Why Performance Metrics Overpromise in Auditory Attention Decoding: an Information-Theoretic Reappraisal

Nicolas Heintz, Simon Geirnaert, Tom Francart, and Alexander Bertrand, *Senior Member, IEEE*

***Abstract*—Auditory attention decoding (AAD) algorithms are predominantly evaluated in a steady state where a listener continuously attends to the same speaker, using metrics such as accuracy and information transfer rate. However, such metrics fail to account for the (in-)dependence of an AAD prediction with respect to previous predictions. In this paper, we argue that failing to take this dependence into account in the algorithm evaluation can lead to severe misrepresentations of the true performance of an AAD algorithm.**

We therefore introduce the relative Incremental Mutual Information (rIMI); the rate at which a new prediction removes the remaining uncertainty about the identity of the attended speaker. This allows us to track how much new, useful information a prediction actually generates on top of the information already obtained from previous predictions. By investigating the rIMI and the behaviour of AAD models around attention switches, we demonstrate that recent direct-classification AAD algorithms are not superior to traditional AAD algorithms based on stimulus reconstruction, despite what accuracy alone may suggest. We also demonstrate how these direct-classification AAD predictions are severely influenced by irrelevant feature drifts, which artificially inflates accuracies by leaking information across windows, and even across trials.

***Index Terms*—auditory attention decoding, information theory, mutual information, performance metrics**

I. Introduction

When multiple people are talking simultaneously, our brain can specifically attend to a single speaker and ignore unattended speech [1]. This allows us to understand speech, even in extremely noisy scenarios. However, people suffering from hearing loss quickly lose this capability [2], [3]. While current hearing aids are increasingly effective at enhancing target speech, they often fail to identify the attended speaker in complex listening environments. A promising solution is to decode the attended speaker from the brain activity, for example based on non-invasive electroencephalography (EEG). This is known as auditory attention decoding (AAD) [4]–[6].

AAD algorithms are typically evaluated based on two key metrics: their time resolution and their decoding accuracy. In general, the resolution of an AAD algorithm is estimated based on the algorithm's decision window length, i.e. the amount of (EEG) data based on which a decision is made.

Traditionally, the auditory attention was decoded using stimulus reconstruction algorithms where an EEG decoder is trained to reconstruct the attended speech envelope from the neural responses in the EEG signals, either with linear or non-linear models. The attended speaker is then identified by computing the correlation between the decoded EEG (i.e., the reconstructed stimulus) and the speech envelope of each speaker. Due to the low SNR of this process, stimulus reconstruction algorithms typically require up to $10\,\mathrm{s}$ to generate a sufficiently accurate prediction of the attended speaker [4], [5], [7]. However, recent AAD research shifts more and more towards spatial or direct-classification algorithms, where the identity of the attended speaker or the locus of attention is directly decoded from EEG without explicitly correlating it to the speech signals [7]–[12]. Although such methods currently suffer from poor generalisation, and regularly overfit on irrelevant confounds such as eye gaze [13]–[15], there remains a strong focus on such approaches within the AAD literature. This focus is driven by their seemingly superior potential to achieve very high accuracies on very short window lengths ($0.1\,\mathrm{s}$ to $2\,\mathrm{s}$).

However, we show in this paper that this potential of direct-classification AAD algorithms to outperform stimulus reconstruction AAD algorithms is largely a mirage, driven by incomplete metrics that mask the true performance of an AAD algorithm. At the core of this mirage is the fact that consecutive direct-classification AAD predictions consistently demonstrate a high conditional self-dependence. This means that a large portion of the variance in the AAD predictions can be explained by previous predictions. We will demonstrate that, when the conditional self-dependence is not accounted for, the performance of an algorithm can be severely overestimated. We argue that this overestimation is further driven by the tendency to evaluate accuracy in sustained-attention experiments, where attention switches are absent or infrequent within an experiment trial. We show that, due to these temporal dependencies, accuracy measured under sustained-attention conditions is a poor predictor of performance in the vicinity of an attention switch.

N. Heintz, S. Geirnaert and A. Bertrand are with KU Leuven, Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics and also with Leuven.AI - KU Leuven institute for AI, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium (e-mail: nicolas.heintz@esat.kuleuven.be, simon.geirnaert@esat.kuleuven.be, alexander.bertrand@esat.kuleuven.be).

T. Francart is with KU Leuven, Department of Neurosciences, Research Group ExpORL, Herestraat 49 box 721, B-3000 Leuven, Belgium (e-mail: tom.francart@kuleuven.be).

The authors acknowledge the financial support of the FWO (Research Foundation Flanders) for project 1S31524N, G026026N, 1242524N, and G081722N, the Flemish Government (AI Research Program), and Internal Funds KU Leuven (project C3/25/017, IDN/23/006 and C14/25/108.

In this paper, we use information theory to study how ignoring the conditional self-dependence can lead to a misleading picture of the performance of an algorithm. We also propose a new metric, which allows us to separate repeated information in consecutive AAD predictions from actual, novel information. This aids us to study the true performance of several AAD algorithms, even in datasets where there are no or limited attention switches to evaluate the true performance at these switch events.

We explore in depth where the conditional self-dependence comes from, how it can affect performance, and how we can correct for it. However, most important is how the performance overestimations can be avoided in the future. We therefore propose a set of concrete guidelines for the evaluation of AAD algorithms, which are summarised below and further justified throughout the paper. While these guidelines emerge naturally from the information-theoretic analysis presented here, we highlight them already in the introduction, as we believe they constitute the paper's most important contribution.

In Section II, we clarify in more detail how ignoring the conditional self-dependence can lead to overestimated performances. In Section III, we formulate the information-theoretic metrics that can correct this overestimation. In Section IV, the core methods to estimate the information-theoretic metrics are explained. We further detail how all experiments were performed in Section V. Finally, the results are displayed and discussed in Section VI.

How to report AAD performances fairly

- **Create a buffer between training and test data.** Supervised training data and test data should be separated by as much time as possible. At least five minutes, but ideally over more than one test session. A LOTO cross-validation is often not enough, as temporal dependencies can sometimes persist across trials.
- **Validate around attention switches.** An AAD model should always be tested on data with attention switches (as many as feasible). The performance in steady state and around an attention switch should be documented separately.
- **Avoid confounds in the data.** Avoid data where confounds such as eye gaze or muscle artifacts can be abused to infer the attention state. Such artifacts cannot be completely removed with preprocessing.
- **Measure the rate of information gain.** The relative Incremental Mutual Information (rIMI) metric proposed in this paper can optionally be used to compare AAD algorithms, in particular when no or limited attention switches are present in the data. It estimates what fraction of the remaining uncertainty about the current attention state a prediction removes, addressing the shortcomings of traditional accuracy metrics when measured on sustained attention data.

II. How Temporal Dependence inflates Performance

At first glance, it may not be immediately evident why conditional self-dependence matters. *"If an algorithm can predict the identity of the attended speaker with high accuracy using a single short window, why would it matter if the prediction in the next window is conditionally dependent?"* To aid with this intuition, we identify three mechanisms where the performance of an algorithm can be incorrectly inflated by ignoring the conditional self-dependence, depending on the underlying cause of the dependencies.

A. An inflated temporal resolution

The first case is the most straightforward, and happens when the algorithm's AAD prediction scores do not change fast enough after an attention switch to be significantly different in consecutive windows. When the listener switches attention, it could take several decision windows before the prediction has fully flipped from the first to the second attended speaker. Since it would take more than 1 window to detect an attention switch, the temporal resolution is lower than the window length would suggest, even if you reach near-perfect accuracy in steady state on a single window. This is comparable to an oversampled signal that is sampled above the Nyquist rate.

A typical example of inflated temporal resolutions is when the auditory attention is predicted on $0.1\,\mathrm{s}$ windows, as often found in recent AAD papers [10], [11]. Even if you could reliably decode the auditory attention using these short windows in a steady state, it is unlikely that you could accurately track constant sub-second attention switches (unless explicitly proven otherwise). The brain cannot switch attention in such a short amount of time. Moreover, a $0.1\,\mathrm{s}$ resolution is impossible by construction, as the bandpass filters that are present in almost all preprocessing pipelines already smear information over multiple $0.1\,\mathrm{s}$ windows.

Another example is the postprocessing that is typically done to smoothen AAD predictions before applying it to a gain control algorithm [6], [16]–[19]. Although these algorithms typically start from short windows (e.g., $1\,\mathrm{s}$), they integrate multiple decisions over time such that it takes much longer to actually detect an attention switch [18], [19]. It would thus be unfair to merely report the accuracy of such postprocessing algorithms and compare them to the accuracies of other AAD algorithms on $1\,\mathrm{s}$ windows.

