Unsupervised Accuracy Estimation for Brain-Computer Interfaces based on Selective Auditory Attention Decoding

Miguel A. Lopez-Gordo*, Simon Geirnaert*, and Alexander Bertrand, *Senior Member, IEEE*

***Abstract*—*Objective*: Selective auditory attention decoding (AAD) algorithms process brain data such as electroencephalography to decode to which of multiple competing sound sources a person attends. Example use cases are neuro-steered hearing aids or communication via brain-computer interfaces (BCI). Recently, it has been shown that it is possible to train such AAD decoders based on stimulus reconstruction in an unsupervised setting, where no ground truth is available regarding which sound source is attended. In many practical scenarios, such ground-truth labels are absent, making it, moreover, difficult to quantify the accuracy of the decoders. In this paper, we aim to develop a completely unsupervised algorithm to estimate the accuracy of correlation-based AAD algorithms during a competing talker listening task. *Methods*: We use principles of digital communications by modeling the AAD decision system as a binary phase-shift keying channel with additive white gaussian noise. *Results*: We show that the proposed unsupervised performance estimation technique can accurately determine the AAD accuracy in a transparent-for-the-user way, for different amounts of training and estimation data and decision window lengths. Furthermore, since different applications demand different targeted accuracies, our approach can estimate the minimal amount of training required for any given target accuracy. *Conclusion*: Our proposed estimation technique accurately predicts the performance of a correlation-based AAD algorithm without access to ground-truth labels. *Significance*: In neuro-steered hearing aids, the accuracy estimates provided by our approach could support time-adaptive decoding, dynamic gain control, and neurofeedback. In BCIs, it could support a robust communication paradigm with accuracy feedback for caregivers.**

***Index Terms*—selective auditory attention decoding, amplitude shift keying, unsupervised performance estimation**

This work was supported by grant B-TIC-352-UGR20, FEDER/Junta de Andalucía-Council for Economic Transformation, Industry, Knowledge and Universities; grant PID2021-128529OA-I00, MCIN/AEI/10.13039/501100011033, by ERDF A way of making Europe; grant PROYEXCEL_00084 funded by Consejería de Universidad, Investigación e Innovación, Junta de Andalucía (2021), a PDM mandate from KU Leuven (for S. Geirnaert, No PDMT1/22/009), a junior postdoctoral fellowship fundamental research of the FWO (for S. Geirnaert, No. 1242524N), FWO project nr. G081722N, the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant No 802895 and No 101138304), and the Flemish Government (AI Research program).

Miguel A. Lopez-Gordo is with the University of Granada, Department of Signal Theory, Telematics and Communications and with the NeuroEngineering and Computing Lab (NECOLab) at the Research Centre for Information and Communication Technologies (CITIC-UGR) (e-mail: malg@ugr.es).

Simon Geirnaert and Alexander Bertrand are with KU Leuven, Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics and with Leuven.AI – KU Leuven institute for AI. Simon Geirnaert is also with KU Leuven, Department of Neurosciences, Research Group ExpORL. (e-mail: simon.geirnaert, alexander.bertrand@esat.kuleuven.be).

*These two co-first authors contributed equally to this work.

I. Introduction

Selective auditory attention decoding (AAD) methods aim to decode the focus of selective auditory attention of a person attending to one out of multiple sounds, for example in a multi-talker scenario (also known as the cocktail party scenario) [1]–[4]. This information is decoded from brain signals, where auditory attention processes are encoded [2], [3]. These brain signals can be recorded with electroencephalography (EEG) [4], magnetoencephalography (MEG) [3], or electrocorticography (ECoG) [2].

AAD algorithms have applications in so-called neuro-steered hearing aids, cochlear implants, or other hearables [1], as they allow the user to steer the hearing device towards the conversation or talker they want to listen to in a cocktail party scenario. The attention information provided by the AAD algorithm can then be used by noise suppression and beamforming algorithms to enhance the attended talker and suppress other talkers considered as background noise.

Another potential application use case for such AAD algorithms is in brain-computer interfaces (BCIs) that establish a communication channel between the brain and the outer world by sending and decoding messages (or codes) from the user by reading brain activity as a correlate of a certain cognitive activity. For this purpose, a communication protocol must be established between the user and the computer to map a set of cognitive tasks to messages. Originally, BCIs were meant for people with severe motor impairment, incapable of using their muscles and thus being unable to produce any communication in any way. Examples are people living with amyotrophic lateral sclerosis, in completely locked-in state (CLIS), Duchenne muscular dystrophy and other forms of neuromuscular degeneration [5], [6]. The BCI literature has previously discussed a variety of applications and paradigms based on attention to auditory stimuli (e.g., multi-talker scenarios [7], [8] with natural or modulated voices [9], [10] and other auditory stimuli [11]).

One of the most popular and established AAD methods is stimulus reconstruction, in which temporal features of the attended speech stimulus (such as the speech envelope) are reconstructed from the brain activity and compared to the original speech stimuli to determine the focus of auditory attention [1], [4]. Stimulus reconstruction relies on a stimulus decoder to reconstruct these features, which is typically trained in a supervised manner where there is access to the ground-truth attention labels. These ground-truth labels are

obtained by instructing the subject to attend to a specific speaker and ignoring the other speaker(s) during a training session. Such a supervised training session takes at least 30-40 min, which is not always desirable (e.g., for plug-and-play hearing devices [12]) and sometimes not even possible. For example, in the case of patients in a minimally conscious state, where the detection of awareness [13], [14] is a prerequisite to establishing a BCI-based communication channel, this is extremely difficult as users cannot cooperate by any means during the calibration or training phase. In [15], the authors used a paradigm based on event-related potentials to study the feasibility of a communication channel with CLIS users. They reported two major difficulties: the uncertainty regarding the patient's levels of awareness and consequent uncertainty about the assessment of the BCI performance, and the lack of communication with the user that precluded knowledge of their volition or intention to cooperate. Other targeted users who cannot collaborate or show intention to collaborate are children living with autism. In these cases, communication skills are severely limited, necessitating the use of rudimentary communication methods such as pictograms to express their bare needs and feelings [16].

For all the reasons above, the development of a completely unsupervised approach for both training and AAD performance assessment is paramount. While an unsupervised (and time-adaptive) AAD training protocol for stimulus decoding has successfully been demonstrated in [12], [17], the assessment of the resulting unsupervised decoder still requires labelled data from the subject on which the decoder was trained. To date, there is no method of assessing the performance of a stimulus decoder in an unsupervised manner. This makes it impossible to, for example, check whether the aforementioned unsupervised decoder performs well enough in case the user cannot or should not give feedback, and thus whether or not more training data is required. Furthermore, an unsupervised accuracy estimation would allow the automatic update of various parameters in an AAD system, such as the speed of adaptation of an unsupervised decoder, gain control parameters, etc. Therefore, the goal of this paper is the development of an *unsupervised* accuracy estimation method for stimulus decoders in the AAD stimulus reconstruction paradigm.

The manuscript is organized as follows. In Section II, we explain the most relevant fundamentals related to (un)supervised stimulus reconstruction for AAD. In Section III, we unravel the principles of digital communications that support the unsupervised estimation of the accuracy and explain our proposed algorithm. In Section IV, we report the details of the dataset and experiments, of which the results are presented in Section V. In Section VI, we discuss the results of the various experiments and showcase applications of our approach in neuro-applications.

II. Stimulus reconstruction for AAD

A. Supervised stimulus reconstruction

The stimulus reconstruction approach for AAD consists of reconstructing temporal features of the attended speech signal, such as the envelope, from the EEG of the listener, and correlating this reconstruction with the actual speech envelopes of the competing speech signals to determine the attended speech signal [1], [4]. To decode or reconstruct the attended speech envelope $\hat{s}_a(t)$, with t the time sample index, we linearly combine time-lagged copies of the EEG channels [1], [4]:

$$\hat{s}_a(t) = \sum_{c=1}^{C} \sum_{l=0}^{L-1} d_c(l) x_c(t+l), \tag{1}$$

where $x_c(t)$ is the value of the c-th EEG channel at time t, $d_c(l)$ is the l-th decoder coefficient for channel c, and C and L are the number of EEG channels and decoder time lags, respectively. As shown in (1), the stimulus decoder $d_c(l)$ is an *anti-causal* filter, where only time lags l ranging from 0 to $L-1$ *after* the current stimulus sample at time t are used. Using vector notations, (1) can be rewritten as:

$$\hat{s}_a(t) = \mathbf{x}(t)^{\mathrm{T}} \mathbf{d},$$

where $\mathbf{x}(t)$ contains all time lags per EEG channel:

$$\mathbf{x}(t) = \begin{bmatrix} x_1(t) \\ \vdots \\ x_1(t+L-1) \\ x_2(t) \\ \vdots \\ x_C(t+L-1) \end{bmatrix} \in \mathbb{R}^{CL},$$

and $\mathbf{d} \in \mathbb{R}^{CL}$ similarly stacks all spatio-temporal decoder coefficients $d_c(l)$.

