Neural tracking to go: auditory attention decoding and saliency detection with mobile EEG

L Straetmans 1, B Holtze 1, S Debener 1,2, M Jaeger 1,3,∗ and B Mirkovic 1

1 Neuropsychology Lab, Department of Psychology, Carl von Ossietzky University of Oldenburg, Oldenburg, Germany
2 Cluster of Excellence Hearing4all, University of Oldenburg, Oldenburg, Germany
3 Fraunhofer Institute for Digital Media Technology, IDMT, Division Hearing, Speech and Audio Technology, Oldenburg, Germany
∗ Author to whom any correspondence should be addressed.

E-mail: manuela.jaeger@uni-oldenburg.de

Keywords: mobile EEG, saliency, P3, speech envelope tracking, selective auditory attention, auditory attention decoding (AAD)

Abstract

Objective. Neuro-steered assistive technologies have been suggested to offer a major advancement in future devices like neuro-steered hearing aids. Auditory attention decoding (AAD) methods would in that case allow for identification of an attended speaker within complex auditory environments, exclusively from neural data. Yet, it is known that ever-present factors like distraction and movement are reflected in the neural signal parameters related to attention.

Approach. Thus, in the current study we applied a two-competing speaker paradigm to investigate performance of a commonly applied electroencephalography-based AAD model outside of the laboratory during leisure walking and distraction. Unique environmental sounds were added to the auditory scene and served as distractor events.

Main results. The current study shows, for the first time, that the attended speaker can be accurately decoded during natural movement. At a temporal resolution of as short as 5 s and without artifact attenuation, decoding was found to be significantly above chance level. Further, as hypothesized, we found a decrease in attention to the to-be-attended and the to-be-ignored speech stream after the occurrence of a salient event. Additionally, we demonstrate that it is possible to predict neural correlates of distraction with a computational model of auditory saliency based on acoustic features.

Significance. Taken together, our study shows that auditory attention tracking outside of the laboratory in ecologically valid conditions is feasible and a step towards the development of future neural-steered hearing aids.

1. Introduction

When we are listening to someone within a noisy environment, our auditory system allows us to follow the attended speaker despite concurrently ongoing sounds (e.g. other conversations) (Cherry 1953). Interestingly, while we are processing speech of interest, the neural signal reflects the attended speech more strongly than ignored speech (Kerlin et al 2010, Mesgarani and Chang 2012). Based on this finding, methods have been developed to decode the attended speaker within multi-speaker environments (Ding and Simon 2012, Alickovic et al 2019). This offers opportunities for promising future applications in assistive devices (Slaney et al 2020) such as neuro-steered hearing aids (Geirnart et al 2021b) or other brain–computer interfaces (Belo et al 2021), especially in complex, uncontrolled natural auditory scenes. However, so far, these methods have been tested exclusively inside the lab, under controlled conditions. It remains unknown if the attended speaker can still be decoded when the neural data has been collected in more ecologically valid situations.

In fact, it is likely that limited cognitive resources, taxed by selective auditory attention to continuous speech, are influenced by cumulative real-life factors such as distracting auditory events, or cognitive and motor processes related to unconstrained movement (Al-Yahiya et al 2011). Such factors should not be ignored if the long-term goal is to include attended speaker decoding methods into assistive technologies (Slaney et al 2020). Neurophysiological studies
show that the interference between cumulative cognitive and motor processes is reflected in significant amplitude differences of attention-related features in electroencephalography (EEG) recordings between mobile and stationary conditions (Debener et al. 2012, Ladouce et al. 2019, Reiser et al. 2020). In a mobile-EEG, auditory oddball study, Debener et al. (2012) found a significant decrease in the P3 event-related potential (ERP) response to target sounds in a walking compared to a sitting condition. ERPs may be considered as neural impulse responses to transient events. The P3 effect was later replicated by de Vos et al. (2014), among others. Recently, Reiser et al. (2020) confirmed the P3 effect and found that further ERP components (i.e. parietal P2, N2, and P3 components and frontal theta power) were affected by movement and cognitive processing as well. Several of these ERP components have previously been linked to stimulus processing and top-down auditory attention (Debener et al. 2005, Aiken and Picton 2008). Additionally, the quality of the EEG signal recorded in mobile conditions is affected by robust gait-related artifacts (Jacobsen et al. 2020) and the choice of artifact attenuation (Reis et al. 2014). These findings give rise to the question whether EEG-based attended speaker decoding works as well during movement as it does while a listener remains stationary.

It is also unclear how sensitive attended speaker decoding is to transient distractions caused by unexpected sounds. Natural acoustic scenes typically consist of several concurrent streams, among which some may distract the listener from focusing on a particular stream of interest (e.g. a car horn while walking along the street and enjoying a conversation with a colleague). Those attention-grabbing events may be best described by their saliency. In a study by Huang and Elhilali (2020) participants were asked to attend to a tone sequence while ignoring a simultaneously presented background scene. The background scene consisted of natural scenarios (e.g. a busy cafeteria) varying in saliency. Behavioral measures showed that attention toward the to-be-attended tone sequence dropped notably during moments of increased saliency in the background scene. Interestingly, highly salient events induced a larger decrease in the neural tracking of the to-be-attended tone sequence. Further, after highly salient events, the authors observed an increase in the neural tracking of the to-be-ignored background scene. Similar findings were reported in a recent study conducted by Holtze et al. (2021) who investigated the neural processing of ignored speech, specifically in moments of distraction. To ensure high saliency of the distracting events, the participants’ own name was embedded into the to-be-ignored speech stream. Similar to Huang and Elhilali (2020), Holtze et al. (2021) observed an increase in the neural tracking of the to-be-ignored speech shortly after presenting one’s own name therein. Interestingly, the neural tracking of the to-be-attended speech stream also increased. The latter may reflect a strong reorienting response to the to-be-attended speech stream after the alerting effect of hearing one’s name. To investigate how distracting auditory events influence selective attention to continuous speech, an objective saliency measure of the respective event is required. Therefore, Kaya and Elhilali (2014) developed a Kalman-filter-based algorithm that computes a saliency vector, estimating the saliency of an auditory scene in each moment. The algorithm has been validated successfully using behavioral measures (Huang and Elhilali 2017). Here, we will use this objective measure to quantify the saliency of the distracting sound.