B. Temporal drift as a confounding feature

Conditionally dependent consecutive predictions are often caused by a temporal dependence in the feature space. This means that feature vectors observed in close succession, are also expected to be close to each other in the feature space. This temporal proximity can be driven by processes that are completely irrelevant to the attention process, such as slowly changing statistics of EEG due to the (background) neural processes, akin to Brownian motion [20].

Since AAD algorithms are almost exclusively evaluated on datasets with none or hardly any attention switches, feature

vectors that are recorded in close succession are extremely likely to belong to the same class. Although the feature drift itself is completely uninformative, it thus artificially clusters features that belong to the same attention class if attention to the same speaker is sustained for a long time [13]. It then suffices to know (a good prediction of) the label of a single feature per cluster to achieve a strong performance, even if the predictions themselves are barely modulated by the actual attention process.

The most blatant example of this process in practice is the usage of random cross-validation to train direct-classification models that directly assign an attention label to an EEG snippet (e.g. attend left versus attend right) without correlating the EEG explicitly with the speech stimuli. Random cross-validation assigns each EEG window randomly to either the train or test set. This gives the classifier access to many ground-truth labels of features recorded closely before and after each test window. If there is a significant amount of temporal drift present (which is the case for EEG-based classification, as we will discuss later), the classifier can then simply abuse the temporal proximity of features to 'copy' the training label to the test label [13], [15].

Another example is the usage of unsupervised, adaptive algorithms. These algorithms iteratively re-train the AAD classifier using generated pseudo-labels [21]–[23]. If the features are conditionally dependent, the AAD classifier gets biased to copy the most recent pseudo-labels in the new prediction. Although this is not problematic as such (the pseudo-labels are generated in an unsupervised way), it does create an indirect form of smoothing, which may decrease the temporal resolution as discussed above.

C. A decreased postprocessing potential

Raw per-window AAD predictions are rarely directly used in gain control algorithms. Instead, they are generally first fused over time using some form of postprocessing [6], [16]–[19]. This significantly decreases the misclassification rate, at the cost of a slightly decreased temporal resolution. However, when consecutive AAD predictions are already conditionally dependent on each other, the potential benefit of such post-processing algorithms is significantly reduced [19].

For example, consider a naive postprocessing where the decision scores of an AAD algorithm are smoothed using a weighted moving average filter. When these predictions are independent, averaging reduces the variance and thus improves the accuracy. However, when there are significant conditional dependencies, this regression to the mean happens much slower. Further improving the accuracy comes thus at a much greater cost in temporal resolution. Therefore, even if an AAD algorithm with independent predictions has a much lower accuracy than its counterpart with dependent predictions, it may still outperform the dependent predictions after postprocessing [19].

III. Using information to merge performance and independence

To avoid all the potential pitfalls of evaluating conditionally dependent predictions on steady state data, we would ideally wish to evaluate each AAD algorithm on a dataset where past data cannot be exploited to predict the current state. In other words, the measured participant should constantly switch attention, choosing a new target speaker at random for every window. However, this is practically infeasible: it would quickly lead to extreme exhaustion, where even the most motivated participant would drop off in mere minutes. Alternatively, it is possible to solely focus on the AAD accuracy achieved on predictions generated just after an attention switch to estimate the temporal resolution, but this would mean that only a very small percentage of a recording can be used for evaluation, again leading to noisy performance metrics. Furthermore, most existing AAD datasets don't have any attention switches, except for a few where switches in attention remain very sparse [24]. Finally, even if datasets with an abundance of attention switches would exist in the future, the exact moment a participant switches attention may always be off by a few seconds relative to the switching cue that defines the ground truth. As AAD algorithms improve, this label uncertainty would become an important source of variance, making exact comparisons hard.

However, it is possible to achieve a similar goal by analysing the performances through the lens of information theory. In general, we can model the AAD problem as a process with a hidden attention state $y(n)$ (the identity of the attended speaker at window n) and scores $\mathbf{f}(n)$. While the scores can be anything, ranging from raw samples of a recorded EEG to classifier outputs (or any feature representation in between), we will (without loss of generality) specifically focus on the classification scores generated by AAD algorithms. In this case, the scores $\mathbf{f}(n)$ can be viewed as a probability vector for each speaker being the attended speaker, or a one-hot vector with a hard decision. Given the hidden state $y(n)$, scores $\mathbf{f}(n)$, and past scores $\mathbf{f}_{past}(n) = [\mathbf{f}^\top(n-\tau)\dots\mathbf{f}^\top(n-1)]^\top,\ \tau > 0$, we will study four key metrics: Mutual Information (MI), Conditional Self-Dependence (CSD), Incremental MI (IMI), and relative IMI (rIMI), which will be explained in the remaining subsections, after explaining some generic information-theroetic concepts in the next subsection.

A. Entropy and information

Entropy is a measure of uncertainty, and is measured in bits. The entropy of a variable x is written as $H(x)$. If $H(x) = 0\,\text{bit}$, the value of x is known with absolute certainty. If $H(x) = 1\,\text{bit}$, there is as much uncertainty about the value of x as about the outcome of a coin toss. Formally, entropy is defined as:

$$H(x) = -\int \mathcal{P}(x)\log_2(\mathcal{P}(x))\mathrm{d}x, \tag{1}$$

where $\mathcal{P}(x)$ is the probability density function of x.

Conditional entropy $H(x|y)$ represents the amount of uncertainty about the value of x if y is known. If x and y are independent, $H(x|y) = H(x)$: knowledge about the value of y does not reduce the uncertainty about the value of x.

When there is a dependence between x and y, the conditional entropy can be computed as:

$$H(x|y) = -\int \mathcal{P}(x,y)\log_2\left(\frac{\mathcal{P}(x,y)}{\mathcal{P}(x)}\right)\mathrm{d}x\mathrm{d}y. \quad (2)$$

Mutual Information (MI) $I(x;y)$ represents how knowledge of y reduces the uncertainty about the value of x, and vice versa. Formally, the mutual information is defined as:

$$I(x;y) = H(x) - H(x|y) \quad (3)$$
$$= H(y) - H(y|x). \quad (4)$$

Conditional Mutual Information (CMI) $I(x;y|z)$ represents how knowledge of y reduces the uncertainty about the value of x (or vice versa) if z is already known. If x or y is strongly correlated with z, the CMI will be low. If x and y are completely independent of z, the CMI is identical to the MI $I(x;y)$. The conditional mutual information is defined as:

$$I(x;y|z) = I(x;y,z) - I(x;z), \quad (5)$$

where $I(x;y,z)$ represents the mutual information between x and the stacked variable (y,z).

B. Mutual Information (MI)

In the context of AAD, we are typically interested in the quantity $I(\mathbf{f}(n);y(n))$, which is the MI between the attention state $y(n)$ and the observation $\mathbf{f}(n)$. It measures how well you can predict the attention state given the observations and vice versa. This metric is very similar to the accuracy of an AAD algorithm and is regularly used to compute the so-called Information Transfer Rate (ITR) [25]. Crucially, $I(\mathbf{f}(n);y(n))$ does **not** take into account whether consecutive observations are dependent, and thus suffers from the same problems as accuracy.

C. Conditional Self-Dependence (CSD)

We define the conditional self-dependence (CSD) as $I(\mathbf{f}(n);\mathbf{f}_{past}(n)|y(n))$, which is the CMI between current observation $\mathbf{f}(n)$ and previous observations $\mathbf{f}_{past}(n)$ given the underlying attention state $y(n)$. The new observation is conditionally independent from previous observations if $I(\mathbf{f}(n);\mathbf{f}_{past}(n)|y(n)) = 0$, i.e., if the previous observations $\mathbf{f}_{past}(n)$ are unrelated to the current observation $\mathbf{f}(n)$, other than their shared attention state $y(n)$. This metric is similar to a partial correlation but also accounts for nonlinear dependencies [26].

D. Incremental Mutual Information (IMI)

We define the incremental mutual information (IMI) as $I(\mathbf{f}(n);y(n)|\mathbf{f}_{past}(n))$, which is the CMI between current observation $\mathbf{f}(n)$ and the attention state $y(n)$ given previous observations $\mathbf{f}_{past}(n)$. This metric is higher when the new observation provides more previously unavailable information about the identity of an attended speaker. This metric therefore allows us to separate novel information on the attention state from information that is carried over from previous predictions. If there is no CSD between previous and current predictions, the IMI is exactly equal to the MI.

E. relative Incremental Mutual Information (rIMI)

We define the rIMI in (6) as the fraction of the remaining uncertainty about the current attention state $y(n)$ that is removed by the current observation $\mathbf{f}(n)$.

$$rIMI \triangleq \frac{I(\mathbf{f}(n);y(n)|\mathbf{f}_{past}(n))}{H(y(n)|\mathbf{f}_{past}(n))}. \quad (6)$$

A perfect model has an rIMI of 1, while a random model or a model that merely repeats previous predictions has an rIMI of 0. This makes the rIMI much easier to interpret than the IMI, which has a variable upper bound dictated by the total amount of remaining state uncertainty $H(y(n)|\mathbf{f}_{past}(n))$.

