EEG-based decoding of the spatial focus of auditory attention in a multi-talker audiovisual experiment using Common Spatial Patterns

Iustina Rotaru[1,2], Simon Geirnaert[1,2,3], Nicolas Heintz[1,2,3], Iris Van de Ryck[1], Alexander Bertrand[2,3], Tom Francart[1,3]

[1]KU Leuven, Department of Neurosciences, ExpORL. Herestraat 49 bus 721, B-3000 Leuven, Belgium

[2]KU Leuven, Department of Electrical Engineering (ESAT), Stadius Center for Dynamical Systems, Signal Processing and Data Analytics. Kasteelpark Arenberg 10, B-3001 Leuven, Belgium

[3]Leuven.AI - KU Leuven institute for AI, Belgium

E-mail: `iustina.rotaru@kuleuven.be, alexander.bertrand@kuleuven.be, tom.francart@kuleuven.be`

Abstract. *Objective.* Auditory attention decoding (AAD) refers to the task of identifying which speaker a person is listening to in a multi-talker setting, based on their neural recordings. The Common Spatial Patterns (CSP) algorithm has previously shown promising potential w.r.t. the state-of-the-art AAD algorithms to create discriminative features from electroencephalography (EEG) signals in a task of spatial AAD (sAAD). However, there has been some skepticism related to the underlying decoding mechanisms of such sAAD methods, as well as their generalization capabilities across subjects and experimental trials. In this study, we aimed to investigate (1) what type of mechanisms (neural vs. non-neural) drive the CSP decoding and (2) how well CSP filters can generalize across trials and subjects. *Approach.* We designed a two-speaker audiovisual sAAD protocol in which we enforced the spatial auditory and visual attention to be either congruent or incongruent, and we recorded EEG data from sixteen participants performing this task. *Main results.* Firstly, we found that the sAAD accuracy with CSP-derived features was significantly higher in scenarios where the target visual and auditory stimuli were co-located, potentially indicating that CSP decoders exploited eye-gaze information. Secondly, CSP decoding remained feasible even without relevant eye-gaze information, i.e., when the location of the attended visual target was continuously shifted to ensure spatial dissociation with the auditory stimulus' location. This finding suggests that CSPs are able to extract neural lateralization patterns reflecting spatial auditory attention independent of the eye-gaze direction. Thirdly, we identified a limitation in the between-trial generalization ability of the CSP feature embeddings, observing strong distribution shifts in the feature space across trials. However, we demonstrated this can be overcome by employing partially-unsupervised classification methods. *Significance.* Collectively, our findings confirm the feasibility of CSP filters in decoding the locus of auditory attention in various AV conditions, while equally emphasizing the need for novel algorithms that are robust to generalization.

Keywords: common spatial patterns, spatial auditory attention decoding (sAAD), audiovisual stimulation

1. Introduction

Auditory attention decoding (AAD) is a well-established term that collectively describes a series of techniques designed to discern what is the acoustic target - among a noisy mixture of acoustic sources, to which a person pays attention. This is made possible with neural recordings such as electroencephalography (EEG), magnetoencephalography (MEG) or electrocorticography (ECoG) (O'Sullivan et al. 2014, de Cheveigné et al. 2018, Mesgarani & Chang 2012).

One of the most sought-after applications of AAD algorithms is their potential to identify the attended speaker in a multi-talker setting, in order to steer the noise suppression algorithms in hearing aids (HAs), which can lead to a novel tier of *neuro-steered HAs*. These devices are currently envisioned to enable effortless control by hearing-impaired users through their brain signals, whereby the attended sounds are automatically recognized and enhanced over non-target sounds. AAD-enhanced HAs are expected to bring an unprecedented benefit to the subjective listening experience because they would significantly improve speech intelligibility and diminish the listening effort expended by the HA users (Geirnaert, Vandecappelle, Alickovic, de Cheveigné, Lalor, Meyer, Miran, Francart & Bertrand 2021, Slaney et al. 2020).

From the rich collection of existing AAD algorithms, here we focus on a novel framework for decoding the *directional* locus of auditory attention, namely the Common Spatial Pattern (CSP) filtering algorithm, initially proposed by Geirnaert, Francart & Bertrand 2021. CSP was one of the first algorithms that targeted the decoding of *spatial* auditory attention (sAAD) in a multi-speaker setting, as opposed to traditional AAD algorithms that use a stimulus reconstruction approach (O'Sullivan et al. 2014, de Cheveigné et al. 2018). In short, a CSP-based attention decoder is optimized to detect the spatial focus of attention from instantaneous neural features which reflect spatial auditory attention patterns. The CSP algorithm holds two major advantages w.r.t. the traditional AAD paradigms based on stimulus reconstruction: (1) it decodes the direction of an attended sound stream directly from the user's EEG, i.e., without requiring access to demixed and clean audio signals and (2) it operates accurately on short time-scales (within 1–5 s) (Geirnaert, Francart & Bertrand 2021). These two positive feats make CSPs particularly suitable for real-time attention decoding systems, whereby HA users are exposed to complex acoustic scenes and spontaneously switch their attention between different acoustic targets, often situated at distinct locations. Yet despite the promising prospects for real-time (s)AAD, the precise decoding mechanisms of the CSP algorithm have not been fully unraveled, nor has its application and generalization been validated across diverse audiovisual scenarios in an sAAD experiment. To address this research gap, we here pursue two major research

objectives.

Firstly, we aim to shed light on which mechanisms (neural or non-neural) drive the CSP decoding performance. As CSP filters are trained on EEG signals during periods of sustained attention to localized speech, it was previously hypothesized they are susceptible to picking up lateralization patterns related to or exhibited by non-neural signals, i.e., eye-gaze, face- or ear-muscle activations, that are correlated to the spatial focus of auditory attention (Geirnaert, Francart & Bertrand 2021, Strauss et al. 2020). A number of recent studies even point towards interactions of the oculomotor system and top-down attention-modulated speech processing. For instance, a phenomenon called *ocular speech tracking* was observed by Gehmacher et al. 2023, claiming that gaze activity tracks acoustic features of an attended natural speech signal more strongly than for a distracting sound. Additionally, they showed evidence that oculomotor activity distinctly contributes to the neural responses of sensors overlapping the auditory processing areas (temporal and parietal). In another study on spatial auditory attention, Popov et al. 2022 showed neurophysiological evidence that alpha power lateralization, an established biomarker for the top-down attention-related mechanism that suppresses distracting input from unattended directions of sound, is closely associated with lateralized oculomotor action, i.e., eye-gaze shifts. However, previous studies employing CSPs in an sAAD task were not able to fully rule out potential contributions from non-neural signals because these were not explicitly measured or controlled for in the evaluated datasets. Thus, we set out to evaluate whether CSPs can accurately capture spatialized auditory attention patterns independent of eye-gaze. To this end, we designed an audiovisual AAD (AV-AAD) experiment where we imposed various degrees of correlation between the spatial directions of to-be-attended *visual* and *auditory* targets, such that they were either co-located, spatially uncorrelated, or the visual stimulus was totally absent.

Leveraging this novel dataset, our second objective is to evaluate the CSPs' ability to generalize across different trials and subjects, as good generalization is crucial for the CSP algorithm to work properly in a real-time context. Across-subject generalization is especially useful and time-efficient because it would allow the creation of plug-and-play AAD algorithms (i.e., pre-trained on EEG data from previous subjects) that can directly generalize to a new subject with a minimal or even without any calibration session. However, previous studies that focused on CSPs in other brain-computer interface (BCI) paradigms, such as, e.g., motor imagery, have highlighted a number of caveats concerning the CSP's generalization. For instance, Reuderink & Poel 2008, Huang et al. 2010 showed that CSPs tend to overfit on small training sets, i.e., the decoding accuracy on new and unseen EEG data degrades as the CSPs are trained on gradually smaller train sets. A few extensions to the standard CSP algorithm, i.e., spatio-spectral CSP filters (Lemm et al. 2005), sparsified CSPs (Farquhar et al. 2006), invariant CSPs (Blankertz et al. 2007) and CSPs with Tikhonov regularization (Lotte & Guan 2010), have been proposed to improve generalization. Moreover, it was hypothesized that different cognitive states or attention levels, coupled with unpredictable and involuntary

artifacts (such as blinking, yawning or frowning), make the EEG signal non-stationary across trials and subjects, which prevents CSP filters from generalizing well to data from trials that were unseen during training (Blankertz et al. 2007). In the sAAD context, Geirnaert, Francart & Bertrand 2021 previously showed that CSP decoding models could generalize across subjects within a subject-independent cross-validation scheme, where the training set consisted of EEG data pooled together from all subjects except the test subject. Additionally, they also showed that the subject-independent CSP filters can closely approach the performance of a subject-specific CSP decoder provided two adaptations are made: (1) the bias term of the classifier is updated based on unlabelled data from the test subject, and (2) the CSP filters are trained and applied on a single frequency band (empirically determined), as opposed to a filterbank approach (Geirnaert, Francart & Bertrand 2021).

The remainder of this paper is structured as follows: Section 2 describes the experimental setup and the novel AAD audiovisual protocol, Section 3 introduces the CSP methodology and the chosen hyperparameters, while Section 4 presents and discusses the obtained results, together with some limitations of the current study. Finally, conclusions are drawn in Section 5.

