# EFFICIENT SOLUTIONS FOR MITIGATING INITIALIZATION BIAS IN UNSUPERVISED SELF-ADAPTIVE AUDITORY ATTENTION DECODING

*Yuanyuan Yao*[1], *Simon Geirnaert*[1,2], *Tinne Tuytelaars*[3], *Alexander Bertrand*[1]

[1]KU Leuven, Department of Electrical Engineering (ESAT),
STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, Belgium
[2]KU Leuven, Department of Neurosciences, Research Group ExpORL, Belgium
[3]KU Leuven, Department of Electrical Engineering (ESAT), PSI, Belgium

## ABSTRACT

Decoding the attended speaker in a multi-speaker environment from electroencephalography (EEG) has attracted growing interest in recent years, with neuro-steered hearing devices as a driver application. Current approaches typically rely on ground-truth labels of the attended speaker during training, necessitating calibration sessions for each user and each EEG set-up to achieve optimal performance. While unsupervised self-adaptive auditory attention decoding (AAD) for stimulus reconstruction has been developed to eliminate the need for labeled data, it suffers from an initialization bias that can compromise performance. Although an unbiased variant has been proposed to address this limitation, it introduces substantial computational complexity that scales with data size. This paper presents three computationally efficient alternatives that achieve comparable performance, but with a significantly lower and constant computational cost. The code for the proposed algorithms is available at `https://github.com/YYao-42/Unsupervised_AAD`.

***Index Terms*—** auditory attention decoding, EEG, unsupervised learning

## 1. INTRODUCTION

Auditory attention decoding (AAD) aims to identify the attended speaker in complex multi-speaker auditory environments from brain signals such as electroencephalography (EEG). This can be useful for neuro-steered hearing aids, where the attended speaker's volume can be enhanced while suppressing unattended speakers [1, 2]. A common approach trains a linear neural decoder to reconstruct features of the attended speech from EEG, and during testing, the attended speaker is identified as the candidate most correlated with the reconstruction [3, 4, 5, 6]. However, training or fine-tuning a neural decoder requires labeled data indicating the attended speaker at each time point, which in practice necessitates a calibration session where the user follows attention instructions. Such sessions are a nuisance, in particular if they have to be repeated frequently.

To address this, Geirnaert et al. [7] proposed unsupervised AAD based on stimulus reconstruction, beginning with training a decoder with random attention labels. Even when using random labels in this initial training phase, it was shown that the resulting decoder can achieve above-chance decoding performance. Each subsequent iteration updates the model using newly predicted labels, which include a higher proportion of true attended labels. This yields a higher-quality decoder that produces more accurate predictions in the next iteration. This bootstrapping effect continues until convergence.

However, this approach suffers from initialization bias, as training and testing on the same data across iterations causes the model to favor initial (possibly wrong) predictions [8]. Heintz et al. addressed this by implementing leave-one-out cross-validation within each iteration [8]: training on $K-1$ segments and predicting the remaining segment, repeated for all $K$ segments. While this cross-validated version improves performance, particularly with limited training data, it requires training the model $K$ times per iteration, creating substantial computational overhead.

In this paper, we propose unsupervised training methods based on canonical correlation analysis (CCA) that are inherently robust to AAD label noise in such self-adaptive iterations. Our approach achieves comparable performance to the cross-validated version from [8] but requires only one model training per iteration instead of $K$. This substantial reduction in computation time makes it particularly well-suited for real-time or time-adaptive implementations [9].

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), the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 101138304), Internal Funds KU Leuven (projects IDN/23/006, C14/25/108, and C3/25/107), and the Flemish Government (AI Research Program). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the granting authorities. Neither the European Union nor the granting authorities can be held responsible for them.
All authors are also affiliated with Leuven.AI - KU Leuven institute for AI, Belgium.