IV. Estimating information-theoretic metrics

In this section, we slowly build towards the practical computation strategies for the four key metrics (MI, CSD, IMI, rIMI) that we will use to evaluate AAD algorithms, being the core focus of this work. To differentiate between general theoretical concepts and implementations that are specific to AAD, we use $\mathbf{x}(n)$ for observations in general, and $\mathbf{f}(n)$ specifically for AAD classification scores for each window n. This allows us to clearly separate theoretical concepts where we make Gaussianity assumptions on the distribution of $\mathbf{x}(n)$ from the practical implementations without this assumption at the end of this section.

A. Gaussian entropy

A key downside of information-theoretic metrics is that they are often hard to estimate. Estimating the entropy $H(\mathbf{x}(n))$ of a variable $\mathbf{x}(n)$ requires knowledge of the distribution $\mathcal{P}(\mathbf{x}(n))$, which is generally unknown. While these distributions can be estimated using binning methods, such approximations are often too inaccurate to provide a reliable estimation of the entropy [26]. However, in some rare cases, the entropy $H(\mathbf{x}(n))$ has a closed-form solution. One important example is when $\mathbf{x}(n)$ is sampled from a K-dimensional multivariate Gaussian distribution, i.e. $\mathbf{x}(n) \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma_{\mathbf{x}})$ [26]:

$$H(\mathbf{x}(n)) = \frac{1}{2}\log_2\left((2\pi e)^K|\Sigma_{\mathbf{x}}|\right). \quad (7)$$

The entropy of a Gaussian variable thus only depends on its dimensionality K and the determinant $|\Sigma_{\mathbf{x}}|$ of its covariance matrix $\Sigma_{\mathbf{x}}$.

This result can easily be expanded to compute the conditional entropy $H(\mathbf{x}(n)|y(n))$ of a variable $\mathbf{x}(n)$ which is sampled from a Gaussian mixture (instead of a single Gaussian), where the discrete state variable $y(n)$ indicates from which Gaussian distribution $\mathbf{x}(n)$ is sampled, i.e. $\mathbf{x}(n)|(y(n)=i) \sim \mathcal{N}(\boldsymbol{\mu}_i, \Sigma_{\mathbf{x},i})$:

$$H(\mathbf{x}(n)|y(n)) = \sum_i p_i H(\mathbf{x}(n)|y(n)=i) \quad (8)$$
$$= \sum_i p_i \frac{1}{2}\log_2\left((2\pi e)^K|\Sigma_{\mathbf{x},i}|\right), \quad (9)$$

with p_i the probability that $y(n)=i$.

While there is no closed-form solution for the unconditioned entropy $H(\mathbf{x}(n))$ of a variable $\mathbf{x}(n)$ sampled from a Gaussian

mixture, several fast algorithms exist that can accurately approximate $H(\mathbf{x}(n))$ for Gaussian mixtures. In this paper, we specifically use the algorithm proposed by Huber et al. [27].

These identities allow us to efficiently estimate the four metrics used in this paper. Using (4), we can compute the **mutual information** (MI) between a mixed Gaussian variable $\mathbf{x}(n)$ and the underlying state $y(n)$:

$$I(\mathbf{x}(n); y(n)) = H(\mathbf{x}(n)) - \sum_i p_i \frac{1}{2} \log_2 \left((2\pi e)^K |\Sigma_{\mathbf{x}_i}| \right). \tag{10}$$

This estimate for the mutual information can then be used to estimate the **incremental mutual information** $I(\mathbf{x}(n); y(n)|\mathbf{x}_{past}(n))$ based on (5):

$$I(\mathbf{x}(n); y(n)|\mathbf{x}_{past}(n)) = I(\mathbf{x}(n), \mathbf{x}_{past}(n); y(n)) - I(\mathbf{x}_{past}(n); y(n)). \tag{11}$$

The first term is the mutual information between the collection of both past and present observations $[\mathbf{x}^\top(n), \mathbf{x}_{past}^\top(n)]^\top$ and the attention state $y(n)$. The second term models how much we already knew about $y(n)$ from past observations.

Using the definition (6), the **relative incremental mutual information** is defined as the fraction between the IMI and the remaining uncertainty $H(y(n)|\mathbf{x}_{past}(n))$ about the state $y(n)$ given previous observations $\mathbf{x}_{past}(n)$. The numerator in (6) is the IMI, which is computed as specified in (11). The denominator of (6) becomes $H(y(n)|\mathbf{x}_{past}(n))$, which can be computed from (4) as:

$$H(y(n)|\mathbf{x}_{past}(n)) = H(y(n)) - I(\mathbf{x}_{past}(n); y(n)) \tag{12}$$

$$H(y(n)) \triangleq -\sum_i p_i \log_2 (p_i), \tag{13}$$

with p_i the prior probability that $y(n) = i$.

Finally, the **conditional self-dependence** $I(\mathbf{x}(n); \mathbf{x}_{past}(n)|y(n))$ between observations $\mathbf{x}(n)$ and $\mathbf{x}_{past}(n)$ given state $y(n)$ can be computed using (9):

$$I(\mathbf{x}(n); \mathbf{x}_{past}(n)|y(n)) \tag{14}$$

$$= H(\mathbf{x}(n)|y(n)) + H(\mathbf{x}_{past}(n)|y(n)) - H(\mathbf{x}(n), \mathbf{x}_{past}(n)|y(n)) \tag{15}$$

$$= \sum_i p_i \frac{1}{2} \log_2 \left(\frac{|\Sigma_{\mathbf{x},i}||\Sigma_{\mathbf{x}_{past},i}|}{|\Sigma_{\mathbf{x}\mathbf{x}_{past},i}|} \right). \tag{16}$$

Here, $H(\mathbf{x}(n), \mathbf{x}_{past}(n)|y(n))$ and $\Sigma_{\mathbf{x}\mathbf{x}_{past},i}$ represent the conditional entropy and the covariance matrix of the joint variable $[\mathbf{x}(n)^\top \mathbf{x}_{past}^\top(n)]^\top$, respectively.

B. The copula trick

Naturally, the scores $\mathbf{f}(n)$ generated by an AAD algorithm are not necessarily sampled from a mixed Gaussian distribution. However, it is possible to transform a non-Gaussian variable $\mathbf{f}(n)$ to a Gaussian variable $\mathbf{f}_\mathcal{G}(n)$ using the Gaussian copula [26]. This transformation works as follows:

- For each discrete label i and each dimension k of $\mathbf{f}$, rank all values $f_k(n)$ where $y(n) = i$ across all samples $n = [0, N]$.
- Replace each value $f_k(n)$ by its ranking.
- Normalise the rankings per dimension, such that they are constrained between $[\frac{1}{R+1}, \frac{R}{R+1}]$, with R the maximal ranking.
- $f_{\mathcal{G},k}(n)$ is then the inverse standard normal cumulative density function of the normalised rank of $f_k(n)$.

Ince et al. showed that both the mutual information $I(\mathbf{f}_\mathcal{G}(n); y(n))$ and the conditional self-dependence $I(\mathbf{f}_\mathcal{G}(n); \mathbf{f}_{\mathcal{G},past}(n)|y(n))$ of the Gaussian transformation provide lower bounds for the mutual information and conditional self-dependence of $\mathbf{f}(n)$ [26]. This allows us to use these fast Gaussian Copula Mutual Information estimators to estimate the key information-theoretic metrics defined in Section IV-A.

V. Experiments

To demonstrate how ignoring conditional self-dependence (e.g., by solely using the accuracy) can lead to misleading results, and how to correct for this using the (r)IMI metric, we outline the details of several experiments in this section. The results of these experiments are discussed in Section VI.

A. Dataset

All AAD algorithms were benchmarked on the audio-visual gaze controlled (AV-GC) dataset presented by Rotaru et al. [13], [24]. This AAD dataset was developed with the purpose of removing potential shortcuts due to eye artifacts in the EEG that are correlated with the locus of auditory attention [13], which is a shortcut that is almost always exploited by spatial or direct-classification AAD algorithms that directly classify the locus of auditory attention (i.e., left or right speaker attended). In this dataset, 13 participants are instructed to listen to one of two competing speakers. The competing speakers were located on the left and right sides of the head. In the meantime, the eye gaze was controlled using various methods, depending on the trial condition:

- **Moving video (MV)**: the participants were instructed to look at a randomly moving video of the attended speaker.
- **Moving target noise (MTN)**: the participants were instructed to look at a randomly moving cross-hair.
- **Static Video (SV)**: the participants were instructed to look at a video of the attended speaker located at the same side as the position of the attended speaker. This introduces a majore gaze-related artifact towards the side of auditory attention that can be used as a shortcut to perform AAD.
- **No visuals (NV)**: the participants were instructed to fixate on an imaginary point in front of them. For some participants, this still results in a slight yet exploitable gaze bias towards the side of auditory attention.

