Unsupervised Self-Adaptive Auditory Attention Decoding

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

***Abstract*—When multiple speakers talk simultaneously, a hearing device cannot identify which of these speakers the listener intends to attend to. Auditory attention decoding (AAD) algorithms can provide this information by, for example, reconstructing the attended speech envelope from electroencephalography (EEG) signals. However, these stimulus reconstruction decoders are traditionally trained in a supervised manner, requiring a dedicated training stage during which the attended speaker is known. Pre-trained subject-independent decoders alleviate the need of having such a per-user training stage but perform substantially worse than supervised subject-specific decoders that are tailored to the user. This motivates the development of a new unsupervised self-adapting training/updating procedure for a subject-specific decoder, which iteratively improves itself on unlabeled EEG data using its own predicted labels. This iterative updating procedure enables a self-leveraging effect, of which we provide a mathematical analysis that reveals the underlying mechanics. The proposed unsupervised algorithm, starting from a random decoder, results in a decoder that outperforms a supervised subject-independent decoder. Starting from a subject-independent decoder, the unsupervised algorithm even closely approximates the performance of a supervised subject-specific decoder. The developed unsupervised AAD algorithm thus combines the two advantages of a supervised subject-specific and subject-independent decoder: it approximates the performance of the former while retaining the 'plug-and-play' character of the latter. As the proposed algorithm can be used to automatically adapt to new users, as well as over time when new EEG data is being recorded, it contributes to more practical neuro-steered hearing devices.**

***Index Terms*—auditory attention decoding, neuro-steered hearing devices, stimulus reconstruction, unsupervised training**

I. Introduction

Auditory attention decoding (AAD) encompasses the process of determining the auditory focus of attention using a person's brain activity. AAD algorithms are a paramount building block of so-called 'neuro-steered hearing devices' [2], [3]. This is because current hearing aids and cochlear implants do not know the speaker or sound source a user intends to attend to. However, this knowledge is crucial to assist the user in cocktail party scenarios, where multiple speakers are simultaneously active. Knowledge of the attended speaker can then be exploited by noise suppression algorithms that suppress unattended speakers and other background activity, effectively enhancing the attended speaker.

Determining the auditory attention directly from the brain activity (e.g., non-invasively recorded using magneto- or electroencephalography (MEG/EEG)) has gained attention due to the fundamental insight that the brain tracks the amplitude envelope of the attended speech signal [4], [5]. Importantly, this neural envelope tracking phenomenon is not only present in normal-hearing subjects but also in hearing-impaired listeners [6]–[8].

The main class of current AAD algorithms exploits this neural envelope tracking by reconstructing the attended speech envelope from the recorded EEG/MEG signals via a stimulus reconstruction decoder [3], [9]. The reconstructed speech envelope can then be compared through the Pearson correlation coefficient with the speech envelopes of the active speakers to determine which speaker is the attended one. Alternatively, the aforementioned backward approach (i.e., reconstructing the speech envelope from the EEG) can be interchanged with a forward approach (i.e., predicting the EEG from the speech envelope). While this has the benefit of interpretability, it performs worse than the backward decoder approach [10], [11]. Originally, the stimulus reconstruction decoder was computed based on a minimum mean-squared-error cost function [9]. Later, this approach was extended to various other linear and nonlinear stimulus reconstruction approaches [3]. Furthermore, other AAD paradigms, such as decoding the spatial focus of attention [12]–[14] (instead of reconstructing the stimulus), have been proposed.

AAD decoders can be used in a subject-specific or subject-independent way [9], trading practical applicability with better performance:

- A *subject-specific* decoder is traditionally trained in a *supervised* manner, requiring a cumbersome a priori training stage in which data from the subject under test are collected to train an AAD decoder. This popular approach is thus less practical to implement on hearing devices. However, it is known that this approach results in the highest AAD performance for a given AAD algorithm [9].
- A *subject-independent* decoder also requires labeled data, but only of subjects other than the subject under test,

This research is funded by an Aspirant Grant from the Research Foundation - Flanders (FWO) (for S. Geirnaert), FWO project nr. G0A4918N, the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 802895 and grant agreement No 637424), and the Flemish Government (AI Research Program). The scientific responsibility is assumed by its authors. *(Corresponding author: Simon Geirnaert.)*

S. Geirnaert and A. Bertrand are with KU Leuven, Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, Kasteelpark Arenberg 10, B–3001 Leuven, Belgium and with Leuven.AI - KU Leuven institute for AI, B–3000, Leuven, Belgium (e-mail: simon.geirnaert@esat.kuleuven.be, alexander.bertrand@esat.kuleuven.be).

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

A Supplementary Material paper corresponding to this paper has been published in [1].

which allows to pre-train it. At test time, this subject-independent decoder can be applied to the incoming data of the new, unseen subject, without a priori requiring information about the attention processing of that particular subject. As such, it could be used in a 'plug-and-play' fashion, pre-installed on each neuro-steered hearing device and thus leading to a generic hearing device. However, this practical applicability comes at the cost of a lower AAD performance, as the decoder fails to capture the subject-specific differences in auditory processing [9].

Moreover, both decoders remain fixed during operation, when new data of the subject under test comes in. They do not adapt to changing conditions and situations and thus result in suboptimal decoding results.

Except for the algorithm in [15], other AAD algorithms [3] are supervised and very often subject-specifically trained. In [15], a dynamic AAD algorithm is proposed, in which a decoder is estimated for each speaker per new incoming segment of data. These decoders are then applied again to that same segment of data to determine the auditory attention. Although some labeled data is required to tune specific hyperparameters, this algorithm is by design unsupervised. However, this algorithm is substantially outperformed by all other traditional (supervised) AAD algorithms [3].

We propose a fully unsupervised subject-specific AAD algorithm, in which a stimulus reconstruction decoder is iteratively updated on the EEG data and speech envelopes. This iterative updating does *not* require ground-truth labels, i.e., knowledge about which is the attended or unattended speaker. Instead, the model updates itself based on its own predictions in the previous iteration. We hypothesize that this results in a self-leveraging effect. As such, it should automatically adapt to a new subject, integrating the two major advantages of a subject-specific and subject-independent decoder:

1) A higher performance than a subject-independent decoder.
2) Retaining the unsupervised 'plug-and-play' feature of a subject-independent decoder, thus without requiring knowledge about the labels during training.

Furthermore, such a self-adaptive algorithm could be applied adaptively in time. As EEG and audio data are continuously recorded, it adapts to changing conditions and situations.

In Section II, we introduce the proposed method to update a stimulus reconstruction decoder in an unsupervised manner. In Section III, the data, preprocessing, and performance evaluation are explained. In Section IV, we provide a recursive mathematical model to track the iterations of the unsupervised algorithm, with the aim to gain some insights into the mechanics of the self-leveraging effect. The proposed method is then tested on two separate datasets in Section V. Applications, future work, and conclusions are discussed in Section VI.

II. Unsupervised self-adaptive AAD

In Section II-A, we concisely revisit the traditional supervised training of a stimulus reconstruction decoder for AAD. The newly proposed unsupervised procedure is explained in Section II-B.

A. Supervised training of a decoder

In the classical approach[1] towards AAD (see, e.g., [3], [9], [10], [16], [17]), a linear spatio-temporal filter $D(l,c)$, referred to as a decoder, reconstructs the attended speech envelope $s_{\text{a}}(t)$ from the C-channel EEG signal $X(t,c)$ by anti-causally integrating EEG samples over L time lags, for each EEG channel $c \in \{1,\dots,C\}$:

$$\hat{s}_{\text{a}}(t) = \sum_{c=1}^{C}\sum_{l=0}^{L-1} D(l,c)X(t+l,c), \tag{1}$$

with l the time lag index and c the channel index.

Equation (1) can be rewritten in vector format as:

$$\hat{s}_{\text{a}}(t) = \mathbf{d}^{\text{T}}\mathbf{x}(t),$$

where

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

contains L lags, for each EEG channel. Similarly, the vector $\mathbf{d} \in \mathbb{R}^{CL\times 1}$ stacks all decoder coefficients $D(l,c)$, across channels and time lags. The decoder $\mathbf{d}$ is then found by minimizing the squared error:

$$\hat{\mathbf{d}} = \underset{\mathbf{d}}{\text{argmin}} \ ||\mathbf{s}_{\text{a}} - \mathbf{X}\mathbf{d}||_2^2, \tag{2}$$

with $\mathbf{s}_{\text{a}} = \begin{bmatrix} s_{\text{a}}(0) & \cdots & s_{\text{a}}(T-1) \end{bmatrix}^{\text{T}} \in \mathbb{R}^{T\times 1}$ and $\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \cdots & \mathbf{X}_C \end{bmatrix} \in \mathbb{R}^{T\times CL}$ a block Hankel matrix, with

$$\mathbf{X}_c = \begin{bmatrix} x_c(0) & x_c(1) & x_c(2) & \cdots & x_c(L-1) \\ x_c(1) & x_c(2) & x_c(3) & \cdots & x_c(L) \\ x_c(2) & x_c(3) & x_c(4) & \cdots & x_c(L+1) \\ \vdots & \vdots & \vdots & & \vdots \\ x_c(T-1) & 0 & 0 & \cdots & 0 \end{bmatrix} \in \mathbb{R}^{T\times L}.$$

Defining the sample autocorrelation matrix $\hat{\mathbf{R}}_{xx} \in \mathbb{R}^{CL\times CL}$ and sample cross-correlation vector $\hat{\mathbf{r}}_{xs_{\text{a}}} \in \mathbb{R}^{CL\times 1}$ as:

$$\hat{\mathbf{R}}_{xx} = \frac{1}{T}\mathbf{X}^{\text{T}}\mathbf{X} \text{ and } \hat{\mathbf{r}}_{xs_{\text{a}}} = \frac{1}{T}\mathbf{X}^{\text{T}}\mathbf{s}_{\text{a}}, \tag{3}$$

the solution of (2) is equal to:

$$\begin{aligned} \hat{\mathbf{d}} &= \left(\mathbf{X}^{\text{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\text{T}}\mathbf{s}_{\text{a}} \\ &= \hat{\mathbf{R}}_{xx}^{-1}\hat{\mathbf{r}}_{xs_{\text{a}}}. \end{aligned} \tag{4}$$

This classical supervised training approach is summarized in Figure 1.

Often, ridge regression is used to avoid overfitting when only a limited amount of training data is available [3], [10], [16], [17], such that the decoder is estimated as:

$$\hat{\mathbf{d}} = \left(\mathbf{X}^{\text{T}}\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^{\text{T}}\mathbf{s}_{\text{a}}, \tag{5}$$

[1] A MATLAB implementation of this AAD approach can be found in [16].