## 2. METHODS

We consider a 2-speaker scenario without loss of generality as all methods can be straightforwardly extended to an $N$-speaker setting. Let $\mathbf{S}_1, \mathbf{S}_2 \in \mathbb{R}^{T \times D_s}$ denote the speech features, such as speech envelopes, of the two candidate speakers, where $T$ is the number of samples and $D_s$ is the feature dimension. Given EEG signals $\mathbf{X} \in \mathbb{R}^{T \times D_x}$ with $D_x$ channels, the goal is to determine which of $\mathbf{S}_1$ or $\mathbf{S}_2$ corresponds to the attended speaker's features (denoted $\mathbf{S}_a$) and which to the unattended speaker's features (denoted $\mathbf{S}_u$). In practice, decoding is performed on a segment basis to obtain time-resolved estimates of attention: a long recording is divided into $K$ segments, yielding $\{\mathbf{X}_k\}_{k=1}^K$, $\{\mathbf{S}_{1k}\}_{k=1}^K$, and $\{\mathbf{S}_{2k}\}_{k=1}^K$, and the attended speaker is identified per segment. For simplicity, we assume the signals are centered, i.e., $\mathbb{E}[\mathbf{X}_k] = \mathbb{E}[\mathbf{S}_{1k}] = \mathbb{E}[\mathbf{S}_{2k}] = \mathbf{0}$, where $\mathbb{E}[\cdot]$ denotes the expectation operator.

### 2.1. Baseline: Single-Encoder Version

The original algorithm in [7] is based on a backward model that reconstructs the attended speaker features from EEG signals using linear regression. Here we present a more general CCA-based variant based on [10]. In the supervised setting, CCA optimizes a decoder $\mathbf{w}_x \in \mathbb{R}^{D_x \times 1}$ (on the EEG side) and an encoder $\mathbf{w}_a \in \mathbb{R}^{D_s \times 1}$ (on the audio side) to maximize the correlation between the transformed EEG signals and the features of the attended speaker:

$$\begin{aligned} \underset{\mathbf{w}_x, \mathbf{w}_a}{\text{maximize}} \quad & \mathbf{w}_x^{\mathrm{T}} \mathbf{X}^{\mathrm{T}} \mathbf{S}_a \mathbf{w}_a \\ \text{subject to} \quad & \mathbf{w}_x^{\mathrm{T}} \mathbf{X}^{\mathrm{T}} \mathbf{X} \mathbf{w}_x = 1, \\ & \mathbf{w}_a^{\mathrm{T}} {\mathbf{S}_a}^{\mathrm{T}} \mathbf{S}_a \mathbf{w}_a = 1. \end{aligned} \tag{1}$$

As opposed to the backward model, CCA allows exploiting both multivariate data modalities to identify a shared subspace that maximizes the correlation between them.

The optimized vectors $\hat{\mathbf{w}}_x$ and $\hat{\mathbf{w}}_a$ are the first canonical components. Higher-order components are obtained iteratively by solving (1) in a subspace where the transformed signals are orthogonal to those from previous iterations. In matrix form, this corresponds to:

$$\begin{aligned} \underset{\mathbf{W}_x, \mathbf{W}_a}{\text{maximize}} \quad & \mathrm{Tr}\left(\mathbf{W}_x^{\mathrm{T}} \mathbf{R}_{xa} \mathbf{W}_a\right) \\ \text{subject to} \quad & \mathbf{W}_x^{\mathrm{T}} \mathbf{R}_{xx} \mathbf{W}_x = \mathbf{I}_Q, \\ & \mathbf{W}_a^{\mathrm{T}} \mathbf{R}_{aa} \mathbf{W}_a = \mathbf{I}_Q, \end{aligned} \tag{2}$$

