platform with microphone inputs we used mBrainTrain's mobile 24-channel SMARTING system (mBrainTrain, Belgrade, Serbia) with a wireless DC amplifier attached to the back of the EEG cap (EasyCap, Hersching, Germany). For experiments with hearing-impaired users and with audio-visual stimulation, we used a 64-channel Biosemi ActiveTwo system.

EEG training data. For closed-loop experiments, offline EEG data for training the system were first obtained from each subject. Subjects were presented with two-talker mixtures (audiobooks, one male, one female) for 30-40 mins. The training session was split into 50 sec long trials and subjects were instructed to attend to one of the two speakers in each trial. The location of the attended target (attend left vs right) was

counterbalanced across trials to avoid spatial bias. In our experiment with audiovisual speech, 10 subjects received additional visual feedback of the two speakers during training, while 10 other subjects were trained using only audio. In separate experiments, we noted that listening to a single speech stream during training data acquisition resulted in similar closed-loop performance. The training data was processed offline with the same processing pipeline as described for the online system below.

Online Implementation. The real-time decoder was implemented in Python assembled as processing blocks (see Fig. 3) using the OpenVibe graphical BCI software platform for real-time code integration (Renard et al., 2010). EEG and audio signals were broadcasted to an OpenViBE acquisition server via labstreaminglayer (Kothe et al., 2014), a library for networking and routing of time-series data. To synchronize the data streams, trigger pulses generated by an Arduino were sent along both paths (EEG and audio), and a receiver module in OpenViBE was then used to quantify and compensate for any delay between the streams. i.e., to maintain time-synchronization between data streams.

Online processing audio. To extract audio envelopes in the online system, each separate audio stream was squared, low-pass filtered at 20 Hz (4th order Butterworth), downsampled to 64 Hz via decimation, and raised to the power of 0.3. The envelopes were further passed through a dyadic filterbank implemented as series of low-pass filters followed by a first order differentiator (de Cheveigné et al., 2018). The low-pass filtering was performed by convolving the signals with square windows of lengths ranging from 1/32 to 1 sec in 10 steps on a logarithmic scale. Differentiation of the low-passed envelopes act as a high-pass filter, effectively creating a filter-bank of bandpass filters in the modulation domain. The audio envelopes were then delayed to account for the difference between stimulus and response by applying a constant shift estimated from the training data. Next, the data were whitened (in the multichannel filterbank domain) and PCA-transformed to reduce dimensionality with transforms determined by cross-validation on the training data. Finally, the multichannel envelopes were multiplied by the CCA weights obtained on the training data.

Online processing EEG. The EEG data were low-pass filtered at 20 Hz (4th order Butterworth) and down-sampled to 64 Hz to match the audio sampling rate. A high-pass filter at 0.1 Hz (2nd order Butterworth) was used to filter out low-frequency drifts, and a common average reference was applied. The data were further filtered to reduce eye-blink activity using a spatial filter obtained on the training data (Wong et al., 2018). The EEG were passed through a multichannel filtering operation identical to the one applied to the audio envelopes. The data were then whitened (jointly in the spatial and multichannel filter domain) and truncated using PCA, similarly as for the audio data. Finally, the data were transformed using the CCA weights obtained on the training data.

Online audio-EEG classification. The temporally aligned CCA-transformed EEG and audio envelopes were concatenated and the correlation between corresponding CC pairs were computed within a continuously updated decoding window. Decoding windows of 8 sec. were used in experiment I and windows of 6.5 sec. were used in experiment II. The correlation scores were z-scored and input to a linear SVM classifier trained on the training data. The classification output was converted to class probabilities for each speech stream being attended based on the training data (Platt et al., 1999).

Audio gain control. The probability outputs of the classifier was used to control the gain balance between speech streams by attenuating ignored streams in the acoustic feedback to the listener. The baseline level of the speech mixture (without BCI feedback) was set to 60 dB HL, and BCI feedback could result in a maximum gain difference between speakers of 10 dB SPL. For hearing-impaired listeners, the baseline level amplified based on their audiometric thresholds to compensate to reduced audibility.