2. Experimental setup

2.1. Participants

Sixteen normal-hearing participants (one male, fifteen females) were recruited to take part in the AV-AAD experiment. The participants' age ranged between 19-27 years (with mean age and standard deviation of 20.72 ± 1.00 years). All subjects were native Flemish speakers. Every subject signed an informed consent form approved by the local Medical Ethics committee. Normal hearing for all subjects was verified by pure-tone audiometry.

2.2. Audiovisual protocol

For stimulation, we used a self-curated playlist of video clips from "Universiteit van Vlaanderen"‡, a popular platform for science-outreach podcasts delivered by various researchers. From a database of more than 100 such videos, we pre-selected 30 male-narrated videos with the best audio and video recording quality, spanning a wide variety of scientific topics. To make the experiment as engaging as possible, we let each participant choose their 10 most and 10 least preferred podcasts (with which they were not previously familiar) from our pre-compiled list. To avoid any stimulus pre-exposure effects in the preference selection stage, the subjects were not provided with the actual videos, but only their titles, which summarized the topic. All presented videos were in the .mp4 format and had an original resolution of 1280×720 px, but were downscaled to a smaller frame size of 640 x 360 px since smaller versions of the

‡ https://www.universiteitvanvlaanderen.be/

videos will be presented at different locations on the screen (as explained further). The audio tracks of each video were separately extracted, and the overall root-mean-square (RMS) intensity was normalized to -27 dB RMS. The silence portions of the audio tracks were not removed nor shortened, in order to ensure precise synchronization with the video. Lastly, all used videos and their audio tracks were cut to a duration of 10 min in order to match the trial length.

In total, the experiment included four conditions, each consisting of two trials of 10 min (i.e., each condition lasted for a total of 20 min). In each trial, two audio stimuli were dichotically presented at 65 dB SPL via the RME Fireface UC soundcard (RME, Haimhausen, Germany) connected to two insertphones of type Etymotic ER10 (Etymotic Research, Inc., IL, USA). To recreate an acoustic spatial impression of sounds coming from the left and right, each pair of presented stimuli was convolved with head-related impulse responses (HRIRs) corresponding to $-90°$ and $+90°$ (the HRIRs were measured on a dummy head in an anechoic room using in-the-ear microphones, cf. Kayser et al. 2009). We randomized the presented stimuli per participant by randomly drawing (without repetition) from the participants' lists of *most* and *least* preferred podcasts one *to-be-attended* and one *to-be-ignored* speech stimulus, respectively, per experimental trial. For all conditions, the subjects had the same task, i.e., to listen to the target speaker as indicated on the computer screen and to ignore the competing speaker. However, depending on each condition's type, the subjects were asked to adhere to a different set of visual instructions, explicitly shown at the beginning of each trial.

An overview of the AV-AAD experimental conditions is presented in Table 1. To probe whether eye-gaze has any influence on the CSP decoding of spatial auditory attention in various AV scenarios, these conditions were designed to have different degrees of consistency between the spatial direction of visual and auditory attention. The visual stimulation in each of the four conditions (further referred to by their acronyms) is described in more detail below:

- In the *Moving Video (MV)* condition, the video of the to-be-attended speaker was presented as randomly moving left-right along a linear horizontal trajectory spanning the entire screen width. The target coordinates of each new video position were randomized along the horizontal axis, and the downscaled video was programmed to move with a constant and moderate speed of 50 px/s between these target points. The random movement was balanced such that the video was presented on each half of the screen for 50% of the time within each trial. Overall, this condition was designed to have complete inconsistency between the spatial visual and auditory attention throughout the whole stimulation duration. Note that despite the lack of spatial correlation, there was still a semantic correlation between the target acoustic and visual stimulus, since the content of the moving video matched the content of the attended speaker.
- The *Moving Target Noise (MTN)* condition visually consisted of a crosshair also randomly moving on a linear horizontal trajectory at 50 px/s, spanning the entire

Presentation order	Condition name	Auditory vs. visual attention	Visual task
1	Moving Video (MV)	Incongruent	Follow the *moving video* of the to-be-attended speaker on a randomized horizontal trajectory.
2	Moving Target Noise (MTN)	Incongruent	Follow the *moving cross-hair* on a randomized horizontal trajectory.
3	No Visuals (NV)	Incongruent	Fixate on an imaginary point in the center of the *black* screen and minimize eye movements.
4	Static Video (SV)	Congruent	Fixate the *static video* presented on the same side of the screen as the to-be-attended speaker.

Table 1: Overview of the AV-AAD experimental conditions

screen width. This condition was intended to be acoustically more challenging than the former, hence background babble noise was added to each insertphone at a Signal-to-Noise Ratio (SNR) of -1 dB (relative to the joint level of both speakers). In addition, by presenting a crosshair instead of a video, the visual semantic cues were removed, in order to enforce the spatial auditory attention as dominant. Altogether, the MTN condition lacked both the spatial and semantic correlation between the audiovisual stimuli.

- In the *No Visuals (NV)* condition, a black screen was presented and the participants were asked to fixate their gaze on an imaginary point within the center of the screen. This was intended as a control condition for visual attention.
- In the *Static Video (SV)* condition, the video of the to-be-attended speaker was fixed on one side, matching the direction of the to-be-attended acoustic stimulus. Hence, there was both spatial and semantic correlation between the audiovisual stimuli. This condition was intended as a proxy for most scenarios in daily life, where visual and auditory attention are spatially aligned (i.e., when a person looks at the audio source they focus on).

Time-wise, the conditions were presented in two separate blocks, i.e., the first trials of each condition were presented sequentially, followed by a second block of trials from each condition. As such, the EEG signals belonging to the two trials of each condition were purposefully measured about 40 min apart, in order to capture the EEG non-stationarity across trials, allowing us to probe the between-trial generalization of CSP filters. The conditions were presented in the same order for all subjects (as specified in Table 1). Following preliminary pilot tests, the SV condition was subjectively rated as the least difficult, hence it was deliberately presented last. The MTN condition was not presented to the first three subjects as it was a later addition to the experiment. Hence, the total EEG recording time was 60 minutes (for 3 subjects) and 80 minutes (for the remainder 13 subjects). After every trial of 10 min, there was a short break, which ensured that the participant's attention levels were constantly refreshed. During each break, a comprehension question related to the content of the stimulus attended in the previous trial was asked and the participants had to answer with a word or a short

phrase. The answers were not further analyzed, as they were only intended to keep up the participants' attention and motivation levels. A longer break of approximately 5 min in-between the two experimental blocks was also offered to the participants.

In addition, one *spatial switch of attention* was introduced in the middle of each trial (at 5 min), by programming both stimuli to swap sides from left (L) to right (R) and vice versa. The starting side of attention was randomized for each trial and subject. In conditions with visual stimulation (MV, MTN and SV), an arrow was continuously displayed on the screen, pointing to the direction of the to-be-attended acoustic stimulus. Additionally, the arrowhead changed from L to R (or reversely) after 5 min, to visually cue the participants about the spatial attention switch. Conversely, in the NV condition, the participants were only verbally cued at the beginning of the trial to which speaker (L or R) they had to listen first. Similarly, switching in the NV condition was only cued in the audio modality at 5 min within each trial, when the two stimuli automatically swapped sides. The participants were *a priori* informed that they were expected to change their locus of attention when the attended stimulus swapped sides in the NV condition.

The entire experiment was conducted in a soundproof, electromagnetically shielded room. 64-channel EEG and 4-channel electrooculography (EOG) were recorded at a sampling rate of 8192 Hz with the BioSemi ActiveTwo system (Amsterdam, The Netherlands). The EOG sensors were placed symmetrically around the eyes (cf. Fig. 4 in Lopez et al. 2016). Two of the EOG sensors were placed approx. 1.5 cm above and below the right eye (aligned with the center of the eye) to measure vertical oculomotor activity (vertical saccades or blinks). The other two EOG sensors were placed approx. 1 cm right of the right eye and 1 cm left of the left eye, respectively, to measure horizontal oculomotor activity. All visual stimuli were presented on a 21.5 inch screen with a resolution of 1920×1080 px by running custom-made Python scripts. Synchronization between the audio and video stimuli was performed with the *pygame* module§, while synchronization between the measured EEG and the corresponding audio stimuli was achieved via squared-pulse triggers presented every second for the entire duration of each trial and recorded using the EEG system.

The range of the visual angle spanned by the participants' eyes was geometrically determined from the screen width (47.6 cm) and the average distance from the participant's head to the screen (45 cm). This yielded an estimation range of $(-20°, +20°)$ relative to the screen center‖.

§ https://www.pygame.org/

‖ A precise measurement was not possible since the participants' heads were not fixated. As such, they could have spontaneously, even unconsciously, changed their head position during the EEG data collection, thus slightly changing the range of the visual angle.

3. Decoding spatial auditory attention with CSP filtering

3.1. CSP filtering

Common spatial patterns (CSP) filtering is a technique widely used in, e.g., BCI motor imagery paradigms to discriminate left- versus right-hand motor imagery (Blankertz et al. 2008, Lotte et al. 2018). Based on previous observations that spatial auditory attention appears to be spatio-temporally encoded in the neural activity (Bednar & Lalor 2018, 2020, Wöstmann et al. 2016, Patel et al. 2018), Geirnaert, Francart & Bertrand 2021 demonstrated the feasibility of CSP filters in an sAAD paradigm to decode the directional focus of auditory attention from EEG in a competing-speaker setting. Below we briefly describe the CSP filtering for the binary sAAD paradigm, where the objective is to optimally discriminate between listening to the left and right. For a detailed description of the general theoretical framework of CSP filters, we refer the reader to the studies of Blankertz et al. 2008, Parra et al. 2005.