where $\mathbf{W}_x = [\mathbf{w}_{x1} \cdots \mathbf{w}_{xQ}]$ and $\mathbf{W}_a = [\mathbf{w}_{a1} \cdots \mathbf{w}_{aQ}]$ contain the first $Q$ canonical components, $\mathbf{R}_{xa} = \mathbf{X}^{\mathrm{T}} \mathbf{S}_a$, $\mathbf{R}_{xx} = \mathbf{X}^{\mathrm{T}} \mathbf{X}$, and $\mathbf{R}_{aa} = {\mathbf{S}_a}^{\mathrm{T}} \mathbf{S}_a$. The solution to (2) can be obtained by solving a generalized eigenvalue decomposition (GEVD) problem [11]:

$$\mathbf{R} \hat{\mathbf{W}} = \mathbf{D} \hat{\mathbf{W}} \mathbf{\Lambda}, \tag{3}$$

where $\mathbf{\Lambda}$ is a diagonal matrix containing the generalized eigenvalues ordered in descending order, and

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_{xx} & \mathbf{R}_{xa} \\ \mathbf{R}_{xa}^{\mathrm{T}} & \mathbf{R}_{aa} \end{bmatrix}, \mathbf{D} = \begin{bmatrix} \mathbf{R}_{xx} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_{aa} \end{bmatrix}, \hat{\mathbf{W}} = \begin{bmatrix} \hat{\mathbf{W}}_x \\ \hat{\mathbf{W}}_a \end{bmatrix}. \tag{4}$$

For a test pair of features $(\mathbf{S}_{1k}, \mathbf{S}_{2k})$, the attended speaker is identified by finding

$$j = \underset{i \in \{1,2\}}{\arg\max}\, \tilde{\rho}_{ik}, \tag{5}$$

where $\tilde{\rho}_{ik}$ is the sum of canonical correlations computed using the trained decoder $\hat{\mathbf{W}}_x$ and encoder $\hat{\mathbf{W}}_a$:

$$\tilde{\rho}_{ik} = \sum_{q=1}^{Q} \frac{\hat{\mathbf{w}}_{xq}^{\top} \mathbf{X}_k^{\top} \mathbf{S}_{ik} \hat{\mathbf{w}}_{aq}}{\sqrt{\hat{\mathbf{w}}_{xq}^{\top} \mathbf{X}_k^{\top} \mathbf{X}_k \hat{\mathbf{w}}_{xq}} \sqrt{\hat{\mathbf{w}}_{aq}^{\top} \mathbf{S}_{ik}^{\top} \mathbf{S}_{ik} \hat{\mathbf{w}}_{aq}}}, \quad i = 1, 2. \tag{6}$$

$\mathbf{S}_{jk}$ is then assigned as $\mathbf{S}_{ak}$, and the other as $\mathbf{S}_{uk}$.

In the unsupervised setting, the true attended speaker labels are unavailable, so the statistics $\mathbf{R}_{aa}$ and $\mathbf{R}_{xa}$ cannot be computed because the per-speaker segments $\{\mathbf{S}_{1k}\}_{k=1}^K$ and $\{\mathbf{S}_{2k}\}_{k=1}^K$ cannot be mapped to attended and unattended sets $\{\mathbf{S}_{ak}\}_{k=1}^K$ and $\{\mathbf{S}_{uk}\}_{k=1}^K$. Following the self-adaptive approach of [7], we begin by randomly assigning one of $\mathbf{S}_{1k}$ or $\mathbf{S}_{2k}$ as $\mathbf{S}_{ak}$ (and the other as $\mathbf{S}_{uk}$). Using these initial random labels, we form $\mathbf{X}, \mathbf{S}_a$, and $\mathbf{S}_u$ by stacking the segments over time. $\mathbf{W}_x$ and $\mathbf{W}_a$ are then estimated by solving (3), and the labels are updated according to (5). This procedure is iterated, each time based on the previously assigned labels, until convergence. A summary of this basic version, here called the single-encoder version, is provided in Algorithm 1. However, as identified in [8], this approach suffers from initialization bias where the model favors assigning the same labels as in the previous iteration, making initial wrong predictions persist. To address this without expensive inner cross-validations as in [8], we propose three variants described in the following sections.