The present study takes a commonly used attended speaker decoding method out of the lab into a more realistic, ecologically valid scenario by addressing two issues: first, we determined whether attended speaker decoding is possible during leisure walking. Second, we investigated how bottom-up distraction impacts the top-down driven neural impulse response to an attended speech stream. We used a well-established two-competitive speaker paradigm (O’Sullivan et al. 2014, Mirkovic et al. 2016) to investigate the dynamics of auditory attention. Participants were instructed to attend to one of two simultaneously presented, spatially separated speech streams. In a third auditory stream various natural environmental sounds served as transient salient events. Participants alternately sat on a chair or walked along an indoor route. Neurophysiological responses were recorded using mobile EEG. The first objective of this study was to compare the decoding accuracy of a representative, commonly used backward auditory attention decoding (AAD) model (O’Sullivan et al. 2014, Crosse et al. 2016) between the mobile and stationary condition. We expected the decoding accuracy to be above chance level in both conditions, yet higher in the sitting than in the walking condition. As artifact attenuation is an integral part of mobile EEG pre-processing, the influence of two different EEG artifact attenuation methods on model performance was investigated as well. The second objective was to investigate the effect of salient distractor events on the neural tracking of the to-be-attended and to-be-ignored speech stream. To investigate the neural correlates of distraction, the neural impulse response (forward AAD model, Crosse et al. 2016) of both speech streams was estimated during periods around distractor events. It is known that salient auditory events in the environment evoke a novelty P3 response associated with bottom-up attention processes (Debener et al. 2005). We hypothesized that salient events would capture the participants’ attention and as such will evoke a novelty P3 response. In addition, we predicted that salient events cause amplitude decrease in components of the neural impulse responses to the to-be-attended speech stream.
2. Materials and methods

2.1. Participants
Twenty-one participants took part in the study. Two participants were excluded due to technical difficulties during data acquisition, resulting in a total of 19 participants (4 male). The age of participants ranged between 20 and 30 years (mean = 24.2, standard deviation = 2.8). All participants were native German speakers, had no past or present psychological condition and reported normal hearing capacities. To account for slight variances in hearing capabilities loudness of the auditory stimuli was set to a comfortable level for the participant. Participants signed an informed consent prior to the study and were paid for their participation. The study was approved by the University of Oldenburg ethics committee. Due to the COVID-19 hygiene protocol, participants and experimenter wore mouth and nose protection masks during the entire experiment.

2.2. Paradigm
Similar to previous studies (O’Sullivan et al. 2014, Mirkovic et al. 2016, Petersen et al. 2017) a two-competitive speaker paradigm was employed. Participants were instructed to attend to one continuous speech stream while ignoring a second, concurrently presented speech stream. Speech stimuli were presented in six approximately five-minute long blocks. During three of these blocks, participants were instructed to walk along a pre-determined route. Participants were instructed to walk in a comfortable speed and no further instructions concerning speed or walking stability was given. The aim was to keep the experiment as natural as possible. During the remaining three blocks participants sat on a chair in front of a white wall (figure 1). Walking and sitting blocks alternated within participants and the starting condition alternated between participants. To ensure participants attended to one stream they had to answer binary choice questions concerning the content of the previously attended story during short breaks between the blocks. The measurement took place in a public cafeteria, providing a large area with tables and chairs around which the participants had to navigate. All measurements took place in the morning when the cafeteria was not open to the public. Thus, while the environment was still natural, background sounds were reduced to a minimum. Most remaining sounds were attenuated due to insert earphones. During all stages of the experiment participants carried a smartphone with applications necessary for neural recording and audio presentation.

2.3. Stimuli
2.3.1. Speech
Six to-be-attended speech stimuli of approximately 5 min duration were used. Each stimulus consisted of a coherent short story of an audio book (Kling 2009, narrated in German by a male speaker. The same stimuli were previously used in Puschmann et al. (2019). The audio was sampled at 48 kHz with 16-bit resolution. For the to-be-ignored speech stimuli, story snippets of the same audio book (each approx. 60 s long and unused in the to-be-attended condition), were concatenated to form six 5 min long stimuli. Each to-be-ignored stimulus was matched with one to-be-attended stimulus and saved into a stereo file. The to-be-attended and to-be-ignored speech streams were spatially separated (see section 2.3.3 for details). The side of the to-be-attended stream alternated across participants. Stimuli were presented to the participants in random order.

2.3.2. Salient events
To investigate the impact of salient events (N = 60), ten different environmental sounds, each lasting up to 2 s, were added to each of the six stereo files (inter-sound intervals between 21 and 54 s, figure 1). No salient event was presented within the first or the last 20 s of a story. The chosen sounds represent a wide dynamical range of acoustic features. While some sounds started gradually in a low, non-audible intensity and a slow build-up, others appear suddenly with higher intensity. The diversity in salient events was chosen to investigate bottom-up attention effects to a variety of qualitatively different, environmental sounds. Content-wise the sounds can be grouped into five categories: animals, traffic, construction, bells, indoor-noise (e.g. banging door).

2.3.3. Stimulus presentation
Audio stimuli were presented via in-ear headphones (Sennheiser, CX300II, Sennheiser electronic GmbH & Co. KG, Wedemark, Germany) which the participants were required to wear during the entire experiment. The to-be-attended and to-be-ignored continuous speech streams, as well as the salient events, were mixed such that they appeared to originate from different locations. This was done to provide spatial cues which are natural to complex listening situations (e.g. social gatherings) and facilitate speech segregation (Shinn-Cunningham and Best 2008). Spatial separation was achieved by convolving the raw sounds with existing head related impulse responses (Kayser et al. 2009). The to-be-attended speech stream was transformed to an azimuth angle of −45° (left side condition) or 45° (right side condition) and a 0° elevation angle. The to-be-ignored speech channel was transformed to the opposite side (right side condition: −45° azimuth angle; left side condition: 45° azimuth angle; 0° elevation angle). The direction from which a salient event appeared was randomly spatialized in 5° steps within a 180° radius in front of the participant. Locations between 30° and 60° and −30° and −60° were left out to avoid overlap with the to-be-attended and to-be-ignored speech streams.
The auditory stimuli were sampled at 48,000 Hz. The presentation application for android (version 1.2.1, Neurobehavioral Systems Inc., Albany, CA, United States) was used to play audio, send event markers, and record behavioral responses in breaks between the stories.

3. Neurophysiological data recordings

EEG was recorded using a wireless 24-channel direct current (DC) amplifier (SMARTING, mBrainTrain, Belgrade, Serbia) attached to the back of the EEG cap (EasyCap GmbH, Hersching, Germany). Participants wore a headband to keep the amplifier on the head. Data from a 24 Ag/AgCl passive electrode set-up (international 10–20 system) were recorded at a sampling rate of 250 Hz. Channel Fz was used as reference electrode. The electrode sites were prepared with 70% alcohol and an abrasive electrolyte gel (Abralyt HiCl, Easycap GmbH, Germany). During data acquisition impedances were kept below 10 kΩ. The EEG signal was wirelessly transmitted to a smartphone ('Sony Xperia', model: C6903; OS: Android 5.1.1) via bluetooth and presentation markers were synchronized and recorded using the lab streaming layer protocol (Christian Kothe, Swartz Center for Computational Neuroscience, Institute for Neural Computation, University of California, San Diego, USA, https://github.com/chkothe/labstreaminglayer) integrated in the Smarting application (version 1.6.0, mBrainTrain, 2016, Fully Mobile EEG Devices) and saved into an .xdf file.

3.1. Objective saliency estimation

The degree of saliency for each time instance of the auditory scene that was presented to participants was estimated using an objective algorithm inspired by a Kalman model of auditory saliency initially described by Kaya and Elhilali (2014). This model gives saliency estimation based exclusively on features of the auditory scene. Specifically, this saliency model was chosen due to the fact that it emphasizes the role of time in auditory processing. Whereas in visual domain previously observed visual scenes do not have a large influence on determining the saliency of objects in the current scene, in auditory scenes previously perceived sounds play an important role in perception of subsequent auditory events (Kaya and Elhilali 2017), e.g. a short, loud tone is not salient if it is continuously repeated. Kalman filter mimics this processing by estimating the current value of an audio feature based on its statistical properties over time, followed by calculating the least square error between the predicted and the actual value of the signal, resulting in 1d array with salient event probability spikes (Kaya and Elhilali 2014).