To train the decoder coefficients $d_c(l)$, assume the availability of K training segments of T time samples, i.e., $\{\mathbf{X}_k, (\mathbf{s}_{1_k}, \mathbf{s}_{2_k})\}_{k=1}^{K}$, with $\mathbf{X}_k = \begin{bmatrix} \mathbf{X}_{k,1} & \cdots & \mathbf{X}_{k,C} \end{bmatrix} \in \mathbb{R}^{T \times CL}$, where $\mathbf{X}_{k,c} \in \mathbb{R}^{T \times L}$ is a Hankel matrix containing the time-lagged EEG data of the c^{th} channel:

$$\mathbf{X}_{k,c} = \begin{bmatrix} x_{k,c}(0) & x_{k,c}(1) & \cdots & x_{k,c}(L-1) \\ x_{k,c}(1) & x_{k,c}(2) & \cdots & x_{k,c}(L) \\ \vdots & \vdots & & \vdots \\ x_{k,c}(T-1) & 0 & \cdots & 0 \end{bmatrix}.$$

For the sake of an easy exposition, we assume only two competing speakers with speech envelopes $\mathbf{s}_{1_k}$ and $\mathbf{s}_{2_k} \in \mathbb{R}^T$, where $\mathbf{s}_{1_k}$ and $\mathbf{s}_{2_k}$ are vectors containing T samples of both speech envelopes. During supervised training, additionally, the attention labels $y_k \in \{1, 2\}$, indicating which speech envelope corresponds to the attended envelope $\mathbf{s}_{a_k} \in \mathbb{R}^T$ in segment k, are available, i.e.,

$$\mathbf{s}_{a_k} = \begin{cases} \mathbf{s}_{1_k} & \text{if } y_k = 1, \\ \mathbf{s}_{2_k} & \text{if } y_k = 2. \end{cases} \tag{2}$$

It is important that both speakers are attended for a significant amount of time to avoid a decoder bias [18].

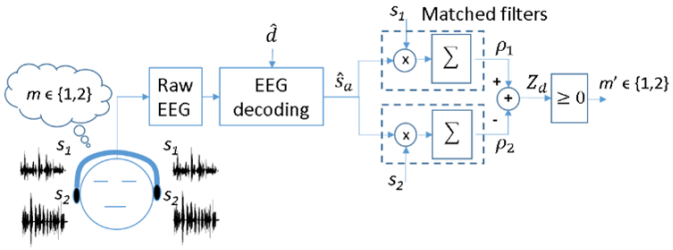


Figure 1: A basic BCI-based communication system based on AAD with stimulus reconstruction.

The stimulus decoder can then be trained by minimizing the squared error between the reconstructed envelope $\mathbf{X}_k\mathbf{d}$ and attended one across all training segments:

$$\hat{\mathbf{d}} = \underset{\mathbf{d}}{\operatorname{argmin}} \sum_{k=1}^{K} ||\mathbf{s}_{a_k} - \mathbf{X}_k\mathbf{d}||_2^2 = \underset{\mathbf{d}}{\operatorname{argmin}} ||\mathbf{s}_a - \mathbf{X}\mathbf{d}||_2^2, \quad (3)$$

with

$$\mathbf{s}_a = \begin{bmatrix} \mathbf{s}_{a_1} \\ \vdots \\ \mathbf{s}_{a_K} \end{bmatrix} \in \mathbb{R}^{KT}, \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_K \end{bmatrix} \in \mathbb{R}^{KT \times CL},$$

the concatenated attended speech envelope and EEG data matrices across all training segments. The solution of (3) can be found by solving the normal equations, leading to:

$$\hat{\mathbf{d}} = \mathbf{R}_{xx}^{-1}\mathbf{r}_{xs},$$

where

$$\mathbf{R}_{xx} = \mathbf{X}^{\mathrm{T}}\mathbf{X} = \sum_{k=1}^{K} \mathbf{X}_k^{\mathrm{T}}\mathbf{X}_k \in \mathbb{R}^{CL \times CL}$$

corresponds to the (unnormalized) EEG autocorrelation matrix and

$$\mathbf{r}_{xs} = \mathbf{X}^{\mathrm{T}}\mathbf{s}_a = \sum_{k=1}^{K} \mathbf{X}_k^{\mathrm{T}}\mathbf{s}_{a_k} \in \mathbb{R}^{CL} \quad (4)$$

to the (unnormalized) cross-correlation vector between the EEG and attended speech envelope.

B. The AAD decision system based on stimulus reconstruction

The trained decoder $\hat{\mathbf{d}}$ can now be used to make AAD decisions (see Figure 1) based on a new segment of EEG data $\mathbf{X}^{(\text{test})} \in \mathbb{R}^{T_{\text{test}} \times CL}$, recorded from a user that is attending to one out of two competing speakers with speech envelopes $\mathbf{s}_1^{(\text{test})}$ and $\mathbf{s}_2^{(\text{test})} \in \mathbb{R}^{T_{\text{test}}}$, while ignoring the other one. First, the stimulus decoder $\hat{\mathbf{d}}$ is used to reconstruct the attended envelope $\hat{\mathbf{s}}_a^{(\text{test})} = \mathbf{X}^{(\text{test})}\hat{\mathbf{d}}$ from the EEG, after which the Pearson correlation coefficients $\rho(\hat{\mathbf{s}}_a^{(\text{test})}, \mathbf{s}_1^{(\text{test})}) = \rho_1$ and $\rho(\hat{\mathbf{s}}_a^{(\text{test})}, \mathbf{s}_2^{(\text{test})}) = \rho_2$ are computed with both competing speech envelopes. The speaker that exhibits the highest correlation with the reconstructed envelope from the EEG is identified as the attended one.

The length of the EEG test segment (T_{test}), often referred to as the decision window length, determines the important trade-off between AAD accuracy and decision speed of the system, as typically the accuracy drastically decreases when moving to shorter decision window lengths as a result of more noisy estimates of the correlation coefficients [19].

C. Unsupervised stimulus reconstruction

In Section II-A, we assumed the availability of the attention labels $\{y_k\}_{k=1}^K$ during training, indicating which one of the competing speech envelopes corresponds to the attended one. In [12], an *unsupervised* training procedure for the stimulus decoder was proposed, removing the necessity of having access to these (ground-truth) attention labels.

From (4), these attention labels are required to compute the cross-correlation vector between the EEG and the attended speech envelope. To still be able to estimate this cross-correlation vector, [12] proposes to use (pseudo-)labels that are iteratively updated. The procedure starts with an initial set of (possibly random) labels and uses these to compute a supervised decoder via (4). The resulting decoder is then used to repredict the pseudo-labels, which then replace the initial pseudo-labels to again construct a new decoder, leading to a new set of pseudo-labels, etc. As shown in [12], this iterative procedure of predicting labels and retraining the decoder results in a self-leveraging effect in which the decoder improves after each iteration and quickly converges to a final decoder, even when the training set is initialized with random labels. A full description and explanation of the algorithm can be found in [12], [17].

III. Unsupervised accuracy estimation

As explained in Section II-C, it is possible to compute a stimulus decoder in an unsupervised fashion without the need for ground-truth labels. However, if such ground-truth labels are not available, we cannot assess the accuracy of this decoder when applied to the EEG data from a particular subject. More generally, there is currently no unsupervised method available to estimate the performance of any stimulus decoder for AAD. Therefore, the objective is to devise a method that allows estimating the performance of a correlation-based stimulus decoder for AAD in an unsupervised setting, i.e., without having access to the ground-truth attention labels. To do so, we draw inspiration from digital communications by modeling the AAD decision system as a Binary Phase-Shift Keying (BPSK) channel with Additive White Gaussian Noise (AWGN) (Section III-A), which requires some assumptions (Section III-B). This allows us to estimate the performance based on the unsupervised estimation of the parameters of the BPSK system (Section III-C). The full algorithm is summarized in Section III-D, while the estimation of confidence intervals is explained in Section III-E.

A. Stimulus reconstruction for AAD as a BPSK channel with AWGN

Drawing inspiration from digital communications, we can view the AAD decision system in Figure 1 as a BPSK communication channel between a sender (the user) and a receiver (the computer, hearing aid, etc.), where the communication

protocol is based on the selective attention listening paradigm, transmitting a binary message that conveys the attention of the listener (speaker 1 or 2).

In a BPSK system, two counter-phased symbols $\{+A, -A\}$ are transmitted over a channel with additive white Gaussian noise (AWGN), such that the received signals can be written as:

$$Z_d = x + n, \tag{5}$$

with the transmitted symbols $x \in \{+A, -A\}$ and the noise $n \sim \mathcal{N}(0, \sigma_d^2)$.