### 2.2. Two-Encoder Version

For the single-encoder version, only the features of the attended speaker $\mathbf{S}_a$ are incorporated in the optimization. We extend this with two encoders: $\mathbf{W}_a$ for the attended features $\mathbf{S}_a$ and $\mathbf{W}_u$ for the unattended features $\mathbf{S}_u$, both sharing the same decoder $\mathbf{W}_x$, to jointly maximize the correlation between EEG and the features of both speakers:

$$\begin{aligned} \underset{\mathbf{W}_x, \mathbf{W}_a, \mathbf{W}_u}{\text{maximize}} \quad & \mathrm{Tr}\left(\mathbf{W}_x^{\mathrm{T}} \mathbf{R}_{xa} \mathbf{W}_a + \mathbf{W}_x^{\mathrm{T}} \mathbf{R}_{xu} \mathbf{W}_u\right) \\ \text{subject to} \quad & \mathbf{W}_x^{\mathrm{T}} \mathbf{R}_{xx} \mathbf{W}_x = \mathbf{I}_Q, \\ & \begin{bmatrix} {\mathbf{W}_a}^{\mathrm{T}} & {\mathbf{W}_u}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \mathbf{R}_{aa} & \mathbf{R}_{au} \\ \mathbf{R}_{au}^{\mathrm{T}} & \mathbf{R}_{uu} \end{bmatrix} \begin{bmatrix} \mathbf{W}_a \\ \mathbf{W}_u \end{bmatrix} = \mathbf{I}_Q, \end{aligned} \tag{7}$$

where $\mathbf{R}_{xu} = \mathbf{X}^{\mathrm{T}} \mathbf{S}_u, \mathbf{R}_{au} = {\mathbf{S}_a}^{\mathrm{T}} \mathbf{S}_u$, and $\mathbf{R}_{uu} = {\mathbf{S}_u}^{\mathrm{T}} \mathbf{S}_u$.

**Algorithm 1** Single-/Two-Encoder Version

**Input:** EEG segments $\{\mathbf{X}_k\}_{k=1}^K$, speaker features $\{\mathbf{S}_{1k}\}_{k=1}^K$ and $\{\mathbf{S}_{2k}\}_{k=1}^K$, number of components $Q$
**Initialize:** For each $k$, draw $j \in \{1,2\}$ uniformly and set $\mathbf{S}_{ak} \leftarrow \mathbf{S}_{jk}$, $\mathbf{S}_{uk} \leftarrow \mathbf{S}_{(3-j)k}$
**while** not converged **do**
  Build $\mathbf{R}$, $\mathbf{D}$ from current labels. For single-encoder,

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_{xx} & \mathbf{R}_{xa} \\ \mathbf{R}_{xa}^{\mathrm{T}} & \mathbf{R}_{aa} \end{bmatrix}, \mathbf{D} = \begin{bmatrix} \mathbf{R}_{xx} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_{aa} \end{bmatrix}.$$

  For two-encoder,

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_{xx} & \mathbf{R}_{xa} & \mathbf{R}_{xu} \\ \mathbf{R}_{xa}^{\mathrm{T}} & \mathbf{R}_{aa} & \mathbf{R}_{au} \\ \mathbf{R}_{xu}^{\mathrm{T}} & \mathbf{R}_{au}^{\mathrm{T}} & \mathbf{R}_{uu} \end{bmatrix}, \mathbf{D} = \begin{bmatrix} \mathbf{R}_{xx} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_{aa} & \mathbf{R}_{au} \\ \mathbf{0} & \mathbf{R}_{au}^{\mathrm{T}} & \mathbf{R}_{uu} \end{bmatrix}.$$

