

Sample-level EEG-based Selective Auditory Attention Decoding with Markov Switching Models

Yuanyuan Yao

*Dept. of Electrical Engineering, STADIUS
KU Leuven
Leuven, Belgium
yuanyuan.yao@kuleuven.be*

Simon Geirnaert

*Dept. of Electrical Engineering, STADIUS
Dept. of Neurosciences, ExpORL
KU Leuven
Leuven, Belgium
simon.geirnaert@kuleuven.be*

Tinne Tuytelaars

*Dept. of Electrical Engineering, PSI
KU Leuven
Leuven, Belgium
Tinne.Tuytelaars@kuleuven.be*

Alexander Bertrand

*Dept. of Electrical Engineering, STADIUS
KU Leuven
Leuven, Belgium
Alexander.Bertrand@kuleuven.be*

Abstract—Selective auditory attention decoding aims to identify the speaker of interest from listeners' neural signals, such as electroencephalography (EEG), in the presence of multiple concurrent speakers. Most existing methods operate at the window level, facing a trade-off between temporal resolution and decoding accuracy. Recent work has shown that hidden Markov model (HMM)-based post-processing can smooth window-level decoder outputs to improve this trade-off. Instead of using a separate smoothing step, we propose to integrate the decoding and smoothing components into a single probabilistic framework using a Markov switching model (MSM). It directly models the relationship between the EEG and speech envelopes under each attention state while incorporating the temporal dynamics of attention. This formulation enables sample-level attention decoding, with model parameters and attention states jointly estimated via the expectation-maximization algorithm. Experimental results demonstrate that this integrated MSM formulation achieves comparable decoding accuracy to HMM post-processing while providing faster attention switch detection. The code for the proposed method is available at <https://github.com/YYao-42/MSM>.

Index Terms—auditory attention decoding, electroencephalography, Markov Switching Model

I. INTRODUCTION

Real-life social situations often involve multiple concurrent speakers, where listeners use selective attention to focus on a single speaker of interest and suppress interfering sound sources [1]. Tracking the attended speaker from neural signals, known as selective auditory attention decoding (sAAD), has been an active research topic in recent years, with potential applications in neuro-steered hearing aids [2].

This research is funded by the Research Foundation - Flanders (FWO) project No G081722N, junior postdoctoral fellowship fundamental research of the FWO (for S. Geirnaert, No. 1242524N), Internal Funds KU Leuven (projects IDN/23/006, C14/25/108, and C3/25/107), and the Flemish Government (AI Research Program).

All authors are also affiliated with Leuven.AI - KU Leuven institute for AI, Belgium.

Electroencephalography (EEG) is a popular choice for recording neural activity in sAAD due to its non-invasiveness and high temporal resolution. However, non-invasiveness comes at the cost of an extremely low signal-to-noise ratio (SNR), making it necessary to aggregate information over time for more reliable decoding. Therefore, the vast majority of sAAD algorithms decode attention based on aggregated EEG data within a pre-defined window length, thereby not fully exploiting the high temporal resolution offered by EEG. A common framework involves reconstructing the attended speech envelope from neural signals using a linear decoder and then comparing it to the speech envelopes of the candidate speakers [2]–[5]. By calculating the correlation between the reconstructed signal and each candidate speaker envelope over a fixed time window, the speaker with the highest correlation is identified as the attended one. More recently, deep learning models have streamlined this by directly mapping windows of neural and speech data to attention metrics such as correlations and probabilities [6]–[8]. However, both approaches face a fundamental trade-off: while shorter windows offer higher temporal resolution, they also produce noisier correlations or probabilities, leading to degraded decoding performance. This motivates methods that can smooth these outputs over time to achieve higher decoding accuracy with minimal impact on the temporal resolution.

To address this, Heintz et al. [9] proposed a post-processing method to refine raw sAAD output metrics, such as correlations, using a hidden Markov model (HMM), which models these outputs as noisy observations conditioned on an underlying attention state. The idea is to exploit the fact that human attention does not switch very frequently, modeling the attention state as a Markov chain where the current focus is likely to persist in the next moment. By assigning a high self-transition probability to the current state, the method can effectively smooth out spurious fluctuations in the raw

sAAD outputs on 1-second windows and achieve comparable decoding performance to that of 30-second windows at a significantly better temporal resolution.