Let us now see how this relates to our AAD system in Figure 1, where the output value $Z_d = \rho_1 - \rho_2$ is treated as a received BPSK signal as in (5). Indeed, if $\mathbf{s}_1$ is the attended speaker, Z_d is positive in expectation ($\sim +A$ transmitted), while it is expected to be negative if $\mathbf{s}_2$ is the attended speaker ($\sim -A$ transmitted). For each test window, we consider the Pearson correlation coefficients $\rho(\hat{\mathbf{s}}_a^{(\text{test})}, \mathbf{s}_a^{(\text{test})}) = \rho_a$ and $\rho(\hat{\mathbf{s}}_a^{(\text{test})}, \mathbf{s}_u^{(\text{test})}) = \rho_u$ between the reconstructed speech envelope from the EEG ($\hat{\mathbf{s}}_a^{(\text{test})}$) and the actual attended ($\mathbf{s}_a^{(\text{test})}$) and unattended ($\mathbf{s}_u^{(\text{test})}$) speech envelope as random variables that are assumed to be normally distributed, i.e., $\rho_a \sim \mathcal{N}(\mu_a, \sigma_a^2), \rho_u \sim \mathcal{N}(\mu_u, \sigma_u^2)$ (see Section III-B). At each decision window, they each correspond to speaker 1 or 2 ($\rho_a = \rho_1$ or $\rho_a = \rho_2$ and vice versa for ρ_u), but this is unknown. The AAD decision variable $\mathrm{Z}_d = \rho_1 - \rho_2$ is then equal to $\rho_a - \rho_u$ (if $\mathbf{s}_1$ is attended/$+A$ is transmitted) or $\rho_u - \rho_a$ (if $\mathbf{s}_2$ is attended/$-A$ is transmitted). In both cases, Z_d is also normally distributed, with a variance equal to $\sigma_d^2 = \sigma_a^2 + \sigma_u^2$, assuming the random variables ρ_a and ρ_u are uncorrelated (see Section III-B). Furthermore, it can be inferred that the transmitted symbols are equal to $x = +A = \mu_a - \mu_u$ ($\mathbf{s}_1$ attended), or $x = -A = \mu_u - \mu_a$ ($\mathbf{s}_2$ attended).

Under the BPSK with AWGN assumption, the performance of the AAD system can be calculated using well-known equations for the bit-error rate (BER) P_e in such a system (see Chapter 14 of [20] for a complete explanation):

$$P_e = Q\left(\frac{A - (-A)}{2\sigma_d}\right) = Q\left(\frac{A}{\sigma_d}\right) = Q\left(\frac{\mu_a - \mu_u}{\sigma_d}\right), \tag{6}$$

where $Q(\cdot)$ represents the Q-function. In the digital communications domain, it is common to define and use the EbNo parameter (also referred as γ_b, or the ratio between the energy per bit and the noise spectral density). In our BPSK modulation system, we have:

$$\gamma_b = \text{EbNo} \triangleq \frac{A^2}{2\sigma_d^2} = \frac{(\mu_a - \mu_u)^2}{2\sigma_d^2}. \tag{7}$$

Using this expression for γ_b in (6) leads to the following final expression for the estimation of P_e:

$$\begin{aligned} P_e = Q\left(\sqrt{2\gamma_b}\right) &= \frac{1}{\sqrt{2\pi}} \int_{\sqrt{2\gamma_b}}^{+\infty} e^{-\frac{x^2}{2}} dx \\ &= \frac{1}{2}\operatorname{erfc}(\sqrt{\gamma_b}) = \frac{1}{2}\operatorname{erfc}\left(\frac{\mu_a - \mu_u}{\sqrt{2}\sigma_d}\right), \end{aligned} \tag{8}$$

where $\operatorname{erfc}(\cdot)$ denotes the complementary error function. This BER corresponds to the cumulative distribution function of $\mathcal{N}(\mu_a - \mu_u, \sigma_d^2)$, evaluated at 0 (Figure 2). Note that $Z_d = \rho_1 - \rho_2$ follows this same distribution if speaker 1 is attended ($+A$ transmitted). In the other case where speaker 2 is attended ($-A$ transmitted), Z_d follows $\mathcal{N}(\mu_u - \mu_a, \sigma_d^2)$. Due to the symmetry between both cases, it is sufficient to only analyze the case $Z_d \sim \mathcal{N}(\mu_a - \mu_u, \sigma_d^2)$.

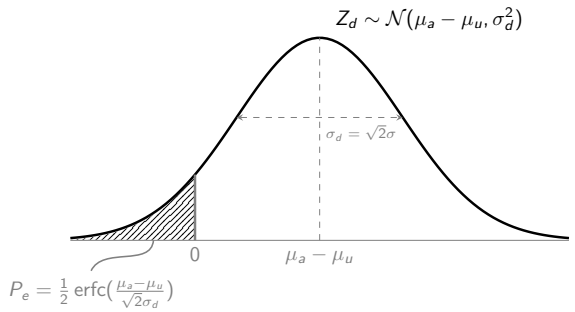


Figure 2: The AAD error rate or BER can be estimated as the cumulative distribution function at 0 of $Z_d \sim \mathcal{N}(\mu_a - \mu_u, \sigma_d^2)$.

The estimation of the BER (i.e., P_e) or AAD accuracy ($1 - P_e$), thus boils down to estimating the parameters $\mu_a - \mu_u, \sigma_d^2$, to plug into (8). In Section III-B, we first outline the different assumptions that are needed in order to estimate these different parameters in an *unsupervised* fashion, resulting in an unsupervised estimation of the accuracy, as explained in Section III-C.

B. Assumptions

To simplify and enable the unsupervised estimation of the aforementioned parameters and corresponding accuracy, we make the following assumptions about the distributions of the correlation coefficients ρ_a, ρ_u.

Assumption 1. The attended and unattended correlation coefficients are normally distributed, i.e., $\rho_a \sim \mathcal{N}(\mu_a, \sigma_a^2), \rho_u \sim \mathcal{N}(\mu_u, \sigma_u^2)$.

While the Pearson correlation coefficient is bounded to $[-1, 1]$ and the normal distribution is unbounded, in practice this is a reasonable assumption. Fisher showed that the z-transformation $z = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right)$ of the Pearson correlation coefficient r is approximately normally distributed even for small sample sizes [21]. However, the typical attended and unattended correlation coefficients $|\rho_a|, |\rho_u|$ using linear stimulus decoders are approximately 0.1 and 0.05, respectively (see, e.g., Figure 5), while the z-transformation is almost equal to the identity function for $|r| < 0.5$. As such, the normality assumption seems reasonable. Furthermore, Geirnaert et al. [12] did not find significant deviations from a normal distribution for the correlation coefficients resulting from stimulus reconstruction, further supporting this assumption.

Assumption 2. ρ_a and ρ_u are uncorrelated.

Both correlation coefficients are generated from the same reconstructed envelope $\hat{\mathbf{s}}_a^{(\text{test})}$, albeit correlated with a different speech envelope in both cases. As a result, there might be some spurious correlation between them due to this common component, such that Assumption 2 might be slightly violated.

Nevertheless, this assumption simplifies the unsupervised estimation of the parameters in Section III-C and was empirically observed to not result in substantially higher estimation accuracies compared to the case where the (small) correlation between ρ_a and ρ_u is taken into account (see Appendix A). Therefore, we will use Assumption 2 in the remaining of the paper for the sake of mathematical tractability.

Assumption 3. ρ_a and ρ_u have the same variance, i.e., $\sigma_a^2 = \sigma_u^2 = \sigma^2$.

This is a reasonable assumption given that the standard deviation on the correlation coefficients is mainly dominated by the noise, i.e., the EEG components that are uncorrelated to the speech envelopes, which have a much higher variance than the speech-following neural responses. Furthermore, in the literature several estimators for the standard deviation of the Pearson correlation coefficients have been proposed. For example, in [22], the standard deviation of Pearson correlation coefficients of two normally distributed variables is approximated by

$$\sigma = \frac{1 - r^2}{\sqrt{n}}, \tag{9}$$

with r the mean correlation coefficient and n the number of samples used in the estimation. Whatever formula is used, all of them have in common that for any two small correlation coefficients (i.e., $r \ll 1$) calculated with the same n, their respective standard deviations are approximately the same. In this regard and considering that typical mean values μ_a and μ_u are approximately 0.10 and 0.05 (see, e.g., Figure 5), their respective standard deviations would differ approximately 0.75%, thus supporting this assumption.

Assumption 4. The attended correlation coefficient is on average larger than the unattended one, i.e., $\mu_a > \mu_u$.

This assumption is grounded in the design of the attended decoder, which minimizes the squared error between the reconstructed and attended envelope (see Section II-A). In [23], it shown that this is equivalent with maximizing the attended correlation.

Assumption 5. The distributions $\rho_a \sim \mathcal{N}\big(\mu_a, \sigma_a^2\big)$ and $\rho_u \sim \mathcal{N}\big(\mu_u, \sigma_u^2\big)$ are stationary, i.e., they do not change over time.

While this assumption might not be true in a long-term setting [17], it can be accommodated by limiting the time span over which the unsupervised accuracy estimation is performed. We will experiment with different estimation times further on.

C. Estimation of parameters

As explained in Section III-A, the unsupervised estimation of the BER/accuracy boils down to the unsupervised estimation of the mean and standard deviation of $Z_d \sim \mathcal{N}\big(\mu_a - \mu_u, \sigma_d^2\big)$, given the assumptions in Section III-B. We assume that M decision windows with correlation coefficients with speaker 1 (ρ_1) and 2 (ρ_2) per decision window are available, i.e., $\{\rho_1^m, \rho_2^m\}_{m=1}^M$, but without knowing which correlation coefficient corresponds to the attended (ρ_a) or unattended one (ρ_u).