  Solve $\mathbf{R}\hat{\mathbf{W}} = \mathbf{D}\hat{\mathbf{W}}\mathbf{\Lambda}$ to obtain $\hat{\mathbf{W}}$ (partitioned as $\hat{\mathbf{W}}_x, \hat{\mathbf{W}}_a$ for single-encoder; $\hat{\mathbf{W}}_x, \hat{\mathbf{W}}_a, \hat{\mathbf{W}}_u$ for two-encoder).
  For each $k$, compute $\tilde{\rho}_{1k}$ and $\tilde{\rho}_{2k}$ using (6), and set $j = \arg\max_{i\in\{1,2\}} \tilde{\rho}_{ik}$, $\mathbf{S}_{ak} \leftarrow \mathbf{S}_{jk}$, $\mathbf{S}_{uk} \leftarrow \mathbf{S}_{(3-j)k}$.
**end while**
**Output:** $\hat{\mathbf{W}}$

In the ideal case with sufficient labeled data, the two-encoder version may suffer performance loss due to its inherent reduced discriminative power as $\mathbf{W}_x$ is encouraged to extract EEG responses to both speakers (not only the attended one). However, in unsupervised self-adaptive settings, this approach may be more robust to label errors, as $\mathbf{W}_x$ will be less attracted to wrongly labeled attended speech responses (which would otherwise bias it towards producing the same wrong labels for the next iteration). This increases the chance of recovering from wrong predictions.

The solution to (7) can again be obtained by solving the GEVD problem (3), with

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_{xx} & \mathbf{R}_{xa} & \mathbf{R}_{xu} \\ \mathbf{R}_{xa}^{\mathrm{T}} & \mathbf{R}_{aa} & \mathbf{R}_{au} \\ \mathbf{R}_{xu}^{\mathrm{T}} & \mathbf{R}_{au}^{\mathrm{T}} & \mathbf{R}_{uu} \end{bmatrix}, \mathbf{D} = \begin{bmatrix} \mathbf{R}_{xx} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_{aa} & \mathbf{R}_{au} \\ \mathbf{0} & \mathbf{R}_{au}^{\mathrm{T}} & \mathbf{R}_{uu} \end{bmatrix}, \quad \mathbf{W} = \begin{bmatrix} \hat{\mathbf{W}}_x^{\mathrm{T}} & \hat{\mathbf{W}}_a^{\mathrm{T}} & \hat{\mathbf{W}}_u^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}. \tag{8}$$

The attended speaker is still identified using (5). When predicting the attended segments, however, only the attended encoder is used (together with $\mathbf{W}_x$). A summary can be found in Algorithm 1.

### 2.3. Soft Version

Instead of making hard assignments of attended and unattended segments, we propose a soft version that assigns weights to each segment based on prediction uncertainty. This approach provides a principled middle ground between the single-encoder and two-encoder versions, maintaining the more optimal single-encoder structure while incorporating information from both attended and unattended segments adaptively when the model is less certain. The formulation follows (1)-(4), but replaces $\mathbf{S}_a$ with its soft version $p_{1k}\mathbf{S}_{1k} + p_{2k}\mathbf{S}_{2k}$, where $p_{1k}$ and $p_{2k}$ are the probabilities of $\mathbf{S}_{1k}$ and $\mathbf{S}_{2k}$ being the attended speaker in segment $k$. The

**Algorithm 2** Soft version

**Input:** EEG segments $\{\mathbf{X}_k\}_{k=1}^K$, speaker features $\{\mathbf{S}_{1k}\}_{k=1}^K$ and $\{\mathbf{S}_{2k}\}_{k=1}^K$, number of components $Q$
**Initialize:** Randomly initialize $\hat{\mathbf{W}}$
**while** not converged **do**
  Compute $\tilde{\rho}_{1k}$ and $\tilde{\rho}_{2k}$ for each $k$ using (6) and estimate parameters $\{\mu_a, \sigma_a^2, \mu_u, \sigma_u^2\}$ as in [12].
  Estimate soft labels $p_{1k}, p_{2k}$ based on (10)-(11).
  Build $\mathbf{R}$, $\mathbf{D}$ using the soft labels:

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_{xx} & \mathbf{R}_{xa} \\ \mathbf{R}_{xa}^{\mathrm{T}} & \mathbf{R}_{aa} \end{bmatrix}, \mathbf{D} = \begin{bmatrix} \mathbf{R}_{xx} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_{aa} \end{bmatrix},$$

  where

$$\mathbf{R}_{xa} = \sum_{k=1}^{K} \mathbf{X}_k^{\mathrm{T}}(p_{1k}\mathbf{S}_{1k} + p_{2k}\mathbf{S}_{2k}),$$

$$\mathbf{R}_{aa} = \sum_{k=1}^{K} (p_{1k}\mathbf{S}_{1k} + p_{2k}\mathbf{S}_{2k})^{\mathrm{T}}(p_{1k}\mathbf{S}_{1k} + p_{2k}\mathbf{S}_{2k}).$$

  Update $\hat{\mathbf{W}}$ by solving GEVD: $\mathbf{R}\hat{\mathbf{W}} = \mathbf{D}\hat{\mathbf{W}}\mathbf{\Lambda}$.
**end while**
**Output:** $\hat{\mathbf{W}}$

statistics $\mathbf{R}_{xa}$ and $\mathbf{R}_{aa}$ become:

$$\begin{aligned} \mathbf{R}_{xa} &= \sum_{k=1}^{K} \mathbf{X}_k^{\mathrm{T}}(p_{1k}\mathbf{S}_{1k} + p_{2k}\mathbf{S}_{2k}), \\ \mathbf{R}_{aa} &= \sum_{k=1}^{K} (p_{1k}\mathbf{S}_{1k} + p_{2k}\mathbf{S}_{2k})^{\mathrm{T}}(p_{1k}\mathbf{S}_{1k} + p_{2k}\mathbf{S}_{2k}). \end{aligned} \tag{9}$$

The probabilities are estimated in a subject-specific, unsupervised manner using the method proposed by Lopez-Gordo et al. [12]. This approach models the (sum of canonical) correlations between EEG and the features of the attended and unattended speakers as two Gaussian distributions: $\mathcal{N}(\mu_a, \sigma_a^2)$ and $\mathcal{N}(\mu_u, \sigma_u^2)$. The parameters of these distributions are estimated from the same training data used to learn the encoder and decoder without knowing the labels (see [12] for details). Let $j$ be the index of the attended speaker. We assume no prior label information, i.e., $p(j=1) = p(j=2) = 0.5$. Using Bayes theorem, $p_{1k}$ and $p_{2k}$ can be estimated as:

$$p_{1k} = \frac{p(\tilde{\rho}_{1k}, \tilde{\rho}_{2k}|j=1)p(j=1)}{\sum_{j=\{1,2\}} p(\tilde{\rho}_{1k}, \tilde{\rho}_{2k}|j)p(j)} = \frac{p(\tilde{\rho}_{1k}; \mu_a, \sigma_a^2)p(\tilde{\rho}_{2k}; \mu_u, \sigma_u^2)}{p(\tilde{\rho}_{1k}; \mu_a, \sigma_a^2)p(\tilde{\rho}_{2k}; \mu_u, \sigma_u^2) + p(\tilde{\rho}_{1k}; \mu_u, \sigma_u^2)p(\tilde{\rho}_{2k}; \mu_a, \sigma_a^2)}, \tag{10}$$

$$p_{2k} = 1 - p_{1k}, \tag{11}$$

where $p(\cdot; \mu, \sigma^2)$ is the probability density function of a Gaussian distribution with mean $\mu$ and variance $\sigma^2$. The soft version is summarized in Algorithm 2.