In a nutshell, the CSP filters $\mathbf{W} \in \mathbb{R}^{C\times K}$ project the (zero-mean) EEG signal $\mathbf{x}(t) \in \mathbb{R}^{C\times 1}$ measured at a time instance t from the original electrode space of C channels into a surrogate subspace $\mathbf{y}_{CSP}(t) = \mathbf{W}^\intercal\mathbf{x}(t) \in \mathbb{R}^{K\times 1}$ of lower dimension $K \ll C$, where the K output channels are uncorrelated and the discrimination between the two classes is maximized.

Mathematically, the CSP algorithm determines the orthogonal spatial filters $\mathbf{w}_k \in \mathbb{R}^{C\times 1}$ (the columns of $\mathbf{W}$) which maximize the output ratio of variance between the instances of the two classes $\mathbf{x}_{1/2}(t) \in \mathbb{R}^{C\times 1}$ in the projected subspace:

$$\mathbf{w} = \arg\max_{\mathbf{w}} \frac{\frac{1}{|C_1|}\sum_{t\in C_1}(\mathbf{w}^\intercal\mathbf{x}_1(t))^2}{\frac{1}{|C_2|}\sum_{t\in C_2}(\mathbf{w}^\intercal\mathbf{x}_2(t))^2} \tag{1}$$

$$= \arg\max_{\mathbf{w}} \frac{\mathbf{w}^\intercal\mathbf{R}_{C_1}\mathbf{w}}{\mathbf{w}^\intercal\mathbf{R}_{C_2}\mathbf{w}} \tag{2}$$

where

$$\mathbf{R}_{C_i} = \frac{1}{|C_i|}\sum_{t\in C_i}\mathbf{x}(t)\mathbf{x}^\intercal(t) \tag{3}$$

is the covariance matrix of class $C_i, i \in \{1, 2\}$ and $|C_i|$ denotes the number of time instances in class C_i. As shown in Blankertz et al. 2008, the solution of Eq. 2 can be found by computing a generalized eigenvalue decomposition of the class-specific covariance matrices:

$$\mathbf{R}_{C_1}\mathbf{w} = \lambda\mathbf{R}_{C_2}\mathbf{w} \tag{4}$$

By plugging Eq. 4 into Eq. 2, we obtain:

$$\mathbf{w} = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^\intercal(\lambda\mathbf{R}_{C_2}\mathbf{w})}{\mathbf{w}^\intercal\mathbf{R}_{C_2}\mathbf{w}} = \arg\max_{\mathbf{w}} \lambda \tag{5}$$

Thus, the best discrimination between the two classes is obtained when the EEG signals are projected onto the generalized eigenvector $\mathbf{w}_1$ corresponding to the largest generalized eigenvalue λ_1. The projection $\mathbf{w}_1^\intercal\mathbf{x}(t)$ will then produce a virtual channel

which will have the maximal relative difference in signal power between the two classes (as targeted in Eq. 1). By reciprocity (i.e., switching the numerator and denominator in Eq. 1), the smallest generalized eigenvector $\mathbf{w}_K$ corresponding to the smallest generalized eigenvalue λ_K will achieve the same for the other class. Therefore, to select the K most informative spatial filters, the generalized eigenvalues λ_k are sorted and the corresponding first $\frac{K}{2}$ and last $\frac{K}{2}$ generalized eigenvectors are then selected as columns of $\mathbf{W}$.

3.2. Classification

CSPs are usually part of a larger classification pipeline to decode directional auditory attention on new EEG data (unseen in the CSP training phase). In the following, we describe the classification procedure which we used in our analysis.

To obtain features for classification, the test EEG data is bandpass-filtered into B pre-defined frequency bands and segmented into smaller time windows of length T, called *decision windows*. The classification task consists in assigning each decision window to either one of the two classes (attended left or right). To this end, the CSP filters (separately trained per frequency band) are applied per frequency band on each decision window of the test data. Finally, the log-energies of each CSP-filtered window and frequency band are computed (using log-energy features is common in CSP-based classifiers). This results in a total of $B \times K$ CSP features $f_{k,b}$ per decision window:

$$f_{k,b} = \log(\sum_{t=1}^{T} y_{k,b}(t)^2), \tag{6}$$

which are stacked together into one feature vector $\mathbf{f} \in \mathbb{R}^{BK\times 1}$. To determine the directional focus of attention, $\mathbf{f}$ is fed to the input of a binary classifier. In this work, we consider two popular classifiers: supervised linear discriminant analysis (LDA) and unsupervised k-means clustering.

3.2.1. Supervised classification with Linear Discriminant Analysis. Fisher's linear discriminant analysis (LDA) is traditionally used in combination with CSP filters (Lotte et al. 2018). LDA optimizes a linear projection $\mathbf{v} \in \mathbb{R}^{BK\times 1}$ within the feature space that maximizes the between-class scatter while minimizing the within-class scatter. This also leads to a generalized eigenvalue problem, which can be solved analytically (Bishop & Nasrabadi 2006):

$$\mathbf{v} = \mathbf{\Sigma_w^{-1}}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1). \tag{7}$$

Here, $\mathbf{\Sigma_w} \in \mathbb{R}^{KB\times KB}$ is the joint covariance matrix of the features $\mathbf{f}$ from both classes and $\boldsymbol{\mu}_{1/2} \in \mathbb{R}^{BK\times 1}$ are the features' means across all decision windows in each class. The LDA decision boundary is given by:

$$D(\mathbf{f}) = \mathbf{v}^\intercal \mathbf{f} + b, \tag{8}$$

where $\mathbf{v}$ is defined in Eq. 7, and b is the classifier's bias or threshold, expressed as the mean of the LDA-projected class means:

$$b = -\frac{1}{2}\mathbf{v}^{\intercal}(\boldsymbol{\mu}_2 + \boldsymbol{\mu}_1) \tag{9}$$

Eventually, $\mathbf{f}$ is classified into class C_1 if $D(\mathbf{f}) > 0$ and into class C_2 if $D(\mathbf{f}) < 0$.

Note that both CSP and LDA are data-driven and supervised, as they require ground-truth-labeled data to be trained on.

3.2.2. Unsupervised classification with k-means. As noted by Lotte et al. 2018, Huang et al. 2010, using CSPs in combination with LDA classifiers in BCIs could suffer from generalization problems across experimental trials and/or subjects. The potential inability to generalize has been generally attributed to EEG non-stationarities which are unavoidable during long and separate measurement sessions due to changes in the subject's brain processes, attention and fatigue levels, different artifactual patterns (Blankertz et al. 2007) (e.g., frowning, blinking, swallowing, or yawning) or changes in the experimental conditions (interfering equipment and noise sources).

To overcome potential problems with train-test generalization generated by EEG data captured at different time points, we here take a different approach and aim to directly classify the test CSP features in an unsupervised way, thus bypassing any train-data biases in the classification step. To this end, we replace the supervised LDA with one of the most popular unsupervised algorithms, i.e., k-means clustering (Bishop & Nasrabadi 2006). In short, k-means finds the optimal partitioning of a given set of observations into K clusters by assigning each observation to the cluster with the nearest centroid (the mean of all points in the cluster) in an iterative process, without needing any *a priori* label/class information.

We note that unsupervised clustering can be performed directly on the test data, i.e., it does not necessarily have to be trained on a separate training set (as opposed to CSP and LDA). As a result, it adds flexibility to adapt the classifier to individual (test) trials. However, without further context, it is impossible to then determine which cluster belongs to which class. In Section 3.4, we explain the workaround we employed in computing accuracies for the k-means clustering analyses.

3.3. Practical implementation

3.3.1. Data pre-processing. The EEG was initially downsampled using an anti-aliasing filter from 8192 Hz to 256 Hz to decrease the processing time. The EEG trials were then filtered between 1-40 Hz using a zero-phase Chebyshev filter (type II, with 80 dB attenuation at 10% outside the passband) and subsequently re-referenced to the common average of all channels. Afterwards, additional downsampling to 128 Hz was performed to speed up the training of attention decoders.

Preprocessing to remove gaze-related artifacts. For regular CSP decoding, no extra artifact removal step is needed during EEG preprocessing, as the CSP filters are

optimized to discard any signal and/or artifacts that are not useful in the current discrimination task, as long as the training set is sufficiently diverse (Blankertz et al. 2008, Geirnaert, Francart & Bertrand 2021). However, our first research question deals with the influence of non-neural signals (particularly eye-gaze) on CSP attention decoding. Hence, we evaluated whether eye-gaze contributions have any effect on sAAD accuracy by comparing the decoding performance of CSP filters when trained with regularly preprocessed EEG (see above) vs. when trained with EEG from which gaze-related signals are explicitly attenuated.

Thus, exclusively for the investigation of eye-gaze confounds, we separately applied three different methods to remove eye-gaze information (blinks and saccades) that could have leaked into the EEG: (1) Independent Component Analysis (ICA), (2) regressing out EOG from EEG with an adaptive least mean squares filter, and (3) filtering the EEG in the beta band (12-30 Hz) to discard slow drifts, including those generated by eye movements.