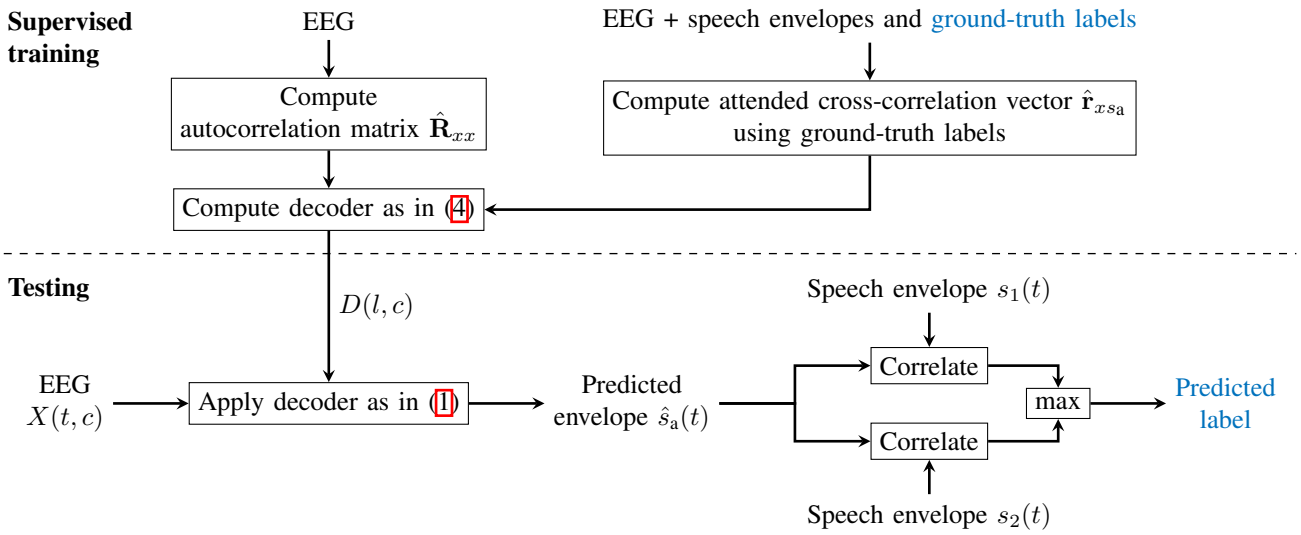


Figure 1: A conceptual overview of the traditional supervised training approach of a stimulus reconstruction decoder and its application to new test data.

where the regularization parameter λ needs to be optimized, e.g., through a cross-validation step. When sufficient training data is available, the regularization can be omitted [17].

In practice, a labeled training set of K segments (for example, corresponding to different trials in an experiment) of EEG data and corresponding speech envelopes of the competing speakers, $\{\mathbf{X}_k, (\mathbf{s}_{1_k}, \mathbf{s}_{2_k}), y_k\}_{k=1}^{K}$, is available. Note that in a practical system, these speech envelopes need to be extracted from the recorded speech mixtures in a hearing device, for which various methods exist [2], [18]–[20]. The labels $y_k \in \{1, 2\}$ indicate whether $\mathbf{s}_{1_k}$ or $\mathbf{s}_{2_k}$ is the attended speech envelope. Per segment k, the attended speech envelope $\mathbf{s}_{a_k}$ thus corresponds to the speech envelope of the set $(\mathbf{s}_{1_k}, \mathbf{s}_{2_k})$ that corresponds to label y_k. Then (5) becomes:

$$\hat{\mathbf{d}} = \underbrace{\left(\sum_{k=1}^{K} \mathbf{X}_k^{\mathrm{T}}\mathbf{X}_k + \lambda\mathbf{I}\right)}_{\hat{\mathbf{R}}_{xx}^{-1}}{}^{-1} \underbrace{\sum_{k=1}^{K} \mathbf{X}_k^{\mathrm{T}}\mathbf{s}_{a_k}}_{\hat{\mathbf{r}}_{xs_a}} \tag{6}$$

It is crucial to realize that the estimation of the decoder in (6) is inherently a *supervised* problem, as the ground-truth label y_k needs to be known to select the attended speech envelope $\mathbf{s}_{a_k}$ in each trial k.

At test time, the estimated decoder $\hat{\mathbf{d}}$ is used to reconstruct the attended speech envelope from a new EEG segment $\mathbf{X}^{(\text{test})}$. Given two speech envelopes $\mathbf{s}_1^{(\text{test})}$ and $\mathbf{s}_2^{(\text{test})}$, corresponding to two competing speakers, the first speaker is identified as the attended one if the sample Pearson correlation coefficient between the reconstructed speech envelope $\hat{\mathbf{s}}_a = \mathbf{X}^{(\text{test})}\hat{\mathbf{d}}$ and the first speaker is larger than with the second speaker, i.e.,

$$\rho\left(\hat{\mathbf{s}}_a, \mathbf{s}_1^{(\text{test})}\right) > \rho\left(\hat{\mathbf{s}}_a, \mathbf{s}_2^{(\text{test})}\right), \tag{7}$$

and vice versa. This is summarized in the 'Testing' part in Figure 1. Note that, for the sake of an easy exposition, we assume that there are two competing speakers, although all proposed algorithms can be generalized to more than two competing speakers.

B. Unsupervised training of a decoder

Assume the availability of a training set of K segments of EEG data and speech envelopes, $\{\mathbf{X}_k, (\mathbf{s}_{1_k}, \mathbf{s}_{2_k})\}_{k=1}^{K}$, but now *without* knowledge of the attended speaker, i.e., the labels y_k are not available. Only the presented competing speech envelopes $(\mathbf{s}_{1_k}, \mathbf{s}_{2_k})$ are known, of which one corresponds to the attended speaker, while the other corresponds to the unattended one. This means that training a decoder to reconstruct the attended speech envelope boils down to an *unsupervised* problem. We thus remove the requirement of subject-specific ground-truth labels. However, we implicitly assume that it is important for the training of the stimulus reconstruction decoder to know which envelope corresponds to the attended speaker and which one to the unattended speaker. In other words, we assume that the attended and unattended speaker are encoded differently in the brain. If this would not be the case, one could simply train the decoder based on the sum of the envelopes of both speakers. Such a training procedure would also be unsupervised and would remove the necessity of determining which speaker is attended during the training process. While the assumption that both competing speakers are encoded distinctly in the brain is already verified in the literature (e.g., see [5]), we also confirm it here in Section IV-B.

Figure 2 shows a conceptual overview of the proposed unsupervised training procedure, in which a decoder is trained in an unsupervised manner by iteratively (re)predicting the labels and updating the decoder. The key idea is thus to replace the ground-truth labels in the supervised training stage (top part of Figure 1), with the *predicted* labels from the testing stage (bottom part of Figure 1), and iterate a few times. Below, we will explain each step of the algorithm, while we refer to Algorithm 1 for a detailed summary.

In the first step, the autocorrelation matrix in (6) is estimated using the subject-specific EEG data. This autocorrelation matrix is independent of the ground-truth labels, which are only required for the cross-correlation vector. It is thus always possible to perform this update. If desired, the estimated and

Algorithm 1 Unsupervised training or adaptation of a stimulus reconstruction decoder

Input: A training set of K segments of EEG data and speech envelopes $\{\mathbf{X}_k, (\mathbf{s}_{1_k}, \mathbf{s}_{2_k})\}_{k=1}^{K}$; initial autocorrelation matrix $\mathbf{R}_{xx}^{(\text{init})}$ and cross-correlation vector $\mathbf{r}_{xs_\text{a}}^{(\text{init})}$; regularization parameter λ and updating hyperparameters α and β; maximal number of iterations $i_{\max}$

Output: A stimulus reconstruction decoder $\hat{\mathbf{d}}$

1: Compute/update the autocorrelation matrix and compute an initial decoder:

$$\begin{cases} \hat{\mathbf{R}}_{xx} = (1-\alpha)\left(\sum_{k=1}^{K} \mathbf{X}_k^\text{T}\mathbf{X}_k + \lambda\mathbf{I}\right) + \alpha\mathbf{R}_{xx}^{(\text{init})} \\ \hat{\mathbf{d}} = \hat{\mathbf{R}}_{xx}^{-1}\mathbf{r}_{xs_\text{a}}^{(\text{init})} \end{cases}$$

2: **while** $i \leq i_{\max}$ and $\hat{\mathbf{d}}$ changes **do**

3: Predict the labels on the training set:

$$\forall\, k \in \{1, \ldots, K\} : \begin{cases} \hat{\mathbf{s}}_k = \mathbf{X}_k\hat{\mathbf{d}} \\ \mathbf{s}_{\text{pred}_k} = \underset{\mathbf{s}_{1_k}, \mathbf{s}_{2_k}}{\text{argmax}} \left(\rho(\hat{\mathbf{s}}_k, \mathbf{s}_{1_k}), \rho(\hat{\mathbf{s}}_k, \mathbf{s}_{2_k})\right) \end{cases}$$

4: Update the cross-correlation vector using the predicted labels and update the decoder:

$$\begin{cases} \hat{\mathbf{r}}_{xs_\text{pred}} = (1-\beta)\sum_{k=1}^{K} \mathbf{X}_k^\text{T}\mathbf{s}_{\text{pred}_k} + \beta\mathbf{r}_{xs_\text{a}}^{(\text{init})} \\ \hat{\mathbf{d}} = \hat{\mathbf{R}}_{xx}^{-1}\hat{\mathbf{r}}_{xs_\text{pred}} \end{cases}$$

5: **end while**

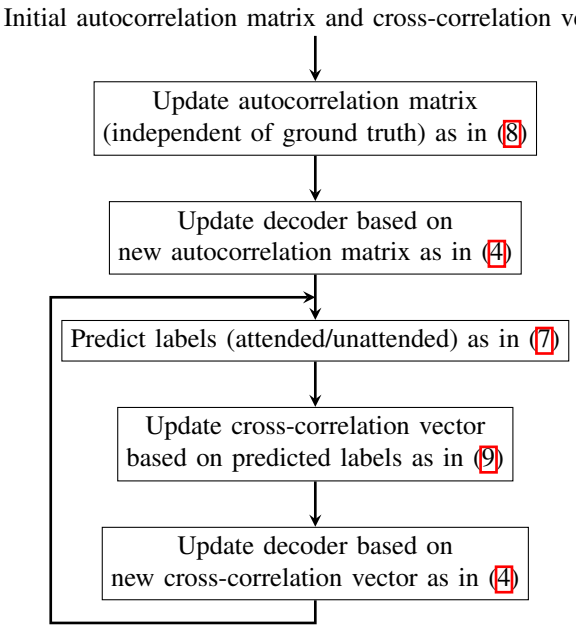


Figure 2: A conceptual overview of the iterative self-adaptive unsupervised training procedure of a stimulus reconstruction decoder.

regularized autocorrelation matrix can be linearly combined with an initially provided autocorrelation matrix $\mathbf{R}_{xx}^{(\text{init})}$, controlled with the user-defined hyperparameter $0 \leq \alpha \leq 1$ (and $1-\alpha$):

$$\hat{\mathbf{R}}_{xx} = (1-\alpha)\left(\sum_{k=1}^{K} \mathbf{X}_k^\text{T}\mathbf{X}_k + \lambda\mathbf{I}\right) + \alpha\mathbf{R}_{xx}^{(\text{init})}. \quad (8)$$

This hyperparameter can be interpreted as the amount of confidence in the a priori available autocorrelation matrix $\mathbf{R}_{xx}^{(\text{init})}$. This initial autocorrelation matrix can be estimated on, for example, subject-independent data and can be considered as an extra regularization term (e.g., as used in Tikhonov regularization). If no such a priori autocorrelation matrix is available, α is simply set to 0. Using the updated autocorrelation matrix, the decoder is estimated in combination with an initially provided cross-correlation vector $\mathbf{r}_{xs_\text{a}}^{(\text{init})}$. This cross-correlation vector can again be estimated in a subject-independent manner but could also be generated fully randomly. It is recommended to normalize the initial autocorrelation matrix and cross-correlation vector such that they have a Frobenius norm equal to the estimated auto-/cross-correlation matrix/vector, improving the interpretability of the hyperparameters.