### 2.4. Sum-Initialized Single-Encoder

Rather than randomly initializing the encoder and decoder, we propose training the single-encoder model with a composite signal—the sum of the features of both speakers—in

the first iteration. This initialization corresponds to setting $p_{1k} = p_{2k} = 0.5$ in the soft version of Section 2.3, which allows the model to capture neural responses that are common to both attended and unattended speakers, providing an informative starting point without a bias to a specific speaker for an individual segment $k$. The underlying hypothesis for this simple heuristic is that the bias found in [8] is mainly driven by the initialization, resulting in a self-sustaining label bias from which the iterations cannot escape.

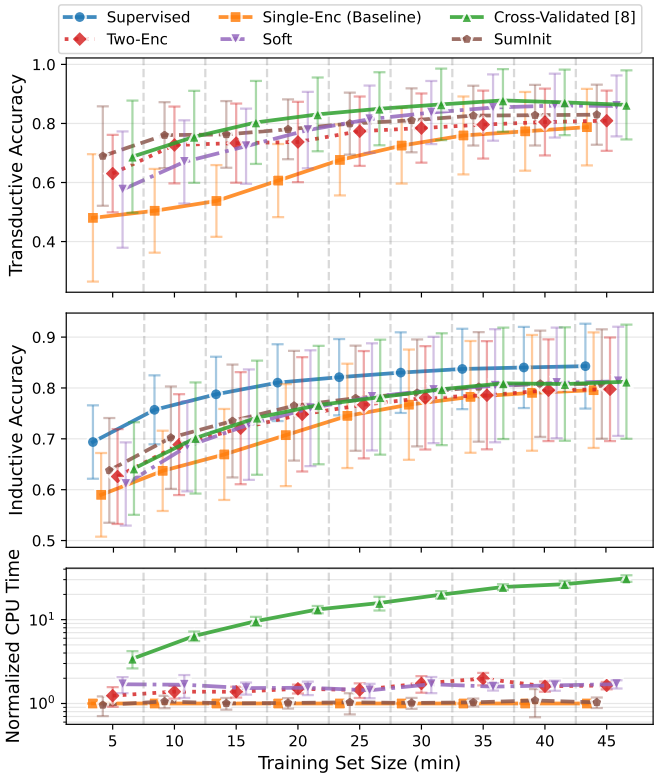


**Fig. 1**: Transductive accuracy, inductive accuracy, and normalized CPU time (w.r.t. baseline) across training set sizes. Dots and bars show mean and standard deviation across subjects and random seeds. Note: the supervised model is not shown in the transductive setting as this would imply using training labels.

## 3. EXPERIMENTS

### 3.1. Dataset and Hyperparameters

We evaluate[1] all methods on a public dataset from [13] used in [7, 8]. It contains 72-min 64-channel EEG recordings (all channels used) from 16 normal-hearing subjects attending to one of two competing speakers at $\pm 90°$ azimuth, with corresponding audio data. Following [7], audio signals are processed using a gammatone filterbank, envelopes are extracted via power-law operation (exponent 0.6) and summed across subbands. Both EEG and speech envelopes are filtered to 1-9 Hz [14], downsampled to 20 Hz, and cut into 60-s segments.

This paper works with CCA-based models, and thus the hyperparameters are slightly different from [7]. The number of components $Q$ is set to 2. For EEG signals, we create time-lagged copies at 0-150 ms (capturing current and future information) and stack them along the channel dimension. For audio envelopes, we create time-lagged copies at -250-0 ms (capturing past and current information) and stack them along the feature dimension.

### 3.2. Results

We evaluate decoding performance in both transductive and inductive settings. In the transductive setting, predictions are generated for the data on which the unsupervised model is trained, while inductive decoding assesses model generalization to unseen data. We also report normalized CPU time, defined as the ratio of each method's computational time (Windows 11, Intel Core i7-13700F, single thread) to that of the baseline single-encoder approach from Section 2.1. To examine the performance under limited training data settings, random 3-fold cross-validation is used, with the training sets subsampled to target durations. Train/test splits are identical across methods.