3.3.2. Hyperparameter choices. The CSP filters are usually trained and applied on EEG data filtered in a single frequency band of interest or across a pre-selected range of frequency bands that are relevant to the analysis at hand. In this work, we adopted the so-called filterbank CSP (as in Geirnaert, Francart & Bertrand 2021), which entails that EEG is first filtered into B different frequency bands and the CSP filters are then trained and applied per frequency band, resulting in a total of B CSP filter matrices $\mathbf{W}_b$, with $b = 1, ..., B$. We considered a filterbank of $B = 14$ overlapping bands spanning a wide frequency range: [1-4], [2-6], ..., [24-28], [26-30] Hz. Thus, we intentionally avoided to manually pick a single frequency band, in favor of allowing the classification algorithm to decide which bands are most relevant for distinguishing between left (L) vs right (R) auditory attention.

The number of CSP filters (per frequency band), i.e., K, is another tunable hyperparameter. Too many filters are not desirable, as the number of filter weights to optimize increases, and with it, the risk of overfitting. Too few filters are neither optimal, as they might not accurately capture the discrimination between the two classes. Thus, following the conventional CSP recipes from the BCI literature (Blankertz et al. 2008), we chose to train a moderate number of $K = 6$ CSP filters per frequency band.

Before the CSP training step (Eq. 4), we regularized the large-dimensional covariance matrices of each class ($\mathbf{R}_{C_{1/2}}$) using diagonal loading, i.e., by computing a weighted combination of the sample covariance matrix (potentially poorly-conditioned, but unbiased) and the identity matrix (well-conditioned, but uninformative). The weights of both matrices were automatically determined via the Ledoit-Wolf criterion (Ledoit & Wolf 2004, Geirnaert, Francart & Bertrand 2021).

Finally, as the generalized eigenvalues represent the ratio between class-specific energies of each spatially filtered signal (cf. Eq. 2), they can become corrupted by outlier segments with a high variance. To avoid this issue, the most discriminative CSP filters were selected based on the ratio of median output energies (RMOE) between

both classes, cf. Blankertz et al. 2008, (instead of the sorted generalized eigenvalues in Eq. 4), taken over all training windows of length equal to the maximal decision window length that is used in the analysis.

3.4. Performance evaluation

The CSP filters were trained and evaluated per experimental condition both within subjects (subject-specifically) as well as across subjects (subject-independently). As evaluation metric, we reported the *decoding accuracy*, i.e., the percentage of correctly classified decision windows averaged across all cross-validation (CV) folds, CV repetitions and tested subjects. We will use several variations of CV schemes (see below). Notably, both the CSP filters and LDA were always trained on the same subset of data.

For the *subject-specific* decoding (further denoted as CSP-SS), different CSP filters and LDA classifiers were trained per subject and per experimental condition with random 5-fold CV. To this end, all the EEG data was split into segments of 60 s, which were randomly shuffled in one of the 5 CV folds. In order to compute the accuracy obtained with different decision window lengths (WLs), the EEG segments of each test fold were further split up into smaller windows ranging from 1 to 60 s. Per subject, average accuracies were computed across the 5 CV test folds, and across 3 repetitions of the random CV scheme. Given that CSP-based classifiers perform comparably well across WLs (Geirnaert, Francart & Bertrand 2021), and that small WLs are required for fast decoding, we used the median accuracies obtained with a WL of 5 s to perform all the statistical analyses presented in Section 4, unless stated otherwise.

We also trained *subject-independent* CSP filters (further denoted as CSP-SI) and LDA classifiers by implementing leave-one-subject-out CV per condition. However, the CSP filters were shown to work less well in the subject-independent setting because of the high variability of signals in different frequency bands across subjects (Geirnaert, Francart & Bertrand 2021). Therefore, exclusively for the CSP-SI analysis, the EEG data was bandpass-filtered into one single broadband frequency range (1-30 Hz), and a bias update was applied to the LDA classifier as a normalization step to improve generalization across subjects (details in Geirnaert, Francart & Bertrand 2021).

Furthermore, we investigated CSP generalization across trials by evaluating the CSP-SS decoder with a 2-fold leave-one-*trial*-out CV scheme per condition, leveraging the fact that the two trials of each condition were recorded roughly 30 min apart, and hence contained heterogeneous EEG data.

For the classification with k-means clustering, we used $K = 2$ clusters, corresponding to the two attended locations. The cluster centroids were initialized with the (non-deterministic) k-means++ algorithm (Arthur & Vassilvitskii 2006), hence we performed 10 repetitions and up to 1000 iterations to re-update the centroids per repetition. From these 10 repetitions, we selected as final classifier the k-means model with the smallest sums of point-to-centroid distances within-cluster. Notably, k-means

is an unsupervised algorithm that assigns arbitrary numerical labels to each cluster (as it is agnostic to the ground-truth labels), hence it is likely that comparing the clustering output labels to the ground-truth labels results in a low accuracy, which can be attributed to an overall label mismatch. To counteract this, both cluster assignments are tested (i.e., cluster 1 = L attended, cluster 2 = R attended, and vice versa), and the assignment that gives the highest accuracy is retained. Note that this implies we use ground-truth labels, and therefore these results based on k-means classification should not be viewed as representative of a realistic sAAD pipeline. In practice, other heuristics can be used to do this label-to-cluster assignment without the use of the ground-truth labels¶, yet this is beyond the scope of this study.

For the reported accuracies obtained with LDA classification, the significance level was determined with the inverse binomial distribution, taking into account the total amount of available test data and a significance threshold of $\alpha = 0.05$ (O'Sullivan et al. 2014, Geirnaert, Francart & Bertrand 2021). As the different CV schemes or different WLs have a distinct number of test samples, this results in different significance levels. For k-means clustering, our manipulation of the L/R attended label assignment to each cluster leads to an inherent bias because it "artificially" pulls all the accuracies above 50%, thus also affecting the significance level. To compensate for this bias, exclusively for the k-means classification accuracy, we compute the significance level as the 97.5th percentile of the inverse binomial distribution (which is mathematically equivalent to the 95th percentile of the *folded* inverse binomial distribution, i.e., the true distribution of the biased accuracies - details omitted).

4. Results and Discussion

4.1. CSP filters achieve the highest AAD accuracy when the target visual and auditory stimuli are spatially aligned

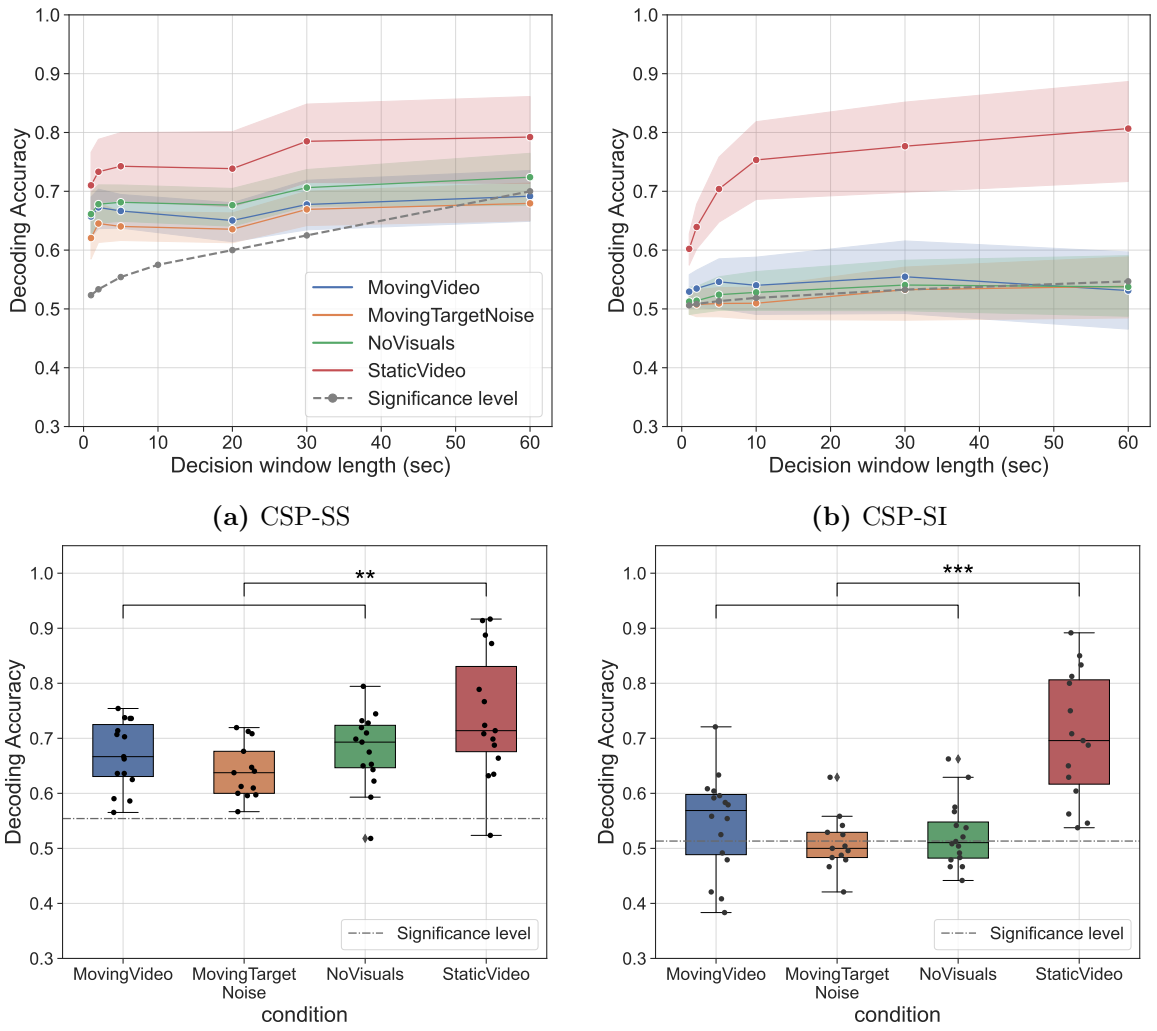
The subject-specific CSP (CSP-SS) decoding accuracies following LDA classification with random 5-fold CV within condition and for various WLs are illustrated in Fig. 1a. For a WL of 5 s, the obtained median accuracies are $66.4\%, 63.8\%, 69.6\%, 71.4\%$ for the MV, MTN, NV and SV conditions, respectively (Fig. 1c). We investigated the effect of condition type on accuracy by means of a Linear Mixed Effects (LME) model where the *condition* was considered as fixed effect and the *subject* as random effect. The LME was fitted to maximize the restricted log-likelihood, and the residuals were checked for normality. Since we were primarily interested in how the audiovisual congruence impacts the decoding accuracy (hence comparing between audiovisual congruent and incongruent conditions), we assigned corresponding contrasts in the model (0.75 for the SV and -0.25 for each of the other conditions). The results revealed that decoding accuracies for the SV condition are significantly higher than for the other conditions

¶ e.g., by combining it with a (slower) stimulus reconstruction approach, or based on speaker localization in combination with speaker activity detection.