Using the updated autocorrelation matrix (8) and the initial cross-correlation vector $\mathbf{r}_{xs_\text{a}}^{(\text{init})}$, we compute an initial decoder $\hat{\mathbf{d}}$ according to (4). This initial decoder acts as a bootstrap to initiate the iterative procedure to update the decoder weights. Starting from this initial decoder, the labels of the training segments are predicted based on the maximal sample Pearson correlation coefficient between the reconstructed envelope and the speech envelopes of the competing speakers. These predicted labels are then used to select the attended speech envelope $\mathbf{s}_{\text{pred}_k}$ in each of the K segments, which is afterwards used to update the cross-correlation vector. Note that it is crucial that the updating is performed not using the reconstructed envelope from the EEG, but with the speech envelope of one of the two competing speakers identified/predicted as the attended one. Again, some prior knowledge can be introduced in the updating of the cross-correlation vector using an initially provided cross-correlation vector $\mathbf{r}_{xs_\text{a}}^{(\text{init})}$ and hyperparameter

$0 \leq \beta \leq 1$:

$$\hat{\mathbf{r}}_{xs_{\text{pred}}} = (1-\beta)\sum_{k=1}^{K} \mathbf{X}_k^{\mathrm{T}} \mathbf{s}_{\text{pred}_k} + \beta \mathbf{r}_{xs_{\text{a}}}^{(\text{init})}. \tag{9}$$

The updated cross-correlation vector can then be used to re-estimate the decoder. Multiple iterations of predicting the labels and updating the decoder can be performed until the decoder has converged or a maximal number of iterations has been reached. It is expected that this iterative process initiates a *self-leveraging* effect, in which the decoder leverages its own predictions to improve. In Section IV, we provide a mathematical analysis that explains the underlying mechanism behind this self-leveraging effect and why it works.

Using the unsupervised updating scheme in Algorithm 1, a stimulus reconstruction decoder can be trained. In Section V, we evaluate this unsupervised algorithm using different hyperparameter settings and compare it to a supervised subject-independent and supervised subject-specific decoder.

III. Experiments and Evaluation Metrics

In this section, we provide all information on the data (Section III-A), preprocessing and decoder settings (Section III-B), and evaluation procedure and metrics (Section III-C) required to replicate and reproduce all experiments and results. All experiments are performed in MATLAB.

A. AAD datasets

We validate the proposed unsupervised AAD algorithm on two separate datasets. The first one (Dataset I) consists of EEG recordings of 16 normal-hearing subjects, attending to one out of two competing speakers [17]. These competing speakers are located at $+/-90^\circ$ along the azimuth direction. Per subject, 72 min of EEG and audio data are available. This dataset is available online [21].

The second dataset (Dataset II) consists of EEG recordings of 18 normal-hearing subjects, attending to one out of two competing speakers, located at $+/-60^\circ$ along the azimuth direction [22]. Per subject, 50 min of EEG and audio data are available. Different acoustic room settings are used: anechoic, mildly reverberant, and highly reverberant. This dataset is available online as well [23]. Both datasets are recorded using a 64-channel BioSemi ActiveTwo system.

B. Preprocessing and decoder settings

The preprocessing of the EEG and audio data is very similar to [17]. The audio signals are first filtered using a gammatone filterbank. From each subband signal, the envelope is extracted using a power-law operation with exponent 0.6, after which one final envelope is computed by summing the different subband envelopes. Both the EEG data and speech envelopes are filtered between 1–9 Hz [24] and downsampled to 20 Hz. Note that we here assume that the clean speech envelopes are readily available and need not be extracted from the microphone recordings [3]. For Dataset II, the 50 s segments are normalized such that they have a Frobenius norm equal to one across all channels.

A maximum of $i_{\max} = 10$ iterations of predicting the labels and updating the decoder is used, which in practice showed to be sufficient (see also Section V).

In the design of the stimulus reconstruction decoder, $L = 250\,\text{ms}$ is chosen [9], such that the filter spans a range of 0–250 ms post-stimulus. Furthermore, the regularization parameter λ in (5), (6), and Algorithm 1 is analytically determined using [25], which is the recommended state-of-the-art method to estimate this regularization parameter [26]. Given data matrix $\mathbf{X} \in \mathbb{R}^{T\times p}$ and sample autocorrelation matrix $\mathbf{S} = \mathbf{X}^{\mathrm{T}}\mathbf{X} \in \mathbb{R}^{p\times p}$, the proposed shrinkage estimator $\hat{\mathbf{S}}$ in [25] of the autocorrelation matrix becomes [27]:

$$\hat{\mathbf{S}} = (1-\eta)\mathbf{S} + \eta \frac{\text{Tr}\,(\mathbf{S})}{p}\mathbf{I}, \tag{10}$$

with

$$\eta = \min\left(\frac{\sum_{t=1}^{T} \left|\left| \mathbf{x}_t\mathbf{x}_t^{\mathrm{T}} - \mathbf{S} \right|\right|_{\text{F}}^2}{T^2\left(\text{Tr}\left(\mathbf{S}^{\mathrm{T}}\mathbf{S}\right) - \frac{\text{Tr}(\mathbf{S})^2}{p}\right)}, 1 \right). \tag{11}$$

Note that in our case, $p = CL$. The shrinkage formula in (10) can easily be rewritten in the form of (5), (6) upon an irrelevant scaling, in which case λ is set as:

$$\lambda = \frac{\eta}{1-\eta}\frac{\text{Tr}\,(\mathbf{S})}{p}.$$

In [25], they show that (10) and (11) lead to a consistent estimator that is asymptotically optimal w.r.t. a quadratic loss function with the underlying unknown autocorrelation matrix.

C. Cross-validation and evaluation

For the *supervised subject-specific* decoder, a random ten-fold cross-validation scheme is used to train and test the decoders. The supervised *subject-independent* decoders are evaluated using a leave-one-subject-out cross-validation scheme where a decoder is trained on the data of all other subjects and tested on the left-out subject. The proposed *unsupervised subject-specific* decoder is tested in a random ten-fold cross-validation manner as well, where the updating happens on the training set (without knowledge of the labels) and the testing on the left-out data. The partitioning of the data is performed on segments of 60 s for Dataset I and 50 s for Dataset II. Per subject, the continuous recordings are thus first split into these segments and then randomly distributed over a training and test set. At test time, the left-out $60/50$ s segments are split into smaller sub-segments, from hereon referred to as 'decision windows'. The accuracy is then defined as the ratio of correctly decoded decision windows across all test folds. These shorter decision windows are only used in the test folds, in order to evaluate the trade-off between the AAD accuracy and the decision window length [3], [28] (longer decision windows provide more accurate correlation coefficients, yielding higher AAD accuracies at the cost of slower decision-making). However, the prediction and updating in Algorithm 1 are always performed on the longer $60/50$ s segments, in order to maximize the accuracy of the unsupervised labels.

To resolve the aforementioned trade-off between accuracy and decision window length, the *minimal expected switch duration* (MESD) was proposed in [28] as a performance metric for AAD. The MESD represents the theoretical expected time it takes to switch the gain in an optimal attention-steered gain control system, following a switch in auditory attention. Such a gain control system is modeled using a Markov chain model, where the time it takes to step from one state (i.e., gain level) to another is represented by the AAD decision window length and where the step size between gain levels is optimized to ensure stable operation within a pre-defined comfort region in the presence of AAD errors. The expected switch duration can then be computed by quantifying the expected number of steps required to switch to the pre-defined comfort region associated with the other speaker. This gain control system/Markov chain model is optimized across decision window lengths to minimize the time it takes to switch the gain from one source to another while assuring a stable operation within the pre-defined comfort region when the attention is sustained. Note that this metric is computed based on a stochastic model of a gain control system and is not evaluated using actual switches in attention. However, it allows to easily and statistically compare different decoders across different decision window lengths based on a single (practically relevant) metric. As such, it resolves the aforementioned accuracy-vs-decision-time trade-off. The underlying mathematical principles and definition of this metric can be found in [28]. To compute the MESD, we used the publicly available MESD toolbox from [29].

IV. Unsupervised Updating Explained: A Mathematical Model

Before extensively testing Algorithm 1 on the different datasets in Section V, we attempt to demystify and explain the hypothesized self-leveraging mechanism through a mathematical analysis of the recursion induced by the algorithm. The busy reader can skip to Section V for the results.