As shown in Fig. 1, trends are similar for both decoding settings, with the effects of removing initialization bias more pronounced in transductive decoding. The sum-initialized single-encoder consistently outperforms the two-encoder method and is particularly strong with limited data (5-15 min). The soft-label method underperforms on small sets but approaches the cross-validated variant with more data. In computational cost, the normalized CPU time for the cross-validated variant scales linearly with training set size, reaching $\sim 30\times$ for 45-min training sets. All our proposed alternative methods eliminate this scaling: two-encoder and soft methods maintain a constant normalized time of $\sim 1.5\times$ regardless of data size, while the sum-initialized method matches the baseline's time cost ($1.0\times$).

## 4. CONCLUSION

We proposed three computationally efficient solutions to mitigate initialization bias in unsupervised self-adaptive AAD. The two-encoder version trains encoders for both attended and unattended features. The soft version replaces hard segment assignments with probabilistic weights. The sum-initialized single-encoder method initializes the model with a composite signal. For smaller datasets, the sum-initialized approach is the top performer while matching the baseline's computational cost. With larger datasets, the soft-label method becomes competitive, approaching the cross-validated variant's accuracy at low cost.

[1] For conciseness, we report results for a single dataset. Similar results for an additional dataset are available in the repository linked in the abstract.

## 5. REFERENCES


[1] Simon Geirnaert, Servaas Vandecappelle, Emina Alickovic, Alain De Cheveigne, Edmund Lalor, Bernd T Meyer, Sina Miran, Tom Francart, and Alexander Bertrand, "Electroencephalography-based auditory attention decoding: Toward neurosteered hearing devices," *IEEE Signal Processing Magazine*, vol. 38, no. 4, pp. 89–102, 2021.

[2] Zexu Pan, Gordon Wichern, François G Germain, Sameer Khurana, and Jonathan Le Roux, "Neuroheed+: Improving neuro-steered speaker extraction with joint auditory attention detection," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11456–11460.

[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] Daniel DE Wong, Søren A Fuglsang, Jens Hjortkjær, Enea Ceolini, Malcolm Slaney, and Alain De Cheveigne, "A comparison of regularization methods in forward and backward models for auditory attention decoding," *Frontiers in neuroscience*, vol. 12, pp. 531, 2018.

[5] Gregory Ciccarelli, Michael Nolan, Joseph Perricone, Paul T Calamia, Stephanie Haro, James O'sullivan, Nima Mesgarani, Thomas F Quatieri, and Christopher J Smalt, "Comparison of two-talker attention decoding from EEG with nonlinear neural networks and linear methods," *Scientific reports*, vol. 9, no. 1, pp. 11538, 2019.

[6] Emina Alickovic, Thomas Lunner, Fredrik Gustafsson, and Lennart Ljung, "A tutorial on auditory attention identification methods," *Frontiers in neuroscience*, vol. 13, pp. 153, 2019.

[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] Nicolas Heintz, Simon Geirnaert, Tom Francart, and Alexander Bertrand, "Unbiased unsupervised stimulus reconstruction for EEG-based auditory attention decoding," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[9] 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, pp. 3767–3778, 2022.

[10] Alain De Cheveigné, Daniel DE Wong, Giovanni M Di Liberto, Jens Hjortkjær, Malcolm Slaney, and Edmund Lalor, "Decoding the auditory brain with canonical component analysis," *NeuroImage*, vol. 172, pp. 206–216, 2018.

[11] Eduardo Bayro Corrochano, *Handbook of geometric computing: applications in pattern recognition, computer vision, neuralcomputing, and robotics*, Springer, 2005.

[12] Miguel A. Lopez-Gordo, Simon Geirnaert, and Alexander Bertrand, "Unsupervised accuracy estimation for brain-computer interfaces based on selective auditory attention decoding," *IEEE Transactions on Biomedical Engineering*, pp. 1–12, 2025.

[13] 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, 2016.

[14] 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, et al., "Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party"," *Neuron*, vol. 77, no. 5, pp. 980–991, 2013.