Although all conditions are used for training, we only use the MTN, NV and MV conditions for validation to ensure that gaze-related artifacts or biases in the EEG data cannot be used to decode the auditory attention.

Each trial lasts $10\,\mathrm{min}$, with an attention switch after $5\,\mathrm{min}$ (either left-to-right or right-to-left). Each condition is repeated twice: once with the left speaker first attended and once with

the right speaker first attended. For most participants, there are thus a total of eight $10\,\mathrm{min}$ trials. However, the MTN trials are not present in the first two participants.

The EEG is recorded with a 64-channel EEG BioSemi ActiveTwo system. The EEG and audio are then preprocessed using the preprocessing framework proposed by Biesmans et al. [28]. For more information about the dataset, we refer to [13].

B. AAD algorithms

To investigate how conditional self-dependence influences the performance of various AAD algorithms, we evaluated several AAD algorithms from the stimulus reconstruction and the direct-classification family. It is noted that the goal of this work is not to provide a benchmark across all existing AAD algorithms, but merely to illustrate how existing performance metrics are flawed when used in solitude, which is why we only selected a representative subset of AAD algorithms from both families.

Concerning the stimulus reconstruction algorithms, we evaluated the Least-Squares (LS) algorithm proposed by O'Sullivan et al. [4], the Canonical Correlation Analysis (CCA) algorithm with linear discriminant analysis (LDA) classifier proposed by de Cheveigné et al. [5], and the non-linear stimulus decoding algorithm AADNet proposed by Nguyen et al. [7].

Concerning the spatial/direct-classification AAD algorithms, we evaluated the linear common spatial patterns (CSP) algorithm with LDA classifier proposed by Geirnaert et al. [8], an adaptive (oracle) version of the latter, and the nonlinear Dual Attention Refinement Network (Darnet) proposed by Yan et al. [29]. The adaptive CSP algorithm uses oracle labels to update the LDA classifier over time, and should therefore be viewed as an upper-bound for the performance of an unsupervised adaptive CSP version[1]. Such an adaptive version of the algorithm would be able to partially compensate for feature drift in the CSP feature space, which is known to severely deteriorate performance [13]. The exact algorithm used to update the LDA classifier is specified in Appendix A.

Each algorithm was trained and tested using leave-one-trial-out (LOTO) cross-validation (CV). We then measure the accuracy, mutual information (MI), relative incremental mutual information (rIMI) and conditional self-dependence (CSD) per trial.

All information-theoretic metrics were derived by first combining the algorithm's classification scores $\mathbf{f}(n)$ across all windows and trials per subject, and then estimating the appropriate distributions for each subject using the Gaussian Copula, as explained in Section IV. All windows n where there was at least one attention switch between $n-\tau$ and n were removed. This ensures that the past data contained in $\mathbf{f}_{past}(n)$ are always relevant to predict $y(n)$, and avoids discrepancies when one would compare the (r)IMI across datasets with different amounts of attention switches.

[1]This could be achieved by using pseudo-labels generated by the unsupervised algorithms in [21] or [30].

The relevant hyperparameters for LS, CCA and CSP are shown in Table I. For all other algorithms, we used the exact hyperparameters proposed by the authors.

Algorithm	**Hyperparameter**	**Value**
All	Window length	$1\,\mathrm{s}$
LS	EEG lags L	$0\,\mathrm{ms}$ to $250\,\mathrm{ms}$
CCA	EEG lags L_{eeg}	$0\,\mathrm{ms}$ to $250\,\mathrm{ms}$
	Envelope lags L_{stim}	$-250\,\mathrm{ms}$ to $0\,\mathrm{ms}$
	Number of transformations K	5
CSP	Frequency bands	Full band ($0\,\mathrm{Hz}$ to $32\,\mathrm{Hz}$)
	Number of transformations K	5

TABLE I: The hyperparameters used in each AAD algorithm.

For LS, we used the difference between the attended and unattended correlation as $\mathbf{f}(n)$. For all other models, we used the one-dimensional classification score obtained at the output of the model to construct $\mathbf{f}(n)$.

C. Estimation of the information metrics

The mutual information and the conditional self-dependence were directly estimated using the GCMI toolbox [26]. The IMI and rIMI are then estimated using the identities specified in Section IV. These identities are also estimated with the GCMI toolbox.

The previous AAD predictions $\mathbf{f}_{past}(n)$ used to compute the (r)IMI and conditional self-dependence, consist of a vector that stacks all scores from all decision windows in the last $5\,\mathrm{s}$ (excluding the current prediction) for window lengths $\leq 5\,\mathrm{s}$. For longer window lengths, only the last AAD prediction was used. This allows us to correct for both short-term and long-term dependencies between scores. Ideally, as many previous predictions as possible should be used to correct for dependencies that persist for longer. However, this would significantly increase the dimensionality of $\mathbf{f}_{past}(n)$, where the curse of dimensionality would negatively affect the quality of the estimated information-theoretic metrics.

D. Experiments

In total, we will study the AAD algorithms through an information-theoretic lens based on four key experiments. However, to facilitate the analysis, and reduce clutter, only the first experiment is performed on all AAD algorithms. The other experiments are performed only with the LS and CSP algorithms. The observed trends for LS and CSP are representative for the other AAD algorithms within their respective classes (i.e., stimulus reconstruction and direct-classification AAD).

a) The relationship between accuracy and rIMI on synthetic data: To facilitate the interpretation of the novel rIMI metric, we first investigate the relationship between the accuracy and the rIMI on synthetic data. We consider two datasets. In the first dataset, 1000 observations $f(n)$ are independently sampled from one of two one-dimensional Gaussian distributions. The first 500 observations are sampled from the first distribution, the second 500 observations from the second. Both distributions have variance $\sigma_{1,2} = 1$. The distance between the class averages $\|\mu_1 - \mu_2\|$ is iteratively increased

from 0 to 4 to artificially improve the decoding accuracy. The first half of the observations receive label $y(n) = 1$, the second half $y(n) = 2$.

The second dataset is directly constructed from the first, with observations $f'(n) = f(n) + 1/3 \sum_{t=0}^{2} f(n-t)$. The current observation thus also depends on the two previous observations.

At each iteration of the dependent and independent datasets, the observations were classified to the class i that minimises $|f(n) - \mu_i|$. We then compute the accuracy and rIMI. The rIMI is conditioned on the past 3 samples. We then compare how the accuracy and rIMI changes for both the first dataset with independent samples, and the second dataset with conditionally dependent samples.

b) An information-theoretic comparison of AAD algorithms: The accuracy and the four information-theoretic metrics are computed on the AV-GC dataset for each AAD algorithm, window length and participant separately. For spatial AAD algorithms, both the leave-one-trial-out cross-validation and random cross-validation results are reported.

c) The accuracy and AAD scores around a switch: To better understand how spatial and stimulus reconstruction AAD algorithms behave around an attention switch, we compute the average evolution of the accuracy and classification scores $\mathbf{f}(n)$ over time around a switch of attention within the data. Ideally, the scores should change immediately after an attention switch and be significantly different before and after the switch. The accuracy should remain stable across the entire trial, with potentially only a short dip in accuracy around the switch, as there is always some time between an instructed switch and the participant actually switching attention. To limit noise, a relatively long $5\,\mathrm{s}$ window is used to compute the classification scores. However, using shorter window lengths did not affect the main conclusions.

The reported accuracies are the average accuracies across participants and trials at one specific moment in time. The scores are first aligned so that the cued attention switch always occurs at the same time, and the left speaker is first attended. When the opposite is true, the scores are multiplied by -1 to ensure alignment. The scores are then averaged over all trials and subjects. Finally, this result is normalised to facilitate the comparison between algorithms.

d) The origin of dependencies: In the last experiment, we study the origin of the prediction dependencies. As will be demonstrated in Section VI, the conditional self-dependency between consecutive predictions is significantly higher for spatial direct-classification AAD algorithms than stimulus reconstruction algorithms. We hypothesise that this is because covariance structure used by stimulus reconstruction algorithms is less dependent than the covariance structure used by direct-classification algorithms.

Stimulus reconstruction algorithms generate predictions based on the interactions between the EEG signal $\mathbf{m}(t)$ and the speaker envelopes $\mathbf{s}(t)$, within a certain window n. For instance, the LS decoder $\mathbf{d}_{LS}$ generates features proportional to $\mathbf{d}_{LS}^{\top}\mathbf{m}(n)s(n)^{\top} \propto \mathbf{d}^{\top}R_{ms}(n)$, with $R_{ms}(n)$ the sample crosscorrelation matrix in a window n. Similarly, CSP decoders $\mathbf{d}_{CSP}$ produce features proportional to $\mathbf{d}_{CSP}^{\top}R_{mm}(n)\mathbf{d}_{CSP}$, with $R_{mm}(n)$ the sample autocorrelation matrix in window n.