A. Mathematical model

Assume that at iteration $i < i_{\max}$ of Algorithm 1, we obtain a decoder with an (unknown) AAD test accuracy of $p_i \in [0, 100]\%$. This means that there is a probability of p_i that the reconstructed envelope using this decoder will have a higher correlation with the attended envelope than with the unattended envelope. Correspondingly, there is a $100\% - p_i$ probability that the unattended envelope will show the highest correlation. Assume for simplicity that $\alpha = 0$ and $\beta = 0$. Due to the linearity of the computation of the cross-correlation vector (see (3)), the updated cross-correlation vector will then be, on average, equal to:

$$\hat{\mathbf{r}}_{xs_{\text{pred}},i+1} = p_i \hat{\mathbf{r}}_{xs_\text{a}} + (1 - p_i)\hat{\mathbf{r}}_{xs_\text{u}}, \tag{12}$$

with $\hat{\mathbf{r}}_{xs_\text{a}}$ the cross-correlation vector using all attended envelopes and $\hat{\mathbf{r}}_{xs_\text{u}}$ the cross-correlation vector using all unattended envelopes. Similarly, and again due to the linearity in the computations, the corresponding updated decoder becomes:

$$\hat{\mathbf{d}}_{i+1} = p_i \hat{\mathbf{d}}_\text{a} + (1 - p_i)\hat{\mathbf{d}}_\text{u}, \tag{13}$$

with $\hat{\mathbf{d}}_\text{a}$ the decoder trained with all attended speech envelopes (which would correspond to the supervised subject-specific decoder with accuracy p_a) and $\hat{\mathbf{d}}_\text{u}$ the unattended decoder that would be trained with all unattended speech envelopes. This unattended decoder has an accuracy equal to p_u on the unattended labels and thus $100\% - p_\text{u}$ on the attended labels. As a result, the reconstructed envelope using this updated decoder is a linear combination of the reconstructed envelope obtained using the (supervised) attended decoder ($\hat{\mathbf{s}}_\text{a}$) and the (supervised) unattended decoder ($\hat{\mathbf{s}}_\text{u}$):

$$\hat{\mathbf{s}}_{\text{pred},i+1} = p_i \hat{\mathbf{s}}_\text{a} + (1 - p_i)\hat{\mathbf{s}}_\text{u}. \tag{14}$$

The goal is now to find the AAD accuracy p_{i+1} of the updated decoder $\hat{\mathbf{d}}_{i+1}$ (13) in iteration $i+1$. We will propose a mathematical model for the function $p_{i+1} = \phi(p_i)$, which determines the accuracy p_{i+1} of the updated decoder as a function of the accuracy p_i of the previous decoder. If $p_{i+1} > p_i$, this implies a self-leveraging effect in which the accuracy improves from one iteration to the next. Given that the speech envelope that exhibits the highest Pearson correlation coefficient with the reconstructed envelope is identified as the attended speaker, this implies that:

$$p_{i+1} = \phi(p_i) = P(\rho(\hat{\mathbf{s}}_{\text{pred},i+1}, \mathbf{s}_\text{a}) > \rho(\hat{\mathbf{s}}_{\text{pred},i+1}, \mathbf{s}_\text{u}))\,, \tag{15}$$

with $\mathbf{s}_\text{a}$ and $\mathbf{s}_\text{u}$ the speech envelopes of the attended and unattended speaker. Using (14) and the definition of the Pearson correlation coefficient of two random variables X and Y:

$$\rho(X, Y) = \frac{\mathbb{E}\{(X - \mu_X)\,(Y - \mu_Y)\}}{\sigma_X \sigma_Y},$$

with the mean $\mu_{X/Y}$ and standard deviation $\sigma_{X/Y}$, (15) becomes:

$$\begin{aligned}\phi(p_i) &= P(p_i\sigma_{\hat{\mathbf{s}}_\text{a}}\rho(\hat{\mathbf{s}}_\text{a}, \mathbf{s}_\text{a}) + (1 - p_i)\,\sigma_{\hat{\mathbf{s}}_\text{u}}\rho(\hat{\mathbf{s}}_\text{u}, \mathbf{s}_\text{a}) \\ &\qquad > p_i\sigma_{\hat{\mathbf{s}}_\text{a}}\rho(\hat{\mathbf{s}}_\text{a}, \mathbf{s}_\text{u}) + (1 - p_i)\,\sigma_{\hat{\mathbf{s}}_\text{u}}\rho(\hat{\mathbf{s}}_\text{u}, \mathbf{s}_\text{u})) \\ &= P(p_i\sigma_{\hat{\mathbf{s}}_\text{a}}\,(\rho(\hat{\mathbf{s}}_\text{a}, \mathbf{s}_\text{a}) - \rho(\hat{\mathbf{s}}_\text{a}, \mathbf{s}_\text{u})) \\ &\qquad > (1 - p_i)\,\sigma_{\hat{\mathbf{s}}_\text{u}}\,(\rho(\hat{\mathbf{s}}_\text{u}, \mathbf{s}_\text{u}) - \rho(\hat{\mathbf{s}}_\text{u}, \mathbf{s}_\text{a}))).\end{aligned} \tag{16}$$

To simplify this expression, and without loss of generality[2], we assume that both speech envelopes have a similar energy content such that it is safe to assume that, on average, $\sigma_{\hat{\mathbf{s}}_\text{a}} = \sigma_{\hat{\mathbf{s}}_\text{u}}$. Furthermore, $\rho(\hat{\mathbf{s}}_\text{a}, \mathbf{s}_\text{a})\,, \rho(\hat{\mathbf{s}}_\text{a}, \mathbf{s}_\text{u})\,, \rho(\hat{\mathbf{s}}_\text{u}, \mathbf{s}_\text{u})\,,$ and $\rho(\hat{\mathbf{s}}_\text{u}, \mathbf{s}_\text{a})$ are independent of p_i and can be considered as random variables $\rho_{\text{aa}}, \rho_{\text{au}}, \rho_{\text{uu}},$ and ρ_{ua}. These random variables represent the correlation coefficients between the reconstructed envelopes using the attended/unattended decoders and the speech envelopes of the attended/unattended speakers, computed over a pre-defined window length. As such, (16) becomes:

$$\phi(p_i) = P\left(\rho_{\text{aa}} - \rho_{\text{au}} > \frac{1 - p_i}{p_i}\,(\rho_{\text{uu}} - \rho_{\text{ua}})\right). \tag{17}$$

Define now the new random variables $R_1 = \rho_{\text{aa}} - \rho_{\text{au}} \sim \mathcal{N}(\mu_1, \sigma)$ and $R_2 = \rho_{\text{uu}} - \rho_{\text{ua}} \sim \mathcal{N}(\mu_2, \sigma)$. We assume

[2]This can always be obtained by normalizing the (reconstructed) envelopes.

that these random variables are normally distributed[3] with known mean and equal standard deviation. These means and standard deviation can be derived a priori from the supervised subject-specific decoders and experiments (note that these are not available in the unsupervised case, yet for analysis and validation purposes, we can use a supervised setting to estimate these). R_1 represents the difference between the correlation coefficients of both competing speakers when using the (supervised) attended decoder, while R_2 would be used when making AAD decisions based on the (supervised) unattended decoder. As the standard deviation of R_1 and R_2 is mostly determined by the noise, which is the same for the attended and unattended decoder, we can assume that they have the same standard deviation σ. This standard deviation can be estimated across the mean-centered $\tilde{R}_1 = R_1 - \mu_1$ and $\tilde{R}_2 = R_2 - \mu_2$ variables.

Finally, we can define $Z = R_1 - \frac{1-p_i}{p_i} R_2$, which is again normally distributed:

$$Z \sim \mathcal{N}\left(\mu_z(p_i), \sigma_z(p_i)\right),$$

with

$$\mu_z(p_i) = \mu_1 - \frac{1-p_i}{p_i}\mu_2 \text{ and } \sigma_z(p_i) = \sigma\sqrt{1 + \frac{(1-p_i)^2}{p_i^2}},$$

assuming that R_1 and R_2 are uncorrelated[4]. Equation (17) then becomes equal to $P\left(Z > 0\right)$, or equivalently:

$$\phi(p_i) = \frac{1}{\sigma_z(p_i)\sqrt{2\pi}} \int_0^{+\infty} e^{-\frac{1}{2}\left(\frac{x-\mu_z(p_i)}{\sigma_z(p_i)}\right)^2} dx. \qquad (18)$$

By numerically evaluating (18) for $p_i \in [0, 100]\%$, we have modeled the AAD accuracy p_{i+1} in iteration $i+1$ as a function of the AAD accuracy p_i in iteration i. Note that p_i and $p_{i+1} = \phi(p_i)$ refer here to the *test* accuracy, as the model parameters will be computed from the correlation coefficients resulting from applying the subject-specific attended/unattended decoders to left-out test data.

Figure 3 shows the modeled curve $\phi(p_i)$ where μ_1, μ_2, and σ are estimated from Dataset I. The modeling is performed per subject based on the correlation coefficients of the attended and unattended decoders tested on 60 s decision windows with ten-fold cross-validation. The modeled curves are then averaged across all subjects to obtain one 'universal' updating curve in Figure 3.

1) Verification of the $\phi(p_i)$ model: The updating curve in Figure 3 can be verified using simulations. Consider an oracle that can produce any mixture $(p_i, 100\% - p_i)$ of correct and incorrect labels. Using this oracle, we can perform a sweep of p_i values and compute a decoder based on this particular ratio of correct and incorrect labels. For each p_i, the corresponding decoder can be applied to the test set to evaluate p_{i+1}, which should be approximately equal to $\phi(p_i)$ if the model is correct. The simulated curve shown in Figure 3 is generated using random ten-fold cross-validation, repeated five times per subject, and averaged over subjects, folds, and runs. As the simulated curve closely resembles the theoretical curve, we can confirm that the assumptions are sensible and that the theoretical updating curve (18) is valid and useful for interpretation and analysis.

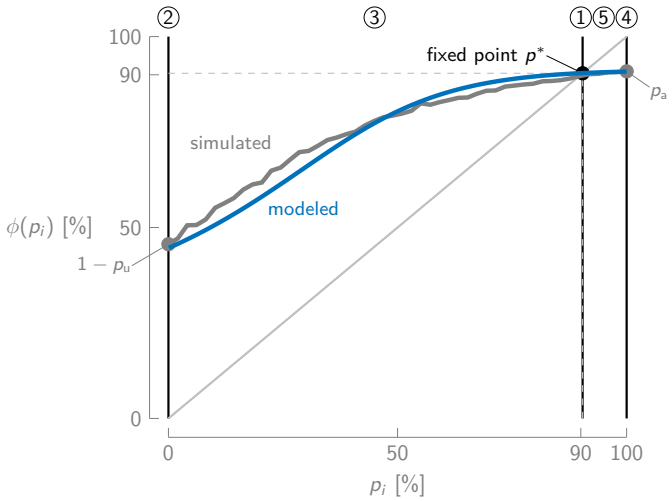


Figure 3: The modeled updating curve, averaged over all subjects of Dataset I, shows the accuracy $\phi(p_i)$ after updating, starting from a decoder with accuracy p_i, and closely corresponds to the simulated curve. As a reference, the identity line is added, where the updated accuracy is equal to the initial accuracy.