In the original computational model by Kaya and Elhilali (2014), five main auditory features (intensity, pitch, spectrogram, bandwidth and temporal modulation) were extracted from short auditory scenes that were presented to participants. The participant’s task was to determine whether each presented scene contained a salient event. Based on the results achieved with speech-related auditory scenes in Kaya and Elhilali (2014) as well as on our own pilot data, we have based our analysis on three of these features that predict the occurrence of salient auditory events reliably: sound intensity, pitch and sound spectrogram based on the model of the cochlear processing (Shamma and Klein 2000). To extract the intensity feature we followed the procedure from Kaya and Elhilali (2014), namely, the absolute of the Hilbert transform of the auditory scene was obtained and then filtered using the Butterworth filter of 6th order and 60 Hz cut-off frequency. Prior to spectrogram extraction, we downsampling the auditory scene to
16 kHz. Since features of above 16 kHz were not of interest for analyzing speech (Stevens 1998), the auditory scene was downsampled to enable faster computation. Using a continuous wavelet transform filter-bank we calculated the spectrogram of the auditory scene from 60 Hz to 8 kHz at 18 voices per octave. Following the procedure of Shamma and Klein (2000), that mimics simplified early auditory processing, the first temporal derivative of each spectrogram channel was performed, followed by half-wave rectification on each spectrogram channel, first derivative over spectrogram channels at each time point (spectral sharpening) and temporal sharpening where only positive local peaks of the spectrogram channels were preserved. Instead of augmenting the feature space with all channels of the spectrogram as in Kaya and Elhilali (2014), we opted to have two features, one representing low frequencies and the other representing high frequency channels of the spectrogram. The low and high frequency channels were separated by averaging over first half and over second half of frequency channels of the spectrogram. Additional to the procedure in Kaya and Elhilali (2014), all features were then downsampled to 250 Hz to correspond to the sampling rate of the recorded EEG. Pitch feature was extracted from the spectrogram, following the template matching approach of Shamma and Klein (2000). Pitch templates were obtained by averaging the pitch estimates over 10 s intervals of the auditory scene. Individual pitch estimates were correlated to the corresponding pitch template and if the correlation exceeded 1.5 standard deviations than the pitch estimate was retained, otherwise it was discarded.

The Kalman model was implemented in Matlab using custom scripts replicating the original procedure of Kaya and Elhilali (2014), with the exception of empirically established values of Kalman filter noise parameters, which were adapted to our pilot data. In further analysis, for saliency estimation we will use the average over all values resulting from the Kalman filter output for the above four features. As we a-priori know at which time instances the salient events were embedded within the speech stream, an average probability spike value in a 1 s window after that time point will be considered a saliency estimate.

4. Neurophysiological data analysis

4.1. Pre-processing

Pre-processing was performed offline using EEGLAB (v2019.1, Delorme and Makeig 2004) and MATLAB (R2017b, Mathworks Inc., Marick, MA). EEG data were filtered between 1 and 40 Hz (low-pass FIR filter of order 826, high pass finite impulse response (FIR) filter of order 84; integrated into EEGLAB, version 1.6.2). Onset and offset markers of salient events ($N = 60$) were added to the EEG file. Timing of salient events was fixed within each audio stream. Control event markers ($N = 60$) were added with a minimum distance of 10 s to salient event markers. Timing of control events was otherwise chosen randomly.

Two artifact attenuation methods were compared to the data without artifact attenuation. As a supervised artifact attenuation method, an extended infomax independent component analysis (ICA) was performed (Delorme and Makeig 2004). Infomax ICA was chosen since it performed reasonably well in previous mobile EEG studies (e.g. Dehener et al 2012, de Vos et al 2014, Ladouce et al 2019, Salvadegoitia et al 2019, Reiser et al 2020). ICA was performed as follows. First data was epoched into consecutive 1 s epochs and epochs containing atypical artifacts were rejected based on probability and kurtosis (SD = 2). Second, stereotypical (e.g. eye blinks, heartbeat etc) and movement-related (e.g. muscle activity) ICA components were identified and removed by back-projecting all remaining components to the continuous, pass-band filtered data. On average 1.79 components in the sitting and 2.47 components in the walking condition were removed. In the second case we used artifact subspace reconstruction (ASR) to attenuate EEG artifacts (Mullen et al 2015; cut-off parameter = 10). ASR is an unsupervised artifact attenuation method that can be performed online. It requires less processing time and computational power compared to ICA (Chang et al 2018). ASR was calibrated using data of a 1 min sitting and standing baseline for the sitting and walking conditions, respectively. The calibration data were recorded in the beginning of the experiment. Subsequently, ASR was performed on the sitting and walking data independently. In total, AAD was performed on three different versions of the dataset (ICA-attenuated, ASR-attenuated, artifact uncorrected).

Speech envelopes of the to-be-attended and to-be-ignored speech signals were extracted following a procedure adapted from Mirkovic et al (2016). The absolute value of the Hilbert transformed data was computed and data were filtered using a 25 Hz low-pass Butterworth filter (3rd order). Then, data were down sampled to 64 Hz. The pre-processed EEG data (ICA, ASR and artifact-uncorrected) were re-referenced to the linked mastoid channels, filtered between 1 and 15 Hz (low pass FIR filter of order 221, high-pass FIR filter of order 827) and down sampled to 64 Hz. To investigate the influence of different trial lengths on AAD performance, EEG data as well as speech envelopes were epoched into consecutive trials of variable length. Trial lengths ranged from 5 (Jaeger et al 2020, Holtze et al 2021) to 60 s (O’Sullivan et al 2014, Mirkovic et al 2016) in 5 s steps. EEG data of different trial lengths were baseline corrected by subtracting the mean of each respective trial. For some trial lengths the total number of trials in two movement conditions differed across participants, depending on the duration of presented stories in sitting and walking condition. To account for this difference, the number of trials was reduced to the minimum
number of available trials in both movement conditions (min 15 (60 trials), max 191 trials (5 s trials)).

4.2. Attended speaker prediction during sitting and walking

A commonly used AAD model for inferring which of the auditory streams is the to-be-attended one, as well as for interpreting the neural impulse response to a continuous stimulus (e.g. speech), is multivariate linear regression (Alickovic et al 2019). The backward multivariate linear regression model, as implemented in the multivariate temporal response function (mTRF) toolbox (version 2.3, Crosse et al 2016), was used to predict which speech stream a participant was attending to. After separating the data into trials, the pre-processed EEG signals and speech envelopes were normalized by the respective standard deviation. For each participant, trial length, artifact attenuation and movement condition, a separate backward model ($g$) was trained by associating the to-be-attended stimulus envelope ($S$) and the concurrently recorded neural response ($R$). To select the optimal regularization parameter, we used an array of 11 values ranging $10^{-5}–10^5$. The models were trained using different time lags from 0 to 350 ms, with a 45 ms moving window and 30 ms overlap. After initial training (equation (1)), the backward models were used to obtain an estimate of the stimulus envelope based on unseen neural data (Crosse et al 2016) in a leave-one-out cross-validation procedure. Model performance was quantified as the percentage of trials in which the reconstructed stimulus envelope correlated more strongly with the original envelope of the to-be-attended than with the envelope of the to-be-ignored stimulus (decoding accuracy, Crosse et al 2016):

$$g = (R^T R + \lambda I)^{-1} R^T S. \quad (1)$$

For each model, decoding accuracies were averaged over participants. This resulted in average model performance for different trial lengths, movement conditions and artifact attenuation methods. Based on the decoding accuracy the regularization parameter and the optimal time lag were selected for the individual models.