We therefore study the conditional dependencies $I(R_{mm}(i,j,n); R_{mm}(i,j,n-1)|y(n))$ and $I(R_{ms}(i,j,n); R_{ms}(i,j,n-1)|y(n))$ for each i and j. If this dependency is high, this means that consecutive predictions are (partially) made based on data that haven't significantly changed. Assuming the prediction model is linear or sufficiently smooth, these dependent data samples will then be translated into conditionally dependent predictions.

VI. Results and Discussion

A. The relationship between accuracy and rIMI on synthetic data

Figure 1 shows the rIMI versus accuracy, each showing an increasing trend when the class separation is increased. The additional smoothing that happens in the dependent dataset strongly affects the accuracy, but not the rIMI. This is desired: the rIMI only measures how much *new* information an observation contains. Temporal smoothing only transfers information that was already available to the current observation, and thus cannot increase the amount of new information that an observation contains, i.e., the rIMI should not change between the two datasets.

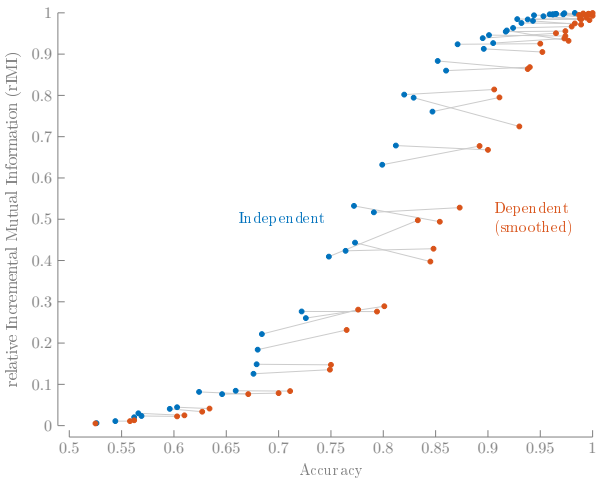


Fig. 1: The relation between the accuracy and rIMI for increasing clas separation in both the independent and dependent dataset. While temporal smoothing can improve the accuracy, it has on average no effect on the rIMI.

B. An information-theoretic comparison of AAD algorithms

Figure 2 shows the accuracy, Mutual Information (MI), Conditional Self-Dependency (CSD) and relative Incremental Mutual Information (rIMI) for each tested model on $2\,\mathrm{s}$ windows. The accuracy and MI (as used in the popular information transfer rate metric) do not take into account that predictions may be dependent.

The CSD measures to what extent consecutive predictions are dependent. A higher dependence means that a larger portion of the prediction is repeated from earlier predictions. Here we see a clear difference between stimulus reconstruction

methods (LS [4], CCA [5] and AADnet [7]), and direct-classification AAD methods such as CSP with static and adaptive classifiers [8], and Darnet [29]. The CSD is near-zero for the stimulus reconstruction algorithms, and much higher for the direct-classification methods. While a high CSD is not necessarily problematic, it does imply that (1) the accuracy cannot be easily improved through postprocessing, such as temporal smoothing [19], (2) the effective temporal resolution may be lower than what the window length suggests, and (3) that the model may be prone to overfitting on temporal fingerprints. This means that the model simply decodes *when* a certain test window was recorded, and infers from nearby training data what its corresponding class is. This is viable strategy when attention switches are rare and there is training data available that was recorded shortly before or after the test window (e.g., when using random cross-validation, adaptive decoders, or when some training trials were recorded shortly before or after the test trial).

The rIMI specifically measures the fraction of remaining uncertainty about the identity of the attended speaker that is removed by a new prediction. A rIMI of 0.02 means that each prediction approximately removes 2% of the remaining uncertainty about the attention state. A model that merely parrots previous predictions has a rIMI equal to 0, no matter its accuracy, as it fails to generate any new information. A good model should obtain both a high accuracy and a high rIMI. When the accuracy is high, but the rIMI is low (such as CSP with adaptive LDA), this indicates that the high accuracy is primarily driven by repeating previous predictions, rather than generating actual new information. This is problematic from the moment there is an actual switch in attention, as we will see later on. We further explore how a model such as adaptive CSP can obtain such a high accuracy, but low rIMI, and why this is a clear sign of overfitting in Section VI-C.

In Figure 3, we also report the accuracy and rIMI in function of the window length for LS and CSP with adaptive and static classifier. The figure demonstrates that there is a limit to reducing the window length in a meaningful way. Although the accuracy stays significantly above chance level for all models on very short decision windows, the rIMI quickly approaches 0. On such short windows, the biological processes that are captured vary too slowly to truly capture new information in consecutive windows. While spatial AAD algorithms are often reported to achieve high accuracies on $0.1\,\mathrm{s}$ windows [10], [11], [29], this does not mean that these algorithms would actually be able to detect several sub-second attention switches. At such time scales, information is already smeared over several windows by various preprocessing filters, even if we ignore the inherent latencies in the biological processes. So while the accuracy and MI may remain high (when the classifier is trained on trial-specific data), the actual rate of new information per window nears 0. Solely reporting the accuracy on short windows in steady state can thus misrepresent the true temporal resolution that the algorithm achieves.

Finally, the high conditional self-dependence between consecutive decisions for direct-classification algorithms also explains why their accuracy does not tend to improve with increasing window length, as shown in Figure 3a and in [8]–[11], [29]. In linear models, using longer windows is mathematically similar to averaging many shorter windows. When predictions are independent, averaging them leads to a reduction in variance, and thus an improved accuracy. However, when the predictions are conditionally dependent, the variance decreases at a much slower rate.

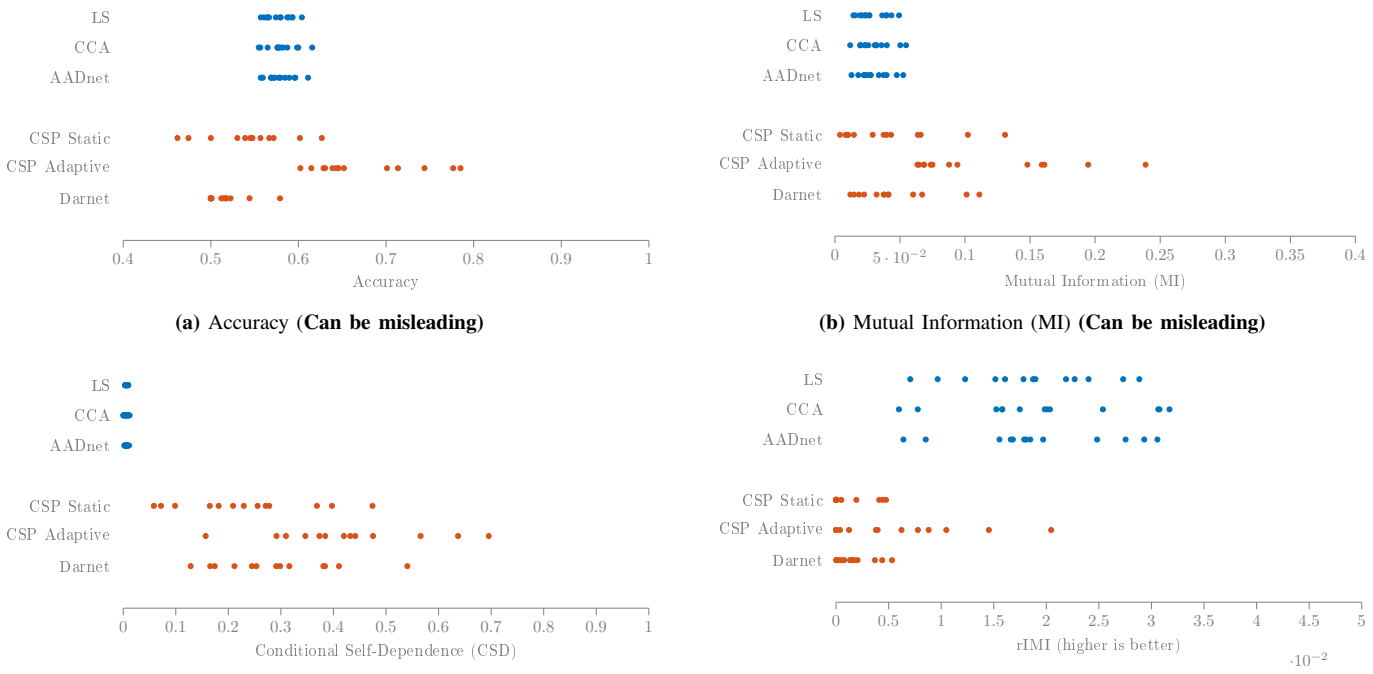
C. The accuracy and AAD scores around a switch

Figure 4 shows the average evolution of the features and classification scores before and after an attention switch. The models were trained such that, for a perfect model, the features and classification scores should be high at the start, and immediately drop after the attention switch[2]. This is, on average, the case for stimulus reconstruction models such as LS: there is a clear, sudden drop after the switch. However, the drop is not instant. This can be expected: we only know when the attention switch was cued. It is normal for a participant to take several seconds before they completed the attention switch, and it can be expected that the brain takes a few seconds to adapt to the new direction of attention.