B. Explaining the updating

1) Analysis of the updating curve: In Figure 3, five points/regions are indicated, which are discussed below:

- Point ① corresponds to $p_i = p^*$, i.e., the cross-over point. For initial accuracy p^*, the updated accuracy remains the same, i.e., $\phi(p^*) = p^*$. This cross-over point thus corresponds to the *fixed/invariant* point of $\phi(p_i)$.
- Point ② corresponds to $p_i = 0\%$, i.e., the decoder is trained using *only* the unattended ground-truth labels and is thus equal to $\hat{\mathbf{d}}_u$. The updated accuracy then corresponds to $100\% - p_u$, as the *unattended* decoder is used to predict *attended* labels. The unattended decoder generally performs worse than the attended decoder, obtaining accuracies below 100%, such that $\phi(0\%) > 0\%$, ergo, an increase in accuracy. This, furthermore, also confirms that unattended speech envelope is encoded differently in the brain than the attended speech envelope.
- Region ③ corresponds to $0\% \leq p_i < p^*$. In this region, the accuracy increases after updating, i.e., $\phi(p_i) > p_i$. Even when using a majority of *unattended* speech envelopes to train the *attended* decoder, the accuracy increases. A possible explanation is that the resulting correlation vector still conveys information about which channels and which time lags are best suited to decode speech from the EEG, albeit unattended speech. It seems that there is still information to gain from unattended

[3]For none of the 16 subjects in Dataset I, the Kolmogorov-Smirnov test indicates a deviation from a normal distribution, which provides empirical support for this assumption, in addition to the validation of the final model that we provide in Section IV-A1.

[4]For none of the 16 subjects in Dataset I, there is a significant correlation between R_1 and R_2, which supports this assumption, in addition to the validation of the final model in Section IV-A1.

speech to compensate for the limited amount of attended speech. However, when p_i increases, the increase in accuracy in general decreases (i.e., the distance to the identity line decreases), possibly because there is less and less information to gain from the unattended speech. Furthermore, it is expected that the cross-correlation of the EEG with the attended speech envelopes ($\hat{\mathbf{r}}_{xs_\mathrm{a}}$) is on average larger than of the EEG with the unattended speech ($\hat{\mathbf{r}}_{xs_\mathrm{u}}$). This reduces the relative weight of the unattended cross-correlation vector (e.g., see (12)) and could make the attended cross-correlation vector more prominent in the estimated one, even when more unattended labels are used, enabling the self-leveraging effect.

- Point ④ corresponds to $p_i = 100\%$, i.e., the decoder corresponds to the supervised subject-specific decoder from Figure 5a, with accuracy p_a. As even the attended decoder is not perfect, $\phi(100\%) < 100\%$, which results in a decrease in accuracy. This could be due to modeling errors (limited capacity of a linear model), the low signal-to-noise ratio of the stimulus-response in the EEG, and a small amount of incorrect ground-truth labels, for example, due to the subject's attention wandering off to the wrong speaker.
- Region ⑤ corresponds to $p^* < p_i < 100\%$, where the accuracy decreases after updating, i.e., $\phi(p_i) < p_i$. The presence of unattended labels does not add information as in region ③, suffering from the same limitations as in point ④.

Lastly, because of the linearity of (3), the point $p_i = 50\%$ reflects the case where one would train the decoder based on the sum of both speech envelopes (i.e., across attended and unattended speaker). As discussed in Section II-B, we implicitly assume that the attended and unattended speech envelopes are encoded differently in the brain. If not, the unsupervised training of a decoder based on the sum of the speech envelopes would result in a similar accuracy as the proposed unsupervised training method. The updating curve in Figure 3, however, shows that $\phi(50\%) < \phi(p^*)$. This indicates that such an unsupervised decoder trained on the sum of the speech envelopes performs worse than the proposed unsupervised method. As such, it confirms the assumption that both speech envelopes are encoded distinctly in the brain and that the inclusion of the unattended envelope misdirects the computation of the cross-correlation vector in (3).

2) A fixed-point iteration algorithm: Using the theoretical model in Figure 3, we can explain the unsupervised AAD algorithm in Algorithm 1 as a *fixed-point iteration* $p_{i+1} = \phi(p_i)$ on this curve. Before analyzing the uniqueness and convergence properties based on the model (18), we first provide an intuitive explanation of why there could only be *one* fixed point p^*. First of all, it is safe to assume that $\phi(0\%) > 0\%$, as the unattended decoder is never perfect. Furthermore, it is very unlikely that regions ③ and ⑤ in Figure 3 would alternate, as this would mean that, when using more attended labels to train the decoder, there is an increase-decrease-increase of AAD accuracy (or the other way around) with respect to the initial accuracy. This implies that there is a unique fixed point. We show in a Supplementary Material paper [1] that, based on the model (18), the existence, uniqueness, and convergence of/to the fixed point are indeed mathematically guaranteed when three reasonable conditions on the accuracy p_a of the (supervised) attended decoder and the accuracy p_u of the (supervised) unattended decoder (on the unattended speech) are satisfied. Furthermore, we also demonstrate in the Supplementary Material paper [1] that these conditions are satisfied for all subjects in both datasets.

These fixed-point iteration properties are also intuitively apparent from Figure 3 and hold in every example we have encountered in practice so far. This means that we could initialize the updating algorithm with *any* decoder, as we would always arrive at the fixed point p^*. As a result, it explains why the updating procedure is possible starting from a random decoder. Figure 4 shows how the fixed-point paths (on average across all folds) follow the theoretical model for three representative subjects of Dataset I, starting from a random decoder.

The fixed point $\hat{p}^*$ based on the theoretical model (where the means and standard deviation in (18) are computed per subject individually) should thus give a good approximation of the unsupervised AAD accuracy p^*. Across all 16 subjects of Dataset I, on 60 s decision windows, the mean absolute error between the predicted and actual unsupervised AAD accuracy is 3.45%. We can thus accurately predict how well the unsupervised updating will perform by computing the fixed point of (18), where the parameters μ_1, μ_2, and σ in (18) can be easily computed from the corresponding *supervised* subject-specific decoders. Furthermore, as mentioned above, the model (18) also allows showing convergence to this fixed point when three reasonable conditions are satisfied (see the Supplementary Material paper [1]).

V. Results and Discussion

In this section, we extensively validate the unsupervised algorithm on the two datasets and compare it with a supervised subject-independent and supervised subject-specific decoder.

A. Random initialization

We first evaluate the proposed unsupervised algorithm using a random initialization and without using any prior knowledge. As such, in Algorithm 1, we set $\alpha = 0$ and $\beta = 0$. The cross-correlation vector $\mathbf{r}_{xs_\mathrm{a}}^{(\mathrm{init})}$ is initialized at random from a multivariate uniform distribution. Figure 5 shows for both datasets the AAD accuracy as a function of decision window length and the MESD values per subject for the supervised subject-specific decoder, the subject-independent decoder, and the proposed unsupervised subject-specific decoder (with random initialization). The significance level in Figure 5a and 5b is computed using the inverse binomial distribution as in [9].

As mentioned in Section I, it is clear that a supervised subject-specific decoder outperforms a subject-independent decoder on both datasets (Figure 5). A Wilcoxon signed-rank test between the MESD values, with a Bonferroni-Holm correction for multiple comparisons, confirms this (Dataset I: $n = 16, p = 0.0022$, Dataset II: $n = 18, p = 0.0030$). On both datasets, the proposed unsupervised subject-specific

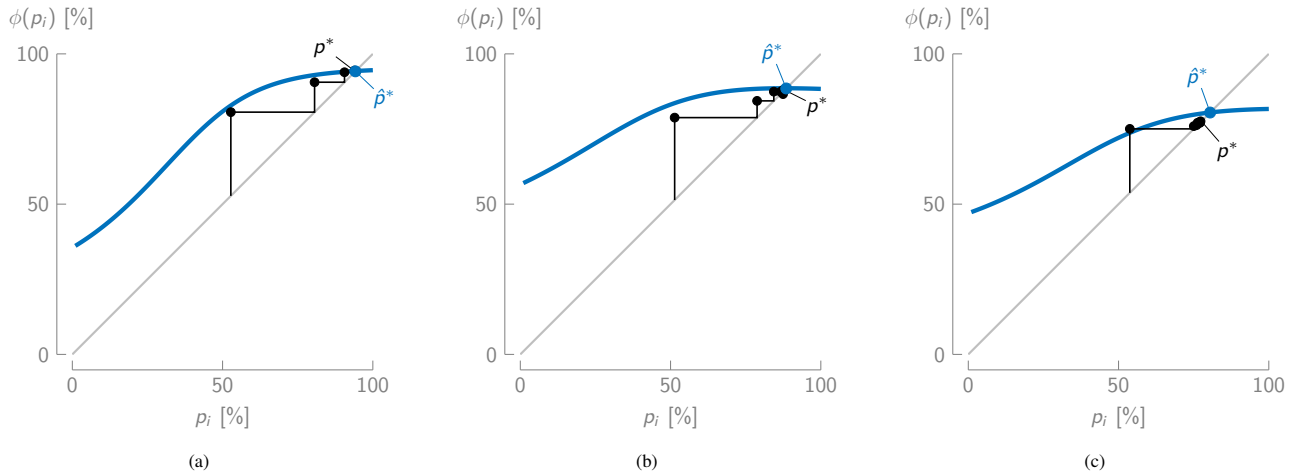


Figure 4: The fixed-point iteration paths followed by three representative subjects (a), (b), (c) from Dataset I closely follow the theoretical model. The predicted fixed point $\hat{p}^*$ from the theoretical model accurately predicts the actual fixed point p^*.