Decoding accuracies in 60 s trials and 5 s trials were statistically evaluated with a $3 \times 2 \times 2$ repeated measures analysis of variance (ANOVA) with factors artifact attenuation (ICA, ASR, uncorrected data), movement (sitting, walking) and trial length (5, 60 s). Trials of 60 s have been investigated within the lab in previous studies and are known to yield high decoding accuracies (O’Sullivan et al 2014, Mirkovic et al 2016, Lesenfants and Francart 2020). However, with the objective to bring AAD further towards an application in assistive devices, like neuro-steered hearing aids, a higher temporal resolution is eventually needed. Based on previous work (Jaeger et al 2020) we therefore chose to investigate AAD performance outside the lab also in 5 s trials.

4.3. Neural correlates of distraction

4.3.1. P3 ERP component to salient events

To analyze the neural response to a salient event, we looked at the novelty P3 in an ERP analysis. ERPs describe time and phase-locked changes in potential evoked by a certain event (Congedo 2018). ICA-attenuated EEG data were re-referenced to linked mastoid channels and low-pass filtered at 10 Hz (FIR filter of order 331). Data were epoched between $-500$ and $1500$ ms relative to the salient and control event onsets and then baseline corrected from $-500$ to 0 ms. This resulted in 60 salient and 60 control event epochs per participant. A latency window of interest was defined between 200 and 430 ms based on the morphology of the grand average ERP at midline channel Cz, where the P3 component is expressed most clearly (Debener et al 2005, Polich 2007). ERPs were computed for sitting and walking conditions separately. In both conditions, the mean amplitude across the window of interest at channel Cz was calculated for each participant. Subsequently, the difference between salient and control events in the two movement conditions was tested in a $2 \times 2$ repeated measures ANOVA with factors saliency (salient, control) and movement (sitting, walking).

Additionally, the relationship between objective auditory saliency, the evoked novelty P3 amplitudes and the change in neural impulse response after individual events was analyzed. To this end, the peak amplitude at channel Cz after every sound event was calculated and averaged across participants, resulting in one novelty P3 amplitude value for each event. The latency window (250–700 ms) was chosen to be wider than in the previous analyses due to variation in response latency to different salient events. Next, the objective saliency estimates, change in neural impulse response and event specific novelty P3 amplitude was correlated for individual events to explore the relationship between saliency estimated based on acoustic features and saliency as indicated by the neural response.

4.3.2. Neural impulse response in proximity to salient events

A forward model was calculated to estimate neuro-physiological processes driving the attended speaker prediction. The forward model (equation (2)) assumes the neural response ($R$) to be a convolution of time-shifted properties of the attended stimulus ($S$) and a neural impulse response ($w$):

$$w = (S^T S + \lambda I)^{-1} S^T R. \quad (2)$$

As we expected the response to the salient events to be transient, we increased the temporal resolution to 5 s intervals—the shortest trial length...
that was found to result in above chance level decoding accuracies (Jaeger et al 2020). For model parameter validation, data of all participants and both movement conditions were concatenated. The model was trained using time lags from −150 to 450 ms between the EEG and the speech envelope. The regularization parameter (λ) was estimated for the to-be-attended condition, using the mTRF toolbox’s mtrfrossval function. The parameter maximizing the correlation between the predicted and the original neural signal was chosen for further analysis (Crosse et al 2016). For each participant, the neural impulse response was obtained for the sitting and walking condition. The global field power (GFP) of the neural impulse responses across channels was calculated using std in Matlab for the to-be-attended and to-be-ignored stream in the sitting and walking condition. GFP was used here as a robust, reference-independent measure of the magnitude of the neural impulse response across channels (Murray et al 2008).

A 2 × 2 repeated measures ANOVA with factors attention (to-be-attended, to-be-ignored) and movement (sitting, walking) was used to compare the effect of attention and movement in the average GFP between 100 and 350 ms. The time range was chosen because the difference between the neural impulse response of the to-be-attended and the to-be-ignored condition was highest within these latencies (figure 4(A)).

Next, a second forward model was trained using 5 s trial windows before and after a distractor to capture the impact of bottom-up distraction immediately after the occurrence of a salient event. The mean GFP corresponding to three 5 s long intervals before and three 5 s long intervals after the salient event were separately calculated for all available conditions. A 2 × 2 × 2 repeated measures ANOVA with factors attention (to-be-attended, to-be-ignored), movement (sitting, walking) and time (before, after) was used to compare a distraction effect in GFP values.

Lastly it was analyzed whether the novelty P3 amplitude at channel Cz was associated with the magnitude of the distraction effect in the neural response. For every participant, the average GFP after a salient event was subtracted from the average GFP before a salient event, resulting in a difference value for the to-be-attended and to-be-ignored condition. Both difference values were statistically correlated with the participants’ average novelty P3 amplitude, respectively.

5. Results

The 88.6% of the content questions were answered correctly supporting the assumption that participants understood the instruction and paid attention to the to-be-attended speech stimuli throughout the experiment.

5.1. Attended speaker prediction during sitting and walking

The objective was to explore how AAD is influenced by factors like movement, artifact attenuation method as well as trial length. First, the optimal time lag interval for speech decoding was identified between 165 and 210 ms as in this interval the correlation between the predicted and the to-be-attended speech envelope was highest. Average decoding accuracy was above chance level in both movement conditions, for all artifact attenuation methods as well as for all trial lengths (figures 2(B)–(D)).

A 3 × 2 × 2 repeated measures ANOVA with factors artifact attenuation (no attenuation, ASR, ICA), movement (sit, walk) and trial length (5, 60 s) was performed. Results revealed significant main effects for artifact attenuation (F(1,12,21) = 14.71, p = 0.001, ηp² = 0.45), movement (F(1,18) = 20.1, p < 10⁻⁴, ηp² = 0.53) and trial length (F(1,18) = 198.5, p < 10⁻¹², ηp² = 0.92).

The interaction between artifact attenuation and trial length (F(1,2,22) = 6.05, p = 0.02, ηp² = 0.25) and movement and trial length (F(1,18) = 4.43, p = 0.049, ηp² = 0.2) was also significant. Follow-up paired sample t-tests revealed significantly lower decoding accuracies in 5 s compared to 60 s long trials within all three artifact attenuation factors (no attenuation: t = −11.9, p < 10⁻¹⁰; ASR: t = −7.6, p < 10⁻⁷; ICA: t = −16.1, p < 10⁻¹²). Within 60 s trials, decoding accuracies were significantly lower in ASR-attenuated compared to artifact-uncorrected data (t = −3.55, p = 0.002) and ICA-attenuated data (t = 3.42, p = 0.003). Decoding accuracy between artifact-uncorrected and ICA attenuated data did not differ significantly (see figure 2(A)). Within 5 s trials ICA attenuation resulted in significantly higher decoding accuracies compared to, both, ASR-attenuated (t = 4.43, p < 0.001) and artifact-uncorrected data (t = 3.04, p = 0.007). ASR attenuation resulted in significantly lower decoding accuracies than no artifact attenuation (t = −3.78, p = 0.001) (figure 2(A)).