1) Estimation of σ_d^2: To estimate the standard deviation, we define the new random variable $Z_s = \rho_a + \rho_u \sim \mathcal{N}\big(\mu_a + \mu_u, \sigma_s^2\big)$. The key ingredient is that we have access to the samples of Z_s (per decision window) in an *unsupervised* way, given that

$$Z_s = \rho_a + \rho_u = \rho_1 + \rho_2 \sim \mathcal{N}\big(\mu_a + \mu_u, \sigma_s^2\big)\,,$$

where it does not matter whether speaker 1 or 2 is attended. This means that we can directly estimate the standard deviation σ_s^2 from the observed values of Z_s, i.e., $\{\rho_1^m + \rho_2^m\}_{m=1}^M$. Using the unbiased estimator of the standard deviation, we find:

$$\sigma_s = \sqrt{\frac{1}{M-1}\sum_{m=1}^{M}(\rho_1^m + \rho_2^m - \mu_s)^2},$$
$$\text{with } \mu_s = \frac{1}{M}\sum_{m=1}^{M}(\rho_1^m + \rho_2^m).$$

Given Assumption 2 and 3, we furthermore find that $\sigma_s^2 = \sigma_a^2 + \sigma_u^2 = 2\sigma^2$, and $\sigma_d^2 = \sigma_a^2 + \sigma_u^2 = 2\sigma^2$, such that $\sigma_d^2 = \sigma_s^2$, resulting in the unsupervised estimation of σ_d^2.

2) Estimation of $\mu_a - \mu_u$: To estimate the mean of Z_d, we define the new random variable $Z_{|d|} = |\rho_a - \rho_u|$. The absolute value is again crucial here to enable unsupervised access to the samples of this variable, i.e., $Z_{|d|} \triangleq |\rho_1 - \rho_2| = |\rho_a - \rho_u| = |\rho_u - \rho_a|$, and therefore it is independent of whether speaker 1 or 2 is attended.

Furthermore, given that $Z_{|d|}$ is defined as the absolute value of a normal distribution, it is distributed as the folded normal distribution, which can be defined in terms of the parameters of the underlying normal distribution [24], i.e., $\mu_a - \mu_u$ and σ_d^2:

$$Z_{|d|} \sim f\big(x; \mu_a - \mu_u, \sigma_d^2\big)$$
$$= \sqrt{\frac{2}{\pi\sigma_d^2}}e^{-\frac{(x^2 + (\mu_a - \mu_u)^2)}{2\sigma_d^2}}\cosh\left(\frac{(\mu_a - \mu_u)x}{\sigma_d^2}\right),$$

where σ_d^2 is already known from Section III-C1. The mean of the underlying normal distribution of Z_d can now be estimated using the method of moments or the maximum likelihood estimator [24]. Given that both estimation methods give similar results, we here choose for the method of moments as it is more intuitive. The first order moment or mean of the folded normal distribution is equal to [24]:

$$\mu_f = \sqrt{\frac{2}{\pi}}\sigma_d e^{-\frac{(\mu_a - \mu_u)^2}{2\sigma_d^2}} + (\mu_a - \mu_u)\,\mathrm{erf}\left(\frac{\mu_a - \mu_u}{\sqrt{2}\sigma_d}\right), \tag{10}$$

with $\mathrm{erf}(\cdot)$ the error-function. By equating (10) to the sample mean of $Z_{|d|}$, we can solve the following equation for x:

$$\sqrt{\frac{2}{\pi}}\sigma_d e^{-\frac{x^2}{2\sigma_d^2}} + x\,\mathrm{erf}\left(\frac{x}{\sqrt{2}\sigma_d}\right) = \frac{1}{M}\sum_{m=1}^{M}|\rho_1^m - \rho_2^m|. \tag{11}$$

Given that the left-hand side of (11) is an even function in x, it has a solution both at $x = \mu_a - \mu_u$ and $x = \mu_u - \mu_a$. Given Assumption 4, we can finally find an estimate of $\mu_a - \mu_u$ by taking the positive solution of (11), which can be found, again, fully unsupervised.

D. Algorithm

The full unsupervised accuracy estimation algorithm is summarized in Algorithm 1[1].

Algorithm 1 Unsupervised accuracy estimation of a correlation-based stimulus decoder

Input: Correlation coefficients between the reconstructed envelope and competing speech envelopes for M decision windows: $\{\rho_1^m, \rho_2^m\}_{m=1}^M$

Output: BER, accuracy

1: Compute

$$\mu_s = \frac{1}{M}\sum_{m=1}^{M}(\rho_1^m + \rho_2^m) \text{ and}$$

$$\sigma_d = \sqrt{\frac{1}{M-1}\sum_{m=1}^{M}(\rho_1^m + \rho_2^m - \mu_s)^2}.$$

2: Find the positive root $x^\star$ of

$$\sqrt{\frac{2}{\pi}}\sigma_d e^{-\frac{x^2}{2\sigma_d^2}} + x\,\mathrm{erf}\left(\frac{x}{\sqrt{2}\sigma_d}\right) - \frac{1}{M}\sum_{m=1}^{M}|\rho_1^m - \rho_2^m| = 0.$$

3: Compute

$$\text{BER} = P_e = \frac{1}{2}\,\mathrm{erfc}\left(\frac{x^\star}{\sqrt{2}\sigma_d}\right) \text{ and}$$

$$\text{accuracy} = 1 - P_e.$$

Note that, if desired, the distributions of ρ_a, ρ_u can be easily derived from the estimated distributions of Z_s and $Z_{|d|}$, by solving the following equations for $\mu_a, \mu_u, \sigma_a^2, \sigma_u^2$:

$$\sigma_s^2 = \sigma_d^2 = 2\sigma^2 = 2\sigma_a^2 = 2\sigma_u^2,$$
$$\mu_s = \mu_a + \mu_u, x^\star = \mu_a - \mu_u.$$

E. Confidence intervals

To quantify the uncertainty on the estimated accuracy using Algorithm 1, bootstrapping can be used to compute the 95%-confidence interval [25]. By resampling the unsupervised set of correlations $\{\rho_1^m, \rho_2^m\}_{m=1}^M$ with replacement (i.e., the Monte Carlo approximation), a bootstrap distribution can be constructed via Algorithm 1 to approximate the distribution of the unsupervised estimate around the true underlying accuracy [25]. From this bootstrapped distribution, the 95%-confidence interval can then be computed to approximate the true 95%-confidence interval. In this paper, however, we use the bias-corrected and accelerated (BCa) bootstrapping method for confidence intervals to deal with the over-coverage issues of the aforementioned basic percentile method, as explained in [26], [27].

IV. Experiments

A. Setup and objectives

We will perform experiments with the following three objectives in mind:

O1: Validation of our proposed approach for the unsupervised estimation of the AAD accuracy of a given (unsupervised) stimulus decoder. For that purpose, we will compare the true accuracy, quantified using the ground-truth attention labels, with the estimated accuracy, using different decision windows to evaluate the P_e (or BER).

O2: Once our approach is validated, we will use different sizes of training sets for the stimulus decoder to evaluate our approach on different versions of (suboptimal) decoders. The purpose is the assessment of the minimum amount of training data and decision window length to achieve a given targeted accuracy. In some critical applications, guaranteeing a certain minimum accuracy in a BCI-based communication system is vital (e.g., driving a wheelchair, controlling a panic button, or expressing living wills).

O3: While we will initially assume abundant availability of test data to estimate the accuracy on, we will evaluate the effect of the amount of estimation data MT_{test} on the accuracy estimation, with in the limit using only one window, i.e., estimating whether a single decision is right or wrong. This is important in a time-adaptive context, where you want to estimate the performance of the AAD system on a short time scale to be able to adapt properties of the system.

While we evaluate the estimation technique on the unsupervised stimulus decoder of Section II-C, we want to stress that the presented methodology is independent from the specific type of stimulus decoder (either supervised, unsupervised, linear, or non-linear), as long as the AAD system employs a correlation-based paradigm as in Section II-B.

B. AAD dataset

The dataset we use for these experiments consists of 72 min of EEG recordings (64-channel BioSemi ActiveTwo system) from 16 normal-hearing participants who participated in a dual-listening task [23], selectively attending to one of two competing talkers coming from the left or right ($\pm 90^\circ$ azimuth). The dataset is available online [28]. No participants were excluded.

The speech and EEG preprocessing, as well as the decoder setup and training, are the same as in [12]. For the sake of clarity, we summarize them in the next two subsections.

C. Speech and EEG preprocessing

Each audio signal was decomposed using gammatone filter bank. A power-law function with exponent 0.6 was applied to the output of each filter to compute the envelope of each subband. Finally, all subband envelopes were summed to obtain the final auditory envelope. Both the EEG data and speech envelopes were then filtered between 1–$9\,\text{Hz}$ and downsampled to $20\,\text{Hz}$. No additional noise or artifact removal was applied.

D. Decoder setup and training

The stimulus decoder was calculated by means of the unsupervised training algorithm from [12] described in Section II-C. Coefficients of the initial decoder were set to random values.

[1]Corresponding MATLAB code can be found online: https://github.com/AlexanderBertrandLab/unsupervised-AAD-accuracy-estimation.

The integration window of the decoder was set to 0–250 ms post-stimulus. As recommended, ten iterations were used to iteratively improve the decoder. During training and testing, the clean speech envelopes (available in the dataset) were used.

While we opt for the unsupervised stimulus decoder in the experiments to showcase the unsupervised accuracy estimation, the choice of the stimulus decoder is independent of the performance estimation, i.e., any other stimulus decoder (e.g., supervised, unsupervised, linear, non-linear) that uses the decision framework of Section II-B would work.

E. Evaluation

Per subject, continuous recordings were partitioned into segments of $T = 5, 10, 20, 40$ and 80 s, from which training and test sets of K, respectively M segments were created. This range of window lengths is considered convenient in real-time BCI applications, including AAD, in which detection of the attention to external stimuli should be carried out in maximally a few tens of seconds. In the experiments for O1 and O2, the training segment length for the unsupervised decoder is chosen equal to the decision window length, i.e., $T = T_{\text{test}}$.

