UNBIASED UNSUPERVISED STIMULUS RECONSTRUCTION FOR EEG-BASED AUDITORY ATTENTION DECODING

Nicolas Heintz[⋆†] *Simon Geirnaert*[⋆†] *Tom Francart*[†] *Alexander Bertrand*[⋆]

[⋆] KU Leuven, Department of Electrical Engineering (ESAT),
STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, Belgium
[†]KU Leuven, Department of Neurosciences, ExpORL, Belgium

ABSTRACT

It is possible to decode auditory attention to speech from electrophysiological brain recordings such as electroencephalography (EEG). Such an auditory attention decoding (AAD) allows, e.g., to determine to which person a listener is attending in a multi-talker scenario. The vast majority of research has focused on developing supervised AAD algorithms in which the decoder is trained based on ground truth labels about the attention to each speaker. However, to work optimally, the trained decoders must be subject-specific and adapt over time to track sudden changes in signal statistics (e.g. electrode failures). Since it is often impractical to regularly retrain these decoders with a dedicated calibration session, an unsupervised algorithm has recently emerged as an alternative.

In this paper, we show that the state-of-the-art unsupervised AAD algorithm is biased by its initialisation, which leads to a sub-optimal convergence. This bias has the largest effect when only a limited amount of data is available to train it, e.g. to train an unsupervised decoder that can quickly adapt to sudden changes. We show that this bias can be easily removed, leading to a better classification accuracy. However, the gain in accuracy reduces as the number of classified segments increases.

***Index Terms*—** Auditory attention decoding, unsupervised learning, electroencephalography, neuro-steered hearing devices

1. INTRODUCTION

As sound travels through the ear and stimulates the auditory nerves, it evokes a neural response that is phase-locked to the envelope of that sound. This neural response differs when the sound is attended or unattended [1, 2], enabling us to decode from the neural response to which speaker a person is attending in a multi-talker scenario [3, 4]. This problem is known as auditory attention decoding (AAD), and it has applications in, e.g., neuro-steered hearing devices [5].

To work optimally, state-of-the-art auditory attention decoders must be trained subject-specifically during a supervised training session of at least 20-30 minutes [5, 6], which is cumbersome for real-life use. Subject-independent decoders can circumvent this mandatory training session, but perform significantly worse [3, 5]. Furthermore, both models are fixed and thus unable to adapt to changes in brain responses, electrodes or listening environments [6].

The aforementioned problems make unsupervised, subject-specific models paramount. They can adapt to changes in signal statistics, have the potential to approximate the performance of supervised subject-specific models [6] and inherit the plug-and-play nature of subject-independent models.

State-of-the-art unsupervised auditory attention algorithms identify which speech stream is attended by reconstructing the envelope of attended speech from electroencephalography (EEG) signals using a least-squares model [3, 4, 6, 7]. The correlation between the reconstructed and the original envelope is then used to classify the speech as attended or unattended. The least-squares model is trained to specifically reconstruct attended envelopes and thus inherently requires ground-truth labels. To mitigate this, the unsupervised algorithm iteratively retrains the least-squares model using its own predictions as training labels, counting on a self-leveraging effect to converge to an optimal point [6, 7].

In this paper, we show that this iterative procedure is biased *against* switching labels over iterations. This bias has a larger deteriorating effect when little training data are present. However, the unsupervised least-squares model is ideally trained on little data to ensure fast adaptability to sudden changes in signal statistics (e.g. a sudden electrode failure) [6]. We show that removing this bias is relatively straightforward, yet can lead up to a 30% improvement in classification accuracy on the updating set when the number of segments is small.

The outline of the paper is as follows. In Section 2, we review and analyse the unsupervised auditory attention decoding algorithm from [6, 7], prove that it is biased against switching labels between iterations and that this bias can be removed with a simple procedure. This theoretical discussion is validated on real data in Section 3, where we discuss the conducted experiments and report the results. We draw conclusions in Section 4.