Different gain control functions were investigated. In experiment I, classification probabilities were mapped to a time-varying gain control signal G_t using a non-linearity corresponding to a modified Weibull cumulative distribution function:

$$G_t = \begin{cases} 1 - e^{-(\frac{2P_t - 1}{\lambda})^k} & \text{for } P_t \geq 0.5 \\ e^{-(\frac{1-2P_t}{\lambda})^k} - 1 & \text{for } P_t < 0.5 \end{cases} \tag{1}$$

where P_t is the time-varying probability outputs of the classifier. In the current two-talker scenario, speaker A is classified as attended when $P_t > 0.5$ and speaker B is classified as attended when $P_t < 0.5$ (and $P_t = 0.5$ indicates the chance-level likelihood). The parameters k and λ determines the shape and gain of the Weibull non-linearity (see Fig. 7). The gain control signal G_t was scaled to the real-valued interval $[-1, 1]$ (dividing by $1 - e^{(1/\lambda)^k}$). $G_t = -1$ then indicates that full gain should be applied to speaker A while speaker B should be attenuated by the maximum gain difference. $G_t = 1$ indicates full gain applied to speaker B while speaker A is attenuated. 0 indicates that the system should not attenuate either speaker. The positive and negative values of the gain control signal were each mapped to dB and used to control the audio levels of the two speech streams.

State tracker. In experiment II, the relative gains were controlled via a state tracker. The state tracker implemented a strategy to control the gain based on the state of the attention decoding system. The likelihood probability of the classifier P_t now controls the *rate* of change in gain. This allows the state to remain stable when the attention classifier yields a chance-level likelihood (i.e., likelihoods toward 0.5), making less prone to detrimental sudden shifts in gain.

The state at time step t is computed as:

$$S_t = S_{t-1} + C\frac{2}{K \cdot F_s} ITR_t - \text{sign}(S_{t-1})\,|S_{t-1}|^{\alpha} \tag{2}$$

where C is the attention state corresponding to the classified speaker:

$$C = \begin{cases} 1 & \text{for } P_t \geq 0.5 \\ -1 & \text{for } P_t < 0.5 \end{cases} \tag{3}$$

with P_t again being the classifier likelihood. ITR_t is the estimated information transfer rate (Wolpaw et al., 1998) in bits per classification window of length K seconds and F_s is the number of classifications per second, here set to 4 Hz. In essence, in absence of the $\text{sign}(S_{t-1})\,|S_{t-1}|^{\alpha}$ term, $\frac{2}{K \cdot F_s} ITR$ allows the state to switch from -1 to 1 (or vice versa) in one second if the information transfer rate is equivalent to 1 bit/s throughout the entire second. The information transfer rate (in bits per classification window) is computed as:

$$ITR_t = \log_2(N) + P_t \log_2(P_t) + (1 - P_t) \log_2(\frac{1 - P_t}{N - 1}) \tag{4}$$

where N is the number of classes (2 in this case). The exponent α in eq. 2 acts as a hyperparameter that control the amplitude of state changes.

Fig.7 illustrates the behavior the different gain control strategies via simulations. The outputs of the state tracker S_t (middle row) and the instantaneous gain function G_t (bottom row) are shown with ideal classifications P_t of an instantaneous attention switch half-way through the trial (left column) or a sinusoidally variation in attention (middle column). The right column illustrates noisy classification of an ideal switch (black) with periods of random fluctuation in attention. The integration time window of the state tracker can be increased to make the state tracker less prone to random fluctuations or noisy classifications but at the expense of slower switch in gain upon actual switches in attention.

Participants. In the demonstration with hearing-impaired listeners, 3 young normal-hearing and 3 older hearing-impaired listeners participated. The hearing-impaired subjects participated without their hearing aids on. Instead, a linear frequency-specific amplification was applied based on their audiometric thresholds to restore baseline audibility. In the audiovisual speech experiment, 20 young normal-hearing listeners participated. Experiments were conducted in accordance with protocol H-16.036.391 approved by the Science Ethics Committee for the Capital region of Denmark.