Inspired by this work, we propose to model sAAD as a Markov switching model (MSM), which is an integrated framework that not only incorporates the temporal structure of attention as in HMMs but also directly models the relationship between EEG and speech envelope under each attention state. The expectation-maximization (EM) algorithm is used to jointly estimate the model parameters and decode the attention state at each time point. Therefore, instead of operating as a post-processing step at the window level, the MSM can be directly applied to the EEG and speech data to decode attention at the sample level. We find that the MSM can achieve comparable decoding accuracy to the HMM post-processing method based on correlations generated by a linear decoder, while achieving even shorter switch detection delay.

II. MARKOV SWITCHING MODEL

The Markov switching model (MSM) was first proposed by Hamilton [10] to model economic time series with regime changes. It assumes each observation y_t is drawn from one of several regimes, each characterized by its own set of parameters. The regime at time t is determined by a hidden state variable S_t that evolves according to a Markov chain. In the context of sAAD, considering a two-speaker setting, S_t indicates which speaker is attended at time t , with $S_t = 1$ for speaker 1 and $S_t = 2$ for speaker 2. Denote the speech envelopes of speaker 1 and speaker 2 at time t as y_{1t} and y_{2t} , respectively. The goal of sAAD is to decode S_t from the recorded EEG signals \mathbf{x}_t at each time point t . We augment each EEG sample \mathbf{x}_t with $L - 1$ time lags to form a CL -dimensional spatial-temporal vector $\hat{\mathbf{x}}_t$, where C is the number of EEG channels:

$$\hat{\mathbf{x}}_t = [x_{1,t}, x_{1,t-1}, \dots, x_{1,t-L+1}, \dots, x_{C,t}, x_{C,t-1}, \dots, x_{C,t-L+1}]^T. \quad (1)$$

We define the observation $y_t = y_{1t} - y_{2t}$ to reduce the problem to a one-dimensional observation. We assume the following model for y_t , conditioned on the hidden state variable S_t :

$$y_t = y_{1t} - y_{2t} = \beta_{S_t}^T \hat{\mathbf{x}}_t + e_t, \quad e_t \sim \mathcal{N}(0, \sigma_{S_t}^2), \quad (2)$$

where β_{S_t} and $\sigma_{S_t}^2$ are the regression coefficients and noise variance corresponding to the attention state S_t . Here, β_{S_t} is akin to the traditional linear decoder often used in sAAD [2]–[4], yet now it decodes the difference between speech envelopes rather than each speech envelope separately¹.

The attention state S_t evolves as a first-order Markov chain with transition probabilities

$$p_{ij} = \Pr(S_t = j | S_{t-1} = i), \quad i, j \in \{1, 2\}, \quad (3)$$

¹One could also use a bivariate model for $[y_{1t} \ y_{2t}]^T$ by making β_{S_t} a $2 \times CL$ matrix. The univariate model (2) can then be viewed as a subtraction of the two rows of this bivariate system. This univariate formulation was observed to behave more stably during training.

with $\sum_{j \in \{1, 2\}} p_{ij} = 1$ for each i . These transition probabilities describe how frequently attention switches between speakers, as in the HMM [9]. Similar to the HMM approach, we set $p_{11} = p_{22}$ as a tunable parameter for controlling the temporal smoothness of the decoded attention state sequence to reduce the number of hyperparameters and ensure symmetric treatment of both speakers.

The attention state sequence and the unknown parameters are to be jointly estimated such that the likelihood of the observed data is maximized. Since the attention state sequence is hidden, we use the expectation-maximization (EM) algorithm to iteratively estimate the parameters and decode the attention states [11], as explained below.