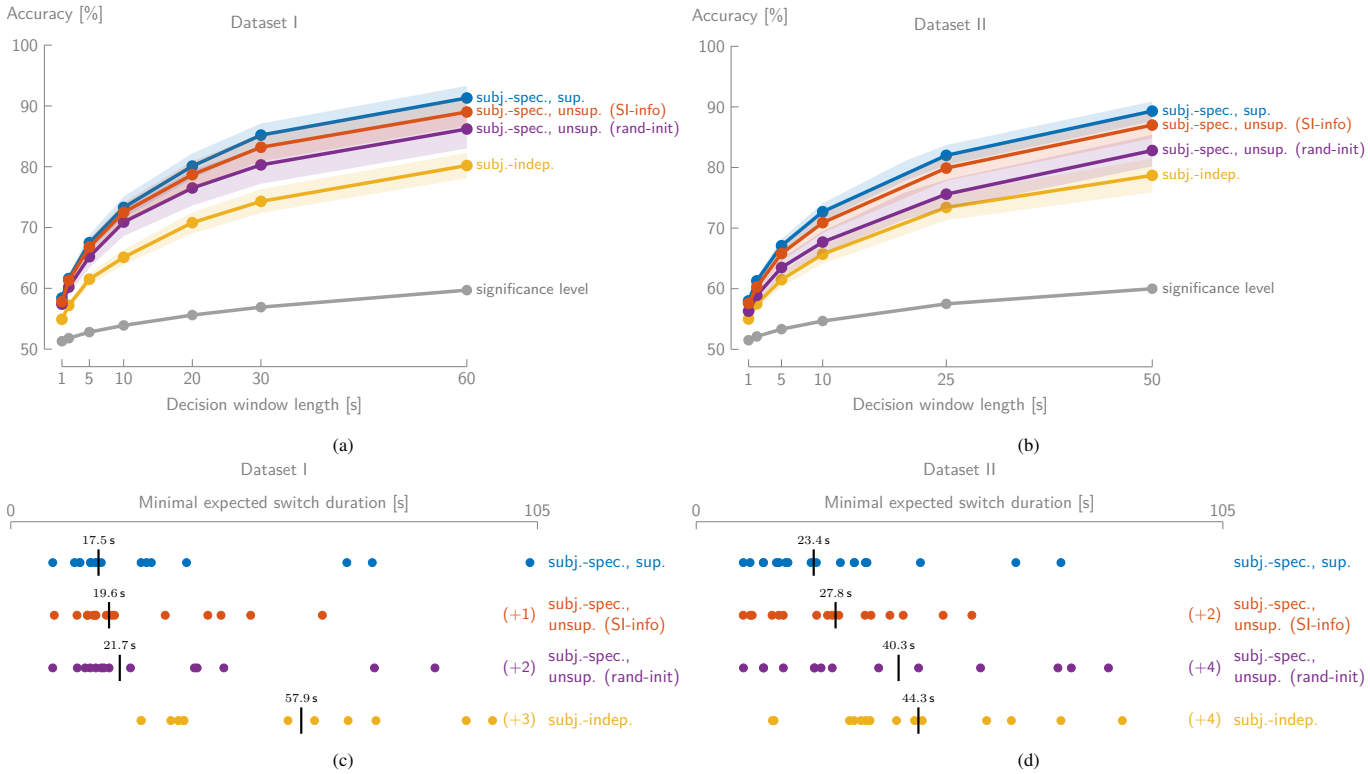


Figure 5: (a) The unsupervised subject-specific decoder, with both types of initialization (random: rand-init, subject-independent information: SI-info) clearly outperforms a subject-independent decoder, while approximating the performance of a supervised subject-specific decoder especially on short decision windows (mean ± standard error of the mean (shading) across subjects). (b) The same trend occurs for Dataset II, although the unsupervised subject-specific decoder with random initialization outperforms the subject-independent decoder less apparent. (c) The per-subject MESD values (each subject = one dot) of Dataset I, with the median indicated with the black bar, confirm that the unsupervised subject-specific decoder outperforms the subject-independent decoder. The number of outlying values that fell off the plot are indicated with (+x) (outliers are still included in the quantitative analysis). (d) The same for Dataset II as (c).

decoder with random initialization outperforms the subject-independent decoder as well (although less clearly on Dataset II). Furthermore, it approximates the performance of the supervised subject-specific decoder, especially for the shorter decision window lengths. However, it does so without requiring ground-truth labels and thus retains the 'plug-and-play' feature of the subject-independent decoder. A Wilcoxon signed-rank test between the MESD values, again with a Bonferroni-Holm correction, shows a significant difference between the unsupervised subject-specific decoder with random initialization and the supervised subject-independent decoder on Dataset I ($n = 16, p = 0.0458$), but not on Dataset II ($n = 18, p = 0.5862$). Lastly, there is a significant difference between the supervised and unsupervised subject-specific decoder with random initialization (Dataset I: $n = 16, p = 0.0034$, Dataset II: $n = 18, p = 0.0010$).

Note that this last result is not per se a negative result: it is not expected that an unsupervised subject-specific decoder, updated starting from a completely random decoder, performs as well as the supervised version. The most important result is that the proposed unsupervised algorithm outperforms a subject-independent decoder, *even* when starting from a random decoder and while not requiring subject-specific ground-truth labels as well. Furthermore, such an unsupervised algorithm could be implemented on a generic hearing device, which trains and adapts itself from scratch to a new user.

Convergence plots: Figure 6 shows the AAD accuracy as a function of the iteration index for all subjects of Dataset I. Computing a decoder with the subject-specific autocorrelation matrix, but with a random cross-correlation vector, seems not to perform better than chance (iteration 0). Surprisingly, even after one iteration of predicting the labels using the decoder after iteration 0, which performs on chance level, and updating the cross-correlation vector, a decoder is obtained that on average performs with $\approx 75\%$ accuracy on 60 s decision windows (see also Figure 3). This implies that even using a random mix of attended and unattended labels results in a decoder that performs much better than chance. In the following iterations, the decoder keeps improving, settling after 4-5 iterations. This matches the fixed-point iteration interpretation of Section IV-B and Figures 3 and 4, explaining the self-leveraging mechanism.

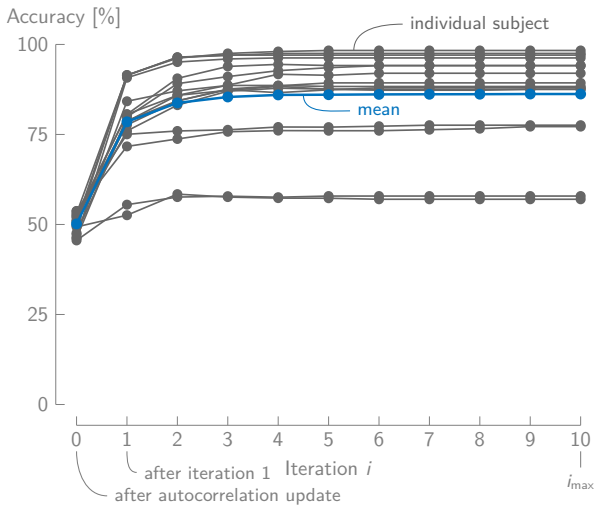


Figure 6: The convergence plots for all subjects of Dataset I using a random initialization, on 60 s decision windows, show that the AAD accuracy converges to the final unsupervised subject-specific accuracy after 4-5 iterations.

B. Subject-independent initialization/information

To use the information in the subject-independent decoder to our advantage, we can put $\alpha \neq 0$ and $\beta \neq 0$ in Algorithm 1. By adding subject-independent information to the estimation of both the autocorrelation matrix and the cross-correlation vector, we can further improve the updating behavior when starting from a random initialization (Section V-A). Especially in the estimation of the cross-correlation vector, the subject-independent cross-correlation vector, which is estimated using ground-truth labels, can compensate for prediction errors.

The initial autocorrelation matrix $\mathbf{R}_{xx}^{(\text{init})}$ and cross-correlation vector $\mathbf{r}_{xs_\text{a}}^{(\text{init})}$ are determined using the (supervised) information of all *other* subjects. The hyperparameters α and β are determined empirically. For Dataset I, $\alpha = 0$ is chosen, i.e., no subject-independent information is used in the autocorrelation estimation. Furthermore, $\beta = \frac{1}{3}$ is chosen, i.e., the subject-independent cross-correlation is half as important as the computed subject-specific one.

The results on Dataset I of this unsupervised subject-specific decoder using subject-independent information are shown in Figure 5a and 5c. Remarkably, the unsupervised procedure here results in a decoder that very closely approximates the supervised subject-specific decoder, *without* requiring subject-specific ground-truth labels. Based on the MESD values, there is no significant difference to be found between the supervised and unsupervised subject-specific decoder with subject-independent information (Wilcoxon signed-rank test with Bonferroni-Holm correction: $n = 16, p = 0.3259$). For 6 subjects, the unsupervised decoder performs even better than the supervised subject-specific one (see also Figure 5c). Furthermore, note that using the subject-independent information with respect to a random initialization and no further information not only fixes poor updating results for some of the outlying subjects but also improves on most other subjects (12 out of 16).

For Dataset II, it turns out that $\alpha = 0.5$ and $\beta = 0.5$, i.e., an equal weight to the subject-specific and subject-independent information, are good choices. Given that the unsupervised subject-specific decoder with random initialization performs worse than in Dataset I, it is not unexpected that a larger weight β of the subject-independent information is required to improve on the unsupervised procedure.

Figure 5b and 5d show the results on Dataset II of the unsupervised procedure with subject-independent information and with the aforementioned choices of the hyperparameters. The usage of subject-independent information results here in an even larger improvement over the random initialization (e.g., both in MESD, for 15 out of 18 subjects, as spread around the median in Figure 5d) and again closely approximates the supervised subject-specific performance, without requiring subject-specific ground-truth labels. However, based

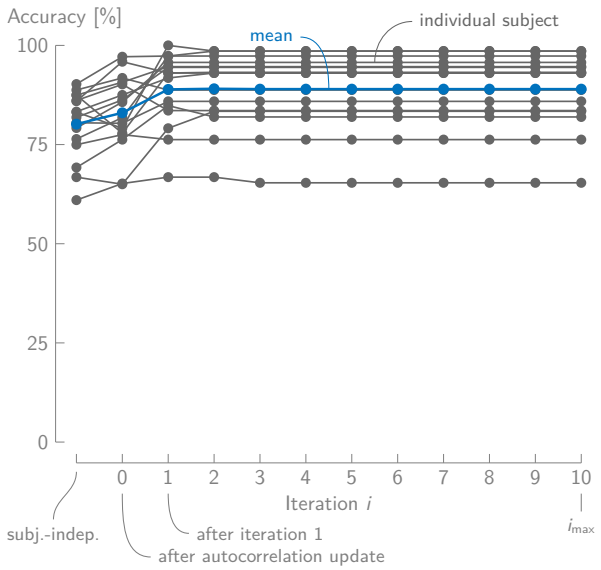


Figure 7: The convergence plots for all subjects of Dataset I using subject-independent information, on 60 s decision windows, show that mostly the autocorrelation update and the first iteration result in a substantial increase in accuracy.

on the MESD values in Figure 5d, there is still a significant difference to be found between the supervised and unsupervised subject-specific performance (Wilcoxon signed-rank test with Bonferroni-Holm correction: $n = 18, p = 0.0498$), albeit very close to the significance level of 0.05. This indicates again that the unsupervised procedure with subject-independent information closely approximates the supervised subject-specific performance *without* ground-truth labels. Furthermore, the unsupervised decoder has a higher performance for four subjects (out of 18) relative to the supervised subject-specific decoder. Lastly, there now is a clear significant difference between the MESD values of the unsupervised procedure and the subject-independent decoder (Wilcoxon signed-rank test with Bonferroni-Holm correction: $n = 18, p = 0.0030$).

Using some information about other subjects, we can thus adapt a stimulus reconstruction decoder that performs almost as well as a supervised subject-specific decoder, but without requiring ground-truth information about the attended speaker during the training procedure.