Upon following up the significant interaction effect between movement and trial length, results confirmed that, within the sitting condition, decoding accuracy for 60 s trials was larger than for 5 s trials. A similar, although weaker effect was found for the difference in decoding accuracies between 60 and 5 s trials within the walking condition (sitting: t = 15.4, p < 10⁻¹², walking: t = 9.77, p < 10⁻⁷). Across artifact attenuation methods, decoding accuracies were significantly higher in the sitting than in the walking condition, within 60 and 5 s long trials (60 s: t = 4.01, p < 0.001; 5 s: t = 3.42, p = 0.003).

5.2. Neural correlates of distraction

5.2.1. P3 ERP component to salient events

As the first, most established neural correlate of distraction, we analyzed whether novelty P3 amplitudes differed in response to salient events compared to
control events where no distractions were presented. Furthermore, we tested whether similar effects could be identified during walking. Results of the $2 \times 2$ repeated measures ANOVA revealed a statistically significant main effect for saliency ($F(1,18) = 27.1$, $p < 0.001$, $\eta^2_p = 0.6$). There was no main effect for movement and no interaction effect. The significant main effect for saliency was followed up by a paired samples $t$-test showing significantly higher P3 amplitudes in response to salient events compared to the control events ($t = 5.2$, $p < 10^{-5}$). The resulting novelty P3 morphology consisted of two peaks, which were more pronounced in the walking condition. When looking into the topographical distribution the early peaks occurred on average over frontal-cerebral areas whereas at later latencies peaks were observed over more parietal areas (figure 3).

Next, the event-specific novelty P3 amplitude within a latency window between 250 and 700 ms after a salient event was related to the degree of saliency based on acoustic features of the event itself and the change in neural impulse response after the salient event. Correlations were tested in Matlab using the Shepherd’s pi correlation procedure, which identifies outliers by bootstrapping the Mahalanobis distance and subsequently removes them from the correlation analysis (Schwarzkopf and de Haas 2012). When testing the correlation between novelty P3 amplitude and estimated degree of saliency, four outliers were detected and excluded. Results of the correlation analysis revealed a significant positive correlation between the estimated degree of saliency and the event-specific novelty P3 (pi(56) = 0.31; $p = 0.04$). The correlation between P3 amplitude and saliency estimation around control events was not significant (pi(57) = 0.11; $p = 0.88$) (see figure 4). Correlations between the event-specific novelty P3 amplitude and the change in neural impulse response in the to-be-attended (pi(59) = 0.14, $p = 0.6$) as well as in the to-be-ignored condition (pi(57) = 0.15, $p = 0.54$) were not significant. The same was true for the change in neural impulse response and the estimated degree of saliency (to-be-attended: pi(55) = −0.11, $p = 0.88$; to-be-ignored: pi(54) = 0.14, $p = 0.59$).

Figure 2. (A) Distribution of decoding accuracy between 165 and 210 ms after speech envelope in 5 s (above) and 60 s (below) trials. Chance levels indicated by gray line. **∗∗∗ $p < 0.001$, **∗ $p < 0.01$, ∗ $p < 0.05$ (B) results for non-attenuated data. Average decoding accuracy for all trial lengths (s) in latencies from 0 to 350 ms relative to speech envelope. Thick line on x-axis represents optimal time lag between 160 and 210 ms. Sitting (green) and walking (yellow). Chance level for each trial length indicated in gray. (C) Results for ASR attenuated data. (D) Results for ICA attenuated data.
5.2.2. Neural impulse response in proximity to salient events
To investigate if the effect of transient distraction on top-down attention is reflected in the neural impulse responses right after the onset of the salient event, high temporal resolution was needed. Therefore, as the first step, we explored the morphology of the neural impulse response waveform, calculated independently of salient events, using 5 s trial length (figure 5(A)). The effect of top-down attention was measured as the difference between the neural impulse response of the to-be-attended and to-be-ignored condition. It was highest in a time lag window between 100 and 350 ms after speech envelope onset. Note that we did not consider latencies to early attention components (<100 ms). GFP was averaged in the same time lag window for sitting and walking condition before running the 2 × 2 repeated measures ANOVA with factors attention (to-be-attended, to-be-ignored) and movement (sitting, walking). Next to a significant main effect for attention ($F(1,18) = 23.96, p < 10^{-4}, \eta_p^2 = 0.57$), results revealed a significant interaction effect between movement and attention ($F(1,18) = 7.7, p = 0.013, \eta_p^2 = 0.3$, figure 5(B)). There was no main effect for movement. The interaction effect was followed up with paired sample $t$-tests. Generally, the GFP for the to-be-attended stream was significantly higher

Figure 3. (A) Grand average ERP in response to salient event at channel Cz in sitting (green) and walking (yellow) condition. Thick lines represent grand average after salient event, thin lines after control event. Shaded areas show ±2 standard error. (B) Topographies visualize spatial distribution of average activation between 250–300 ms, 300–350 ms, and 350–400 ms after event onset (from left to right) for both movement conditions.

Figure 4. Correlation between amplitude of event specific P3 component and estimated degree of saliency for the respective event.
than the GFP for the to-be-ignored stream ($t = 4.9$, $p < 10^{-4}$) (see figure 5). Within the sitting condition, results revealed significantly higher GFP values in the to-be-attended compared to the to-be-ignored condition ($t = 5.95$, $p < 10^{-5}$). A similar, although less pronounced, difference between GFP in the to-be-attended and to-be-ignored condition was found within walking condition ($t = 2.81$, $p = 0.012$; figure 5(B)). To further describe the robustness of the attention effect in both sitting and walking condition, we computed the attention effect sizes for both movement conditions (figure 5(A)). Average effect sizes in the previously mentioned time window were high in both sitting and walking condition, although higher in the sitting condition (sitting: Hedge’s $g = 1.05$; walking: Hedge’s $g = 0.65$).