A. Forward-backward algorithm

Let $\psi_t = \{y_1, \dots, y_t\}$ denote the observed data up to time t , and $\Theta = \{\beta_1, \beta_2, \sigma_1^2, \sigma_2^2\}$ denote the set of to-be-estimated model parameters. In the EM algorithm, we need to compute the posterior probabilities of the hidden states given the observed data and current parameter estimates, i.e., $p(S_t | \psi_T; \Theta^k)$, $t = 1, \dots, T$, where Θ^k denotes the parameter estimates at the k -th iteration and is omitted for brevity in the following equations. This can be computed using the forward-backward algorithm [12], with

- *Forward pass:* Computing the posterior probabilities $p(S_t | \psi_t)$ with two iterative steps:

$$\begin{aligned} \Pr(S_t | \psi_t) &= \Pr(S_t | \psi_{t-1}, y_t) = \frac{f(S_t, y_t | \psi_{t-1})}{f(y_t | \psi_{t-1})} \\ &= \frac{f(y_t | S_t) \Pr(S_t | \psi_{t-1})}{\sum_{S_t} f(y_t | S_t) \Pr(S_t | \psi_{t-1})}, \end{aligned} \quad (4)$$

$$\Pr(S_t | \psi_{t-1}) = \sum_{S_{t-1}} \Pr(S_t | S_{t-1}) \Pr(S_{t-1} | \psi_{t-1}), \quad (5)$$

where $f(y_t | S_t)$ is the likelihood of observation y_t given state S_t that can be computed from (2), and $\Pr(S_t | S_{t-1})$ is given by (3). $\Pr(S_0 | \psi_0) = \Pr(S_0)$ is the initial state distribution and can be set to a uniform distribution if no prior knowledge is available.

- *Backward pass:* Given $\Pr(S_T | \psi_T)$ at the last iteration of the forward pass, we can compute the smoothing probabilities $\Pr(S_t | \psi_T)$ iteratively backward in time:

$$\begin{aligned} \Pr(S_t | \psi_T) &= \sum_{S_{t+1}} \Pr(S_t, S_{t+1} | \psi_T) \\ &= \sum_{S_{t+1}} \Pr(S_t | S_{t+1}, \psi_t) \Pr(S_{t+1} | \psi_T) \\ &= \sum_{S_{t+1}} \frac{\Pr(S_t, S_{t+1} | \psi_t) \Pr(S_{t+1} | \psi_T)}{\Pr(S_{t+1} | \psi_t)} \\ &= \sum_{S_{t+1}} \frac{\Pr(S_t, S_{t+1} | \psi_t) \Pr(S_{t+1} | \psi_T)}{\sum_{S_t} \Pr(S_t, S_{t+1} | \psi_t)} \\ &= \sum_{S_{t+1}} \frac{\Pr(S_t | \psi_t) \Pr(S_{t+1} | S_t) \Pr(S_{t+1} | \psi_T)}{\sum_{S_t} [\Pr(S_t | \psi_t) \Pr(S_{t+1} | S_t)]}, \end{aligned} \quad (6)$$

where $\Pr(S_t|\psi_t)$ is provided by the forward pass (4)-(5) and $\Pr(S_{t+1}|S_t)$ by (3).

Note that here we assume the full observed data are available to improve the inference of S_t using future information. However, the method can easily be made causal by performing only the forward pass.

B. EM iterations

The EM algorithm performs two steps iteratively until convergence. In the *E-step*, we formulate the Q-function, which is the expectation of the complete data log-likelihood w.r.t. the posterior distribution of the hidden states given the observed data and current parameter estimates:

$$Q_k(\Theta) = E_{\xi_T|\psi_T; \Theta^k} [\log f(\xi_T, \psi_T; \Theta)], \quad (7)$$

where $\xi_T = \{S_1, \dots, S_T\}$ is the sequence of hidden states. The likelihood of the complete data can be factorized as

$$\begin{aligned} f(\xi_T, \psi_T; \Theta) &= f(\psi_T|\xi_T; \Theta) \Pr(\xi_T) \\ &= \prod_t f(y_t|S_t; \Theta) \prod_t \Pr(S_t|S_{t-1}), \end{aligned} \quad (8)$$

and thus the Q-function can be written as

$$Q_k(\Theta) = \sum_t \sum_{S_t} \Pr(S_t|\psi_T; \Theta^k) \log f(y_t|S_t; \Theta) + C, \quad (9)$$

where C is a constant independent of Θ coming from the factor $\prod_t \Pr(S_t|S_{t-1})$ in (8).