Closed-loop test procedures. In test sessions with the real-time system, 24 trials (16 closed-loop, 8 open-loop) of 90 sec. were collected per participant in the demonstrations with hearing-impaired listeners, and 36 trials (24 closed-loop, 12 open-loop) of 50 sec. were collected in the audiovisual speech experiment. Pre-recorded two-talker speech mixtures (one male, one female) were presented. Subjects were prompted to attend to one speaker at the beginning of the trial, and to switch attention to the other speaker half-way through the trial.

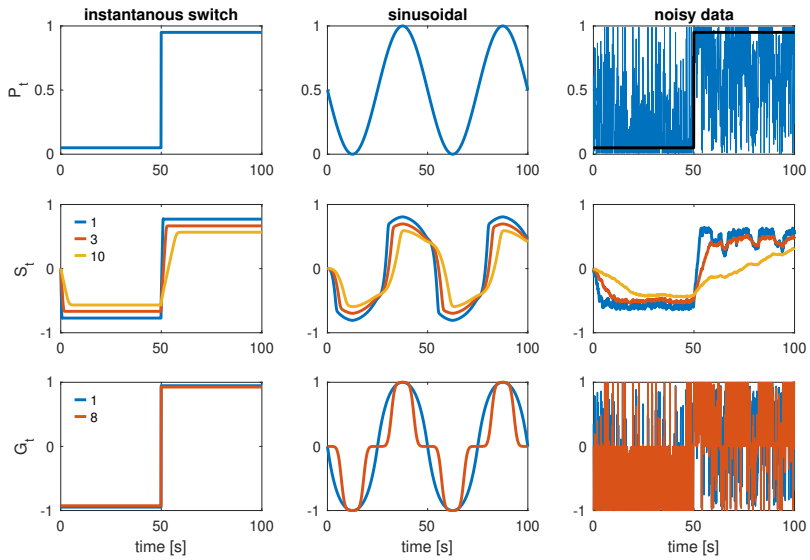

Figure 7: Simulations illustrating the behavior of the different audio gain control strategies. Top row: simulated attention classification probabilities P_t. Middle row: outputs of the state tracker S_t for different durations of the decoding window (in secs). Bottom row: outputs of the instantaneous gain control function G_t for different values of the shape parameter k. Three different simulations are considered here. The first simulation (left column) assumes an ideal scenario where the classifier correctly identifies an attention switch after 50 s. In the second simulation (middle column) classification probabilities fluctuates sinusoidally. The third simulation (right column) assumes a target attention switch at 50 s but the classification is noisy. As can be seen, noisy classification results in a fluctuating gain for the instaneous mapping function G_t (bottom) but a smooth gain control for the state tracker (middle).

Code. Code for the decoding pipeline is available via https://cocoha.org/the-cocoha-matlab-toolbox/

Acknowledgments

This work was supported by the EU H2020-ICT grant 644732 (COCOHA), and grants ANR-10-LABX-0087 IEC and ANR-10-IDEX-0001-02 PSL. It draws on work performed at the 2016 Telluride Neuromorphic Engineering workshop. Alain de Cheveigné is funded in part by the FrontCog grant ANR-17-EURE-0017. Søren A Fuglsang is funded in part by the grant NNF17OC0027872 from the Novo Nordisk Foundation. Søren A Fuglsang and Jens Hjortkjær are funded in part by the Center for Auditory Neuroscience supported by the William Demant Foundation.

References

Akram, S., Presacco, A., Simon, J. Z., Shamma, S. A., and Babadi, B. (2016). Robust decoding of selective auditory attention from meg in a competing-speaker environment

via state-space modeling. *NeuroImage*, 124:906–917.

Arns, M., De Ridder, S., Strehl, U., Breteler, M., and Coenen, A. (2009). Efficacy of neurofeedback treatment in adhd: the effects on inattention, impulsivity and hyperactivity: a meta-analysis. *Clinical EEG and neuroscience*, 40(3):180–189.

Aroudi, A., Mirkovic, B., De Vos, M., and Doclo, S. (2016). Auditory attention decoding with eeg recordings using noisy acoustic reference signals. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 694–698. IEEE.

Aroudi, A., Mirkovic, B., De Vos, M., and Doclo, S. (2019). Impact of different acoustic components on eeg-based auditory attention decoding in noisy and reverberant conditions. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(4):652–663.