Confirming the expected neural impulse response morphology and attention effect, the impact of distraction on the neural impulse response to the to-be-attended and to-be-ignored speech signal was explored using 15 s of data before and 15 s of data after each salient event. The temporal evolution of the average GFP amplitude is shown in figure 6. In general, before the onset of a salient event, the effect of attention was higher than after the salient event (15 s before: Hedges $g' = 0.72$; 10 s before: Hedges $g' = 0.65$; 5 s before: Hedges $g' = 0.82$; 5 s after: Hedges $g' < 0.01$; 10 s after: Hedges $g' = 0.12$; 15 s after: Hedges $g' = 0.11$). Following up on this result, we evaluated the period directly before (5 s) and directly after (5 s) salient events for both sitting and walking conditions (figures 6(B) and (C)). Results of the $2 \times 2 \times 2$ repeated measures ANOVA with factors attention (to-be-attended, to-be-ignored), movement (sitting, walking) and time (5 s before salient event, 5 s after salient event) revealed significant main effects for attention ($F(1,18) = 13.7$, $p = 0.002$, $\eta^2_p = 0.43$) and time ($F(1,18) = 43.32$, $p < 10^{-6}$, $\eta^2_p = 0.71$). There was no main effect for movement. There was a significant interaction effect between attention and time ($F(1,18) = 7.77$, $p = 0.012$, $\eta^2_p = 0.3$). To follow up the interaction effect, paired sample $t$-tests were performed. The results confirmed a significantly higher GFP for the to-be-attended than for the to-be-ignored stream before the occurrence of a salient event ($t = 3.5$, $p = 0.003$). This significant effect disappeared after the occurrence of the salient event ($t = 0.36$, $p = 0.72$). Further, a significantly higher GFP before compared to after a distraction was found in the to-be-attended condition ($t = 4.85$, $p < 10^{-4}$). Interestingly, GFP values of the to-be-ignored condition also decreased significantly after a salient event, although to a lesser extent ($t = 2.6$, $p = 0.018$).
Lastly the relation between participants’ average novelty P3 amplitude and the average change in neural tracking of the speech envelopes, relative to the occurrence of a salient event, was analyzed (similar to the procedure in Holtze et al 2021). One outlier was identified and removed from the data. Results showed a positive correlation between the change in the to-be-attended stream and the average novelty P3 amplitude ($pi(18) = 0.59, p = 0.02$). For the to-be-ignored stream two outliers were identified and removed. The correlation for the to-be-ignored stream was not significant ($pi(17) = 0.16, p = 1$).

6. Discussion

In the current study AAD is taken out of the lab and applied within a realistic, ecologically valid scenario. For the first time, it was shown that it is possible to decode the attended speaker based on data acquired during free walking. Additionally, we showed that distraction is reflected in decreased neural tracking of the attended speaker not only for data acquired while participants were seated but also when they were walking.

6.1. Attended speaker decoding during sitting and walking

It was explored whether a commonly applied AAD model (O’Sullivan et al 2014, Mirkovic et al 2016) achieves high accuracy, even if neural information was recorded in a non-stationary setting. With the objective to bring AAD further towards an everyday application, artifacts were attenuated in two independent ways. First, an ICA was applied. ICA was chosen as a robust and efficient artifact attenuation method (Delorme and Makeig 2004). Second, artifacts were attenuated using ASR. ASR offers an unsupervised way to attenuate artifacts in an online fashion and is therefore potentially attractive for BCI applications (Chang et al 2018). Like in previous findings (O’Sullivan et al 2014, Mirkovic et al 2016), we identified an optimal decoding window between 165 and 210 ms to result in highest prediction performance. As a trial length of 60 s was shown to yield
reliably high AAD performances (O’Sullivan et al 2014, Mirkovic et al 2016, Lesenfants and Francart 2020), we chose to use 60 s long trials to investigate AAD in walking. Our results revealed high decoding accuracies for ICA-attenuated and uncorrected data. ASR attenuation resulted in lowest accuracies, although still above chance level. This pattern was consistent for both, sitting and walking data. It is promising that we were able to decode the attended speaker, even when people are walking freely. Future listening devices may benefit from neural tracking functionality that is robust to movement (Slaney et al 2020). However, for those devices to adapt to daily life listening demands, attended speaker decoding with good temporal resolution would be required. Previous research has shown that a resolution as short as 5 s results in decoding accuracies above chance level (Jaeger et al 2020). Increasing temporal resolution further resulted in decoding accuracies below chance accuracy levels. Additionally, using a resolution of 5 s, Holtze et al (2021) were able to observe an effect of transient distraction on AAD performance. Considering these findings, we additionally investigated decoding accuracies in trials of 5 s. We found that for shorter trials ICA attenuation led to higher decoding accuracies compared to uncorrected, or ASR attenuated data. Again, this pattern was found for sitting as well as walking data.

It is conceivable that decoding accuracies in the walking conditions were influenced by remaining movement artifacts in EEG data. Therefore, we have performed an additional exploratory analysis where we correlated differences in gyroscope data between sitting and walking condition with differences in decoding accuracies between the two movement conditions. Gyroscope data were recorded from the amplifier and can be related to the gait rhythm of an individual. In 5 and 60 s trials, correlations were not significant for all three artifact attenuation methods. In a future study it would be interesting to repeat a similar analysis with accelerometer data. Accelerometer gives information about walking speed and stability which could more strongly relate to differences in decoding accuracies.

In sum, ICA improved performance of attended speaker decoding in 5 s trials, whereas in 60 s trials no difference in decoding accuracy was found between ICA-attenuated and uncorrected data. Thus, longer trials may provide sufficient data to decode the attended speaker accurately, even without any attenuation of transient artifacts. Further, in both 60 and 5 s trials, AAD based on ASR-attenuated data was outperformed by ICA-attenuation and uncorrected data. A potential explanation for these inferior decoding accuracies could be that ASR not only attenuated artifacts, but also relevant neural activity (Jacobsen et al 2020). In the current study, the ASR threshold was chosen to be quite strict (cut-off = 10) leading to a rather aggressive data attenuation approach (Chang et al 2018). Since ASR offers many benefits concerning the implementation of AAD in assistive listening devices, it would be of interest to investigate AAD based on ASR-attenuated data with different cut-off parameters to optimize the decoding outcome. Further, the duration of ASR calibration data was quite short in the current study (1 min). Since the ASR artifact attenuation method is based on artifact free calibration data it is likely that the calibration duration and its data quality influence the outcome of the later artifact attenuation procedure. A systematic investigation of these aspects would be well suited to optimize ASR performance further.

Generally, decoding accuracies in sitting condition were higher than those in walking condition. This observation supports previous findings, suggesting that cognitive and motor processes associated with walking, have an influence on available attentional resources, which was found to be reflected in neural correlates of auditory attention processes (Debener et al 2012, Ladouce et al 2019, Reiser et al 2020). In the current study this is, for the first time, found for neural impulse responses reflecting selective auditory attention to continuous speech. Although we found significant differences in decoding accuracies between sitting and walking condition, in both movement conditions model performance was found to be above chance level in 60 s as well as 5 s long trials. These results suggest that AAD can indeed be a viable method for neuro-steered assistive devices, that is, when people are in motion.

7. Auditory attention decoding (AAD)—distraction

Bottom-up stimulus processing in response to a unique environmental salient event is reflected in a significant novelty P3 response shortly after the salient event was presented. In the current study, we observed a double-peak morphology in the grand average P3 at channel Cz. When looking at the topographies an earlier central positivity approximately between 250 and 300 ms is followed by a more central-posterior positivity at latencies between 300 and 400 ms. This pattern in response to unique environmental sounds has been observed before (Debener et al 2005). While the early peak is suggested to reflect a P3 response linked to involuntary, bottom-up attention-orienting processes, the later peak is suggested to reflect a P3b response. P3b has been associated with updating the stimulus representation in working memory (Debener et al 2005, Polich 2007).