In the *M-step*, we set the derivatives of the Q-function w.r.t. the model parameters to zero and solve for the parameters that maximize the Q-function. It is straightforward to derive the update rules for the regression coefficients and noise variances [12]:

$$\begin{aligned} \beta_i^{k+1} &= \\ &[\sum_t \hat{\mathbf{x}}_t \hat{\mathbf{x}}_t^\top \Pr(S_t = i|\psi_T; \Theta^k)]^{-1} [\sum_t \hat{\mathbf{x}}_t y_t \Pr(S_t = i|\psi_T; \Theta^k)], \end{aligned} \quad (10)$$

$$(\sigma_i^{k+1})^2 = \frac{\sum_t (y_t - \hat{\mathbf{x}}_t^\top \beta_i^{k+1})^2 \Pr(S_t = i|\psi_T; \Theta^k)}{\sum_t \Pr(S_t = i|\psi_T; \Theta^k)} \quad (11)$$

for $i \in \{1, 2\}$. The inference of the hidden states $\Pr(S_t|\psi_T; \Theta^k)$ in each iteration is performed using the forward-backward algorithm described in the previous subsection.

C. Initialization

Initialization is important for the convergence of the EM algorithm, especially when working with low SNR data such as EEG. Therefore, although the proposed method can in principle be initialized with random parameters and work in a purely unsupervised manner, we find that it greatly accelerates the convergence and leads to better decoding performance if we initialize the parameters with a pretrained decoder β^* that is trained, e.g., on a separate dataset using linear regression as in [2]–[4], where the attended speaker is used

as the target. Intuitively, in our formulation (2), the state-dependent decoders β_1 and β_2 can be initialized as β^* and $-\beta^*$, respectively, assuming $\hat{\mathbf{x}}_t$ is mainly dominated by the neural response to the attended speaker². The noise variances σ_1^2 and σ_2^2 can have the same initialization, e.g., the mean squared error of the pretrained decoder β^* on the training data, if we assume the speech envelopes are properly normalized and the reconstruction error is similar regardless of which speaker is attended. We note that this symmetry is only used for initialization, but it is not enforced within the EM algorithm during training, as the underlying assumptions could be violated in practice.

III. EXPERIMENT

We compare the proposed MSM method with the HMM post-processing method (non-causal version) applied to neural tracking correlations generated by a linear decoder as proposed by Heintz et al. [9] on a dataset [4] that is publicly available at [13]. The dataset contains 72-min audio data with two competing speakers and the corresponding 64-channel EEG data from 16 normal-hearing participants instructed to attend to one of the speakers in each trial. The speech envelopes are extracted using a gammatone filterbank as described in Biesmans et al. [4]. The audio and EEG data are bandpass filtered in the 0.1–4 Hz range [14] and downsampled to 10 Hz. To artificially increase the number of speaker switches, we divide the data into 1-minute segments and randomly swap the assignment of speaker 1 and speaker 2 in each segment. For MSM, the performance is evaluated directly on the per-sample decoding results without any window-based post-processing.

A. Metrics and parameters

We define the decoding accuracy as the percentage of time points where the predicted attended speaker matches the true attended speaker. The switch detection time is defined as the time difference between the true attention switch and the nearest time point where the decoded attention state switches correctly, i.e., the posterior probability for the newly attended speaker exceeds 50%. Since both methods are non-causal, meaning that this can happen before the actual switch, we take the absolute value of the switch detection time and average it across all true attention switches. For the HMM post-processing method, switch detection time is by definition larger than the chosen window length as window-based processing has this inherent algorithmic delay. If the computed switch detection time is larger than the time interval between two true attention switches, this is considered a missed switch and the switch detection time is set to the time interval between two true attention switches.