Bleichner, M. G. and Debener, S. (2017). Concealed, unobtrusive ear-centered eeg acquisition: ceegrids for transparent eeg. *Frontiers in human neuroscience*, 11:163.

Bronkhorst, A. W. (2000). The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acustica united with Acustica*, 86(1):117–128.

Ceolini, E., Hjortkjær, J., Wong, D. D., O'Sullivan, J., Raghavan, V. S., Herrero, J., Mehta, A. D., Liu, S.-C., and Mesgarani, N. (2020a). Brain-informed speech separation (biss) for enhancement of target speaker in multitalker speech perception. *NeuroImage*, 223:117282.

Ceolini, E., Kiselev, I., and Liu, S.-C. (2020b). Evaluating multi-channel multi-device speech separation algorithms in the wild: a hardware-software solution. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1428–1439.

Ceolini, E. and Liu, S.-C. (2019). Combining deep neural networks and beamforming for real-time multi-channel speech enhancement using a wireless acoustic sensor network. In *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America*, 25(5):975–979.

Chun, M. M., Golomb, J. D., and Turk-Browne, N. B. (2011). A taxonomy of external and internal attention. *Annual review of psychology*, 62:73–101.

Ciccarelli, G., Nolan, M., Perricone, J., Calamia, P. T., Haro, S., O'Sullivan, J., Mesgarani, N., Quatieri, T. F., and Smalt, C. J. (2019). Comparison of two-talker attention decoding from eeg with nonlinear neural networks and linear methods. *Scientific reports*, 9(1):1–10.

Das, N., Bertrand, A., and Francart, T. (2018). Eeg-based auditory attention detection: boundary conditions for background noise and speaker positions. *Journal of neural engineering*, 15(6):066017.

Dau, T., Kollmeier, B., and Kohlrausch, A. (1997). Modeling auditory processing of amplitude modulation. i. detection and masking with narrow-band carriers. *The Journal of the Acoustical Society of America*, 102(5):2892–2905.

Dau, T., Maercher Roersted, J., Fuglsang, S., and Hjortkjær, J. (2018). Towards cognitive control of hearing instruments using eeg measures of selective attention. *The Journal of the Acoustical Society of America*, 143(3):1744–1744.

de Cheveigné, A., Slaney, M., Fuglsang, S. A., and Hjortkjaer, J. (2021). Auditory stimulus-response modeling with a match-mismatch task. *Journal of Neural Engineering*, 18(4):046040.

de Cheveigné, A., Wong, D. D. E., Di Liberto, G. M., Hjortkjaer, J., Slaney, M., and Lalor, E. (2018). Decoding the auditory brain with canonical component analysis. *NeuroImage*, 172:206–216.

de Taillez, T., Kollmeier, B., and Meyer, B. T. (2020). Machine learning for decoding listeners' attention from electroencephalography evoked by continuous speech. *European Journal of Neuroscience*, 51(5):1234–1241.

Debettencourt, M. T., Cohen, J. D., Lee, R. F., Norman, K. A., and Turk-Browne, N. B. (2015). Closed-loop training of attention with real-time brain imaging. *Nature neuroscience*, 18(3):470–475.

Ding, N. and Simon, J. Z. (2012a). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences*, 109(29):11854–11859.

Ding, N. and Simon, J. Z. (2012b). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of neurophysiology*, 107(1):78–89.

Esterman, M., Noonan, S. K., Rosenberg, M., and DeGutis, J. (2013). In the zone or zoning out? tracking behavioral and neural fluctuations during sustained attention. *Cerebral cortex*, 23(11):2712–2723.

Fiedler, L., Wöstmann, M., Graversen, C., Brandmeyer, A., Lunner, T., and Obleser, J. (2017). Single-channel in-ear-eeg detects the focus of auditory attention to concurrent tone streams and mixed speech. *Journal of neural engineering*, 14(3):036020.

Fuglsang, S. A., Dau, T., and Hjortkjær, J. (2017). Noise-robust cortical tracking of attended speech in real-world acoustic scenes. *Neuroimage*, 156:435–444.

Fuglsang, S. A., Märcher-Rørsted, J., Dau, T., and Hjortkjær, J. (2020). Effects of sensorineural hearing loss on cortical synchronization to competing speech during selective attention. *Journal of Neuroscience*, 40(12):2562–2572.