(p = 0.002, b = 0.08, CI = [0.03 - 0.13]). This suggests that CSP filters can exploit eye-gaze-related signal components to infer the location of the attended speaker. On the other hand, the fact that CSPs are able to achieve a significant performance in the audiovisual incongruent conditions implies that it is also able to find informative signal components that are independent of the eye-gaze.

Additionally, the subject-independent (CSP-SI) decoding results with LDA classification are depicted in Fig. 1b. The obtained median accuracies with a leave-one-subject-out CV and a WL of 5 s are 56.9%, 50%, 51.1%, 69.5% for the MV, MTN, NV and SV conditions respectively (Fig. 1d). In stark contrast to the CSP-SS decoder, the CSP-SI decoder scores below significance in the visual-incongruent conditions (MV, MTN and NV). In general, poorer performance for a subject-independent model is somewhat expected, as it is more difficult to generalize across subjects than within subject, given the heterogeneous EEG and idiosyncratic CSP feature distributions. Nevertheless, the significant accuracy in the SV condition seems to suggest that CSP is able to capture a dominant subject-independent signal component that is probably related to the eye-gaze direction. An LME model fitted on the CSP-SI accuracies with the same fixed and random effects as for the CSP-SS revealed a similar trend: the accuracy in the audiovisual-congruent SV condition is significantly higher than in the other conditions ($p \leq 0.001$, b = 0.18, CI = [0.13 - 0.23]). Remarkably, the CSP-SS and CSP-SI accuracies in the SV condition are *not* significantly different (Wilcoxon signed rank test: W=31.5, N=15, p=0.12), suggesting that the eye-gaze signal components are strongly captured by both the subject-specific and subject-independent decoders. Still, one must interpret this non-significant result with caution, as the underlying amount of training data for the CSP-SI decoders is much higher (300 min, aggregated across subjects) than for the CSP-SS (15 min per train CV fold). As such, it would not be unreasonable to expect that with more training data available, the CSP-SS would score even better than the CSP-SI for the SV condition.

The beneficial effect of the eye-gaze directional information observed in the SV condition could have three possible reasons. Firstly, this result is in line with previous studies, where the spatial alignment of eye-gaze and auditory attention was found to enhance the auditory percept, the behavioral performance in auditory target detection tasks and the discrimination of interaural time and level differences, which are crucial cues for sound localization (Maddox et al. 2014, Pomper & Chait 2017, Best et al. 2007, Andersen et al. 2009). A recent study (Best et al. 2023) showed that gaze direction alone had a strong effect on speech intelligibility: the recall of digit sequences in a competing multi-talker spatial acoustic scene was significantly better when the look direction coincided with the auditory target direction. Moreover, Gehmacher et al. 2023 showed evidence that *ocular speech tracking* (i.e., the phenomenon in which eye-gaze tracks prioritized acoustic features such as the acoustic envelope and acoustic onsets) is more pronounced for an acoustically attended target than for a distractor, thus advocating for a joint network of auditory selective attention and eye movement control. Extrapolating these findings to our sAAD task, it is probable that a visual



(a) CSP-SS

(b) CSP-SI

(c) CSP-SS, WL = 5 s

(d) CSP-SI, WL = 5 s

Figure 1: Attention decoding accuracy with subject-specific (CSP-SS) and subject-independent (CSP-SI) decoders peaks for the spatially congruent audiovisual presentation (SV condition). Top: median decoding accuracies for all tested window lengths (WL), with shaded areas representing the 95% confidence interval. Bottom: decoding results for WL = 5 s, with each dot representing the accuracy for one subject. Notably, the significance level is reduced for the CSP-SI as more test data was available than for the CSP-SS.

target spatially aligned with an attended acoustic stimulus enhances the neural auditory attention patterns exploited by CSPs.

Secondly, the AV incongruence, which is artificially enforced in all conditions except in SV, might have made the attention task much harder for the participants to follow. In essence, the incongruent AV conditions are a dual task, with a different visual and auditory spatial focus, which could thus partly explain the significantly lower accuracies observed both for CSP-SS and CSP-SI decoders.

Lastly, it is remarkable that the CSP-SS and CSP-SI accuracies both culminate in the SV condition, and particularly that CSP-SI accuracies are only significant in

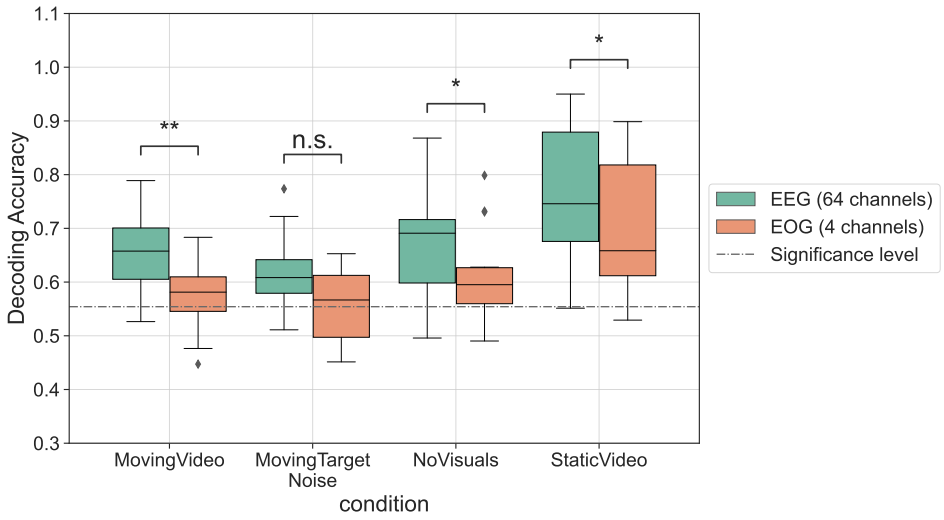


Figure 2: Accuracy scores (with LDA classifier) for CSP-SS decoding with 64 EEG channels vs 4 EOG channels. To enable a direct comparison between the EEG- and EOG-based CSP decoders, we trained only two CSP filters per frequency band (instead of 6). Statistical significance for a Wilcoxon signed-rank test is marked with (*) for $p \in (0.01, 0.05]$ and with (**) for $p \in (0.001, 0.01]$. n.s. = not significant.

the SV condition, despite a comparable amount of training data for all conditions. These results could lead to the suspicion that CSP decoding is predominantly driven by signal components that originate from the motion of the eyeballs (i.e., EOG-related components), and therefore have no neurological component whatsoever. Assuming this is true, when the look direction does not match with the direction of auditory attention (used as ground truth), the CSP decoding accuracies are expected to drop, which is what we observe in the AV incongruent conditions. In an additional analysis, we further evaluated this hypothesis by training new CSP-SS filters exclusively on the four external EOG channels (using a similar 5-fold CV scheme). According to Fig. 2, the median decoding accuracies based on EOG channels (with two CSP filters) are consistently lower than those obtained on the standard, 64-channel EEG, regardless of condition. This is confirmed by a non-parametric Wilcoxon signed-rank test (obtained p-values are $0.001, 0.15, 0.03$ and 0.04 for MV, MTN, NV and SV respectively). While the AV incongruent conditions with only EOG channels lead to barely significant accuracies (Fig. 2), as expected, the EOG accuracies are above the significance level in the SV condition. This supports the hypothesis that CSP filters based on EEG do leverage the explicit eye-gaze directivity patterns as captured by the EOG sensors. However, in the AV-congruent SV condition, the fact that the EOG-based accuracy is significantly lower than the EEG-based accuracy implies that CSP can still exploit some neurological components, despite the presence of the eye-gaze confound. We will further elaborate on this finding in the next subsection.