2. INITIALISATION BIAS IN UNSUPERVISED AAD

The linear AAD algorithm introduced in [3] uses a spatio-temporal decoder which reconstructs the envelope of the attended speech $s(t)$ from EEG signals. Given an EEG segment $X_k \in \mathbb{R}^{C\times T}$ with C the number of EEG channels, k the segment index and T the length of the EEG segment, the corresponding reconstructed speech envelope is then:

$$\hat{s}_k(t) = \sum_{c=1}^{C} \sum_{l=0}^{L-1} D(c,l) X_k(c, t+l), \tag{1}$$

This research is funded by Aspirant Grant 1S31522N (for N. Heintz) from the Research Foundation - Flanders (FWO), PDM madate PDMT1/22/009 from KU Leuven (for S. Geirnaert), the Research Foundation - Flanders (FWO) project No G0A4918N and G081722N, the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 802895 and grant agreement No 637424), and the Flemish Government (AI Research Program). The scientific responsibility is assumed by its authors.

N. Heintz, S. Geirnaert and A. Bertrand are also affiliated with Leuven.AI - KU Leuven institute for AI, Belgium.

where $D \in \mathbb{R}^{C \times L}$ is the decoder matrix, and L is the number of post-stimulus time lags that are used in the reconstruction.

Equation (1) can be written as an inner vector product:

$$\begin{aligned} \hat{s}_k(t) &= \mathbf{d}^\top \mathbf{x}_k(t), \\ &\text{with } \mathbf{x}_k(t) = [X_k(1,t) \dots X_k(1,t+L-1), \dots, \\ &\qquad X_k(C,t) \dots X_k(C,t+L-1)]^\top, \end{aligned} \tag{2}$$

and similarly for rewriting D as $\mathbf{d}$.

In a supervised model, $\mathbf{d}$ is optimised such that it minimises the squared distance between $\hat{s}_k(t)$ and the attended speech envelope $s_{a,k}(t)$ across all segments k, i.e.

$$\hat{\mathbf{d}} = \arg\min_{\mathbf{d}} \sum_{k=1}^{K} \sum_{t=0}^{T-L} (\mathbf{d}^\top \mathbf{x}_k(t) - s_{a,k}(t))^2, \tag{3}$$

with K the total number of segments. (3) is the well-known least-squares problem, with as solution

$$\hat{\mathbf{d}} \propto R_{xx}^{-1} \mathbf{r}_{xs}, \tag{4}$$

$$\begin{aligned} \text{with } R_{xx} &= \sum_{k=1}^{K} \sum_{t=0}^{T-L} \mathbf{x}_k(t) \mathbf{x}_k(t)^\top \equiv \sum_{k=1}^{K} R_{xx,k} \\ \mathbf{r}_{xs} &= \sum_{k=1}^{K} \sum_{t=0}^{T-L} \mathbf{x}_k(t) s_{a,k}(t) \equiv \sum_{k=1}^{K} \mathbf{r}_{xs_a,k}, \end{aligned} \tag{5}$$

where the former represents the spatio-temporal autocorrelation matrix of the EEG data, and the latter represents the cross-correlation vector between the EEG signal $\mathbf{x}(t)$ and the attended speech envelope $s_a(t)$, for L time lags.

For each test segment, the speech envelope with the highest Pearson correlation to the reconstructed envelope is classified as attended.

Note that (4) requires knowledge of which of the two stimuli is attended. This information is not available in an unsupervised setting. In [7], an iterative procedure is proposed to (re-)train the decoder without the use of ground truth labels. In this method, the decoder is first trained using random labels, which then reconstructs an initial envelope. This generates a set of K labels $\mathbf{l} \in \mathbb{B}^{K \times 1}$ with $l_k \in \{0, 1\}$, which are used to iteratively recompute the decoder $\hat{\mathbf{d}}$ and regenerate the set of labels $\mathbf{l}$ until convergence. The complete unsupervised algorithm from [7] is summarised in Algorithm 1.

To show that the classification in Algorithm 1 is biased, we will further simplify (6) by assuming without loss of generality that $r_{ss_1} = r_{ss_2} = 1$. In this case, a label $l_k^{(i+1)} = 1$ at iteration $i+1$ if:

$$\begin{aligned} &\mathbf{r}_{xs}^\top R_{xx}^{-1} \mathbf{r}_{xs_1,k} > \mathbf{r}_{xs}^\top R_{xx}^{-1} \mathbf{r}_{xs_2,k} \\ \iff &\sum_{j=1}^{K} \left(l_j^{(i)} \mathbf{r}_{xs_1,j} + (1 - l_j^{(i)}) \mathbf{r}_{xs_2,j} \right)^\top R_{xx}^{-1} \mathbf{r}_{xs_1,k} > \\ &\sum_{j=1}^{K} \left(l_j^{(i)} \mathbf{r}_{xs_1,j} + (1 - l_j^{(i)}) \mathbf{r}_{xs_2,j} \right)^\top R_{xx}^{-1} \mathbf{r}_{xs_2,k}. \end{aligned} \tag{7}$$