Geirnaert, S., Francart, T., and Bertrand, A. (2020). Fast eeg-based decoding of the directional focus of auditory attention using common spatial patterns. *IEEE Transactions on Biomedical Engineering*, 68(5):1557–1568.

Geirnaert, S., Vandecappelle, S., Alickovic, E., de Cheveigne, A., Lalor, E., Meyer, B. T., Miran, S., Francart, T., and Bertrand, A. (2021). Electroencephalography-

based auditory attention decoding: Toward neurosteered hearing devices. *IEEE Signal Processing Magazine*, 38(4):89–102.

Golumbic, E. M. Z., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., Goodman, R. R., Emerson, R., Mehta, A. D., Simon, J. Z., et al. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". *Neuron*, 77(5):980–991.

Hosseini, M., Celotti, L., and Plourde, E. (2021). Speaker-independent speech enhancement with brain signals.

Itani, M., Chen, T., Yoshioka, T., and Gollakota, S. (2023). Creating speech zones with self-distributing acoustic swarms. *Nature Communications*, 14(1):5684.

Kelly, S., Lalor, E., Reilly, R., and Foxe, J. (2005). Independent brain computer interface control using visual spatial attention-dependent modulations of parieto-occipital alpha. In *Conference Proceedings. 2nd International IEEE EMBS Conference on Neural Engineering, 2005.*, pages 667–670. IEEE.

Kiselev, I., Ceolini, E., Wong, D., De Cheveigne, A., and Liu, S.-C. (2017). Whisper: Wirelessly synchronized distributed audio sensor platform. In *2017 IEEE 42nd Conference on Local Computer Networks Workshops (LCN Workshops)*, pages 35–43. IEEE.

Kochkin, S. (2010). Marketrak viii: Consumer satisfaction with hearing aids is slowly increasing. *The Hearing Journal*, 63(1):19–20.

Kothe, C. et al. (2014). Lab streaming layer (lsl).

Lesica, N. A. (2018). Why do hearing aids fail to restore normal auditory perception? *Trends in neurosciences*, 41(4):174–185.

Looney, D., Kidmose, P., Park, C., Ungstrup, M., Rank, M. L., Rosenkranz, K., and Mandic, D. P. (2012). The in-the-ear recording concept: User-centered and wearable brain monitoring. *IEEE pulse*, 3(6):32–42.

Luo, Y. and Mesgarani, N. (2018). Tasnet: time-domain audio separation network for real-time, single-channel speech separation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 696–700. IEEE.

May, T., Kowalewski, B., and Dau, T. (2020). Scene-aware dynamic-range compression in hearing aids. In *The Technology of Binaural Understanding*, pages 763–799. Springer.

Mesgarani, N. and Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, 485(7397):233–236.

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

Mirkovic, B., Debener, S., Jaeger, M., and De Vos, M. (2015). Decoding the attended speech stream with multi-channel eeg: implications for online, daily-life applications. *Journal of neural engineering*, 12(4):046007.

Monastra, V. J., Lynn, S., Linden, M., Lubar, J. F., Gruzelier, J., and La Vaque, T. J. (2006). Electroencephalographic biofeedback in the treatment of attention-deficit/hyperactivity disorder. *Journal of neurotherapy*, 9(4):5–34.

Musk, E. et al. (2019). An integrated brain-machine interface platform with thousands of channels. *Journal of medical Internet research*, 21(10):e16194.

O'Sullivan, J. A., Crosse, M. J., Power, A. J., and Lalor, E. C. (2013). The effects of attention and visual input on the representation of natural speech in eeg. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2800–2803. IEEE.

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., Slaney, M., Shamma, S. A., and Lalor, E. C. (2015). Attentional selection in a cocktail party environment can be decoded from single-trial eeg. *Cerebral cortex*, 25(7):1697–1706.

Overby, N., Dau, T., and May, T. (2023). A data-driven distance metric for evaluating the effects of dynamic range compression in adverse conditions. *Trends in Hearing*, 27:23312165231192302.

O'Sullivan, A. E., Crosse, M. J., Di Liberto, G. M., and Lalor, E. C. (2017a). Visual cortical entrainment to motion and categorical speech features during silent lipreading. *Frontiers in human neuroscience*, 10:679.