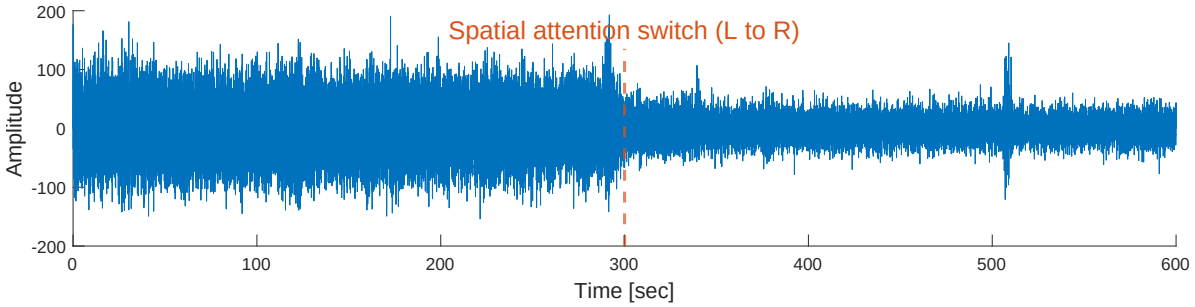


Figure 3: Prominent amplitude changes in CSP-filtered signals reflect the discriminative properties of CSP log-energy features between L and R auditory attention. For illustration, the first channel of the CSP-filtered EEG from subject 12 (trial 1 of the MovingVideo condition) is depicted. The red dashed line indicates the moment when the spatial attention switch occurred.

4.2. CSP filters yield significant sAAD accuracies independent of eye-gaze

While the previous results make it clear that EOG-related components allow CSP to achieve higher accuracies in AV-congruent settings, it remains nevertheless remarkable that decoding accuracies above the significance threshold also occur in conditions without any spatial overlap between the visual and auditory attention (MV, MTN and NV), albeit only in the subject-specific case (Fig. 1a, 1c). This suggests that neural patterns purely reflecting the spatial lateralization of *auditory* attention could be driving the decoding performance in visually-incongruent conditions. Altogether, even though spatially-matched eye-gaze does significantly benefit auditory attention decoding, it is still possible for CSP filters to decode spatial auditory attention in the absence of a spatially-matched visual stimulus (at least in the subject-specific case). Below we present some additional strands of evidence that support this claim.

Firstly, the amplitude of CSP-filtered EEG signals can provide further insights into the CSP discrimination mechanism (Blankertz et al. 2008). Fig. 3 depicts the EEG signal of a representative subject after being spatially filtered by the first CSP component $\mathbf{w}_1$ (maximally discriminative for the L-attention). For visualization purposes, the CSP filters were trained on the entire batch of broadband-filtered EEG (1-30 Hz) from the respective condition and subject. The figure shows clearly-differentiated amplitude levels between the time spans of L vs R spatial auditory attention. Similarly salient L-R amplitude differences were found for other subjects and conditions (not shown), confirming that CSP filters can sharply discriminate between the two directions of spatial attention even in an AV incongruent condition.

Another piece of evidence is linked to the fact that CSP filters achieve similar decoding performances (across conditions) even when contributions of gaze-related information are suppressed from the EEG. This is depicted in Fig. 4: the baseline (CSP-SS) results are plotted along with the accuracies obtained for training the CSP filters with EEG data cleaned with 3 different methods of ocular artifact rejection (cf. Section 3.3.1). Per condition, these results are compared pairwise with a Wilcoxon signed-rank

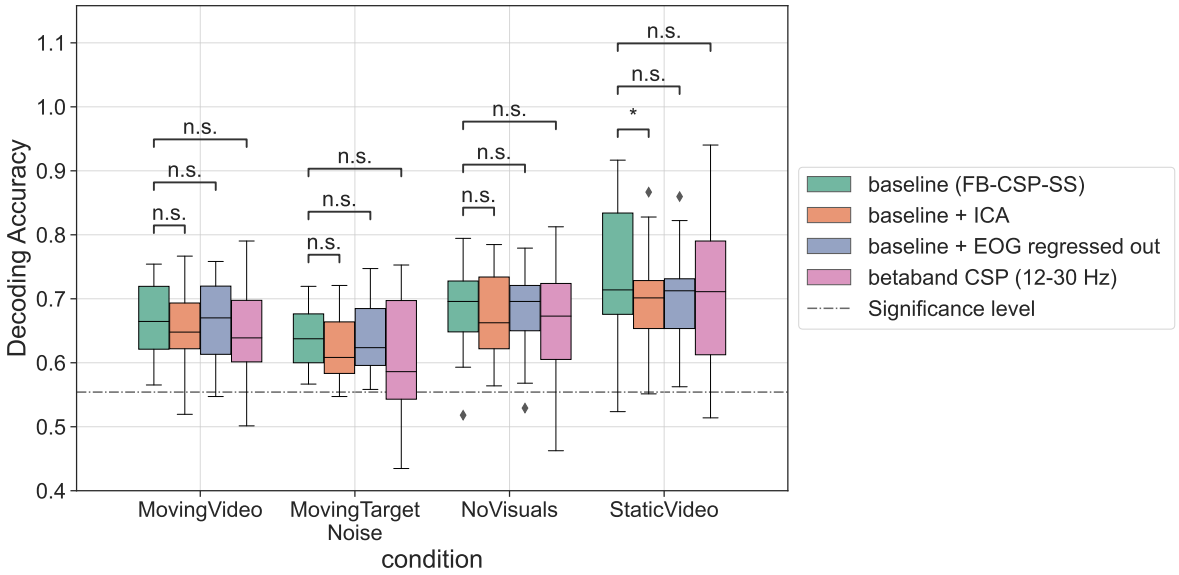


Figure 4: The effect of removing eye artifacts on the CSP decoding accuracies (5 fold CV within-subject and -condition with an LDA classifier, WL=5 sec). The horizontal line in each boxplot denotes the median accuracy across subjects and the diamonds denote outlier accuracies.

test and Bonferroni-corrected for 3 comparisons. Remarkably, regardless of condition type or artifact rejection method (with a single exception), no significant difference in performance could be found between baseline decoders and decoders trained on EEG deprived of gaze information. Furthermore, Table 2 displays relatively narrow 95% confidence intervals of the pairwise differences in accuracies for each condition and pair of compared decoders. These findings reinforce the fact that CSPs are able to capture attention patterns relevant for discriminating L vs R auditory attention independently of eye-gaze. Nevertheless, we acknowledge there is no guarantee that the employed methods fully removed all gaze-related information from the EEG (although visual inspection of randomly picked EEG samples confirmed that eye-blinks and saccade-induced slow drifts were removed).

Lastly, the significantly higher accuracies obtained for EEG decoding compared to EOG decoding in the SV condition (cf. Fig. 2) also support the idea that even if eye-gaze activity can leak into the EEG signals, the CSP filters could pick up additional neural patterns meaningful for the spatial discrimination task, i.e., spatial patterns related to the locus of auditory attention, which render an extra benefit in decoding, and hence lead to elevated accuracies.

4.3. Poor generalization of CSP and LDA across trials can be fixed using unsupervised classification

The results of the CSP generalization analysis (with LDA) are depicted in Fig. 5. The median accuracies for the leave-one-trial out (LOTO) analysis are largely non-significant

	Baseline-EOGregressed	Baseline-ICA	Baseline-Betaband
Moving Video	(-0.012, 0.014)	(-0.003, 0.027)	(-0.008, 0.05)
Moving Target Noise	(-0.015, 0.022)	(-0.012, 0.036)	(-0.009, 0.072)
No Visuals	(-0.014, 0.016)	(-0.012, 0.031)	(-0.005, 0.046)
Static Video	(-0.0002, 0.07)	(0.008, 0.077)	(-0.018, 0.071)

Table 2: The 95% confidence intervals of the difference in accuracies between the baseline CSP-SS decoder and the CSP-SS decoders trained on data preprocessed with 3 types of ocular artifact rejection methods (cf. Section 3.3.1). For reference, the accuracy values for each type of decoder are in the range 0–1.

(44.6%, 44.26%, 41.45% and 43.3% for MV, MTN, NV and SV conditions respectively), suggesting that either the CSP or LDA cannot generalize across two separate trials of the same condition and subject. One factor potentially explaining the low accuracies could be the insufficient amount of training data. As each train fold only contains data from one single trial (i.e., 10 min), it is rather plausible that a single trial does not display diverse enough EEG signal patterns to enable CSPs to generalize to a similar experimental trial recorded at a later time (despite the fact that the amount of time instances with L-R attention is balanced across trials).

4.3.1. CSP features reveal trial biases The difficulties with generalization warrant a closer look at the CSP feature space. A couple of exemplifying CSP feature distributions from the LOTO evaluation are illustrated in Fig. 6 (showing a 2D principal component analysis projection of the full CSP feature space). It is noteworthy that the trained LDA boundary manages to separate the CSP features of the train trial, but fails to do so for the CSP features of the test trial, as they are shifted or rotated with reference to the features of the train trial. These distinct feature distributions across trials may partly explain the suboptimal generalization results across trials observed in Fig. 5.

Yet on a closer look, both the train and test feature clouds, if taken separately, still exhibit distinct, well-separable clusters for each class. By making abstraction of the shift between the train and test features, one can in principle directly classify the test features in an *unsupervised* way. Hence, we probed whether the change of the classifier impacts the decoding accuracy in generalizing across trials by replacing the supervised LDA classifier with the unsupervised k-means algorithm. The LOTO decoding results depicted in Fig. 7 show that k-means is able to restore the median accuracies above the significance threshold, vastly outperforming LDA (the median accuracies obtained with k-means clustering are 65.2%, 77%, 71.9% and 72.9% for MV, MTN, NV and SV conditions, respectively).