The comparison is performed under three settings based on how the pretrained decoder β^* is obtained:

- *Supervised user-specific (Sup-US):* β^* is trained for each participant with attention labels (on a held-out part of

²One could also subtract a pre-trained unattended-envelope decoder from β^* to make the initialization more compatible with (2), but this did not improve the results.

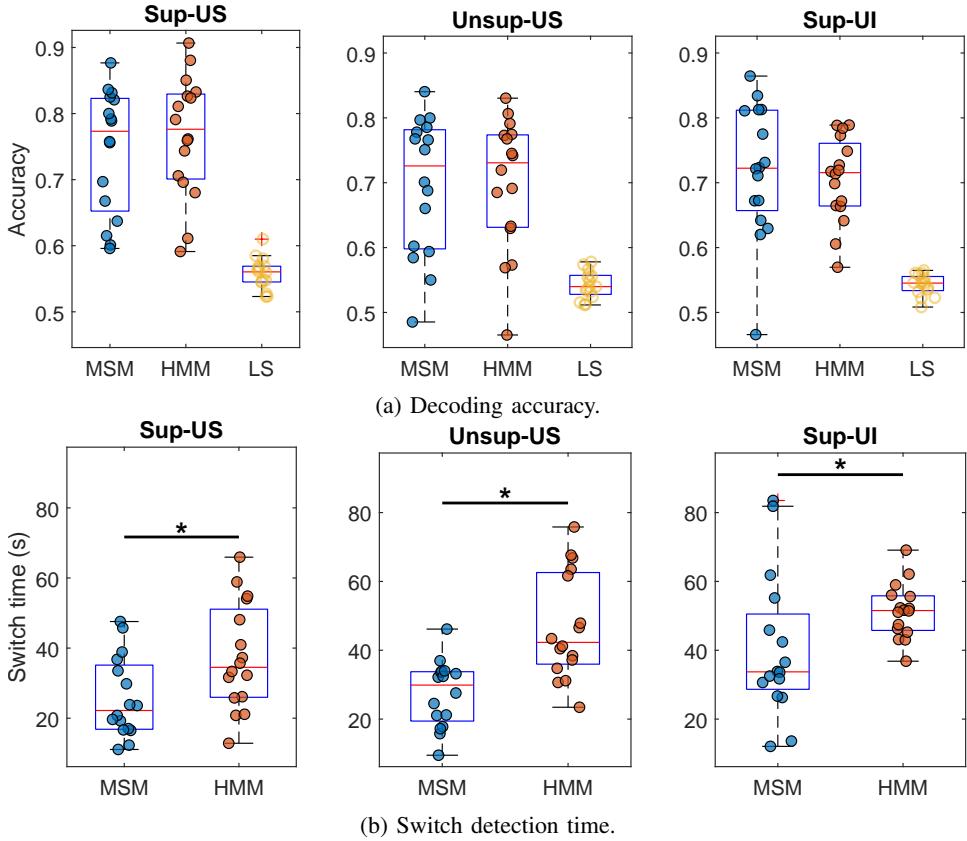


Fig. 1: Performance comparison of the proposed MSM method and the HMM post-processing method (operating on 1-second windows) under different settings. The dots mark the cross-validated outcomes for each participant. Box-and-whisker plots are used to display the median and quartiles, with the whiskers encompassing the full data spread excluding outliers. “*” indicates the results of MSM are significantly lower than HMM (Wilcoxon signed-rank test, $p < 0.05$). The decoding accuracy of the LS decoder without post-processing is also shown for reference. Its switch detection time is not reported, as the noisy outputs produce a highly fragmented state sequence, which does not allow computing a meaningful switch detection time.

the data of the participant) using least squares (LS), as commonly done in the literature [2]–[4].

- *Unsupervised user-specific (Unsup-US)*: β^* is trained for each participant in an unsupervised manner as proposed by Geirnaert et al. in [15].
- *Supervised user-independent (Sup-UI)*: β^* is trained as in the *Sup-US* setting but with data from all participants (except the test participant) concatenated together.