For O1, we followed a holdout (ho) approach in which 75% of the segments of the full dataset, corresponding to 54 min, were randomly selected for the unsupervised training of the decoder and 25% (ho $= 25\%$), i.e., 18 min, to test the accuracy estimation. We randomly repeated this process of training and unsupervised accuracy estimation several times to guarantee a minimum of $M = 1000$ detections per decision window length to ensure sufficient estimation data is available. For the confidence intervals, 1000 resamplings were used.

For O2, we repeated the whole process described for O1 three times with different ho-ratios to decrease the amount of training data. Each time, the size of the training dataset used for the unsupervised training was 50%, 25% and 10% (ho equals $50\%, 75\%$ and 90%, respectively). This corresponds to training set sizes of $36, 18$ and 7 min, respectively.

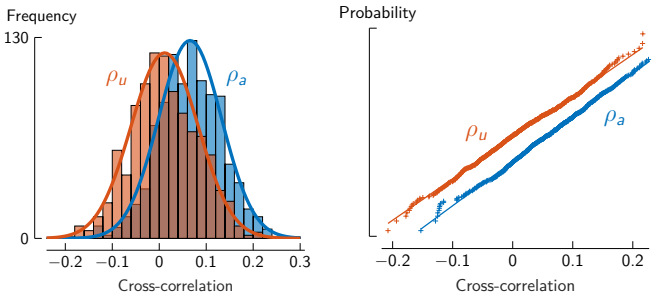
For O3, a fixed amount of training data, i.e., 75% of segments or 54 min, is taken. As opposed to in O1 and O2, the training segment length is kept constant on $T = 80$ s and is thus decoupled from the decision window length T_{test}, which again varied from 5 to 80 s. While the amount of training data is fixed, the testing/estimation time is, however, varied from $1, 5, 10, 30$ to 60 min. The holdout procedure is repeated a few times to accumulate enough decisions to achieve the desired estimation time. Per subject, this whole procedure is repeated 5 times to obtain a stable estimate of the mean absolute difference between measured (with ground-truth) and estimated (without ground-truth) accuracy.

V. Results

A. Validation of assumptions

In this section we present the results related to the assumptions described in Section III-B.

1) Assumption 1: normality: As an illustrative example, the left plot in Figure 3 shows a typical distribution of the attended and unattended cross-correlation values for an intermediate performer, whereas the right plot shows the corresponding normplot curve.



(a) (b)

Figure 3: Illustrative example of distributions of ρ_a and ρ_u (ho $= 25\%, T =$ 20 s) of an intermediate performer (subject 6). Both (a) the histogram on the left and (b) the normplot on the right suggest normality.

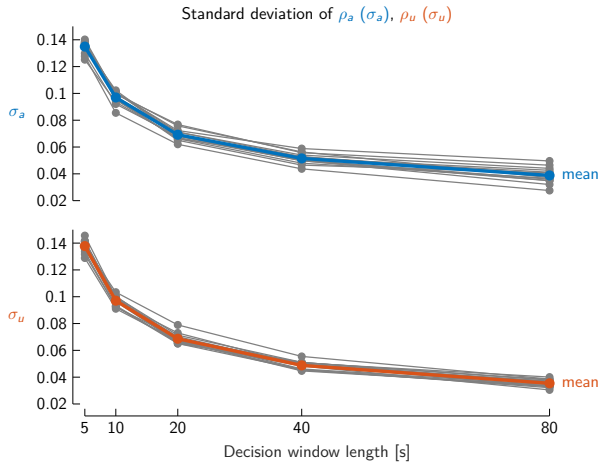


Figure 4: Upper and bottom plot show the standard deviations of $\rho_a(\sigma_a)$ and $\rho_u(\sigma_u)$ for different decision window lengths (ho $= 25\%$). Each line represents one subject. The colored line represents the mean across subjects.

2) Assumption 2: uncorrelatedness: When testing the significance of the correlation between the attended and unattended correlation coefficients per subject, we find a significant correlation (p-value < 0.05) for 6 out of 16 subjects, when using 20 s decision windows, ho $= 25\%$ and 1000 detections. This implies that this assumption is not (always) satisfied. However, we refer to Section III-B and Appendix A for further elaboration and the practical impact of this assumption.

3) Assumption 3: equal variance: Figure 4 shows the standard deviation of $\rho_a(\sigma_a)$ and $\rho_u(\sigma_u)$ for different decision window lengths (upper and bottom plots respectively). The maximum absolute difference between σ_a and σ_u across subjects and decision window lengths never exceeded 0.01.

4) Assumption 4: $\mu_a > \mu_u$: Figure 5 shows the mean values of $\rho_a(\mu_a)$ and $\rho_u(\mu_u)$ (upper and bottom plots respectively) with different values of the decision window length, for all subjects.

B. Objective O1

Table I shows the measured (Meas) and estimated (Est) AAD accuracies for decision windows of $5, 10, 20, 40$ and 80 s for all subjects (ho $= 25\%$). Accuracies in the 'Meas' columns were calculated using the ground-truth attention labels, whereas those in the 'Est' columns were estimated by means of our

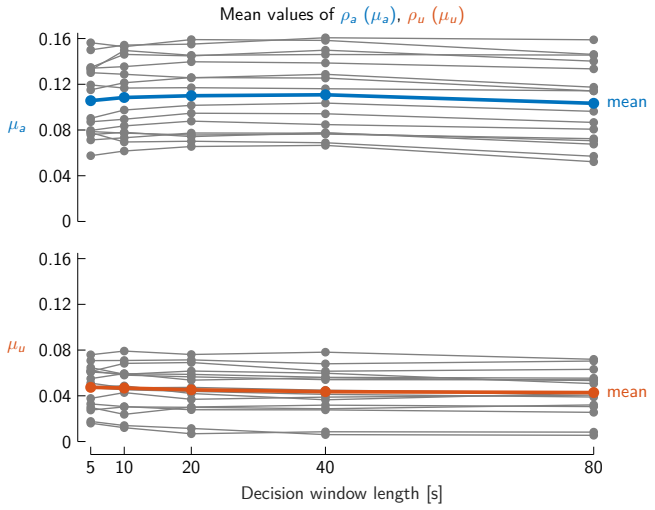


Figure 5: Upper and bottom plot show the mean values of $\rho_a(\mu_a)$ and $\rho_u(\mu_u)$ for different decision window lengths (ho $= 25\%$). Each line represents one subject. The colored line represents the mean values subjects.

User	Decision window length [s]									
	80		40		20			10		5	
	Meas	Est	Meas	Est	Meas	Est	95%-CI	Meas	Est	Meas	Est
1	49.7	50.0	46.4	50.0	50.1	50.0	[50.0,60.5]	49.3	50.0	48.5	55.7
2	97.3	93.7	90.4	90.1	81.3	82.7	[80.3,85.0]	74.6	74.6	65.1	61.7
3	95.6	97.8	93.3	95.0	84.1	88.3	[85.7,90.3]	71.3	81.7	68.0	73.9
4	90.6	82.9	85.0	82.4	75.5	74.1	[69.7,77.9]	67.9	72.6	64.7	64.8
5	93.5	95.3	89.1	89.4	85.0	78.7	[74.7,81.5]	75.0	73.3	68.6	66.3
6	84.1	80.6	80.5	77.8	69.2	71.8	[66.5,76.1]	64.9	63.5	58.3	56.0
7	90.3	92.3	85.4	89.6	78.1	78.4	[74.6,81.3]	70.8	68.9	61.3	67.3
8	61.2	58.5	64.9	68.8	62.5	57.6	[49.9,66.4]	56.4	50.0	57.6	50.0
9	72.6	79.3	69.0	73.6	60.4	68.2	[60.8,72.8]	57.8	60.1	55.3	53.0
10	88.7	92.7	80.0	87.3	73.7	78.0	[73.8,81.3]	65.7	64.7	60.1	51.9
11	83.6	85.4	77.8	80.7	72.2	67.1	[59.4,72.4]	65.8	64.3	61.6	50.0
12	69.1	50.0	66.0	63.3	63.5	56.2	[49.9,64.7]	58.4	49.9	54.0	62.0
13	81.5	84.2	79.5	76.5	69.5	69.0	[61.5,73.8]	65.6	61.6	62.8	60.1
14	99.5	98.9	97.4	97.2	92.4	92.7	[91.2,94.0]	84.1	84.4	71.2	71.5
15	89.6	94.6	86.3	90.6	80.7	79.5	[76.1,82.5]	68.1	64.8	64.6	60.6
16	91.5	89.9	86.8	84.0	76.9	74.7	[70.4,78.4]	68.3	66.3	63.1	68.3
Mean	83.7	82.9	79.9	81.0	73.4	72.9	[68.4,77.4]	66.5	65.7	61.6	60.8
Max abs err	19.1		7.3		7.8			10.4		11.6	
Min abs err	0.3		0.2		0.1			0.01		0.1	
Mean abs err	4.1		2.9		3.1			3.1		4.8	

Table I: Comparison between measured and estimated accuracies (ho $= 25\%$). For readability, only the 95%-confidence intervals (CIs) for 20 s decision windows are added. Gray rows correspond to selected subjects in Figure 6 and 7.

unsupervised approach. The last three rows represent the maximum, minimum and mean absolute difference across subjects between the estimated and the measured accuracies. The 95%-confidence intervals for 20 s decision windows are reported between brackets (other confidence intervals are not reported to not overload the table).