Let us now consider the case where the resulting label for segment k in iteration $i+1$ is $l_k^{(i+1)} = 1$, i.e., speaker 1 is identified as attended. In case $l_k^{(i)}$ was equal to 0 in the previous iteration i, (7) becomes:

$$\begin{aligned} &\sum_{j=1, j \neq k}^{K} \left(l_j^{(i)} \mathbf{r}_{xs_1,j} + (1 - l_j^{(i)}) \mathbf{r}_{xs_2,j} \right)^\top R_{xx}^{-1} (\mathbf{r}_{xs_1,k} - \mathbf{r}_{xs_2,k}) > \\ &\mathbf{r}_{xs_2,k}^\top R_{xx}^{-1} \mathbf{r}_{xs_2,k} - \mathbf{r}_{xs_2,k}^\top R_{xx}^{-1} \mathbf{r}_{xs_1,k}. \end{aligned} \tag{8}$$

However, if the label was already equal to $l_k^{(i)} = 1$ in iteration i, $l_k^{(i+1)} = 1$ in the next iteration if:

$$\begin{aligned} &\sum_{j=1, j \neq k}^{K} \left(l_j^{(i)} \mathbf{r}_{xs_1,j} + (1 - l_j^{(i)}) \mathbf{r}_{xs_2,j} \right)^\top R_{xx}^{-1} (\mathbf{r}_{xs_1,k} - \mathbf{r}_{xs_2,k}) > \\ &\mathbf{r}_{xs_1,k}^\top R_{xx}^{-1} \mathbf{r}_{xs_2,k} - \mathbf{r}_{xs_1,k}^\top R_{xx}^{-1} \mathbf{r}_{xs_1,k}. \end{aligned} \tag{9}$$

Note that the lefthand sides of (8) and (9) are equal, and should be $\mathbf{r}_{xs_2,k}^\top R_{xx}^{-1} \mathbf{r}_{xs_2,k} + \mathbf{r}_{xs_1,k}^\top R_{xx}^{-1} \mathbf{r}_{xs_1,k} - 2\mathbf{r}_{xs_1,k}^\top R_{xx}^{-1} \mathbf{r}_{xs_2,k} > 0$ larger if $l_k^{(i)} = 0$ compared to when $l_k^{(i)} = 1$ to get $l_k^{(i+1)} = 1$. Therefore, the threshold for identifying speaker 1 as attended for segment k in iteration $i+1$ is larger in case $l_k^{(i)} = 0$ than in case $l_k^{(i)} = 1$. This causes the model to be biased towards keeping $l_k^{(i)}$ and $l_k^{(i+1)}$ identical, even if $l_k^{(i)}$ is chosen completely randomly. Moreover, we will show in Section 3 that $\mathbf{r}_{xs_{1/2},k}^\top R_{xx}^{-1} \mathbf{r}_{xs_{1/2},k} >> \mathbf{r}_{xs_{1/2},j \neq k}^\top R_{xx}^{-1} \mathbf{r}_{xs_{1/2},k}$, which makes it very difficult for an unsupervised model with a small amount of segments K to overcome this bias. This promotes fast convergence to local, suboptimal minima, especially when K is small (i.e. for small datasets).

To remove this bias, the cross-correlation vector $\mathbf{r}_{xs}$ should be estimated without the current segment k, as done in Algorithm 2. This removes the inherent overfitting that caused the bias. This can be done with minimal additional computational cost, e.g. by subtracting $\mathbf{r}_{xs_{1/2},k}$ from the sum of all K cross-correlation matrices.

Although Algorithm 2 focuses on a two-speaker problem, this algorithm can easily be expanded to scenarios with more than two speakers. In this case, only the speech envelope that is the most correlated with the reconstructed envelope is classified as attended. Similar to before, only envelopes classified as attended are used to estimate the cross-correlation matrix in the next iteration.