For all settings, β^* is a spatial-temporal decoder that captures EEG information from 0 – 500 ms after stimulus onset. The obtained pretrained decoder β^* is used to initialize the MSM as described in Section II-C, and the HMM post-processing method is applied to the outputs of the pretrained decoder. In the *Sup-US* and *Unsup-US* settings, the pretrained decoder is cross-validated on the same participant’s data (3 folds), while in the *Sup-UI* setting, we use leave-one-participant-out cross-validation. In both cases, β^* is trained on the available training data (either 2 folds or $N - 1$ participants). While the test set is used solely for evaluation in the HMM method, the MSM method utilizes the test data for two purposes: first, to fit the MSM parameters via the EM algorithm, and subsequently to

evaluate decoding performance. Note that no attention labels are needed during EM fitting, so this does not introduce label leakage.

As pointed out by Heintz et al. [9], the transition probabilities p_{ij} and the window length are important parameters that influence the decoding accuracy and switch detection time. Therefore, we follow their setting and evaluate on correlations obtained from 1-second windows with $p_{12} = p_{21} = 10^{-3}$. For a fair comparison, we set the MSM’s transition probabilities as $p_{12} = p_{21} = 10^{-4}$ such that the expected number of switches per unit time is equal for both methods, given the 10 Hz sampling frequency. In the original HMM post-processing method, the distributions of the correlation between the reconstructed speech envelope and the ground-truth speech envelope under attended and unattended states are estimated in a supervised manner. These distributions are then used when computing the emission probabilities. To ensure a fair comparison with the MSM, which does not require attention labels, we modify the HMM’s emission distribution estimation to be unsupervised: we pool the correlations from both states and fit a two-component Gaussian mixture model where the

component with the higher mean is assigned to the attended state.

B. Results

We present the decoding accuracy and switch detection time for the proposed MSM and the HMM post-processing methods across the three initialization settings (*Sup-US*, *Unsup-US*, and *Sup-UI*) in Fig. 1. For reference, the decoding accuracy using raw correlations on 1-second windows without HMM post-processing (denoted as “LS”) is included in Fig. 1a. Under all settings, the accuracy of the raw LS decoder ranges between 50% and 60%, which is near chance-level. With HMM post-processing, the median decoding accuracy improves to 77.6%, 73.1%, and 71.6% for the *Sup-US*, *Unsup-US*, and *Sup-UI* settings, respectively. The MSM achieves comparable results, with median accuracies of 77.3%, 72.6%, and 72.2%, respectively. No significant difference in decoding accuracy was observed between the two methods across all settings (Wilcoxon signed-rank test, $p > 0.05$ in all cases).

Regarding switch detection time (Fig. 1b), the LS decoder is excluded because its noisy outputs lead to highly fragmented state sequences, making the switch detection time poorly defined. The MSM achieves significantly shorter switch detection times than the HMM post-processing method across all settings ($p = 0.0004$, 0.0003 , and 0.0233 for *Sup-US*, *Unsup-US*, and *Sup-UI*, respectively). Specifically, the median switch detection time drops from 34.5 s to 22.2 s in the *Sup-US* setting, from 42.3 s to 29.9 s in the *Unsup-US* setting, and from 51.5 s to 33.7 s in the *Sup-UI* setting.

In principle, the choice of pretrained decoder should have a smaller impact on the MSM since it only serves as an initialization, whereas the HMM post-processing directly operates on the decoder’s output. However, the results do not show that the MSM is substantially more robust in the *Unsup-US* and *Sup-UI* settings, where the pretrained decoder is typically weaker. This suggests that the MSM may be susceptible to local optima during EM iterations, meaning the initialization remains a significant factor in final performance. This is consistent with the low SNR of EEG data and the high dimensionality of the MSM parameter space.