Figure 6 shows the estimated and reference measured curves of the BER (P_e) as a function of the unsupervised estimated EbNo for specific subjects (rows with dark background in Table I), corresponding to the worst and best performers (S1 and S14 respectively) and one intermediate performer (S13).

Holdout [%]	Training [min]	Decision window length [s]									
		80		40		20		10		5	
		Meas	Est	Meas	Est	Meas	Est	Meas	Est	Meas	Est
25	54	83.7	82.9	79.9	81.0	73.4	72.9	66.5	65.7	61.6	60.8
50	36	79.2	79.8	73.6	75.2	69.3	68.3	64.7	62.7	60.9	59.8
75	18	71.9	72.3	66.6	67.9	63.2	61.8	59.5	57.8	57.2	56.7
90	7	66.6	65.5	62.2	63.4	58.9	59.6	56.9	56.7	54.3	54.3

Table II: Comparison between the mean measured and estimated accuracies across subjects for different holdouts. Along the white and grey counter-diagonal lines, similar mean accuracies can be found

C. Objective O2

Analogously to Figure 6, Figure 7 shows the measured and the estimated curves of BER/P_e of selected subjects for holdouts 50% (36 min training data), 75% (18 min training data) and 90% (7 min training data) and decision window lengths of $5, 10, 20, 40$ and 80 s.

Analogously to Table I, Table II shows the mean values of the measured and estimated accuracies across subjects for holdouts $25\%, 50\%, 75\%$ and 90%. Along the white and gray counter-diagonal lines, similar mean accuracies can be found.

D. Objective O3

Figure 8 shows the mean absolute difference across all subjects when varying the amount of estimation data that is available, for different decision window lengths. The amount of training data was kept constant on 54 min, corresponding to 75% of the data, and the training window length was equal to 80 s.

VI. Discussion

A. General assumptions

In our approach, we assumed that the accuracy could be estimated in an unsupervised fashion by applying principles from BPSK modulation in the presence of AWGN. Figures 3, 4, 5 and the satisfying performance in Tables I and II suggest that the underlying assumptions are (approximately) satisfied. More specifically, in Section III-B, we outlined the different assumptions, which were evaluated in Section V-A.

1) Assumption 1: normality: From Figure 3, no substantive departures from normality can be identified in the norm-plot, while the histogram clearly reveals a Gaussian shape. These curves support the AWGN assumption. This confirms the results in [12], where no substantial deviation from the normal distribution of the correlations was found with the Kolmogorov-Smirnov test, on the same dataset.

2) Assumption 2: uncorrelatedness: This is the most debatable assumption, given that for 6 out of the 16 subjects, a significant correlation between the attended and unattended correlation coefficients was identified. However, as mentioned in Section III-B, this assumption is necessary to simplify and enable an easy unsupervised accuracy estimation. More specifically, it allows to estimate σ_d^2 via σ_s^2. When removing

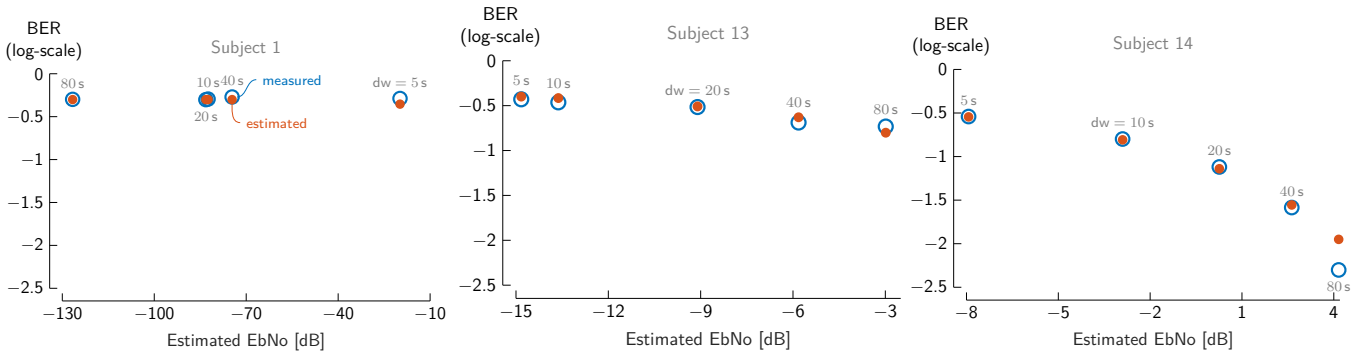


Figure 6: This figure shows the estimated (full circles) and reference measured (open circles) BER vs estimated EbNo of selected participants from Table I (the worst (S1), an intermediate (S13), and the best (S14) performer in terms of measured reference performance). Each circle corresponds to a decision window length (dw) of $5, 10, 20, 40$ or 80 s. The x-axis shows the unsupervised estimated EbNo using (7) for these decision windows. Our proposed unsupervised estimated accuracy closely approximates the reference measured accuracy.

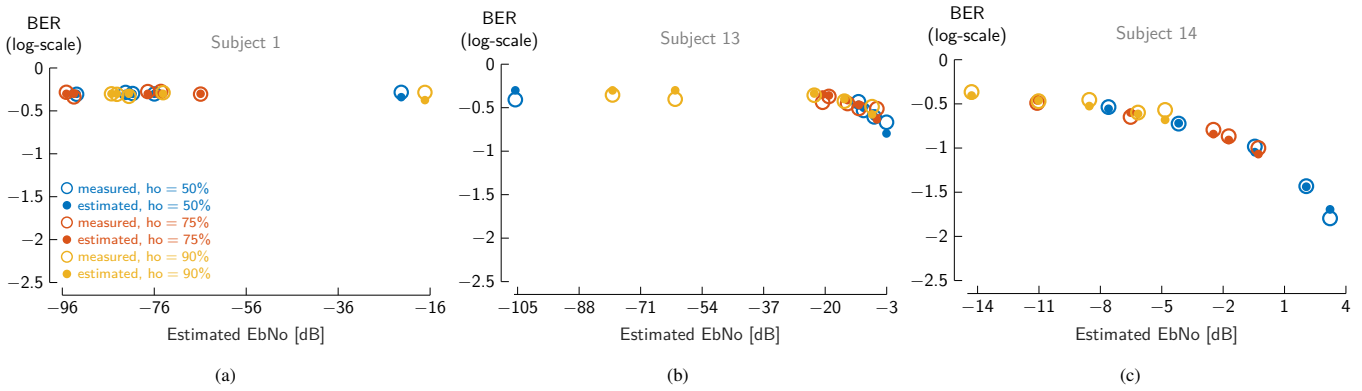


Figure 7: This figure shows the estimated (full circles) and reference measured (open circles) BER vs estimated EbNo of selected participants (the worst (S1), an intermediate (S13), and the best (S14) performer) for holdouts (ho) $50\%, 75\%$ and 90% (in blue, orange and yellow respectively) and decision windows (dw) of $5, 10, 20, 40$ and 80 s. Each plot contains 15 pairs (3 ho $\times 5$ dw lengths). This combination generates a pattern of BERs that matches the expected curves of BERs under a wide range of EbNo.

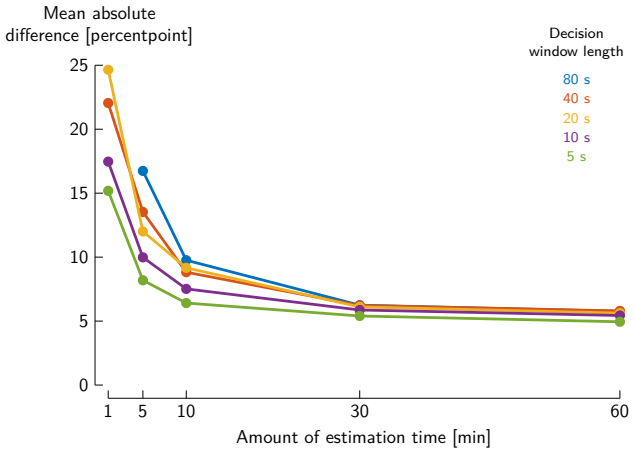


Figure 8: This figure shows the mean absolute difference as a function of the amount of data available to estimate the accuracy, for different decision window lengths. It reveals the trade-off between amount of estimation data (speed of estimation) and estimation accuracy (reliability), which is relevant in a time-adaptive, non-stationary context.

this assumption, the variance of the sum needs to be corrected using this correlation to be able to estimate the variance of the difference. In Appendix A, we show that, if we had access to this information, this can slightly improve the estimated accuracy, thus showing that this assumption is probably the most violated in practice – although still giving reasonable performance.

3) Assumption 3 and 4: Figure 5 shows that the mean values of $\rho_a(\mu_a)$ and $\rho_u(\mu_u)$ for any participant remain constant for all decision window lengths (ho $= 25\%$). Furthermore, Figure 4 clearly shows how the standard deviation decreases for longer decision windows. This suggests that the decision window length does not affect the decoder performance in terms of average cross-correlation values (approximately 0.10 and 0.05 for the attended and unattended correlation (Figure 5), confirming Assumption 4), but rather influences the estimation noise and thus the variance on the estimated correlation coefficients.