However, there is no clear effect of an attention switch on the feature values of CSP. Some features (such as f_1) display inconsistent changes across trials (resulting in a flat slope on average), while other features (such as f_3) display clear slopes that are consistent across trials. Although we cannot say with certainty what causes these slopes, we hypothesise they are caused by temporal effects such as tiredness, or the aging of the electrode electrolyte, which causes the β-band filtered CSP features [8] to change over time. However, it is certain that the slopes are not influenced by attention, as the attention switch has no effect on the direction or slope of the drift.

Depending on the classifier, these feature drifts can be abused in different ways, as shown in Figure 4b. The adaptive LDA classifier has access to the (oracle) pseudo-labels of the most recent CSP features. Due to the feature drifts, a new feature vector will lay most closely to the most recent features [13], and thus be assigned to the same class. Since attention switches are rare, this is most often correct, which allows the model to obtain a high accuracy, even if the features are in no way influenced by attention. Indeed, immediately after the attention switch, most feature vectors are still assigned to the first class. It is only after a sufficient amount of the most recent feature vectors in the training set belong to the second class, that new test features are also assigned to that class, leading to a high accuracy but a slow reaction to attention switches. The CSP model with adaptive LDA thus abuses the slow drifts in the features (which are not attention-related) to merely parrot the pseudo-labels provided to the LDA classifier. Since the pseudo-labels are in this case perfect (oracle labels), the accuracy is very high. But the only new information in the scores is driven by 'improving' the adaptive LDA such that it

[2]All models were trained such that a feature value and classification score should be maximised for the left speaker, and minimised for the right speaker. For trials where the attention switches from the right speaker to the left speaker, the feature values and scores are multiplied by -1. This ensures that for every trial, the features and scores should start high, and drop to low values after the switch. The features and classification scores were then normalised such that that the average is 0 and the standard deviation is 1 for each model.



(a) Accuracy (**Can be misleading**)

(b) Mutual Information (MI) (**Can be misleading**)

(c) Conditional Self-Dependence (CSD)

(d) relative Incremental Mutual Information (rIMI)

Fig. 2: The Accuracy, Mutual Information (MI), conditional self-dependence (CSD) and relative Incremental Mutual Information (rIMI) for various stimulus reconstruction and spatial AAD algorithms on 2 s windows. While CSP with adaptive classifier appears to perform better at first glance, based on traditional metrics like accuracy and MI (used in ITR), its performance is severely overestimated due to the conditional self-dependence between consecutive predictions. The rIMI corrects for this dependence by only measuring how much new information a model generates per window. This is significantly higher for all stimulus reconstruction algorithms.

better reflects the feature drift within the trial. This causes the low, but non-zero rIMI.

Remarkably, while LOTO cross-validation is often considered to be insensitive to overfitting effects due to within-trial feature drift [13]–[15], we observe in Figure 4b that the static CSP scores also show a trend that seems to partially overlap with the change in attention (although the trend starts way before the switch such that it can not be attributed to attention)., and vice versa. Although this is far from consistent across trials, it suffices to obtain an accuracy slightly above chance level (see Figure 3), despite its near-zero rIMI and the absence of any notable effect of attention on the features.

This demonstrates how even the accuracy obtained with LOTO cross-validation can be driven by a subtle form of information leakage from train to test set. With sufficiently complex models, such leakage can be exploited as soon as there is any predictable pattern in class labels across trials (which is typically the case because datasets are designed to be balanced across trials).

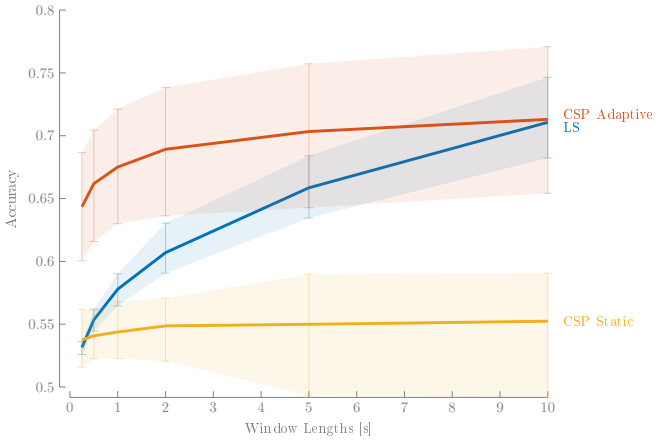
D. The origin of the dependencies

In this last experiment, we wish to investigate why spatial and direct-classification AAD algorithms have an increased conditional dependency, which historically has led to their overestimated performances. This is done by investigating their raw inputs; the per-window autocorrelation matrix $R_{mm}(n)$ for spatial and direct-classification AAD algorithms and crosscorrelation matrix $R_{my}(n)$ for stimulus reconstruction algorithms.

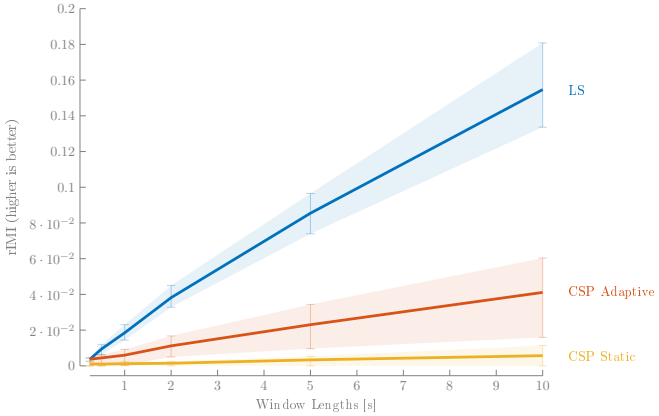
On average, the conditional dependency $I(R_{mm}(i,j,n); R_{mm}(i,j,n-1)|y(n))$ of the individual elements of the sample autocorrelation matrix $R_{mm}(n)$ is $0.04\,\text{bit}$. For the sample crosscorrelation matrix $R_{my}(n)$, the conditional dependency $I(R_{my}(i,j,n); R_{my}(i,j,n-1)|y(n))$ is on average $0.02\,\text{bit}$. This confirms that there is (on average) more temporal dependency in the statistics used by spatial algorithms (the sample autocorrelation matrix) compared to stimulus reconstruction algorithms.

However, the distinctions become more apparent when the individual CSDs $I(R_{mm}(i,j,n); R_{mm}(i,j,n-1)|y(n))$ and $I(R_{my}(i,j,n); R_{my}(i,j,n-1)|y(n))$ are inspected in closer detail. In the crosscorrelation matrix, the dependencies are roughly equally distributed, and vary between $0.01\,\text{bit}$ to $0.03\,\text{bit}$. On the other hand, there is a clear pattern in the autocorrelation matrix, as shown in Figure 5: the more frontal a channel is located, the more its autocorrelation and crosscorrelations with other frontal channels are temporally dependent. The conditional self-dependency ranges between $0.1\,\text{bit}$ to $0.2\,\text{bit}$ for all frontal channels, an order of magnitude larger than the conditional dependency present in other channels. If these frontal channels are excluded, the average conditional dependency $I(R_{mm}(i,j,n); R_{mm}(i,j,n-1)|y(n)),\ i,j \notin F$ drops to $0.025\,\text{bit}$, with F containing the channels with prefix 'F'.

This suggests that the majority of the conditional dependencies present in the features from spatial AAD algorithms stem from slow-moving processes in the frontal lobe. While eye

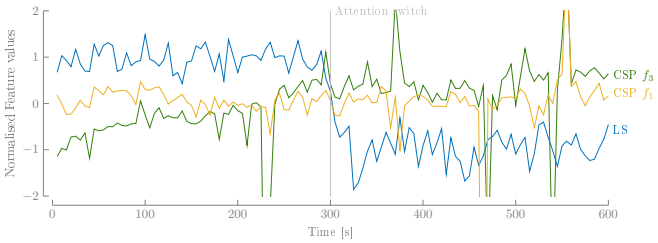


(a) Accuracy (**Can be misleading**)

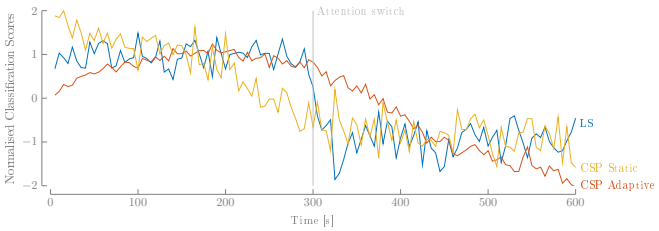


(b) relative Incremental Mutual Information (rIMI)

Fig. 3: The accuracy and rIMI per window length. Although the spatial CSP model with adaptive classifier seems to outperform the LS stimulus reconstruction model, it generates significantly less new information per window than LS.