To test if salient events did in fact introduce bottom-up stimulus processing, we investigated whether the salient events would impact the neural tracking of the to-be-attended speech envelope. In line with previous findings (Hambrock and Tata 2019, Huang and Elhilali 2020) we observed a significant drop in the neural tracking of the to-be-attended
stream after the occurrence of a salient event. Different results were found by Holtze et al. (2021) who observed a significant increase in the to-be-attended stream after the occurrence of the participant's own name, which is assumed to be of high semantic relevance. One explanation for this finding could be that hearing one's name alerts the participants to refocus more quickly on the task at hand, while distraction associated with the environment (e.g. car honk) might not have the same alerting effect. Further studies are needed to investigate the effect of qualitatively different salient events. In addition, previous studies also reported an increase in the neural tracking of the to-be-attended speech stream after presenting one's own name. Contrary to these findings we did not observe a significant increase but a decrease in the neural tracking of the to-be-ignored stream. This may result from the fact that in both mentioned studies, salient events originated exclusively from the to-be-ignored stimuli which would explain a transient orientation towards the direction of occurrence. In the current study, the salient events were neither embedded in the to-be-ignored nor in the to-be-attended stream but originated from a different location. This lends support to the interpretation, that attention is transiently directed to the source of distraction, and therefore, in case of the current study, away from the to-be-attended and the to-be-ignored speech stream. Further studies are needed to test this prediction.

Lastly, the relation between individual participants’ average P3 amplitude and the change of neural tracking of the to-be-attended and to-be-ignored speech streams was investigated. After excluding one outlying participant, results revealed a significant correlation between the P3 amplitude and the drop in magnitude of impulse response in the to-be-attended condition on an average level. These results may indicate a link between subjective susceptibility to external salient sounds and the ability to follow an auditory top-down attention task despite ongoing distraction.

For neuro-steered hearing aids it would be of importance to have an accurate objective estimate of saliency within the auditory scene. Such measure would provide valuable information about possible distractions and sudden important auditory events directly from the auditory scene. We adapted a Kalman-filter based saliency estimation from Kaya and Elhilali (2014) and investigated if neural measures of distraction correlate with the objective saliency estimates. A significant correlation between across subject averaged, sound specific P3 responses and objective saliency estimates of corresponding events further validate this measure on a neural level. This is an important finding, given that previous validation studies required explicit attention to salient events (Kaya and Elhilali 2014, Huang and Elhilali 2017), which somehow contradicts the concept of saliency. We did not observe significant correlations between the objective estimates of saliency and AAD measures, which may be due to the fact that the latter measure reflects top-down regulation and may not be strongly influenced by acoustic features of the distractors used. However, given the complexity of real-life salient event estimation the moderate correlation between the novelty P3 and objective measures of saliency as reported here for the first time demonstrates a promising approach for future algorithmic improvements.

Despite using a smartphone for stimulus presentation and the same smartphone in combination with a relatively small wireless EEG amplifier for data acquisition, the present study relied on a cap for mounting electrodes on the scalp. A well-established around-the-ear EEG sensor (cEEGrid, Debener et al 2015, Bleichner and Debener 2017) has already proven to yield above chance level AAD results in a stationary experimental set-up (MirKovic et al 2016). Future studies should investigate whether unobtrusive ear-EEG systems that do not require a cap for mounting electrodes are sufficient to implement AAD in mobile scenarios.

Further, including electromyography (EMG) recordings may help to improve artifact attenuation (Gabsteiger et al 2014). Yet, in previous studies it was possible to sufficiently attenuate movement-related artifacts using ICA without additional information from EMG channels (Debener et al 2012, Salvidegotia et al 2019, Scanlon et al 2020). Since one objective of the current study was to keep the set-up compact and to avoid additional cables, we decided to leave out EMG electrodes.

In future studies the challenge will be to combine ecological validity with new algorithms, to bring hearables like neuro-steered hearing aids further towards application. So far, in most studies AAD models had access to clean speech streams. Yet, there is effort to decode the attended speakers from speech-mixtures (van Eyndhoven et al 2017) and from noisy speech signals (Aroudi et al 2016, Han et al 2019). In an approach suggested by Geirnart et al (2021a) AAD model is not based on the speech envelope of the attended speaker but uses information about the directional focus of auditory attention reflected in hemispheric differences. In doing so, the authors achieved above chance accuracy in decoding the directional focus of attention within trial lengths as short as 1 s using a subset of EEG channels located around the ear. It is of interest whether this approach achieves similarly high decoding accuracies in a non-stationary setting.
8. Conclusion

Our study demonstrates for the first time successful AAD while listeners were walking freely. Even with AAD evaluation periods as short as 5 s, predicting the attended speaker was possible. This finding holds for artifact attenuated as well as uncorrected data. Furthermore, we successfully confirmed the effect of transient salient events on sustained attention, using the neural impulse responses to attended speech. Limited attentional resources appear to be recruited by salient events, before they can be redirected to the task at hand, and this appears to be the case in stationary as well as mobile scenarios.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: doi:10.18112/openneuro.ds003801.v1.0.0.

Acknowledgments

This project was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—432063183. We would like to thank Marc Rosenkranz for his support in preparing the stimuli and during the collection of the pilot data. Further we would like to thank Alexander Scholz (speaker) for lending his voice to the auditory stimuli. This paper was partially funded by the Bundesministerium für Bildung und Forschung in the project a2i: Augmented Auditive Intelligence (16SV8596).

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORCID iDs

L Straetmans https://orcid.org/0000-0002-6727-9638
B Mirkovic https://orcid.org/0000-0001-5912-3740