O'Sullivan, J., Chen, Z., Herrero, J., McKhann, G. M., Sheth, S. A., Mehta, A. D., and Mesgarani, N. (2017b). Neural decoding of attentional selection in multi-speaker environments without access to clean sources. *Journal of neural engineering*, 14(5):056001.

Pedersen, N. F., Dau, T., Hansen, L. K., and Hjortkjær, J. (2022). Modulation transfer functions for audiovisual speech. *PLOS Computational Biology*, 18(7):e1010273.

Platt, J. et al. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.

Popov, T., Gips, B., Weisz, N., and Jensen, O. (2023). Brain areas associated with visual spatial attention display topographic organization during auditory spatial attention. *Cerebral Cortex*, 33(7):3478–3489.

Power, A. J., Foxe, J. J., Forde, E.-J., Reilly, R. B., and Lalor, E. C. (2012). At what time is the cocktail party? a late locus of selective attention to natural speech. *European Journal of Neuroscience*, 35(9):1497–1503.

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., et al. (2021). Speechbrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*.

Renard, Y., Lotte, F., Gibert, G., Congedo, M., Maby, E., Delannoy, V., Bertrand, O., and Lécuyer, A. (2010). Openvibe: An open-source software platform to design,

test, and use brain–computer interfaces in real and virtual environments. *Presence*, 19(1):35–53.

Slaney, M., Lyon, R. F., Garcia, R., Kemler, B., Gnegy, C., Wilson, K., Kanevsky, D., Savla, S., and Cerf, V. G. (2020). Auditory measures for the next billion users. *Ear and hearing*, 41:131S–139S.

Smallwood, J. and Schooler, J. W. (2015). The science of mind wandering: empirically navigating the stream of consciousness. *Annual review of psychology*, pages 487–518.

Steinmetz, N. A., Aydin, C., Lebedeva, A., Okun, M., Pachitariu, M., Bauza, M., Beau, M., Bhagat, J., Böhm, C., Broux, M., et al. (2021). Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. *Science*, 372(6539):eabf4588.

Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M., and Zhong, J. (2021). Attention is all you need in speech separation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 21–25. IEEE.

Thornton, M., Mandic, D., and Reichenbach, T. (2022). Robust decoding of the speech envelope from eeg recordings through deep neural networks. *Journal of Neural Engineering*.

Van Eyndhoven, S., Francart, T., and Bertrand, A. (2016). Eeg-informed attended speaker extraction from recorded speech mixtures with application in neuro-steered hearing prostheses. *IEEE Transactions on Biomedical Engineering*, 64(5):1045–1056.

Vandecappelle, S., Deckers, L., Das, N., Ansari, A. H., Bertrand, A., and Francart, T. (2021). Eeg-based detection of the locus of auditory attention with convolutional neural networks. *Elife*, 10:e56481.

Viana, P. F., Duun-Henriksen, J., Glasstëter, M., Dümpelmann, M., Nurse, E. S., Martins, I. P., Dumanis, S. B., Schulze-Bonhage, A., Freestone, D. R., Brinkmann, B. H., et al. (2021). 230 days of ultra long-term subcutaneous eeg: seizure cycle analysis and comparison to patient diary. *Annals of clinical and translational neurology*, 8(1):288–293.

Weissman, D. H., Roberts, K., Visscher, K., and Woldorff, M. (2006). The neural bases of momentary lapses in attention. *Nature neuroscience*, 9(7):971–978.

Wolpaw, J. R., Ramoser, H., McFarland, D. J., and Pfurtscheller, G. (1998). Eeg-based communication: improved accuracy by response verification. *IEEE transactions on Rehabilitation Engineering*, 6(3):326–333.

Wong, D. D., Fuglsang, S. A., Hjortkjær, J., Ceolini, E., Slaney, M., and De Cheveigne, A. (2018). A comparison of regularization methods in forward and backward models for auditory attention decoding. *Frontiers in neuroscience*, 12:531.

Yost, W. A. (1997). The cocktail party problem: Forty years later. *Binaural and spatial hearing in real and virtual environments*, pages 329–347.