Algorithm 1: Biased unsupervised AAD

```
Compute $R_{xx}$ as in (5).
while $\mathbf{l}$ changes do
    Estimate the cross-correlation vector:
        $$\mathbf{r}_{xs} = \sum_{k=1}^{K} l_k^{(i)} \mathbf{r}_{xs_1,k} + (1 - l_k^{(i)}) \mathbf{r}_{xs_2,k}.$$
    for $k \in [1 \dots K]$ do
        Compute the Pearson correlation coefficient between $\hat{s}_k(t) = \mathbf{r}_{xs}^\top R_{xx}^{-1} \mathbf{x}_k(t)$ and the two speech envelopes $s_{1,k}(t)$, $s_{2,k}(t)$:
            $$\rho_{1/2,k} \equiv \frac{\mathbf{r}_{xs}^\top R_{xx}^{-1} \mathbf{r}_{xs_{1/2},k}}{\sqrt{\mathbf{r}_{xs}^\top R_{xx}^{-1} R_{xx,k} R_{xx}^{-1} \mathbf{r}_{xs}} \sqrt{r_{ss_{1/2},k}}}, \tag{6}$$
        with $\mathbf{r}_{xs_{1/2},k}$ as defined in (5) and $r_{ss_{1/2},k} = \sum_{t=0}^{T-L} s_{1/2,k}(t)^2$.
        $l_k^{(i+1)} = 1$ if $\rho_{1,k} > \rho_{2,k}$, else $l_k^{(i+1)} = 0$.
```

Algorithm 2: Unbiased unsupervised AAD

```
Compute $R_{xx}$ as in (5).
while $\mathbf{l}$ changes do
    for $k \in [1 \dots K]$ do
        $$\mathbf{r}_{xs} = \sum_{j=1, j \neq k}^{K} l_j^{(i)} \mathbf{r}_{xs_1,j} + (1 - l_j^{(i)}) \mathbf{r}_{xs_2,j}$$
        $\rho_{1/2,k} = \mathbf{r}_{xs}^\top R_{xx}^{-1} \mathbf{r}_{xs_{1/2},k}$,
        $l_k^{(i+1)} = 1$ if $\rho_{1,k} > \rho_{2,k}$, else $l_k^{(i+1)} = 0$,
        with $\mathbf{r}_{xs_{1/2},j}$ as defined in (5).
```

3. EXPERIMENTS

In this section, we compare the proposed unbiased unsupervised algorithm with the (biased) unsupervised algorithm from [7]. We will also compare both unsupervised algorithms with a supervised subject-specific and subject-independent least-squares model proposed in [4].

3.1. Dataset

We validate the algorithms on a publicly available dataset [8]. In this dataset, a 64-channel EEG signal of 16 Flemish speaking subjects is measured as they are attending to one of two competing Flemish stories. All stories are narrated by male speakers in 12 trials of 6 minutes each. The speakers are located at the left and right side of the subject. 72 minutes of data are recorded in total per subject with a BioSemi ActiveTwo system. For more details, we refer to [4, 8].

The EEG signals and the auditory stimuli are preprocessed according to the preprocessing framework proposed in [4]. The auditory stimulus $y(t)$ is first split in frequency bands $y_b(t)$ using a gammatone filter bank with center frequencies between $150\,\text{Hz}$ and $4000\,\text{Hz}$. The auditory envelope is then extracted from each subband by applying the power-law operation $s_b(t) = |y_b(t)|^{0.6}$. Finally all subband envelopes are added with equal weight. This process is repeated for both the attended and unattended stimulus.

The EEG signals and the envelopes are then bandpass filtered between $1\,\text{Hz}$ and $9\,\text{Hz}$ and resampled to a $64\,\text{Hz}$ sample frequency. They are then all cut in segments of length T. Finally, the mean is subtracted from each segment and each envelope segment is normalised such that $r_{ss_{1/2},k} = 1$.

3.2. Hyperparameters

Unless stated otherwise, all 72 minutes of EEG are used for each subject, each segment is $T = 60\,\text{s} * 64\,\text{Hz}$ long and the different EEG lags l are selected between 0 and $L = 0.25\,\text{s} * 64\,\text{Hz}$. No regularisation is used to estimate the correlation matrices.