IV. CONCLUSION

We proposed a MSM for sample-level sAAD that integrates the relationship between EEG and speech envelopes with the temporal dynamics of attention. The MSM achieves decoding accuracy comparable to recently proposed HMM post-processing methods while yielding significantly shorter switch detection latency. Furthermore, it provides per-sample attention decisions rather than using a traditional window-based approach, eliminating the need to pre-select a window length. While the current formulation relies on a linear regression framework, it can, in principle, be extended to non-linear architectures such as deep neural networks. However, the estimation of such non-linear models is inherently challenging given the low SNR of EEG data and may require more advanced inference techniques, such as variational inference

or Markov Chain Monte Carlo methods. We leave these explorations for future work.

REFERENCES

- [1] E. C. Cherry, “Some experiments on the recognition of speech, with one and with two ears,” *Journal of the acoustical society of America*, vol. 25, pp. 975–979, 1953.
- [2] S. Geirnaert, S. Vandecappelle, E. Alickovic, A. de Cheveigné, E. C. Lalor, B. T. Meyer, S. Miran, T. Francart, and A. Bertrand, “Electroencephalography-Based Auditory Attention Decoding: Toward Neurosteered Hearing Devices,” *IEEE Signal Processing Magazine*, vol. 38, no. 4, pp. 89–102, 2021.
- [3] J. A. O’sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, and E. C. Lalor, “Attentional selection in a cocktail party environment can be decoded from single-trial eeg,” *Cerebral cortex*, vol. 25, no. 7, pp. 1697–1706, 2015.
- [4] 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, 2016.
- [5] E. Alickovic, T. Lunner, F. Gustafsson, and L. Ljung, “A tutorial on auditory attention identification methods,” *Frontiers in neuroscience*, vol. 13, p. 153, 2019.
- [6] G. Ciccarelli, M. Nolan, J. Perricone, P. T. Calamia, S. Haro, J. O’sullivan, N. Mesgarani, T. F. Quatieri, and C. J. Smalt, “Comparison of two-talker attention decoding from eeg with nonlinear neural networks and linear methods,” *Scientific reports*, vol. 9, no. 1, p. 11538, 2019.
- [7] C. Puffay, B. Accou, L. Bollens, M. J. Monesi, J. Vanthornhout, H. Van Hamme, and T. Francart, “Relating eeg to continuous speech using deep neural networks: a review,” *Journal of Neural Engineering*, vol. 20, no. 4, p. 041003, 2023.
- [8] N. Nguyen Thanh Duc, H. Phan, S. Geirnaert, K. Mikkelsen, and P. Kidmose, “AADNet: An End-to-End Deep Learning Model for Auditory Attention Decoding,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 33, pp. 2695–2706, 2025.
- [9] N. Heintz, T. Francart, and A. Bertrand, “Post-processing of eeg-based auditory attention decoding decisions via hidden markov models,” *arXiv preprint arXiv:2506.24024*, 2025.
- [10] J. D. Hamilton, “Analysis of time series subject to changes in regime,” *Journal of econometrics*, vol. 45, no. 1-2, pp. 39–70, 1990.
- [11] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the em algorithm,” *Journal of the royal statistical society: series B (methodological)*, vol. 39, no. 1, pp. 1–22, 1977.
- [12] C.-J. Kim and C. R. Nelson, *State-space models with regime switching: classical and Gibbs-sampling approaches with applications*. MIT press, 2017.
- [13] N. Das, T. Francart, and A. Bertrand, “Auditory attention detection dataset kuleuven,” Aug. 2019. [Online]. Available: <https://doi.org/10.5281/zenodo.4004271>
- [14] T. Li, S. Geirnaert, D. Van de Broek, E. Bellon, B. De Smedt, and A. Bertrand, “Temporal variation in the acoustic dynamic range is a confounding factor in EEG-based tracking of absolute auditory attention to speech,” in *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2025, pp. 1–4.
- [15] S. Geirnaert, T. Francart, and A. Bertrand, “Unsupervised Self-Adaptive Auditory Attention Decoding,” *IEEE Journal on Biomedical and Health Informatics*, vol. 25, no. 10, pp. 3955–3966, 2021.