(a) The average evolution of LS scores and the first and third CSP features after normalisation.



(b) The average evolution of LS and CSP classification scores with static and adaptive LDA classifier after normalisation.

Fig. 4: The average evolution of LS and CSP features and classification scores across a trial (for LS, we use the correlation coefficient both as the feature and classification score). At the attention switch, there is a clear jump in feature values for stimulus reconstruction algorithms such as LS. This is absent for direct-classification algorithms such as CSP. Instead, there is only a slow drift that can be misused by the classifier to obtain accuracies above chance level, despite the absence of any attention-driven effect.

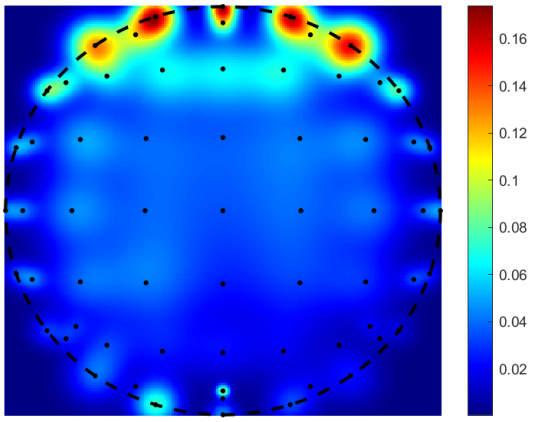


Fig. 5: The distribution of the conditional self-dependency (CSD) of EEG across channels. More frontal channels display a significantly higher CSD.

gaze is one potential source of these dependencies, such long-range temporal correlations (LRTC) also occur when eye gaze is controlled, or eyes are closed, and are thus likely driven by complex neural interactions [31]. These dependencies remain significant for up to 200 s, and thus enables information leakage across trials when they are recorded in quick succession.

Because the CSD of consecutive predictions in CSP is caused by the underlying CSD of the sample autocorrelation matrix $R_{mm}(n)$, the overestimated performances obtained by ignoring CSD are unlikely to be limited to the algorithms discussed in this paper. Instead, it is expected to affect all or almost all spatial or other direct-classification AAD algorithms. While this hypothesis should be rigorously verified in future work, there are clear signs that this phenomenon is indeed inherent to the broader framework of spatial and direct-classification algorithms. For instance, the reported accuracies of such direct-classification AAD algorithms typically barely increase for longer window lengths [6], [8], [9], [11], [32]–[34]. They also consistently drop in performance when tested on LOTO CV [13]–[15]. As clarified previously, this strongly indicates that the performance is largely driven by irrelevant feature drift (akin to Brownian motion), rather than an attention-driven component. In contrast, the accuracy of stimulus reconstruction algorithms consistently increases with increasing window length, no matter whether the stimulus was reconstructed using linear or non-linear models [5], [7], [12], [35]. Note that hybrid AAD algorithms, which directly classify

a combination of EEG and envelopes by minimising classification loss, also behave like a direct-classification model, and are equally affected by these dependencies. An example of such a hybrid model is the model proposed by Cicarelli et al. [32]. Similar to EEG-based direct-classification algorithms, the reported accuracies of these models barely increase for longer window lengths.

Finally, although the high conditional dependencies in EEG are an important driver of the conditional dependencies in spatial AAD predictions, they are not necessarily the only driver. Various model architectures, such as LSTM's, state space machines and recurrent neural networks can implicitly include a 'memory' component, which also create conditional dependencies across consecutive predictions. Although these dependencies are often desired, it is important to compare such algorithms fairly to AAD algorithms where such components are not (yet) added, and to take note of both the steady state performance and the temporal resolution around an attention switch, or use the rIMI metric when insufficient attention switches are available.

VII. Conclusion

AAD algorithms are currently (almost) uniquely validated by measuring their accuracy or ITR in function of the window length during a steady state, where a listener continuously listens to the same speaker without or with limited intra-trial attention switches. However, such metrics can be misleading, as they implicitly assume conditional independence between consecutive AAD predictions. This assumption is often violated, especially by direct-classification AAD algorithms that directly classify the EEG without explicitly relating it to the speech signal. When these conditional dependencies are not accounted for, the true temporal resolution of an algorithm can be severely overestimated.

To correct for these dependencies, we have proposed a new metric called relative Incremental Mutual Information (rIMI), which measures what fraction of the remaining uncertainty about the attention state a new prediction removes, given all previous predictions. This metric rewards predictions that are both accurate and provide new information not previously available. It demonstrates that direct-classification AAD algorithms remove uncertainty at a much lower rate than stimulus reconstruction algorithms in steady state. By also investigating the behaviour of these AAD models around an attention switch, we have discovered that the higher apparent performance of direct-classification AAD higher accuracy is a mirage obtained by exploiting temporal drift, which leaks information from training to test data. This information leakage is even possible when leave-one-out cross-validation is used, a cross-validation scheme that is generally assumed to avoid inflated accuracies due to overfitting to feature drift.

The conditional self-dependence present in direct-classification algorithms is caused by Brownian-like drift of the EEG statistics, which is naturally present in EEG recordings. The overestimated performances reported in this paper can thus affect any EEG decoding paradigm where the decoded states rarely change.

Appendix A
The adaptive CSP algorithm

The adaptive CSP algorithm uses the same CSP filters as the static CSP algorithm [8], which is trained in a supervised fashion using leave-one-trial-out cross-validation. However, the LDA classifier is adaptively retrained with oracle labels. The adaptive algorithm thus merely serves as an upper bound of what a practical unsupervised adaptive CSP algorithm could achieve. Given a CSP feature vector $\mathbf{x}(n)$ that belongs to either to class $\mathcal{C}_+$ or $\mathcal{C}_-$, we update the estimated class averages $\boldsymbol{\mu}_\pm$, class covariances $\Sigma_\pm$, LDA decoder $\mathbf{d}$ and threshold θ as follows:

$$
\begin{aligned}
&\text{if } n \in \mathcal{C}_+ : \\
&\quad \boldsymbol{\mu}_+ \leftarrow \lambda \boldsymbol{\mu}_+ + (1-\lambda)\mathbf{x}(n) \\
&\quad \tilde{\mathbf{x}}(n) = \mathbf{x}(n) - \boldsymbol{\mu}_+ \\
&\quad \Sigma_+ \leftarrow \lambda \Sigma_+ + (1-\lambda)\tilde{\mathbf{x}}(n)\tilde{\mathbf{x}}^\top(n) \\
&\quad \boldsymbol{\mu}_- \leftarrow \boldsymbol{\mu}_- \\
&\quad \Sigma_- \leftarrow \Sigma_- \\
&\text{else if } n \in \mathcal{C}_- : \\
&\quad \boldsymbol{\mu}_- \leftarrow \lambda \boldsymbol{\mu}_- + (1-\lambda)\mathbf{x}(n) \\
&\quad \tilde{\mathbf{x}}(n) = \mathbf{x}(n) - \boldsymbol{\mu}_- \\
&\quad \Sigma_- \leftarrow \lambda \Sigma_- + (1-\lambda)\tilde{\mathbf{x}}(n)\tilde{\mathbf{x}}^\top(n) \\
&\quad \boldsymbol{\mu}_+ \leftarrow \boldsymbol{\mu}_+ \\
&\quad \Sigma_+ \leftarrow \Sigma_+ \\
&\mathbf{d} = (\Sigma_+ + \Sigma_-)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\
&\theta = \mathbf{d}^\top(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2
\end{aligned}
$$

The classification score $f(n+1)$ of the next window is then computed as:

$$f(n+1) = \mathbf{d}^\top \mathbf{x}(n+1) - \theta. \tag{17}$$

The class averages $\boldsymbol{\mu}_\pm$ and covariances $\Sigma_\pm$ are initialised using all left-out trials. The forgetting factor λ is set to $\frac{N_{eff}/\tau - 1}{N_{eff}/\tau + 1}$, with N_{eff} the equivalent sliding window length equal to $600\,\mathrm{s}$, and τ the decision window length.

References

[1] E. C. Cherry, "Some Experiments on the Recognition of Speech, with One and with Two Ears." *Journal of the Acoustical Society of America*, vol. 25, pp. 974–979, 1953.