As mentioned above, one particular limitation of the leave-one-trial-out generalization is the limited amount of training data (only 2 trials of 10 min are available per condition, of which only one at a time is considered for training). One way to increase the amount of training data is to perform a leave-one-*condition*-out

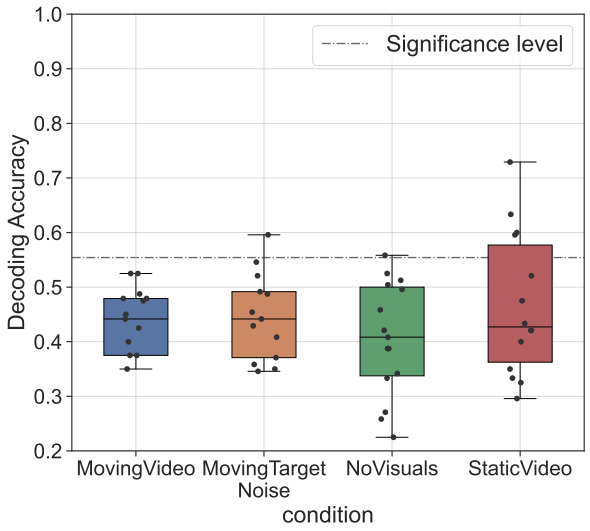


Figure 5: Leave-one-trial-out generalization results of the CSP-SS filters and LDA classifiers (WL = 5 s).

evaluation (i.e., by pooling together train data from all but the test condition). In a separate analysis, we did investigate the generalization across conditions by performing 4-fold leave-one-condition-out CV. In this case, the CSP-SS decoding accuracies with both LDA classifiers and k-means-based classifiers remained under the significance level (results not shown). Given the results from Section 4.1, this is not surprising, as CSPs might rely on different decoding mechanisms depending on the condition type. Hence, pooling train data across different audiovisual conditions (each with its own distinct (neural) patterns reflecting auditory attention) could result in less class-distinctive feature clusters, in turn leading to non-significant decoding accuracies.

Nonetheless, the results from Fig. 7 confirm that despite the limited amount of train data per condition, as well as the shift and rotation of the features across trials, the trained CSP filters are still able to create class-discriminative clusters when trained on one trial from a specific condition and applied to a distinct (test) trial from the same condition, which can be leveraged by an unsupervised clustering algorithm. Remarkably, it appears that separately clustering the CSP features of each test trial in the LOTO analysis outperforms the LDA classification when generalizing within condition (Fig. 7 vs. Fig. 1c). We suspect this occurs because in the latter, LDA was trained by pooling data across trials (cf. 5-fold CV within condition), hence the decision boundary could become corrupted by potential shifts between train features belonging to different trials, thus preventing robust generalization to unseen data from the test fold. In contrast, unsupervised classification of the CSP features of each test trial completely bypasses any potential biases from the train data.

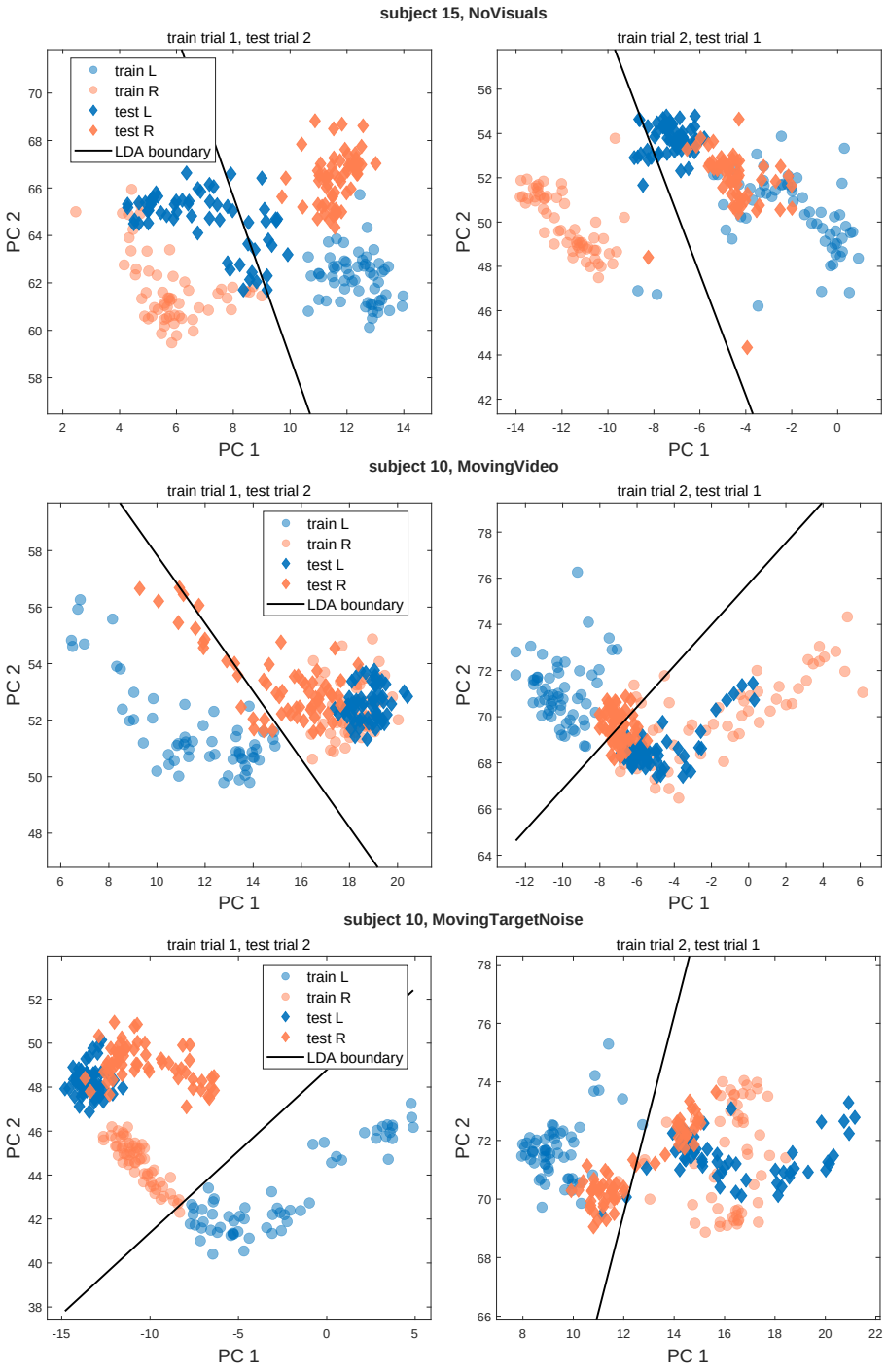


Figure 6: Representative examples of CSP-SS feature shifts and rotations across trials. For 2D visualization, Principal Component Analysis (PCA) was applied on the original CSP features to obtain the first two principal components (PCs) explaining the most variance in the original feature space. Each data point thus corresponds to the first two PCs of every decision window of 5 s. The data points belonging to the train and test sets are marked with circles and diamonds, respectively. The color denotes attention to the left (L) or right (R) speaker. The plotted LDA boundary (solid black line) was obtained after training LDA on the PCs obtained from the CSP train features.

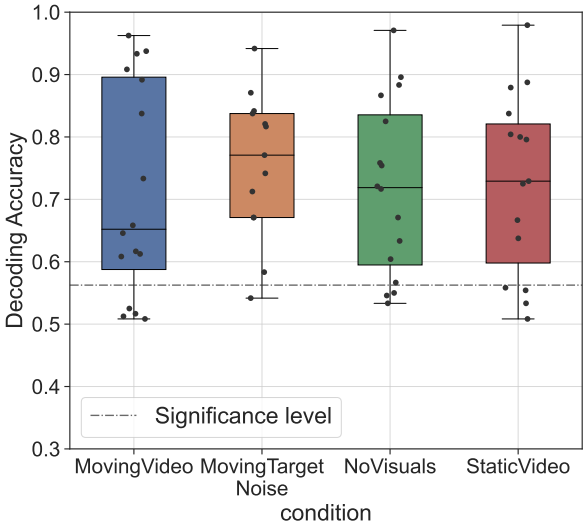


Figure 7: Leave-one-trial-out generalization results of the CSP-SS filters and classifiers based on k-means clustering (WL = 5 s).

4.4. General limitations

Concerning our experimental AV-AAD paradigm, we note that the gaze manipulation we enforced was not completely naturalistic. In real-life, the eye movements have more degrees of freedom and would span a wider range of the visual scene, not only limited to the horizontal direction. However, we only presented the moving stimuli along the horizontal direction because it is the most representative for spatial localization of speech sources and hence, for probing spatial AAD and possible non-neural confounds. Besides that, we only presented the moving stimuli across a rather small range in the horizontal direction, limited by the physical characteristics of our setup (the screen width and the participants' distance to the screen). Still, we hypothesize that a matched gaze direction would have had a similarly beneficial effect on sAAD with CSP filters even when the eye-gaze would span a wider and more realistic moving trajectory.

Regarding CSP decoding, it is important to point out that we only performed binary classifications, i.e., we discriminated auditory attention between only two spatial locations (-90° and +90°). To reduce the gap between the lab setup and the real-life acoustic conditions, future work should focus on validating spatial decoding in experimental trials with more than 3 simultaneously active speakers, while varying the amount of spatial separation between them. Based on the current results, we project that spatial decoding for more than 3 speakers will similarly tend to be more accurate in congruent audiovisual settings.