3.3. Experiments

In a first experiment, we assess the influence of the bias on the final decision. As explained in Section 2, this bias promotes labels in the next iteration to be identical to the labels in the current iteration. First, we will assess the magnitude of this bias term in comparison with the average magnitude of the other terms. We will also study how this bias influences the performance of the unsupervised algorithms in function of the number of segments K in the dataset. For every value of $K \in [2, 72]$, we compute the accuracy of both the original, biased and the proposed unbiased unsupervised AAD algorithm. Both iterative algorithms are run until they converge to a fixed point [7]. The accuracy is computed by comparing their predicted labels to the ground-truth. This experiment is repeated 10 times with a different subset of segments when possible. Since both algorithms are unsupervised and inherently give a classification as output, no cross-validation procedure is used. Note that [7] validates Algorithm 1 on new data. This equates to adding a single iteration of Algorithm 2 after Algorithm 1, lowering the effect of the bias.

In a second experiment, we compare the performance of both unsupervised algorithms with a supervised subject-specific and a subject-independent algorithm for various segment lengths using the full dataset. The unsupervised algorithms are validated as explained above. The supervised subject-specific algorithm [4] is validated differently, using leave-one-out cross-validation, i.e. segment k is classified by a model trained on all segments $j \neq k \in [1, K]$. The subject-independent algorithm is validated using leave-one-subject-out cross-validation.

All statistical significance is tested with a non-parametric and paired Wilcoxon signed rank test (significant if $p < 0.05$).

3.4. Results and discussion

The average magnitudes of the bias term and the other terms are shown in Table 1. The bias term is on average 30 times larger than any other term. This makes it very difficult for the original unsupervised algorithm to correct wrong initial labels when the number of segments K is small. Indeed, Figure 1 shows that the bias causes the algorithm to perform at chance level at $K = 20$, whereas the unbiased algorithm only performs 5% worse at $K = 20$ than at $K = 72$. Remarkably, the unbiased algorithm still obtains 61% accuracy at $K = 2$, where the segment is decoded using a decoder trained on just 1 neighbouring segment.

Bias: $\mathbf{r}_{xs_{1/2},k}^\top R_{xx}^{-1} \mathbf{r}_{xs_{1/2},k}$	**Other:** $\mathbf{r}_{xs_{1/2},j \neq k}^\top R_{xx}^{-1} \mathbf{r}_{xs_{1/2},k}$
$1.4\text{e}-4 \pm 4.5\text{e}-5$	$5.0\text{e}-6 \pm 7.4\text{e}-6$

Table 1: The mean and standard deviation of the bias term in comparison to the other terms from (9) ($T = 60\,\text{s}$).

Figure 2 shows that the unbiased unsupervised model performs significantly better than the original algorithm for long segments with $T > 10\,\text{s}$ when the full dataset is used. Since the amount of data is kept constant in this experiment, the number of segments increases as the segments shorten. This causes the bias to be less influential.

4. CONCLUSION

We have shown that the unsupervised algorithm from [7] is biased by its initialisation, which leads to poor convergence when only a limited number of segments are available (either due to limited data or long segment length). Removing the bias is relatively straightforward and leads up to a 30% higher classification accuracy on the updating set itself when only 20 segments are available. However, as more segments become available, the influence of the bias steadily diminishes. This makes the unbiased unsupervised AAD algorithm well suited for use in applications where only a limited amount of

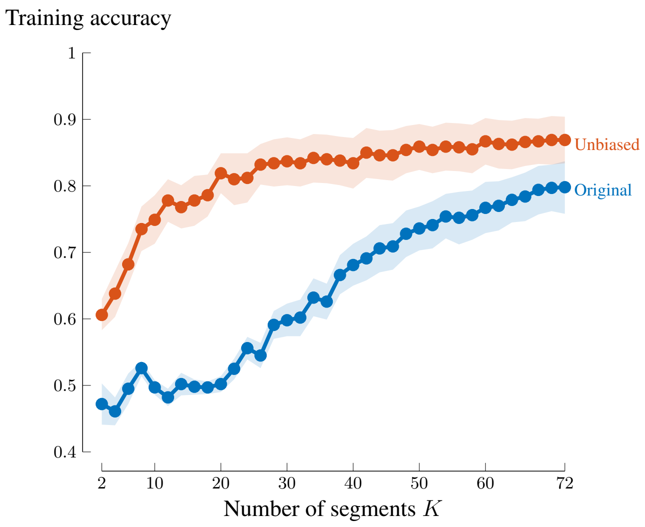

Fig. 1: The unbiased algorithm significantly outperforms the original algorithm for any number of segments. The difference in accuracy on the updating set is maximal around $K = 20$, where the original unsupervised algorithm does not perform better than chance, whereas the unbiased unsupervised algorithm has barely dropped in performance. The shaded area represents the standard error of the mean.