Convergence plots: Figure 7 shows the AAD accuracy as a function of the different steps of Algorithm 1 for all subjects of Dataset I. It appears that fully replacing (i.e., $\alpha = 0$) the autocorrelation matrix in the subject-independent decoder with the subject-specific information, which is a fully unsupervised step, already results in a substantial increase in accuracy, despite the resulting mismatch between the auto- and cross-correlation matrix/vector ('after autocorrelation update' versus 'subj.-indep.' in Figure 7). Further updating the cross-correlation vector with the predicted labels while using subject-independent information with $\beta = \frac{1}{3}$ results in a self-leveraging effect, leading to a further increase in accuracy, which converges after a few iterations similarly to Figure 6.

VI. Outlook and Conclusions

A. Applications and future work

The proposed unsupervised self-adaptive algorithm paves the way for further extensions and applications. We presented a batch-version of the algorithm, i.e., the updating is performed on a large dataset of EEG and audio data. This enables the 'plug-and-play' capabilities of a stimulus reconstruction decoder for a new hearing device user. However, Algorithm 1 could be extended to an adaptive version, tailored towards the application of neuro-steered hearing devices, where EEG and audio data are continuously recorded. As a result, the stimulus reconstruction decoder could automatically update itself in an unsupervised manner when new data comes in and adapt to changing conditions and situations (e.g., non-stationarities in neural activity, changing electrode-skin contact impedances, ...). The development of such an efficient, adaptive version of the unsupervised procedure is left open as future work. Furthermore, similarly to the supervised stimulus reconstruction decoder and other AAD algorithms, the practical applicability in more realistic listening scenarios, using the demixed and potentially noise-corrupted speech envelopes, and using wearable and miniaturized EEG devices, needs to be further investigated. For a literature overview on the state-of-the-art on those challenges, we refer to [3].

Note that the deployed stimulus reconstruction approach performs worse on short decision window lengths (see Figure 5), making this algorithm less suitable for real-time decoding of the auditory attention [3], [28]. However, the proposed unsupervised updating of a stimulus reconstruction decoder can still be used on a longer time scale to generate reliable labels to train another, potentially more accurate, algorithm on short decision windows (e.g., [13], [14]).

The aforementioned adaptive implementation of the unsupervised procedure also potentially enables and improves the success of neurofeedback effects in a closed-loop implementation, of which preliminary studies have stressed the importance for AAD [30]. The interplay of the subject and the adaptive updating algorithm in a closed-loop system could further improve the AAD performance, as the subject learns to control the updating procedure.

B. Conclusion

We have shown that it is possible to train a subject-specific stimulus reconstruction decoder for AAD using an unsupervised procedure, i.e., without requiring information about which speaker is the attended or unattended one. Training such a decoder on the data of a particular subject from scratch, even starting from a random decoder and without any prior knowledge, leads to a decoder that outperforms a subject-independent decoder. Unsupervised adaptation of a subject-independent decoder, trained on other subjects, to a new subject even leads to a decoder that closely approximates the performance of a supervised subject-specific decoder. The proposed updating algorithm thus combines the two main advantages of a supervised subject-specific and subject-independent decoder:

1) It substantially outperforms a subject-independent decoder, approximating the performance of a supervised subject-specific decoder.
2) It can be used in a 'plug-and-play' fashion, without requiring ground-truth labels and potentially automatically adapting to changing conditions without external intervention.

Using a mathematical model for the updating procedure, the unsupervised algorithm can be interpreted as a fixed-point algorithm. This interpretation explains why there is a self-leveraging effect, even when starting from a random decoder. Furthermore, using this mathematical model, we are able to accurately predict the accuracy of the unsupervised decoder starting from the results of the supervised subject-specific decoder.

The proposed unsupervised self-adaptive algorithm can be used in an online and adaptive manner in a practical neuro-steered hearing device, allowing the decoder to automatically adapt to the non-stationary brain and changing environments and conditions. Furthermore, it avoids having a cumbersome a priori training stage for each new hearing device user, as it automatically adapts to the new user. Lastly, the developed method potentially enables stronger neurofeedback effects when using a closed-loop system, which is paramount for the successful application of AAD.

REFERENCES

[1] S. Geirnaert, T. Francart, and A. Bertrand, "Unsupervised Self-Adaptive Auditory Attention Decoding: Supplementary Material," *Zenodo*, Apr. 2021. [Online]. Available: https://doi.org/10.5281/zenodo.4721018

[2] S. Van Eyndhoven, T. Francart, and A. Bertrand, "EEG-Informed Attended Speaker Extraction from Recorded Speech Mixtures with Application in Neuro-Steered Hearing Prostheses," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 5, pp. 1045–1056, 2017.

[3] S. Geirnaert *et al.*, "EEG-based Auditory Attention Decoding: Towards Neuro-Steered Hearing Devices," *arXiv*, 2020.

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

[5] 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.

[6] E. B. Petersen, M. Wöstmann, J. Obleser, and T. Lunner, "Neural tracking of attended versus ignored speech is differentially affected by hearing loss," *Journal of Neurophysiology*, vol. 117, no. 1, pp. 18–27, 2017.

[7] L. Decruy, J. Vanthornhout, and T. Francart, "Hearing impairment is associated with enhanced neural tracking of the speech envelope," *Hearing Research*, vol. 393, p. 107961, 2020.

[8] S. A. Fuglsang, J. Märcher-Rørsted, T. Dau, and J. Hjortkjær, "Effects of Sensorineural Hearing Loss on Cortical Synchronization to Competing Speech during Selective Attention," *Journal of Neuroscience*, vol. 40, no. 12, pp. 2562–2572, 2020.

[9] J. A. O'Sullivan *et al.*, "Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG," *Cerebral Cortex*, vol. 25, no. 7, pp. 1697–1706, 2014.

[10] D. D. E. Wong, S. A. Fuglsang, J. Hjortkjær, E. Ceolini, M. Slaney, and A. De Cheveigne, "A Comparison of Regularization Methods in Forward and Backward Models for Auditory Attention Decoding," *Frontiers in Neuroscience*, vol. 12, p. 531, 2018.

[11] N. Das, J. Vanthornhout, T. Francart, and A. Bertrand, "Stimulus-aware spatial filtering for single-trial neural response and temporal response function estimation in high-density EEG with applications in auditory research," *NeuroImage*, vol. 204, p. 116211, 2020.

[12] S. Vandecappelle, L. Deckers, N. Das, A. H. Ansari, A. Bertrand, and T. Francart, "EEG-based detection of the attended speaker and the locus of auditory attention with convolutional neural networks," *bioRxiv*, 2020.

[13] 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.

[14] S. Geirnaert, T. Francart, and A. Bertrand, "Riemannian geometry-based decoding of the directional focus of auditory attention using EEG," *arXiv*, 2020.

[15] S. Miran, S. Akram, A. Sheikhattar, J. Z. Simon, T. Zhang, and B. Babadi, "Real-time tracking of selective auditory attention from M/EEG: A Bayesian filtering approach," *Frontiers in Neuroscience*, vol. 12, p. 262, 2018.

[16] M. J. Crosse, G. M. Di Liberto, A. Bednar, and E. C. Lalor, "The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli," *Frontiers in Human Neuroscience*, vol. 10, p. 604, 2016.

[17] 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.

[18] C. Han, J. A. O'Sullivan, Y. Luo, J. Herrero, A. D. Mehta, and N. Mesgarani, "Speaker-independent auditory attention decoding without access to clean speech sources," *Science Advances*, vol. 5, no. 5, p. eaav6134, 2019.

[19] A. Aroudi and S. Doclo, "Cognitive-Driven Binaural Beamforming Using EEG-Based Auditory Attention Decoding," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 862–875, 2020.

[20] N. Das, J. Zegers, H. Van hamme, T. Francart, and A. Bertrand, "Linear versus deep learning methods for noisy speech separation for EEG-informed attention decoding," *Journal of Neural Engineering*, vol. 17, no. 4, p. 046039, 2020.

[21] N. Das, T. Francart, and A. Bertrand, "Auditory Attention Detection Dataset KULeuven," *Zenodo*, 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3997352

[22] S. A. Fuglsang, T. Dau, and J. Hjortkjær, "Noise-robust cortical tracking of attended speech in real-world acoustic scenes," *NeuroImage*, vol. 156, pp. 435–444, 2017.

[23] S. A. Fuglsang, D. D. E. Wong, and J. Hjortkjær, "EEG and audio dataset for auditory attention decoding," *Zenodo*, 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1199011

[24] 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. 5, p. 056014, 2016.

[25] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of Multivariate Analysis*, vol. 88, no. 2, pp. 365–411, 2004.

[26] F. Lotte *et al.*, "A review of classification algorithms for EEG-based brain-computer interfaces: a 10 year update," *Journal of Neural Engineering*, vol. 15, no. 3, p. 031005, 2018.

[27] Y. Chen, A. Wiesel, Y. C. Eldar, and A. O. Hero, "Shrinkage algorithms for MMSE covariance estimation," *IEEE Transactions on Signal Processing*, vol. 58, no. 10, pp. 5016–5029, 2010.

[28] 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.

[29] S. Geirnaert, T. Francart, and A. Bertrand, "MESD Toolbox," *GitHub*, 2019. [Online]. Available: https://github.com/exporl/mesd-toolbox

[30] R. Zink, S. Proesmans, A. Bertrand, S. Van Huffel, and M. De Vos, "Online detection of auditory attention with mobile EEG: closing the loop with neurofeedback," *bioRxiv*, 2017.

Unsupervised Self-Adaptive Auditory Attention Decoding: Supplementary Material

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

In this supplementary material corresponding to the paper 'Unsupervised Self-Adaptive Auditory Attention Decoding' by Geirnaert et al. [1], we show convergence to a unique fixed point of the fixed-point iteration on the updating model (Equation (16) in the original paper [1])[1]. We hypothesize that under three reasonable conditions on the accuracies of the attended and unattended decoder, there exists a unique fixed point p^* to which the fixed-point iteration $p_{i+1} = \phi(p_i)$ converges, starting from any (possibly random) decoder. In Section I, we first show that there always exists such a fixed point, while in Section II we check the uniqueness of and convergence to this fixed point under the hypothesized conditions.

I. Existence

Consider the following fixed-point theorem, also known as Brouwer's fixed-point theorem [2]:

Theorem 1 (Brouwer's fixed point theorem [2]). *Any continuous self map of a nonempty compact convex subset of a Euclidean space has a fixed point.*

As the function $\phi(p_i) : [0, 100]\% \rightarrow [0, 100]\%$ in (16) [1] is a continuous function that maps its domain onto itself and $[0, 1]$ is a closed (thus, compact) convex subset of $\mathbb{R}$, Brouwer's fixed point theorem assures that there exists at least one fixed point.

II. Uniqueness and convergence

We evaluate the model in (16) [1] in a relevant range of the parameters μ_1, μ_2, and σ, obeying three reasonable conditions, to show the convergence to a unique fixed point.