References

Aiken J S and Picton T W 2008 Human cortical responses to the speech envelope Ear. Hear. 29 139–57
Al-Yahya E, Dawes H, Smith L, Dennis A, Howells K and Cockburn J 2011 Cognitive motor interference while walking: a systematic review and meta-analysis Neurosci. Biobehav. Rev. 35 715–28
Alickovic E, Lunner T, Gustafsson F and Ljung L 2019 A tutorial on auditory attention identification methods Front. Neurosci. 13 153
Aroudi A, Mirkovic B, de Vos M and Doclo S 2016 Auditors attention decoding with EEG recordings using noisy acoustic reference signals IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2016 pp 694–8
Belo J, Clerc M and Schön D 2021 EEG-based auditory attention detection and its possible future applications for passive BCI Front. Comput. Sci. 3 661178
Bleicher M G and Debener S 2017 Concealed, unobtrusive, ear-centered EEG acquisition: aEEGGrids for transparent EEG Front. Hum. Neurosci. 11 163
Chang C-Y, Sheng-Hsio H, Pion-Tonachini L and Jung T-P 2018 Evaluation of artefact subspace reconstruction for automatic EEG artifact removal 40th Annual International Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC) (July 2018) pp 1242–5
Cherry E 1953 Some experiments on the recognition of speech, with one and with two ears J. Acoust. Soc. Am. 25 975–3
Congedo M 2018 Computational EEG-Analysis. Methods and Applications (Singapore: Springer)
Crosse M J, di Liberto G M, Bednar A and Lalor E C 2016 The multivariate temporal response function (mTRF) toolbox: a MATLAB toolbox for relating neural signals to continuous stimuli Front. Hum. Neurosci. 10 604
de Vos M, Gandras K and Debener S 2014 Towards a truly mobile auditory brain–computer interface: exploring the P500 to take away Int. J. Psychophysiol. 91 46–53
Debener S, Emkes R, de Vos M and Bleicher M G 2015 Unobtrusive ambulatory EEG using a smartphone and flexible printed electrodes around the ear Sci. Rep. 5 16743
Debener S, Makeig S, Delorme A and Engel A K 2005 What is novel in the novelty oddball paradigm? Functional significance of the novelty P3 event-related potential as revealed by independent component analysis Cogn. Brain Res. 22 309–21
Debener S, Minow F, Emkes R, Gandras K and de Vos M 2012 How about taking a low-cost, small, and wireless EEG for a walk? Psychophysiology 49 1449–53
Delorme A and Makeig S 2004 EEGLAB: an open-source toolbox for analysis of single-trial EEG dynamics including independent component analysis J. Neurosci. Methods 134 9–21
Ding N and Simon J 2012 Emergence of neural encoding of auditory objects while listening to competing speakers Proc. Natl Acad. Sci. 109 11854–9
Gabsteiger F, Leuthesser H, Reis P, Lochmann M and Eskofier B M 2014 ICA-based reduction of electromyographic artifacts in EEG data: comparison with and without EMG data Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. 2014 1–4
Geirnart S, Francart T and Bertrand A 2021a Fast EEG-based decoding of the directional focus of auditory attention using common spatial patterns IEEE Trans. Biomed. Eng. 68 1557–68
Geirnart S, Vandecappelle S, Alickovic E, de Cheveigné, Lalor E, Meyer B T, Miran S, Francart T and Bertrand A 2021b EEG-based auditory attention decoding—towards neuro-steered hearing devices IEEE Signal Process. Mag. (arXiv:2008.04569) [https://doi.org/10.1109/MSP.2021.3075332]
Hambrock D A and Tata M S 2019 The effects of distractor set-size on neural tracking of attended speech Brain Lang. 190 1–9
Han C, O’Sullivan J, Luo Y, Herrero J, Mehta A D and Mesgarani N 2019 Speaker-independent auditory attention decoding without access to clean speech sources Sci. Adv. 2019 5
Holze B, Iager M, Debener S, Adiboglu K and Mirkovic B 2021 Attention capture is reflected in the neural tracking of attended and ignored speech Front. Neurosci. 15 643705
Huang N and Elhilali M 2017 Auditory salience using natural soundscapes J. Acoust. Soc. Am. 141 2163–76
Huang N and Elhilali M 2020 Push-pull competition between bottom-up and top-down auditory attention to natural soundscapes eLife 9 1–22

IOP Publishing J. Neural Eng. 18 (2021) 066054 I. Straetmans et al
Jacobsen N, Blum S, Witt K and Deebener S 2020 A walk in the park? Characterizing gait-related artifacts in mobile EEG recordings Eur. J. Neurosci. 2020 1–20
Jaeger M, Mirkovic B, Bleicher M G and Deebener S 2020 Decoding the attended speaker from using evaluation intervals captures fluctuations in attentional listening Front. Neurosci. 14 603
Kaya E M and Elhilali M 2014 Investigating bottom-up auditory attention Front. Hum. Neurosci. 8 1–12
Kaya E M and Elhilali M 2017 Modelling auditory attention Phil. Trans. R. Soc. 372 1–10
Kaysr S, Ewert H, Anemüller J, Rohdenburg T, Hohmann V and Kollmeier B 2009 Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses EURASIP J. Adv. Signal Process. 2009 157–67
Kerlin J R, Shahin A J and Lee M M 2010 Attentional gain control of ongoing cortical speech representations in a ‘cocktail party’ J. Neurosci. 30 620–8
Kling M-U 2009 Die Känguru-Chroniken: Ansichten eines vorlauten Beuteltiers (München: Ulstein Verlag)
Ladouce S, Donaldson D I, Dudchenko P A and Ietswaart M 2019 Mobile EEG identifies the re-allocation of attention during real-world activity Sci. Rep. 9 13851
Lesenfants D and Francart T 2020 The interplay of top-down focal attention and the cortical tracking of speech Sci. Rep. 10 6922
Mesgarani N and Chang E F 2012 Selective cortical representation of attended speaker in multi-talker speech perception Nature 485 233–6
Mirkovic B, Bleicher M G, de Vos M and Deebener S 2016 Target speaker detection with concealed EEG around the ear Front. Neurosci. 10 349
Mullen T R, Kothe C A E, Chi Y M, Ojeda A, Kerth T, Makeig S, Jung T-P and Cauwenberghs G 2015 Real-time neuroimaging and cognitive monitoring using wearable dry EEG IEEE Trans. Bio-Med. Eng. 62 2553–67
Murray M, Brunet D and Michel C 2008 Topographic ERP analyses: A step-by-step tutorial review Brain Topogr. 20 249–69
O’Sullivan J A, Power A J, Mesgarani N, Siddharth R, Fouse J J, Shinn-Cunningham B G, Slaney M, Shamma S A and Lalor E C 2014 Attentional selection in a cocktail party environment can be decoded from single-trial EEG Cereb. Cortex 25 1697–706
Petersen E, Wöstmann M, Obleser J and Lunner T 2017 Neural tracking of attended versus ignored speech is differently affected by hearing loss J. Neurophysiol. 117 18–27
Polich J 2007 Updating P300: an integrative theory of P3a and P3b Clin. Neurophysiol. 118 2128–48
Puschmann S, Daeglow M, Strohpal M, Mirkovic B, Rosemann S, Thiel C M and Deebener S 2019 Hearing impaired listeners show increased audiovisual benefit when listening to speech in noise Neuralmage 162 261–8
Reis P M R, Hebenstreit F, Gabsteiger F, von Tscharner V and Lochmann M 2014 Methodological aspects of EEG and body dynamics measurements during motion Front. Hum. Neurosci. 8 156
Reiser J E, Wascher G, Rinkenauer E and Arnau S 2020 Cognitive-motor interference in the wild: assessing the effect of movement complexity on task switching using mobile EEG Eur. J. Neurosci. 2020 1–21
Salvidgeotia M P, Jacobsen N, Bauer A-K R, Griffiths B, Hanslmayr S and Deebener S 2019 Out and about: subsequent memory effect captured in a natural outdoor environment with smartphone EEG Psychophysiology 56 5
Scanlon J, Jacobsen N S J, Maack M C and Deebener S 2020 Does electrode amplification style matter? A comparison of active and passive EEG system configurations during standing and walking Eur. J. Neurosci. 2020 1–15
Schwarzkopf D and de Haas B 2012 Better ways to improve standards in brain-behaviour correlation analysis Front. Hum. 6 1–6
Shamma S and Klein D 2000 The case of the missing pitch templates: how harmonic templates emerge in the early auditory system J. Acoust. Soc. Am. 107 2331
Shinn-Cunningham B and Best V 2008 Selective attention in normal and impaired hearing Trends Amplif. 12 283–99
Slaney M, Lyon R F, Garcia R, Klemm B, Gnegy C, Wilson K, Kanevsky D, Savila S and Cerf V 2020 Auditory measures for the next billion users Ear Hear. 41 131–9
Stevens K N 1998 On the quantal of speech J. Phonet. 17 3–45
van Eyndhoven S, Francart T and Bertrand A 2017 EEG-informed attended speaker extraction from recorded speech mixtures with application in neuro-steered hearing prostheses IEEE Trans. Biomed. Eng. 64 1045–56