Furthermore, the limited amount of 20 min of data per condition and subject is another inherent limitation of the acquired dataset. Even though the EEG data of each condition were collected in two distinct trials well separated in time, we suspect this low amount might not be sufficient to train highly accurate and robust CSP decoders. Geirnaert, Francart & Bertrand 2021 obtained a mean accuracy of 80% when evaluating

CSP-SS decoders (with 5 s decision windows in a random 10-fold CV scheme) on an extensive EEG dataset (Das et al. 2020), where the amount of training data per each fold was around 64 min. This result is substantially higher than the maximum median accuracy of 71.4% we obtained in the SV condition by training a CSP-SS decoder with 15 min of data per CV fold (cf. Fig. 1c). It is thus expected that with more training data available per subject and condition, the CSP-SS accuracy scores would be much higher, as the decoders would be exposed to a more diverse repertoire of spatial patterns relevant to auditory attention.

Finally, although the CSP feature shifts and rotations can be handled by means of an unsupervised classifier such as k-means, we re-iterate that we have used ground-truth labels to assign the two clusters to each of the two spatial directions (L and R). Practical ways to do such a cluster assignment without knowledge of the ground truth labels should be further explored.

5. Conclusion

In summary, we designed an audiovisual AAD protocol to probe whether non-neural signals contribute to the decoding of spatial auditory attention with CSP filters from EEG signals. This study yielded three key findings. Firstly, we found that CSP filters achieve distinct performances in decoding spatial auditory attention in audiovisual conditions with various degrees of spatial correlation between visual and auditory attention. Specifically, spatial AAD with CSP-derived features performs significantly better when the visual and auditory stimuli are co-located (i.e., in congruent audiovisual scenarios). Secondly, CSP decoding is also possible without any relevant eye-gaze information, i.e., in conditions where the location of the attended visual target was randomized and purposefully uncorrelated with the auditory stimulus. This suggests that CSPs can extract neural lateralization patterns reflecting directional auditory attention even in the absence of co-located visual stimuli. Thirdly, our extended analysis showed that an AAD pipeline consisting of the CSP algorithm and LDA classification is rather inadequate to generalize to data from other trials or listeners. However, we showed this shortcoming can be bypassed with unsupervised classification methods (with some minor supervision of the cluster-to-attended-direction assignments). Altogether, we have confirmed the applicability of decoding the locus of auditory attention with CSP filters in a variety of audiovisual conditions. Our results also highlight the need for novel, generalization-robust (CSP) algorithms tailored to real-time applications, where attention needs to be continuously decoded to keep track of attention switches.

6. Acknowledgements

The authors are grateful to all the participants in this study, and to Anouck Jaspers and Koen van den Eeckhout for their help with data collection. The authors would also like to thank Debora Fieberg for the early brainstorming sessions and for providing the audio

stimuli used in this study. Financial support was provided by the Research Foundation Flanders (FWO) (SBO mandate 1S14922N for I. Rotaru, SBO mandate 1S31522N for N. Heintz, SBO mandate 1S34821N for I. Van de Ryck and FWO project G0A4918N), by KU Leuven through a PDM mandate (for S. Geirnaert, No. PDMT1/22/009), and by the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (Grant Agreements No. 637424 and 802895 for T. Francart and A. Bertrand, respectively). The authors declare no conflicts of interest.

References

Andersen, T. S., Tiippana, K., Laarni, J., Kojo, I. & Sams, M. (2009), 'The role of visual spatial attention in audiovisual speech perception', *Speech Communication* **51**(2), 184–193.

Arthur, D. & Vassilvitskii, S. (2006), k-means++: The advantages of careful seeding, Technical report, Stanford.

Bednar, A. & Lalor, E. C. (2018), 'Neural tracking of auditory motion is reflected by delta phase and alpha power of EEG', *NeuroImage* **181**, 683–691.

Bednar, A. & Lalor, E. C. (2020), 'Where is the cocktail party? Decoding locations of attended and unattended moving sound sources using EEG', *NeuroImage* **205**(116283).

Best, V., Boyd, A. D. & Sen, K. (2023), 'An effect of gaze direction in cocktail party listening', *Trends in Hearing* **27**, 23312165231152356.

Best, V., Ozmeral, E. J. & Shinn-Cunningham, B. G. (2007), 'Visually-guided attention enhances target identification in a complex auditory scene', *JARO - Journal of the Association for Research in Otolaryngology* **8**(2), 294–304.

Bishop, C. M. & Nasrabadi, N. M. (2006), *Pattern recognition and machine learning*, Vol. 4, Springer.

Blankertz, B., Kawanabe, M., Tomioka, R., Hohlefeld, F., Müller, K.-r. & Nikulin, V. (2007), 'Invariant common spatial patterns: Alleviating nonstationarities in brain-computer interfacing', *Advances in neural information processing systems* **20**.

Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M. & Müller, K. R. (2008), 'Optimizing spatial filters for robust EEG single-trial analysis', *IEEE Signal Processing Magazine* **25**(1), 41–56.

Das, N., Francart, T. & Bertrand, A. (2020), 'Auditory attention detection dataset kuleuven', *Zenodo* .

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

Farquhar, J., Hill, N. J., Lal, T. N. & Schölkopf, B. (2006), *Regularised CSP for sensor selection in BCI*.

Gehmacher, Q., Schubert, J., Schmidt, F., Hartmann, T., Reisinger, P., Roesch, S., Schwarz, K., Popov, T., Chait, M. & Weisz, N. (2023), 'Eye movements track prioritized auditory features in selective attention to natural speech', *bioRxiv* pp. 2023–01.

Geirnaert, S., Francart, T. & Bertrand, A. (2021), '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 Cheveigné, A., Lalor, E., Meyer, B. T., Miran, S., Francart, T. & Bertrand, A. (2021), 'Electroencephalography-based auditory attention decoding: Toward neurosteered hearing devices', *IEEE Signal Processing Magazine* **38**(4), 89–102.

Huang, G., Liu, G., Meng, J., Zhang, D. & Zhu, X. (2010), 'Model based generalization analysis of common spatial pattern in brain computer interfaces', *Cognitive neurodynamics* **4**(3), 217–223.

Kayser, H., Ewert, S. D., Anemüller, J., Rohdenburg, T., Hohmann, V. & Kollmeier, B. (2009), 'Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses', *EURASIP Journal on advances in signal processing* **2009**, 1–10.

Ledoit, O. & Wolf, M. (2004), 'A well-conditioned estimator for large-dimensional covariance matrices', *Journal of multivariate analysis* **88**(2), 365–411.

Lemm, S., Blankertz, B., Curio, G. & Muller, K.-R. (2005), 'Spatio-spectral filters for improving the classification of single trial eeg', *IEEE transactions on biomedical engineering* **52**(9), 1541–1548.

Lopez, A., Ferrero, F. J., Valledor, M., Campo, J. C. & Postolache, O. (2016), A study on electrode placement in eog systems for medical applications, *in* '2016 IEEE International symposium on medical measurements and applications (MeMeA)', IEEE, pp. 1–5.

Lotte, F., Bougrain, L., Cichocki, A., Clerc, M., Congedo, M., Rakotomamonjy, A. & Yger, F. (2018), 'A review of classification algorithms for EEG-based brain-computer interfaces: A 10 year update', *Journal of Neural Engineering* **15**(3).

Lotte, F. & Guan, C. (2010), 'Regularizing common spatial patterns to improve bci designs: unified theory and new algorithms', *IEEE Transactions on biomedical Engineering* **58**(2), 355–362.

Maddox, R. K., Pospisil, D. A., Stecker, G. C. & Lee, A. K. (2014), 'Directing eye gaze enhances auditory spatial cue discrimination', *Current Biology* **24**(7), 748–752.

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

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., Slaney, M., Shamma, S. A. & Lalor, E. C. (2014), 'Attentional

Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG', *Cerebral Cortex* **25**(7), 1697–1706.

Parra, L. C., Spence, C. D., Gerson, A. D. & Sajda, P. (2005), 'Recipes for the linear analysis of eeg', *Neuroimage* **28**(2), 326–341.

Patel, P., Long, L. K., Herrero, J., Mehta, A. D. & Mesgarani, N. (2018), 'Joint Representation of Spatial and Phonetic Features in the Human Core Auditory Cortex', *Cell Reports* **24**(8), 2051–2062.e2.

Pomper, U. & Chait, M. (2017), 'The impact of visual gaze direction on auditory object tracking', *Scientific Reports* **7**(1), 1–16.

Popov, T., Gips, B., Weisz, N. & Jensen, O. (2022), 'Brain areas associated with visual spatial attention display topographic organization during auditory spatial attention', *Cerebral Cortex* **1**, 12.

Reuderink, B. & Poel, M. (2008), 'Robustness of the common spatial patterns algorithm in the bci-pipeline', *University of Twente, Tech. Rep* .

Slaney, M., Lyon, R. F., Garcia, R., Kemler, B., Gnegy, C., Wilson, K., Kanevsky, D., Savla, S. & Cerf, V. G. (2020), 'Auditory Measures for the Next Billion Users', *Ear and hearing* **41**, 131S–139S.

Strauss, D. J., Corona-Strauss, F. I., Schroeer, A., Flotho, P., Hannemann, R. & Hackley, S. A. (2020), 'Vestigial auriculomotor activity indicates the direction of auditory attention in humans', *eLife* **9**(e54536).

Wöstmann, M., Herrmann, B., Maess, B. & Obleser, J. (2016), 'Spatiotemporal dynamics of auditory attention synchronize with speech', *Proceedings of the National Academy of Sciences of the United States of America* **113**(14), 3873–3878.