Furthermore, Figure 4 shows that the respective variances of the attended and unattended correlation coefficients are approximately the same (also across all subjects), thereby confirming Assumption 3. The standard deviations approximately decay with a $\frac{1}{\sqrt{2}}$ factor every time the decision window doubles, consistent with the theoretical formula for the standard deviation of the correlation coefficient in (9). This justifies our assumption that both ρ_a and ρ_u were affected by

the same amount of AWGN.

4) Assumption 5: While this assumption is necessary to accumulate sufficient test data to be able to estimate the performance, it is not explicitly tested in this paper. Furthermore, the dataset we use has been recorded in one relatively short session, thereby limiting the (longer-term) non-stationarities that could occur in EEG recordings [17]. Exactly to test our approach when this assumption over longer time periods is removed, we introduced O3, limiting the amount of estimation time to cope with the longer-term non-stationarities in a time-adaptive context (see Section VI-D).

B. Objective O1

Table I evidences the capability of our approach for the unsupervised estimation of the AAD accuracy of a stimulus decoder. Across all subjects, the absolute difference between the mean measured (supervised) and estimated (unsupervised) accuracy varies in the range of 0.5–1.1 percentage points (pp) across decision window lengths. Even on an individual level, the mean absolute difference remains low (between $2.9 - 4.8$ pp, see Table I, last row). These results reveal the level of accuracy that our approach can achieve for the accuracy estimation. Furthermore, in 61.3% of the cases (median 75%), the measured accuracy lies within the estimated 95%-confidence interval using the bootstrapping method in Section III-E.

For the three selected participants (S1, S13 and S14), Figure 6 represents the P_e (or BER) as a function of the EbNo calculated according to (8). Again, it can be seen that our unsupervised approach to estimate the accuracy nicely approximates the measured accuracy. The only exception is for the best performer (S14) with the longest decision window ($80\,\mathrm{s}$). In this case, there is a visually large error between the measured and the estimated P_e, despite the quantitatively negligible difference of 99.5% vs 98.9% (see Table I). This large visual error can be attributed to the steep slope of the log-function close to 0, given that the measured P_e is approximately 0.50%. For the worst performer (S1), it can be seen that the BER-curve remains flat at chance level regardless of the decision window. Interestingly, for all selected subjects (except for S1), the graphs show that above a certain minimum amount of EbNo (approximately $> -12\,\mathrm{dB}$), it increases by approximately $3\,\mathrm{dB}$ each time the decision window is doubled. This is consistent with the linear model that we assumed and can be attributed to the fact that mean correlations do not change across decision window lengths (Figure 5), while standard deviations vary as in (9) (see also Section VI-A). The implication is that it is possible from our model in (8) to predict the AAD accuracy for any decision window length once the parameters are estimated for only one specific decision window length.

Our approach is in line with previous studies that also used principles of digital constellations, specifically BPSK [10] and PSK [8], to characterize a BCI-based communication channel, although in those studies, the detection of the attended stimulus was performed by supervised media and voice tampering was required to embed the attention labels.

Lastly, note that our approach flexibly allows using any type of stimulus decoder that uses a correlation-based decision scheme as in Section II-B (either linear or non-linear, with supervised or non-supervised training).

C. Objective O2

Analogously to the results of O1, Table II presents the measured (supervised) and estimated (unsupervised) accuracies for different training set sizes. The objective is to assess whether it is possible to determine the minimum amount of training data and corresponding decision window length to guarantee or at least give an indication towards a given targeted accuracy.

As expected, a decay of the AAD accuracy can be observed with decreasing training set sizes. We also observe that our proposed approach keeps yielding accurate estimations of the accuracy. However, a certain degree of degradation occurs with decreasing the size of the training set (e.g., ho $= 90\%$). However, even for ho $= 90\%$ ($7\,\mathrm{min}$ of training data) the estimation error shows to be more than acceptable (see Table II).

Figure 7 shows that, for the selected subjects and combinations of training data sizes (36 (ho $= 50\%$), 18 (75%), and $7\,\mathrm{min}$ (90%)) and decision window lengths ($5, 10, 20, 40$ and $80\,\mathrm{s}$), the estimated P_e's closely match the measured P_e's similarly to Figure 6 for ho $= 25\%$. In practical terms, this fact implies that a certain targeted accuracy for a BCI application can be approximated by trading off the size of the training set and decision window length. This is reflected in Table II, when comparing accuracies across the white and gray counter-diagonal lines. For instance, if a specific application would require an accuracy of 65%, it could be obtained with any of these combinations that belong to the same diagonal (training size $= 7\,\mathrm{min}, \mathrm{dw} = 80\,\mathrm{s}$), (training size $= 18\,\mathrm{min}, \mathrm{dw} = 40\,\mathrm{s}$), (training size $= 36\,\mathrm{min}, \mathrm{dw} = 20\,\mathrm{s}$) and (training size $= 54\,\mathrm{min}, \mathrm{dw} = 10\,\mathrm{s}$). Moreover, through the proposed algorithm, this trade-off can be made fully unsupervised, enabling applications where supervised training is impossible (see Section VI-E). Finally, the 95%-confidence interval cannot only be used as a measure of uncertainty, useful in itself, but the lower bound can also be used as a conservative estimate for the accuracy, in case stricter guarantees about the performance are required.

A procedure for the determination of the minimum amount of training data and the decision window length for a targeted accuracy would be to collect data, train an unsupervised decoder and unsupervised estimate γ_b with our proposed algorithm for one or several decision window lengths. In case only one decision window length is used, the estimated γ_b in (7) can be extrapolated to other decision window lengths by adapting the estimated standard deviation using (9) (i.e., by changing n), while the estimated mean can be kept constant, as discussed in Section VI-A. With these values, the P_e could be estimated by means of (8). The optimal combination depends on functional elements, such as the usability, time to prepare, collect data and train the decoder, or the maximum latency (i.e., minimum rate) at which detections should be performed.

D. Objective O3

Figure 8 clearly shows that the fewer estimation data available to assess the AAD performance, the worse the estimation

(as expected), following an exponential-like trend. In a time-adaptive context, this introduces a clear trade-off: the faster an estimate is required, the cruder it will be. Below $5\,\mathrm{min}$ of estimation data, the errors become unacceptably large. However, it is also clear that this effect is more pronounced for longer decision window lengths. This can be explained because we take time as a starting point to quantify the amount of estimation data, such that the shorter the decision window length becomes, the more decisions are available in Algorithm 1 (i.e., M is larger). The effect of the interplay between estimation time and decision window length becomes apparent from the moment that less than $30\,\mathrm{min}$ of estimation data is available.

This experiment becomes especially relevant when the stationarity assumption (Assumption 5) no longer holds. In a practical BCI or AAD application, there will be longer-term non-stationarities in the data arising from the EEG equipment (impedances that change, electrodes that move, etc.) or from the user and its environment. As shown in [17], these non-stationarities impact the AAD performance, necessitating time-adaptive stimulus decoders that update on a shorter time period. This means that, in practice, the (estimated) distributions of ρ_a, ρ_u change (shift) over time. In a practical application, it is thus paramount to take these shifts into account, for example, by reducing the estimation time while taking the trade-off between reliability and speed of estimation presented in Figure 8 into account. Note that this non-stationarity problem is not fully taken into account in Figure 8: the dataset does not contain many of these longer-term non-stationarities, given that it was recorded in a lab-controlled environment in a single experiment. Introducing these non-stationarities could change the presented trade-off in Figure 8, leading to a more optimal trade-off at shorter estimation times.

E. Applications in neuro-engineering

1) Impact in neuro-steered hearing devices: The developed unsupervised estimation of the AAD accuracy of a stimulus decoder has several applications in a neuro-steered hearing device. For example, when targeting a time-adaptive, unsupervised stimulus decoder as in [17], a time-adaptive estimate of the performance could be used to dynamically change the updating window of the adaptive decoder. It allows to more accurately control the trade-off between stability/accuracy of the decoder and time-adaptivity/speed of adaptation in this time-adaptive decoder.

Furthermore, a time-adaptive estimate of the AAD accuracy can be employed to dynamically update the AAD decision system (e.g., decision window length) and the properties of a gain control system, acting upon the AAD decisions. For example, when a certain performance needs to be guaranteed, the decision window length might need to be adapted over time, based on the estimated performance we provide. Furthermore, in [19], it is shown that the optimal parameters for an AAD-based gain control system depend on the AAD accuracy. A dynamic unsupervised estimate of this accuracy could therefore be used to optimally update the gain control system properties.

Thirdly, this unsupervised estimate allows monitoring the AAD system's performance in a hearing device, flagging potential problems to the user and/or expert and informing appropriate interventions.

Lastly, the proposed approach in Section III gives additional insights into the distribution of the correlation coefficients, informing a clear strategy to improve future stimulus decoders, i.e., by increasing the separation between the average attended and unattended correlation. This is directly apparent from (8), as the erfc-function is a monotonically decreasing function. Furthermore, (8) allows inferring how well the decoder needs to separate the competing talkers to achieve a certain desirable performance. Note that these insights do not only inform a better algorithm design for stimulus decoders, but also give rise to a neurofeedback training program for neuro-steered hearing device users. For example, the estimate of the mean difference of correlations, $\mu_a - \mu_u$, could be used as a training metric for users to learn to maximize.