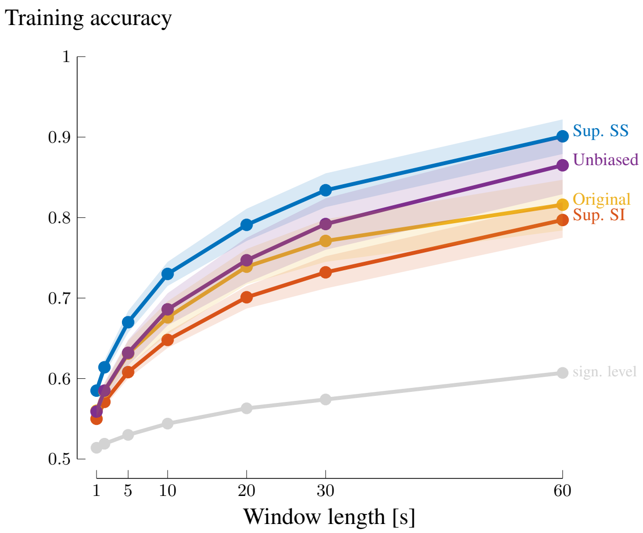

Fig. 2: The unbiased unsupervised AAD algorithm performs on average better than the original unsupervised AAD algorithm. The improvement is only significant for long segments, since in this case there are less segments and the bias has thus more effect. The shaded area represents the standard error of the mean.

unlabelled recording is available, or in situations where the algorithm must quickly adapt to changes in the signal statistics based on a buffer with previous data of limited length.

5. REFERENCES

[1] Nai Ding and Jonathan Z. Simon, “Emergence of neural encoding of auditory objects while listening to competing speakers,” *Proceedings of the National Academy of Sciences*, vol. 109, no. 29, pp. 11854–11859, 2012.

[2] Elana M. Zion Golumbic, Nai Ding, Stephan Bickel, Peter Lakatos, Catherine A. Schevon, Guy M. McKhann, Robert R. Goodman, Ronald Emerson, Ashesh D. Mehta, Jonathan Z. Simon, David Poeppel, and Charles E. Schroeder, “Mechanisms underlying selective neuronal tracking of attended speech at a ”cocktail party”,” *Neuron*, vol. 77, no. 5, pp. 980–991, 2013.

[3] James A. O’Sullivan, Alan J. Power, Nima Mesgarani, Siddharth Rajaram, John J. Foxe, Barbara G. Shinn-Cunningham, Malcolm Slaney, Shihab A. Shamma, and Edmund C. Lalor, “Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG,” *Cerebral Cortex*, vol. 25, no. 7, pp. 1697–1706, 2015.

[4] Wouter Biesmans, Neetha Das, Tom Francart, and Alexander 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.

[5] Simon Geirnaert, Servaas Vandecappelle, Emina Alickovic, Alain de Cheveigné, Edmund Lalor, Bernd T. Meyer, Sina Miran, Tom Francart, and Alexander Bertrand, “Neuro-Steered Hearing Devices: Decoding Auditory Attention From the Brain,” *IEEE Signal Processing Magazine*, vol. 38, no. 4, pp. 89–102, 2021.

[6] Simon Geirnaert, Tom Francart, and Alexander Bertrand, “Time-adaptive Unsupervised Auditory Attention Decoding Using EEG-based Stimulus Reconstruction,” *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 8, 2022.

[7] Simon Geirnaert, Tom Francart, and Alexander Bertrand, “Unsupervised Self-Adaptive Auditory Attention Decoding,” *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 10, pp. 3955–3966, 2021.

[8] Neetha Das, Wouter Biesmans, Alexander Bertrand, and Tom Francart, “The effect of head-related filtering and ear-specific decoding bias on auditory attention detection,” *Journal of Neural Engineering*, vol. 13, no. 5, pp. 56014, 2016.