A. Three conditions for convergence

Consider the supervised subject-specific attended decoder $\hat{\mathbf{d}}_{\text{a}}$ with accuracy p_{a} (on the attended labels) and supervised subject-specific unattended decoder $\hat{\mathbf{d}}_{\text{u}}$ with accuracy p_{u} (on the unattended labels). We then a priori postulate the following three intuitive and reasonable conditions on the accuracies p_{a} and p_{u} (which will turn out to be satisfied for all subjects in both datasets in [1]):

- $p_{\text{a}} - p_{\text{u}} > 5\%$, i.e., the attended decoder needs to perform 5% better (on the attended labels) than the unattended decoder (on the unattended labels). Given that the attended speech envelope is typically better represented in the EEG, we indeed expect a difference in performance between both decoders. Moreover, this condition can be linked to the expectation that the cross-correlation between the EEG and attended speech envelope is on average larger than with the unattended speech envelope, serving as a possible explanation for the self-leveraging effect (see Section IV-B in the original paper [1]).
- $p_{\text{u}} < 85\%$, i.e., the *unattended* decoder may not perform better than 85% (on the unattended labels). If the unattended decoder performs too well, then, again, the self-leveraging effect may not be present for the same reason as mentioned in the previous condition.
- $p_{\text{a}} > 100\% - p_{\text{u}}$, i.e., the attended decoder is better at predicting attended labels than the unattended decoder. This assures that the starting point of the model curve $\phi(0\%) = 100\% - p_{\text{u}}$ (e.g., see Figure 2 in the original paper [1]) is below the end point $\phi(100\%) = p_{\text{a}}$.

In the following sections, we will use the model in (16) [1] to show that there is convergence to a unique fixed point when these three conditions are satisfied. However, it is noted that these postulated conditions are conservative in the mathematical sense, i.e., they are 'sufficient' but *not* 'necessary' conditions. When they are not satisfied, there can still be convergence to a unique fixed point.

Moreover, the three conditions are also intuitive and very reasonable from a practical point of view, as they are satisfied for all subjects in both datasets [1]; the minimum across all subjects of $p_{\text{a}} - p_{\text{u}} = 8.3\% > 5\%$, the maximum across all subjects of $p_{\text{u}} = 76.7\% < 85\%$, and the minimum across all subjects of $p_{\text{a}} + p_{\text{u}} = 124\% > 100\%$.

B. Convergence to a unique fixed point

Consider the following fixed-point theorem that provides sufficient conditions for convergence to a unique fixed point of the fixed-point iteration $p_{i+1} = \phi(p_i)$ [3]:

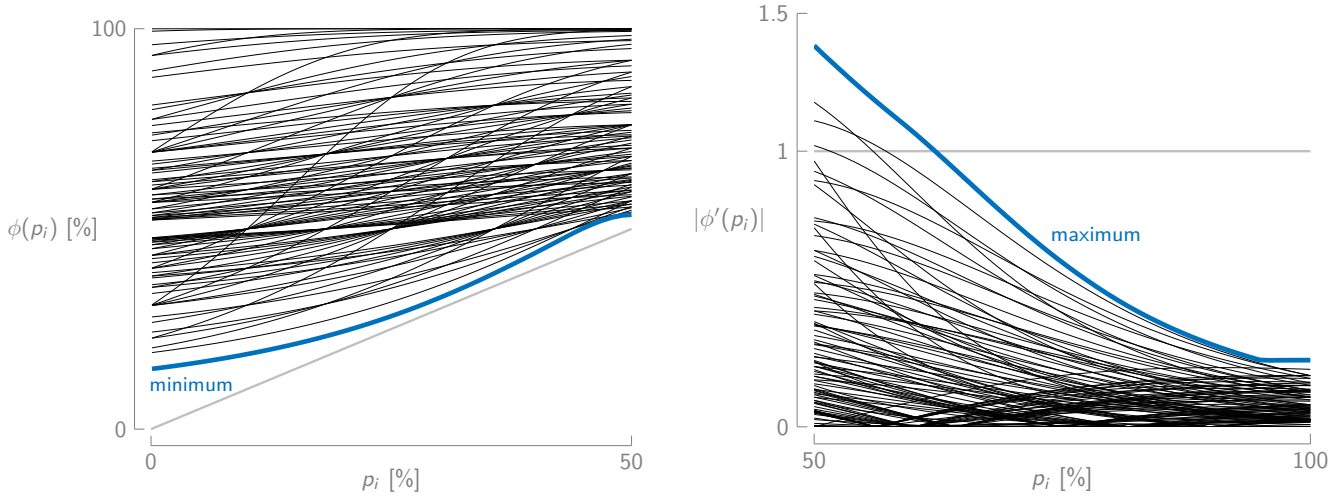
Theorem 2. *Let ϕ be a continuous function on $[a, b]$, such that $\phi(p_i) \in [a, b], \forall\, p_i \in [a, b]$, and suppose that ϕ' exists $\forall\, p_i \in [a, b]$ and that a constant $0 < \alpha < 1$ exists such that:*

$$|\phi'(p_i)| \leq \alpha, \forall\, p_i \in [a, b],$$

then there is exactly one fixed point $p^* \in [a, b]$ and the fixed-point iteration $p_{i+1} = \phi(p_i)$ will converge to this unique fixed point in $[a, b]$.

We now evaluate the model $\phi(p_i)$ in (16) [1] and its derivative $\phi'(p_i)$ to show convergence to a unique fixed point based on Theorem 2 for the case where the conditions in Section II-A are satisfied.

[1] This supplementary material has also been peer-reviewed together with the original article in [1].



(a) (b)

Figure 1: (a) A subset of the evaluated $\phi(p_i)$ for $p_i \in [0, 50]\%$ and the minimum over all evaluated (μ_1, μ_2, σ) that obey the conditions are all above the identity line, where $\phi(p_i) = p_i$, which shows that $\phi(p_i) > p_i, \forall\, p_i \in [0, 50]\%$. (b) A subset of the evaluated $|\phi'(p_i)|$ for $p_i \in [50, 100]\%$, together with the maximum over all evaluated (μ_1, μ_2, σ) that obey the conditions.

The derivative $\phi'(p_i)$ of the model in (16) can be computed by hand or by using any symbolic math software and is equal to:

$$\phi'(p_i) = \frac{p_i \sigma_z(p_i)^2 \mu_2 + (1 - p_i)\sigma^2 \mu_z(p_i)}{\sqrt{2\pi} p_i^3 \sigma_z(p_i)^3} e^{-\frac{1}{2}\left(\frac{\mu_z(p_i)}{\sigma_z(p_i)}\right)^2}. \quad (1)$$

To evaluate (16) [1] and its derivative (1), we take 300 equidistant samples of $\mu_1 \in [-2, 2]$, 300 equidistant samples of $\mu_2 \in [-2, 2]$, and 100 equidistant samples of $\sigma \in\,]0, 4]$. These intervals contain the complete range of parameters concerning the difference in correlation coefficients R_1 and R_2. From this parameter range, we select all combinations of (μ_1, μ_2, σ) for which the three conditions of Section II-A are satisfied. The connection between p_a and p_u (as used in the three conditions) and the model parameters (μ_1, μ_2, σ) is given by:

$$p_\text{a} = P(R_1 > 0) = \frac{1}{\sigma\sqrt{2\pi}} \int_0^{+\infty} e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2} dx \text{ and}$$

$$p_\text{u} = P(R_2 > 0) = \frac{1}{\sigma\sqrt{2\pi}} \int_0^{+\infty} e^{-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma}\right)^2} dx,$$

using the assumptions in Section IV-A in the original paper [1]. These connections can be derived from the updating model in Equation (16) from the original paper [1] by setting $p_i = 100\%$, resp. $p_i = 0\%$, resulting in the decoder accuracy of the supervised attended, resp. unattended decoder.

Figure 1a now shows a subset of $\phi(p_i)$ for $p_i \in [0, 50]\%$, for all evaluated (μ_1, μ_2, σ) that obey the three conditions, together with the minimum over all these $\phi(p_i)$. Similarly, Figure 1b shows a subset of $|\phi'(p_i)|$ for $p_i \in [50, 100]\%$, for all evaluated (μ_1, μ_2, σ) that obey the three conditions, together with the maximum over all these $|\phi'(p_i)|$. Both results are required to show convergence to a unique fixed point using Theorem 2:

- **Result 1**: From Figure 1a, it can be seen that $\phi(p_i) > p_i, \forall\, p_i \in [0, 50]\%$. This implies that there is no fixed point within this interval and that the fixed-point iteration will always diverge to the $p_i \in [50, 100]\%$ interval. This is because $\forall\, p_i \in [0, 50]\% : p_{i+1} = \phi(p_i) > p_i$, i.e., the new accuracy in the fixed-point iteration is always larger than the previous one, such that, inevitably, at a certain iteration, $p_{i+1} > 50\%$. It thus suffices to show that there is convergence to a unique fixed point for $p_i \in [50, 100]\%$, which is shown in the next result.
- **Result 2**: From Figure 1b, there are two possible cases, which both individually can be shown to guarantee convergence to a unique fixed point:
  1) $|\phi'(p_i)| < 1, \forall\, p_i \in [50, 100]\%$. For all these cases, we then numerically confirmed that $\phi(p_i) \in [50, 100]\%, \forall\, p_i \in [50, 100]\%$ such that all conditions of Theorem 2 are fulfilled to show convergence to a unique point.
  2) $\exists\, x \in [50, 100]\% : \phi'(p_i) \geq 1, \forall\, p_i \in [50, x]\%$ and $|\phi'(p_i)| < 1, \forall p_i \in [x, 100]\%$. Since $\phi(50\%) > 50\%$ (see Result 1) and since the derivative is positive, it is guaranteed that $\phi(p_i) > p_i, \forall\, p_i \in [50, x]\%$, i.e., there is no fixed point and the fixed-point iteration diverges to the $p_i \in [x, 100]\%$ interval (using a similar reasoning as in Result 1). Furthermore, it can again be numerically checked that $\phi(p_i) \in [x, 100]\%, \forall\, p_i \in [x, 100]\%$ to show that there is a unique point to which there is convergence in this interval (see Theorem 2).

References

[1] S. Geirnaert, T. Francart, and A. Bertrand, "Unsupervised Self-Adaptive Auditory Attention Decoding," *IEEE Journal of Biomedical and Health Informatics*, 2021.

[2] J. M. Borwein and A. S. Lewis, *Convex Analysis and Nonlinear Optimization: Theory and Examples*, 2nd ed. Springer-Verlag New York, 2006.

[3] Walter Gautschi, *Numerical Analysis*. Birkhäuser Basel, 2012.