[2] B. G. Shinn-Cunningham and V. Best, "Selective Attention in Normal and Impaired Hearing," *Trends in Amplification*, vol. 12, no. 4, pp. 283–299, 2008.

[3] A. Presacco, J. Z. Simon, and S. Anderson, "Speech-in-noise representation in the aging midbrain and cortex: Effects of hearing loss," *PLOS ONE*, vol. 14, no. 3, p. e0213899, Mar. 2019.

[4] J. A. O'Sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, and E. C. Lalor, "Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG," *Cerebral Cortex*, vol. 25, no. 7, pp. 1697–1706, 2015.

[5] A. de Cheveigné, D. E. Wong, G. M. Di Liberto, J. Hjortkjær, M. Slaney, and E. Lalor, "Decoding the auditory brain with canonical component analysis," *NeuroImage*, vol. 172, pp. 206–216, 2018.

[6] S. Geirnaert, S. Vandecappelle, E. Alickovic, A. de Cheveigné, E. Lalor, B. T. Meyer, S. Miran, T. Francart, and A. Bertrand, "Electroencephalography-Based Auditory Attention Decoding: Toward Neurosteered Hearing Devices," *IEEE Signal Processing Magazine*, vol. 38, no. 4, pp. 89–102, 2021.

[7] N. D. T. Nguyen, H. Phan, S. Geirnaert, K. Mikkelsen, and P. Kidmose, "AADNet: An End-to-End Deep Learning Model for Auditory Attention Decoding," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 33, pp. 2695–2706, 2025.

[8] S. Geirnaert, T. Francart, and A. Bertrand, "Fast EEG-based decoding of the directional focus of auditory attention using common spatial patterns," *IEEE Transactions on Biomedical Engineering*, 2020.

[9] ——, "Riemannian Geometry-Based Decoding of the Directional Focus of Auditory Attention Using EEG," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2021, pp. 1115–1119.

[10] Z. Zhang, G. Zhang, J. Dang, S. Wu, D. Zhou, and L. Wang, "EEG-based Short-time Auditory Attention Detection using Multi-task Deep Learning," in *Interspeech 2020*, 2020, pp. 2517–2521.

[11] E. Su, S. Cai, L. Xie, H. Li, and T. Schultz, "STAnet: A Spatiotemporal Attention Network for Decoding Auditory Spatial Attention From EEG," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 7, pp. 2233–2242, Jul. 2022.

[12] G. Sridhar, S. Boselli, M. A. Skoglund, B. Bernhardsson, and E. Alickovic, "Improving auditory attention decoding in noisy environments for listeners with hearing impairment through contrastive learning," *Journal of Neural Engineering*, vol. 22, no. 3, p. 036041, Jun. 2025.

[13] I. Rotaru, S. Geirnaert, N. Heintz, I. Van de Ryck, A. Bertrand, and T. Francart, "What are we really decoding? Unveiling biases in EEG-based decoding of the spatial focus of auditory attention," *Journal of Neural Engineering*, vol. 21, no. 1, p. 016017, Feb. 2024.

[14] G. Ivucic, S. Pahuja, F. Putze, S. Cai, H. Li, and T. Schultz, "The Impact of Cross-Validation Schemes for EEG-Based Auditory Attention Detection with Deep Neural Networks," in *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Jul. 2024, pp. 1–4.

[15] Y. Yan, X. Xu, H. Zhu, S. Li, B. Wang, X. Wu, and J. Chen, "Overestimated performance of auditory attention decoding caused by experimental design in EEG recordings," in *Proc. Interspeech 2025*, 2025, pp. 1053–1057.

[16] A. Presacco, S. Miran, B. Babadi, and J. Z. Simon, "Real-Time Tracking of Magnetoencephalographic Neuromarkers during a Dynamic Attention-Switching Task," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. Institute of Electrical and Electronics Engineers Inc., Jul. 2019, pp. 4148–4151.

[17] J. Hjortkjær, D. D. E. Wong, A. Catania, J. Märcher-Rørsted, E. Ceolini, S. A. Fuglsang, I. Kiselev, G. Di Liberto, S.-C. Liu, T. Dau, M. Slaney, and A. de Cheveigné, "Real-time control of a hearing instrument with EEG-based attention decoding," *Journal of Neural Engineering*, vol. 22, no. 1, p. 016027, Feb. 2025.

[18] N. Heintz, S. Geirnaert, I. Van de Ryck, T. Francart, and A. Bertrand, "Probabilistic Gain Control in a Multi-Speaker Setting Using EEG-Based Auditory Attention Decoding," in *2024 32nd European Signal Processing Conference (EUSIPCO)*, Aug. 2024, pp. 892–896.

[19] N. Heintz, T. Francart, and A. Bertrand, "Post-processing of EEG-based Auditory Attention Decoding Decisions via Hidden Markov Models," *arXiv preprint arXiv:2506.24024*, Jun. 2025.

[20] X. Xu, B. Wang, B. Xiao, Y. Niu, Y. Wang, X. Wu, and J. Chen, "Beware of Overestimated Decoding Performance Arising from Temporal Autocorrelations in Electroencephalogram Signals," *arXiv preprint arXiv:.2405.17024*, May 2024.

[21] S. Geirnaert, T. Francart, and A. Bertrand, "Unsupervised Self-Adaptive Auditory Attention Decoding," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 10, pp. 3955–3966, 2021.

[22] ——, "Time-adaptive Unsupervised Auditory Attention Decoding Using EEG-based Stimulus Reconstruction," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 8, 2022.

[23] N. Heintz, T. Francart, and A. Bertrand, "Unsupervised EEG-based decoding of absolute auditory attention with canonical correlation analysis," *arXiv preprint arXiv:2504.17724*, Apr. 2025.

[24] I. Rotaru, S. Geirnaert, A. Bertrand, and T. Francart, "Audiovisual, Gaze-controlled Auditory Attention Decoding Dataset KU Leuven (AV-GC-AAD)," Apr. 2024.

[25] S. Geirnaert, T. Francart, and A. Bertrand, "An Interpretable Performance Metric for Auditory Attention Decoding Algorithms in a Context of Neuro-Steered Gain Control," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 1, pp. 307–317, 2020.

[26] R. A. Ince, B. L. Giordano, C. Kayser, G. A. Rousselet, J. Gross, and P. G. Schyns, "A statistical framework for neuroimaging data analysis based on mutual information estimated via a gaussian copula," *Human Brain Mapping*, vol. 38, no. 3, pp. 1541–1573, 2017.

[27] M. F. Huber, T. Bailey, H. Durrant-Whyte, and U. D. Hanebeck, "On entropy approximation for Gaussian mixture random vectors," in *2008 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*. Seoul, Korea (South): IEEE, Aug. 2008, pp. 181–188.

[28] W. Biesmans, N. Das, T. Francart, and A. Bertrand, "Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 5, pp. 402–412, 2017.

[29] S. Yan, C. fan, H. Zhang, X. Yang, J. Tao, and Z. Lv, "DARNet: Dual Attention Refinement Network with Spatiotemporal Construction for Auditory Attention Detection," *arXiv preprint arXiv:2410.11181*, Nov. 2024.

[30] N. Heintz, S. Geirnaert, T. Francart, and A. Bertrand, "Unbiased Unsupervised Stimulus Reconstruction for EEG-Based Auditory Attention Decoding," in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Institute of Electrical and Electronics Engineers (IEEE), 2023.

[31] K. Linkenkaer-Hansen, V. V. Nikouline, J. M. Palva, and R. J. Ilmoniemi, "Long-Range Temporal Correlations and Scaling Behavior in Human Brain Oscillations," *The Journal of Neuroscience*, vol. 21, no. 4, pp. 1370–1377, Feb. 2001.

[32] G. Ciccarelli, M. Nolan, J. Perricone, P. T. Calamia, S. Haro, J. O'Sullivan, N. Mesgarani, T. F. Quatieri, and C. J. Smalt, "Comparison of Two-Talker Attention Decoding from EEG with Nonlinear Neural Networks and Linear Methods," *Scientific Reports*, vol. 9, 2019.

[33] S. Vandecappelle, L. Deckers, N. Das, A. H. Ansari, A. Bertrand, and T. Francart, "EEG-based detection of the locus of auditory attention with convolutional neural networks," *eLife*, vol. 10, p. e56481, Apr. 2021.

[34] I. R. Khan, S.-L. Peng, R. Mahajan, and R. Dey, "Short-window EEG-based auditory attention decoding for neuroadaptive hearing support for smart healthcare," *Neuroscience Informatics*, vol. 5, no. 3, p. 100222, Sep. 2025.

[35] J. O'Sullivan, Z. Chen, J. Herrero, G. M. McKhann, S. A. Sheth, A. D. Mehta, and N. Mesgarani, "Neural decoding of attentional selection in multi-speaker environments without access to clean sources." *Journal of neural engineering*, vol. 14, no. 5, p. 056001, 2017.