2) Impact in BCIs: In the BCI realm, there is an interesting opportunity for the development of the unsupervised approach presented in this contribution for applications intended for users with severe motor or cognitive impairments. In the literature, there are antecedents of using a dichotic listening task to build a communication channel. For instance, in [8], [10] the envelopes of two to six simultaneous speakers were synthetically modified to evoke a constellation of BPSK and 4-PSK symbols respectively [29]. However, in these studies, the training and performance calibration was supervised, that is, the active collaboration of the user was required to obtain knowledge of the classification labels, so those techniques could not be used with severely affected BCI users. In another similar approach [30], signals in the gamma-band of ECoG signals were used to detect the attention to one out of two speech signals with a mean accuracy of 77% and a decision window of $10\,\mathrm{s}$. Despite the high accuracy obtained with this proposal, this was a strongly invasive approach intended for patients implanted with an array of electrodes that required *active* collaboration from the user. In comparison with the examples presented in this paragraph, our non-invasive and unsupervised approach represents an alternative to scenarios that are not suited for those approaches.

VII. Conclusion

In this paper, we have proposed a novel method to estimate the AAD accuracy of a correlation-based stimulus decoder in an *unsupervised* manner. This estimation technique is inspired by principles of digital communications, modelling the AAD decision system as a BPSK channel with AWGN.

Using our approach, we have been able to very closely approximate the actual average performance of an unsupervised stimulus decoder and achieve a mean absolute difference of 3.6 pp on a per-subject level. Moreover, with our unsupervised approach, we have demonstrated that the optimal combination of minimum training duration and decision window length can be established in an unsupervised manner, and that the amount of estimation data can be reduced to a few tens of minutes to achieve a reliable estimation.

The proposed unsupervised estimation of AAD accuracy opens up many applications, both in neuro-steered hearing devices (e.g., in time-adaptive decoding, dynamic gain control, or neurofeedback) and other BCIs (e.g., to establish a communication system for users with severe motor or cognitive impairments). As such, it represents an important piece of the puzzle in AAD-based systems.

APPENDIX A

As discussed in Section III-B, we assume that the attended and unattended correlation coefficients are uncorrelated to be able to equate σ_s^2 with σ_d^2. However, the results in Section V-A show that this is untrue for 6 out of 16 subjects, given that there is a significant correlation r_{au} present for these subjects. Note that the sign of this correlation between the attended and unattended correlation coefficient changes per subject: for 11 subjects, there is a positive correlation (mean 0.05/standard deviation 0.04), while for 5 subjects, this is a negative correlation (mean -0.07/standard deviation 0.04).

If we were able to access or estimate this correlation r_{au}, we would be able to correct σ_s^2 to find σ_d^2 as follows, using Assumption 3:

$$\sigma_d^2 = \sigma_s^2 \frac{1 - r_{au}}{1 + r_{au}}$$

However, we did not find a straightforward way of estimating this correlation in an unsupervised manner. We have tried directly estimating σ_d^2 from $Z_{|d|}$ using the method moments or maximum likelihood estimator [24] similar to Section III (note that this leads to an iterative procedure of updating $\mu_{Z_{|d|}}$ and σ_d^2), but this did not improve the estimation technique w.r.t. Algorithm 1. This is most likely due to the necessity of iteratively estimating both parameters from a single distribution $Z_{|d|}$, allowing estimation errors of one parameter to leak into the other estimated parameter. If we would have access to this (ground-truth) correlation, the mean absolute difference would, however, decrease to 1.26% ($20\,\mathrm{s}$ decision windows, 1000 detections, 25% ho). The maximum absolute difference decreases to 3.15%. This shows the potential added benefit of estimating this correlation, which is left open as a future challenge.

REFERENCES

[1] S. Geirnaert, S. Vandecappelle, E. Alickovic, A. de Cheveigné, E. C. 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.

[2] N. Mesgarani and E. F. Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature*, vol. 485, pp. 233–236, 2012.

[3] N. Ding and J. Z. Simon, "Emergence of neural encoding of auditory objects while listening to competing speakers," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 29, pp. 11 854–11 859, 2012.

[4] J. A. O'Sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. 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, 2014.

[5] Y. Maruyama, N. Yoshimura, A. Rana, A. Malekshahi, A. Tonin, A. Jaramillo-Gonzalez, N. Birbaumer, and U. Chaudhary, "Electroencephalography of completely locked-in state patients with amyotrophic lateral sclerosis," *Neuroscience Research*, vol. 162, pp. 45–51, 2021.

[6] N. Birbaumer, N. Ghanayim, T. Hinterberger, I. Iversen, B. Kotchoubey, A. Kübler, J. Perelmouter, E. Taub, and H. Flor, "A spelling device for the paralysed," *Nature*, vol. 398, pp. 297–298, 1999.

[7] M. A. Lopez-Gordo, H. Pomares, M. Damas, A. Prieto, and E. M. de la Plaza Hernandez, "Use of Kohonen Maps as Feature Selector for Selective Attention Brain-Computer Interfaces," in *Bio-inspired Modeling of Cognitive Tasks*. Springer, 2007, pp. 407–415.

[8] M. A. Lopez-Gordo, F. Pelayo, E. Fernandez, and P. Padilla, "Phase-shift keying of EEG signals: Application to detect attention in multitalker scenarios," *Signal Processing*, vol. 117, pp. 165–173, 2015.

[9] M. A. Lopez-Gordo, E. Fernandez, S. Romero, F. Pelayo, and A. Prieto, "An auditory brain–computer interface evoked by natural speech," *Journal of neural engineering*, vol. 9, no. 3, p. 036013, 2012.

[10] M. A. Lopez-Gordo and F. Pelayo, "A binary phase-shift keying receiver for the detection of attention to human speech," *International Journal of Neural Systems*, vol. 23, no. 4, p. 1350016, 2013.

[11] J. Höhne and M. Tangermann, "Towards User-Friendly Spelling with an Auditory Brain-Computer Interface: The CharStreamer Paradigm," *PLoS One*, vol. 9, no. 6, p. e98322, 2014.

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

[13] A. M. Owen, N. D. Schiff, and S. Laureys, "A new era of coma and consciousness science," *Progress in Brain Research*, vol. 177, pp. 399–411, 2009.

[14] M. Khalili-Ardali, S. Wu, A. Tonin, N. Birbaumer, and U. Chaudhary, "Neurophysiological aspects of the completely locked-in syndrome in patients with advanced amyotrophic lateral sclerosis," *Clinical Neurophysiology*, vol. 132, no. 5, pp. 1064–1076, 2021.

[15] G. Pires, S. Barbosa, U. J. Nunes, and E. Gonçalves, "Visuo-auditory stimuli with semantic, temporal and spatial congruence for a P300-based BCI: An exploratory test with an ALS patient in a completely locked-in state," *Journal of Neuroscience Methods*, vol. 379, p. 109661, 2022.

[16] J. A. Pereira, D. Macêdo, C. Zanchettin, A. L. I. de Oliveira, and R. do Nascimento Fidalgo, "Pictobert: Transformers for next pictogram prediction," *Expert Systems with Applications*, vol. 202, p. 117231, 2022.

[17] S. Geirnaert, T. Francart, and A. Bertrand, "Time-Adaptive Unsupervised Auditory Attention Decoding Using EEG-Based Stimulus Reconstruction," *IEEE Journal on Biomedical and Health Informatics*, vol. 26, no. 8, pp. 3767–3778, 2022.

[18] N. Das, W. Biesmans, A. Bertrand, and T. Francart, "The effect of head-related filtering and ear-specific decoding bias on auditory attention detection," *Journal of Neural Engineering*, vol. 13, no. 056014, 2016.

[19] 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.

[20] A. B. Carlson, P. B. Crilly, and J. C. Rutledge, *Communication systems*, 4th ed. New York: McGraw-Hill, 2002.

[21] R. Fisher, "On the 'Probable Error' of a Coefficient of Correlation Deduced from a Small Sample," *Metron*, vol. 1, pp. 1–32, 1921.

[22] A. L. Bowley, "The Standard Deviation of the Correlation Coefficient," *Journal of the American Statistical Association*, vol. 23, no. 161, pp. 31–34, 1928.

[23] 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.

[24] M. Tsagris, C. Beneki, and H. Hassani, " On the Folded Normal Distribution ," *Mathematics*, vol. 2, no. 1, pp. 12–28, 2014.

[25] B. Efron, "Bootstrap Methods: Another Look at the Jackknife," in *Breakthroughs in Statistics: Methodology and Distribution*. Springer New York, 1992, pp. 569–593.

[26] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York: Chapmann & Hall, 1993.

[27] T. J. DiCiccio and B. Efron, "Bootstrap confidence intervals," *Statistical Science*, vol. 11, no. 3, pp. 189–228, 1996.

[28] N. Das *et al.*, "Auditory Attention Detection Dataset KULeuven," Zenodo, 2019. [Online]. Available: https://zenodo.org/record/3997352

[29] J. Minguillon, M. A. Lopez-Gordo, and F. Pelayo, "Detection of attention in multi-talker scenarios: A fuzzy approach," *Expert Systems with Applications*, vol. 64, pp. 261–268, 2016.

[30] K. V. Dijkstra, P. Brunner, A. Gunduz, W. Coon, A. L. Ritaccio, J. Farquhar, and G. Schalk, "Identifying the attended speaker using electrocorticographic (ECoG) signals," *Brain-Computer Interfaces*, vol. 2, no. 4, pp. 161–173, 2015.